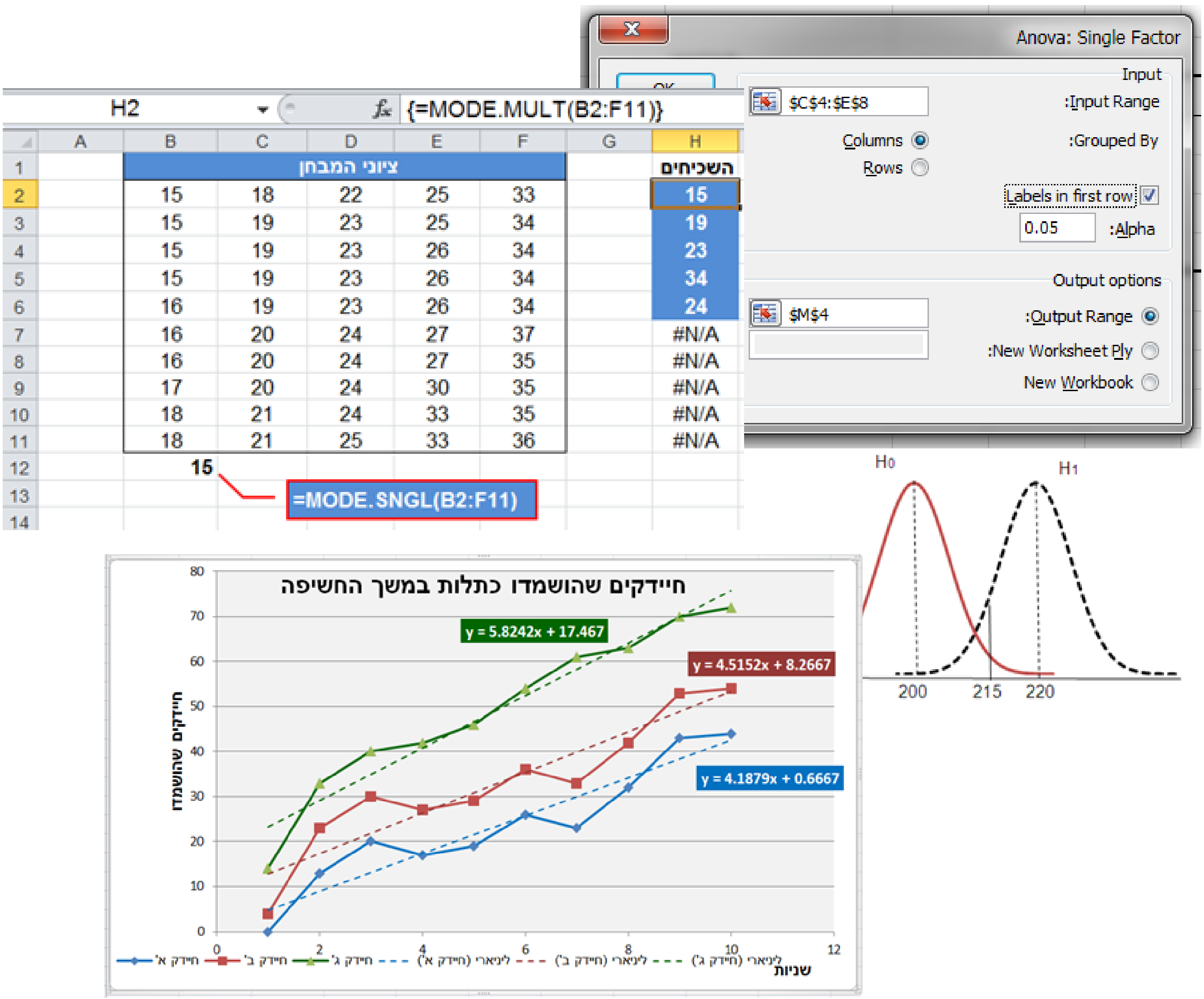
יוסי גודלניק

יישומי סטטיסטיקה

בגיליון אלקטרוני EXCEL



לשם שטף הקריאה כתובה חוברת זו בלשון זכר בלבד. החוברת מיועדת לגברים ולנשים כאחד ואין בכוונתנו להפלות או לפגוע בציבור המשתמשים/ות.

הפצה:

הוצאת הוד-עמי

09-9564716

info@hod-ami.co.il   www.hod-ami.co.il

ISBN 978-965-361-415-4 מסת"ב

מוקדש באהבה לנכדיי

שירה, נועה, גיא ועומר

תוכן עניינים

פרק 1 – פתח דבר

אקסל הוא גיליון אלקטרוני המאפשר בניית מסד נתונים, ייבוא מסד נתונים, ביצוע חישובים הנוגעים לנתונים, לרבות חישובים סטטיסטיים, הצגת הנתונים בחתכים שונים, הצגה גרפית של הנתונים וכפי שכתוב באתר MICROSOFT OFFICE "ניתוח, שיתוף וניהול של מידע לקבלת החלטות מושכלות יותר".

ישנן תוכנות סטטיסטיות, חלקן חינמיות כמו PSPP [התוכנה מהווה תחליף חינמי לתוכנת SPSS ודומה לה מאוד] ו- IDAMS [התוכנה פותחה ע"י UNESCO] וחלקן בתשלום כמו SAS ו- SPSS המשמשות גופים רבים לרבות אוניברסיטאות. תוכנות אלו "חזקות" הרבה יותר מאשר האקסל מציע בנושא הסטטיסטיקה, אולם האקסל זמין יותר לכול וגם נותן מענה הולם ללומדים בקורסי טכנאים, הנדסאים ואפילו לסטודנטים במדעי החברה.

החומר שיובא, במסגרת חוברת זו, מניח שלמשתמש ידע בסיסי ב- OFFICE [פתיחת קובץ, סגירת קובץ וכו'] וכן ידע בסיסי באקסל – למרות זאת מובאת חזרה קצרה לגבי פונקציות בסיסיות באקסל. ההתייחסות תהיה לגרסאות EXCEL 2007-2010.

חשוב להדגיש כי חוברת זו **אינה** מהווה ספר לימוד לנושא הסתברות וסטטיסטיקה ורצוי לפנות לספר לימוד מתאים, להשתתף בהרצאות ובמקביל לשפר את מיומנויות החישוב בעזרת חוברת זו.

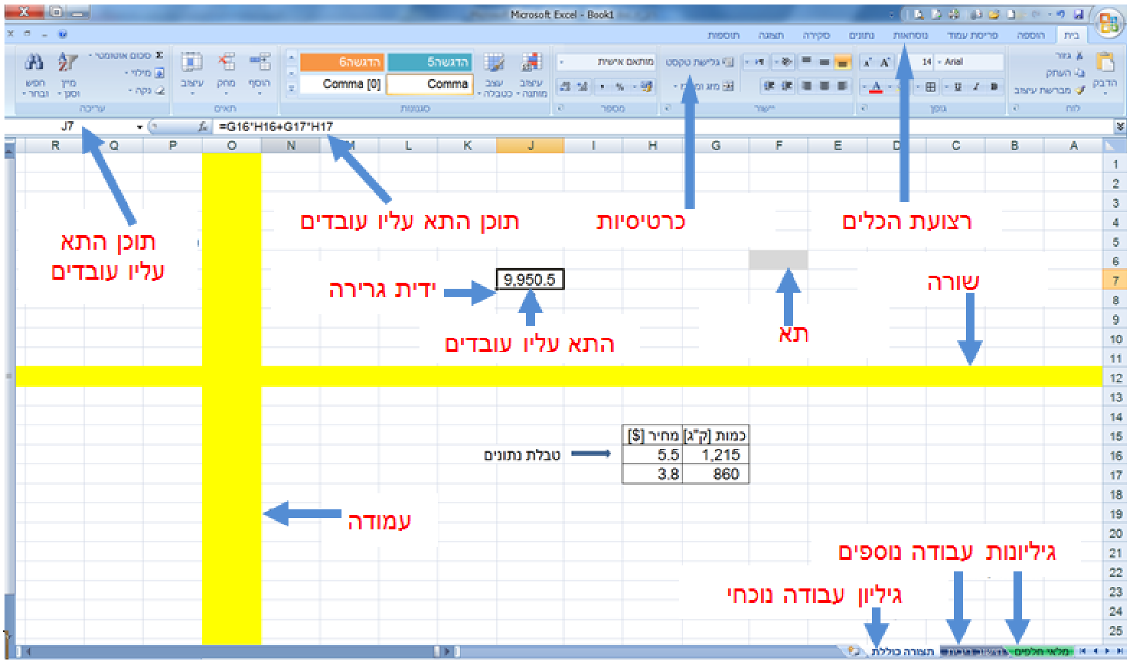
החוברת פונה לסטודנטים הלומדים בקורסי טכנאות והנדסאות לסוגיהם אך גם מתאימה לכל מי שלומד אקסל וסטטיסטיקה.

סדר פרקי החוברת אינו שגרתי והוא משלב, לפי הצורך, את הפונקציות האקסליות והסטטיסטיות.

פרק 2 – הכרת אקסל

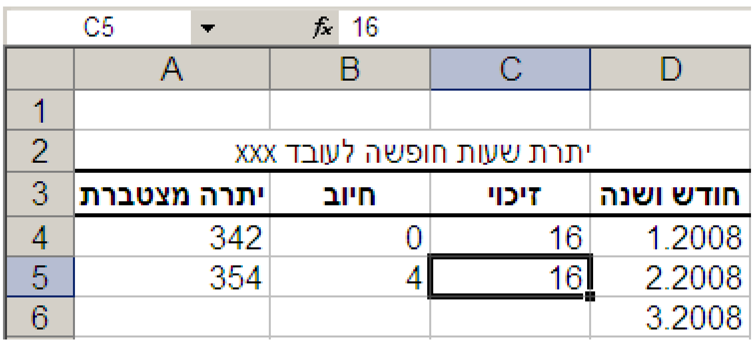
קובץ אקסל נקרא בשם "חוברת עבודה" או BOOK באנגלית. כל חוברת עבודה כזו בנויה מכמה גיליונות (SHEETS). כל גיליון הוא עצמאי אך ניתן לקשר בינו לבין הגיליונות האחרים באותה חוברת העבודה. המעבר בין הגיליונות השונים נעשה באמצעות בחירת הלשונית הנושאת את שם הגיליון.

באיור הבא מבנה כללי של חוברת עבודה הפתוחה בגיליון "תצורה כוללת":



העתקה

להעתקת תוכן של תא [שבו טקסט, מספר או נוסחה] נבחר בתחילה את התא שאותו אנו רוצים להעתיק. בחירת התא נעשית על ידי "הצבעה" על התא בעזרת העכבר ולחיצה [קליק] על המקש השמאלי שלו. התא אותו בחרנו מכונה "התא הפעיל" והוא יודגש [כברירת מחדל] בקו שחור עבה כשבפינתו השמאלית התחתונה [כאשר עובדים עם גיליון שבו העמודה A היא ראשונה מימין – להלן גיליון עברי] או בפינתו הימנית התחתונה [כאשר עובדים עם גיליון שבו העמודה A היא ראשונה משמאל – להלן גיליון עבודה אנגלי] "ידית גרירה".

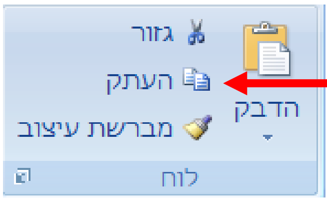


C5 | 16

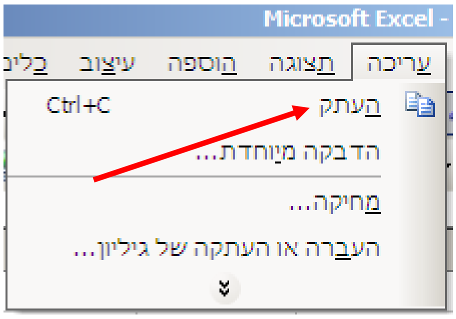
	A	B	C	D
1				
2	יתרת שעות חופשה לעובד xxx			
3	**יתרה מצטברת**	**חיוב**	**זיכוי**	**חודש ושנה**
4	342	0	16	1.2008
5	354	4	16	2.2008
6				3.2008

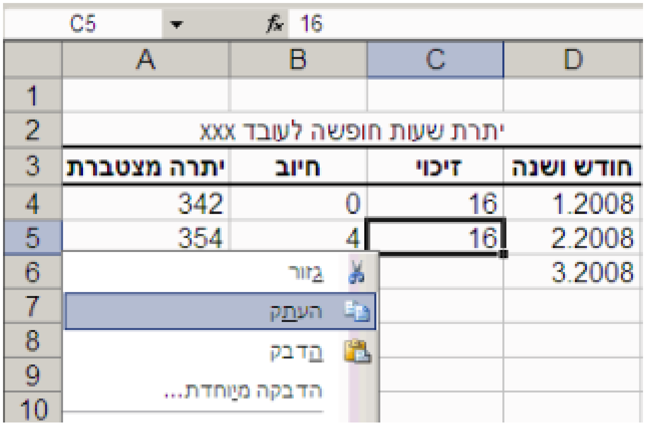
אנחנו נמצאים בתא C5 ומתכוונים להעתיק את תוכנו לתא C6. עומדות בפנינו 3 אפשרויות לעשות זאת:

- ע"י הקשה על האייקון **העתק**



- ע"י הקשה של CTRL+C [המשמעות היא לחיצה על CTRL ביחד עם לחיצה על האות **C**]
- ע"י פתיחת תפריט "עריכה" ואחר כך **העתק**, או ע"י לחיצה על קליק ימני בעכבר אז נפתחת תיבת תפריט ושם לחיצה על **העתק**





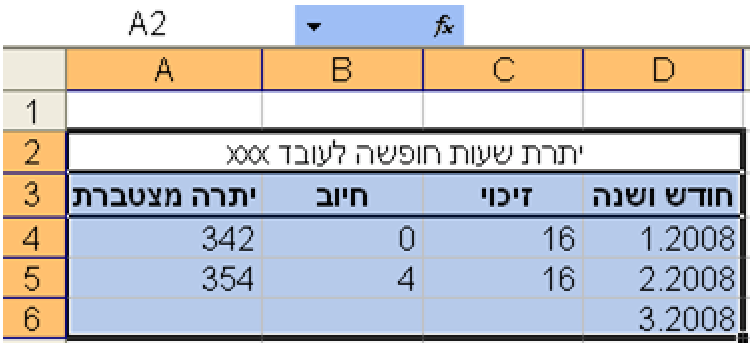
לאחר שביצענו את ההעתקה באחת מתוך 3 האפשרויות לעיל, מגיעים לתא שאליו רוצים להעתיק את הנתונים [C6 או כל תא אחר] הופכים אותו לפעיל ומבצעים **הדבקה** שגם אותה אנחנו יכולים לבצע ב- 3 דרכים בהתאמה ל- 3 אפשרויות **העתקה** שצוינו לעיל – הקשה על CTRL+V, הקשה על האייקון **הדבק** ופתיחת תפריט **עריכה** ואחר כך **הדבק**.

- האפשרות הרביעית מתאימה רק למצב של העתקה לתא סמוך [שמעל, שמתחת, שמימין או שמשמאל] לתא הפעיל, ע"י אחיזה בידית הגרירה [בעזרת העכבר], גרירה לתא הסמוך אליו אנו רוצים להעתיק ואם שחרור הלחיצה בעכבר מתבצעת העתקה.

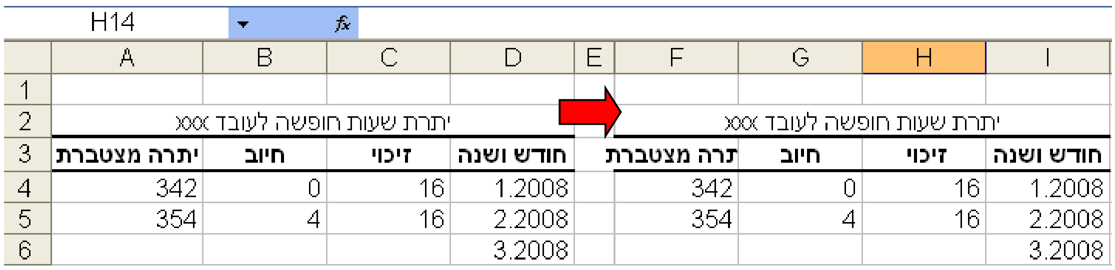
העתקת תחום של נתונים

העתקת תוכן של תחום דומה בעיקרה להעתקת תוכן של תא בודד ובכל זאת יש כמה דברים שאנו צריכים לתת עליהם את הדעת:

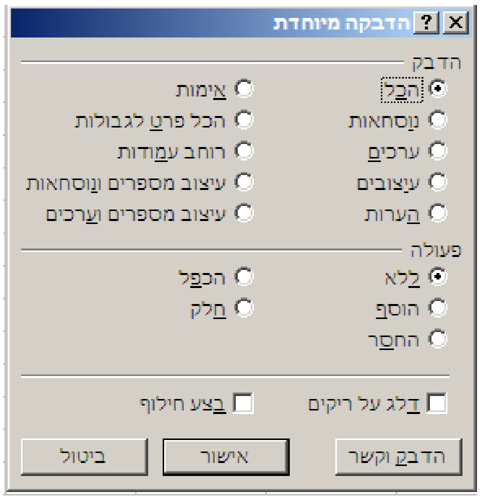
- התחום צריך להיות רציף.
- עלינו לסמן את התחום וע"י הימצאות באחת הפינות שלו כתא פעיל וללא שחרור המקש השמאלי של העכבר לגרור ולסמן את התחום הרצוי. עם סיום הגרירה האזור יהיה צבוע, להוציא את התא הפעיל שיישאר בצבע לבן [בדוגמה הבאה התאים A2:D2 הם בלבן וזאת משום שהתאים מוזגו].



עכשיו עלינו להימצא בתא הפעיל החדש שיהווה את הפינה השמאלית העליונה של האזור בו תבוצע ההעתקה [במקרה שלנו התא F2] ולבצע הדבקה ע"י CTRL+V [למשל].



נשים לב לנקודות הבאות:



- אם נלחץ בתפריט על **עריכה** ולאחר מכן על **הדבקה מיוחדת** נקבל את התפריט שאנו רואים. אם נסמן בו **הכל** [כפי שזה מופיע] פעולת ההעתקה תכלול את כל המרכיבים שבקבוצת **הדבק** [עיצובים כוללים עיצובי תא אך לא רוחב עמודה וגובה שורה ולכן רואים שעמודה F אינה מקבלת את הרוחב של עמודה A].
- בגיליון עבודה "אנגלית" התחום יודבק ימינה ולמטה יחסית לתא הפעיל שנבחר כ"תא הפינה". בגיליון עבודה "עברית" התחום יודבק שמאלה ולמטה יחסית לתא הפעיל שנבחר כ"תא פינה".

העתקה "מקושרת"

בשיטות ההעתקה אותן סקרנו עד כה, לאחר ביצוע ההעתקה – שינויים בנתוני המקור [התאים עליהם הפעלנו את הפקודה **העתק**] לא ישפיעו על הנתונים המועתקים [התאים עליהם הפעלנו את הפקודה **הדבק**]. נוכל ליצור קשר בין נתוני המקור לנתונים המועתקים בדרך הבאה:

	A	B	C	D	E	F	G
1							
2		10%	25%	15%	25%	25%	משקל
3	ציון כולל	שיעורי בית	מבחן 3	בוחן	מבחן 2	מבחן 1	התלמיד
4	64	40	48	90	65	72	ריש לקיש
5	80	60	72	92	80	88	יונה צור
6	71	80	62	84	78	60	רונית שרייבר
7	86	40	82	100	88	98	יותם פרץ
8	89	60	84	96	90	100	לינה מתיאו

מורה המלמד מספר מקצועות בכיתה מנהל את ציוני התלמידים בקובץ אקסל. לכל מקצוע הוא מכין גיליון עבודה ובו הוא מרכז את ציוני התלמידים. בדוגמה מובאים ציוני התלמידים במקצוע מתמטיקה המרוכזים בגיליון עבודה הנושא את השם "מתמטיקה". לקראת סוף הסמסטר אנו רוצים לרכז בטבלה פשוטה את הציון הכולל בכל אחד מהמקצועות אותם מלמד המורה.

C4 =מתמטיקה!A4

	A	B	C	D
1				
2				
3	מכניקה	שרטוט	מתמטיקה	התלמיד
4	54	76	64	ריש לקיש
5	78	84	80	יונה צור
6	82	90	71	רונית שרייבר
7	88	68	86	יותם פרץ
8	84	90	89	לינה מתיאו

ראשית עלינו לפתוח גיליון עבודה חדש הנושא את השם "ריכוז ציונים". בהמשך נוכל להעתיק כל נתון שנרצה מגיליון של כל מקצוע. כאן אנו רואים איך נעשית העתקת ה"ציון כולל" במתמטיקה, אנו נמצאים בתא הפעיל C4 כותבים את הסימן "=", עוברים לגיליון "מתמטיקה" מגיעים לתא A4 [בו נמצא ה"ציון כולל" של התלמיד ריש לקיש] מסמנים אותו ומקישים על הסימן √. מייד לאחר הלחיצה המחשב חוזר לגיליון "ריכוז ציונים" ובתא C4 אנו מקבלים את ה"ציון כולל" של ריש לקיש במתמטיקה. עכשיו אנו יכולים להעתיק את התא C4 לתאים C5:C8. אם ריש לקיש יערער על אחד מציוניו והערעור יתקבל וכתוצאה ישתנה ה"ציון כולל" שלו – הציון החדש יתוקן מייד גם בגיליון "ריכוז ציונים".

DATE =מתמטיקה!A4

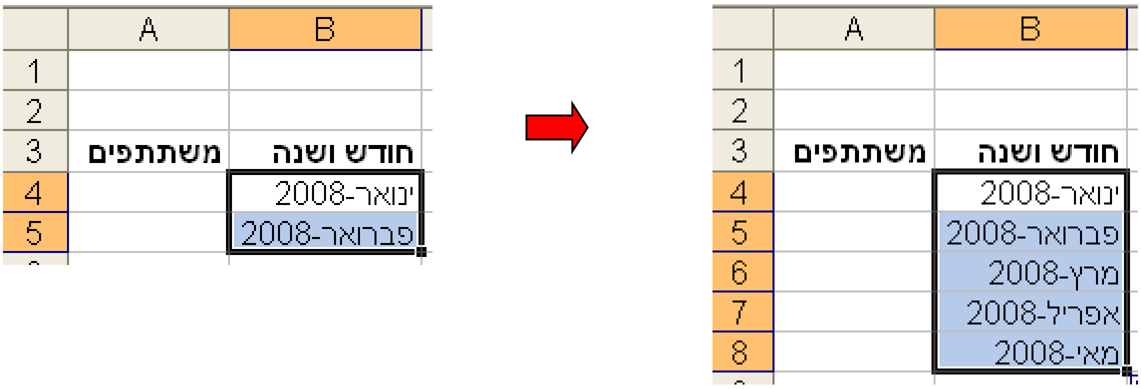
	A	B	C	D
1				
2		10%	25%	15%
3	ציון כולל	שיעורי בית	מבחן 3	בוחן
4	64	40	48	90
5	80	60	72	92
6	71	80	62	84
7	86	40	82	100
8	89	60	84	96

יצירת סדרה חשבונית או סדרה לוגית בעזרת העתקה

אנו יכולים ליצור סדרה כזאת באופן הבא:

נרשום את 2 האיברים הראשונים של הסדרה בתאים רצופים [בדוגמה שלנו 1 ו- 3], נחזור לתא B2 ונלכוד אותו ביחד עם תא B3 בעזרת העכבר ונשחרר את העכבר. עכשיו תופסים את ידית הגרירה ומושכים כלפי מטה [ככול שאנו צריכים] ומקבלים סדרת מספרים המקיימת תנאים של סדרה חשבונית ששני האיברים הראשונים שלה הם כפי שרשמנו.

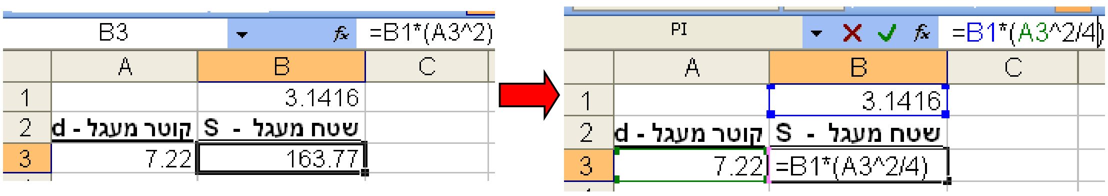
באופן דומה אנו יכולים לבנות סדרה לוגית כבדוגמה הבאה:



שינוי תוכן של תא

כדי לשנות תוכן של תא עלינו להימצא בתא [כך שיהיה תא פעיל] ואז או לבצע את השינוי באזור הסרגל בתא הנקרא "תוכן התא עליו עובדים" [נקרא גם "שורת הנוסחאות"], או בתוך התא עצמו לאחר שלחצנו על מקש הפקודות F2 [FUNCTION KEY] או על ידי הקשה כפולה [ברצף] של המקש השמאלי בעכבר. לאחר ביצוע השינוי יש להקיש ENTER או TAB או ללחוץ על הסימן √ כדי שהשינוי ייקלט.

בדוגמה שלפנינו תרגיל לחישוב שטח מעגל [S] כשנתון קוטרו [d]. הנוסחה שהוקלדה [משמאל] מתאימה לחישוב שטח מעגל לפי רדיוסו $S = \pi \cdot r^2$ ובהינתן d יש להשתמש בנוסחה $S = \pi \cdot d^2 / 4$. התיקון נעשה בתוך שורת הנוסחאות.



סימני פעולה

נוסחאות אריתמטיות עושות שימוש בארבע סימני פעולה [אופרטורים OPERATORS] בסיסיים: חיבור [+], חיסור [-], כפל [x] וחילוק [÷]. ביצירת [כתיבת] נוסחה נשתמש בעוד שלושה אופרטורים: העלאה בחזקה, סימון לשליליות ואחוזים. בכתיבת נוסחאות באקסל לסימני פעולה אלו סימונים ומשמעות כבטבלה:

סימן הפעולה	המשמעות
+	חיבור
-	חיסור, הפחתה*
*	כפל
/	חילוק, חלוקה
^	העלאה בחזקה
%	אחוזים

* כאשר הסימן בא לצידו הימני של מספר מבטא הדבר שהמספר שלילי

בנוסף לסימני הפעולה, בעיבוד נתונים, אנו עושים שימוש בסימני פעולה לוגיים. באקסל לסימני פעולה אלו סימונים ומשמעות כבטבלה:

סימן הפעולה	המשמעות
=	**שווה ל-**
>	**גדול מ-**
>=	**גדול או שווה ל-**
<	**קטן מ-**
<=	**קטן מ- או שווה ל-**
<>	**שונה מ-**

סדר הפעולות

סדר הפעולות כפי שאנו מכירים אותו מלימוד החשבון הוא הסדר שלפיו עובד המחשב בעת ביצוע חישובים בתוכנת האקסל:

- הפעולות בתוך סוגריים [ובתוך הסוגריים שוב לפי אותו סדר המוזכר כאן]
- הפעולות המעריכיות [חזקות]
- פעולות הכפל והחילוק

• כל פעולות החיבור והחיסור

<u>הערות</u>:

אנחנו יכולים גם להשתמש בסוגריים בתוך סוגריים, ואז מה שבתוך הסוגריים הפנימיים יחושב קודם. במקרה שלא מוסיפים סוגריים בצורה נכונה, כלומר אין מס' זהה של סוגריים שמאליים ")" [פותחים] וסוגריים ימניים "(" [סוגרים] או כאשר יש מספר זהה של סוגריים כאלו אבל הם אינם במקום מבחינה לוגית – המחשב יציג הודעת שגיאה. כמו כן, הוא עשוי לנסות להוסיף סוגר בעצמו, לכן רצוי לבדוק היטב את הסוגריים כדי שהחישוב ייצא בדיוק כפי שהתכוונו. דבר נוסף שעלינו לזכור הוא שלסוגריים מרובעים או אחרים – מאלו המוזכרים [")" ו- "("] אין משמעות באקסל.

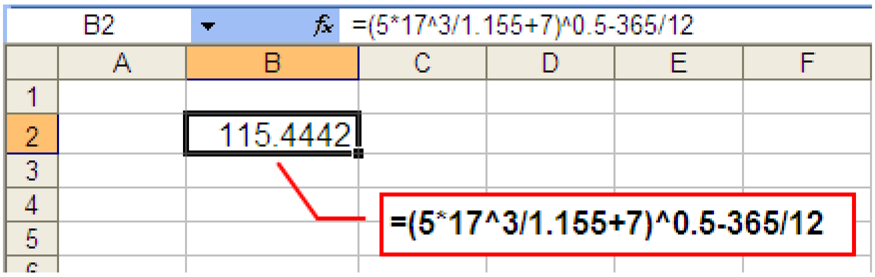
להוצאת שורש נשתמש בזהות הידועה $\sqrt[Y]{X} = X^{1/y}$, כלומר במקום להוציא שורש Y של מספר נעלה את המספר בחזקת הערך ההופכי [רציפרוק] של Y.

ביצוע חישוב

כדי לבצע פעולה חשבונית פשוטה, או מורכבת, בתוך תא פעיל כלשהו יש לרשום את סימן השווה בהתחלת הנוסחה החשבונית ולאחריו את הנוסחה. ניקח לדוגמה את התרגיל הבא:

$$Y = \sqrt{5 \cdot \frac{17^3}{1.155} + 7} - \frac{365}{12}$$

אם נרצה למצוא את הערך של Y בעזרת אקסל, נרשום בתא פעיל [במקרה שלנו B2] את הנוסחה כבאיור ולאחר הקשת ENTER נקבל בתא את הערך 115.4442 שהוא הערך של Y.

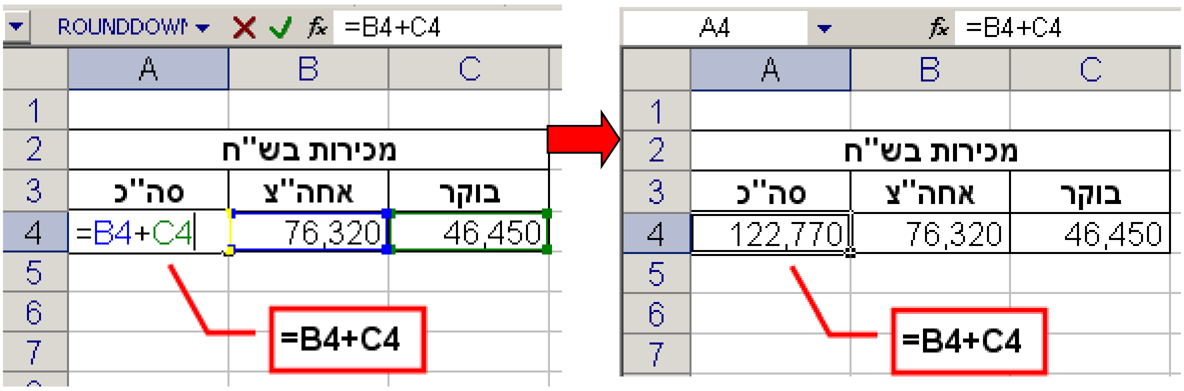


כתיבת והעתקת נוסחה

לביצוע פעולה חשבונית באקסל, כפי שתואר לעיל, יש יתרון בכך שאנו רואים את הנוסחה ויכולים לשנות אותה על ידי עריכתה כדי להתאימה לתנאים חדשים. למשל החלפת הערך 1.155 [המייצג מקדם הקשור לשיעור המע"מ בערכו הנוכחי] בערך 1.14 [המייצג את שיעור המע"מ שאולי יוחל בעת הקרובה].

אולם, היתרון הגדול של האקסל הוא ביכולת לבצע פעולות חשבוניות, מיון, סינון של נתונים שהוקלדו לקובץ או יובאו אליו.

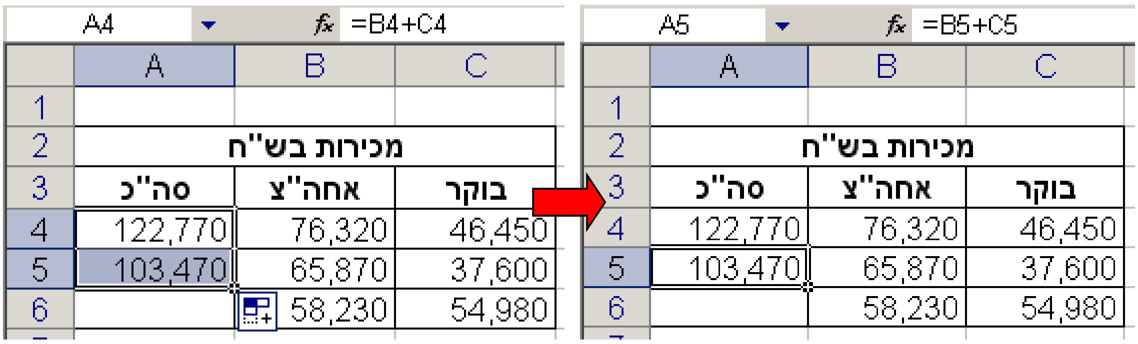
בחברת מכירות, הפועלת באמצעות מוקדנות טלפונית, עובדים בשתי משמרות. בטבלה הבאה מוצגות המכירות, של יום מסוים, בתאים B4 ו- C4. כדי לדעת את סה"כ המכירות באותו יום נכתוב בתא A4 את הנוסחה =B4+C4 ולאחר מכן נקיש ENTER. תתקבל התשובה 122,770.



בטבלה הבאה נתוני מכירות של יומיים נוספים, נרצה לחשב גם עבורם את סה"כ המכירות. יש לנו 3 אפשרויות:

- להימצא בתא A5 ולהקיש את הנוסח =B5+C5 [וזו האפשרות הפחות טובה]
- להימצא בתא A4 ולהפעיל פעולת **העתק** [CTRL+C], אח"כ לעבור לתא A5 ולבצע פעול **הדבק** [CTRL+V]
- להימצא בתא A5, לאחז, באמצעות העכבר, את ידית הגרירה ולגרור אותה לתא A5 ולשחרר

בשני המקרים האחרונים תופיע בתא A5 הנוסחה =B5+C5 [כפי שרואים באיור השמאלי].

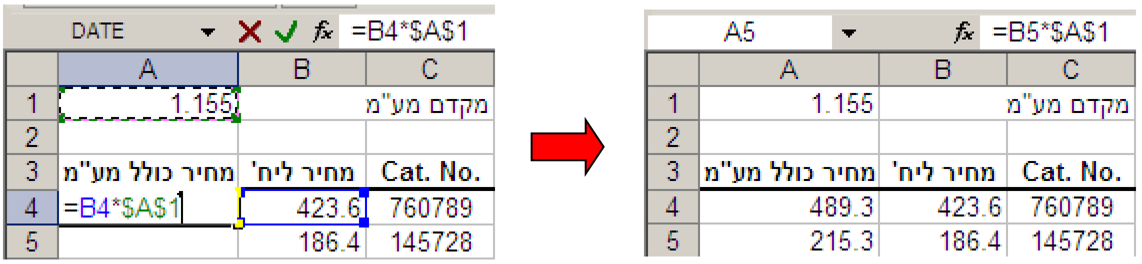


הפניה [כתובת] יחסית [RELATIVE]

בדוגמה האחרונה עשינו שימוש בנוסחה המבוססת על המיקום היחסי של התאים היוצרים את הנוסחה [תאים B4 ו- C4]. בעת העתקת הנוסחה, מתא A4 לתא A5 ההפניה [לתאים B5 ו- C5] מותאמת באופן אוטומטי. להפניה שכזאת קוראים הפניה יחסית.

הפניה [כתובת] מוחלטת [ABSOLUTE]

ישנם מקרים בהם נרצה שבנוסחה תופיע כתובת מדויקת של תא כך שאם נעתיק את הנוסחה מתא אחד לתא אחר הנוסחה תמשיך לקחת את הערך שנמצא בתא הכתובת המדויקת, ללא קשר למיקום התא המכיל את הנוסחה. נראה זאת בדוגמה הבאה:

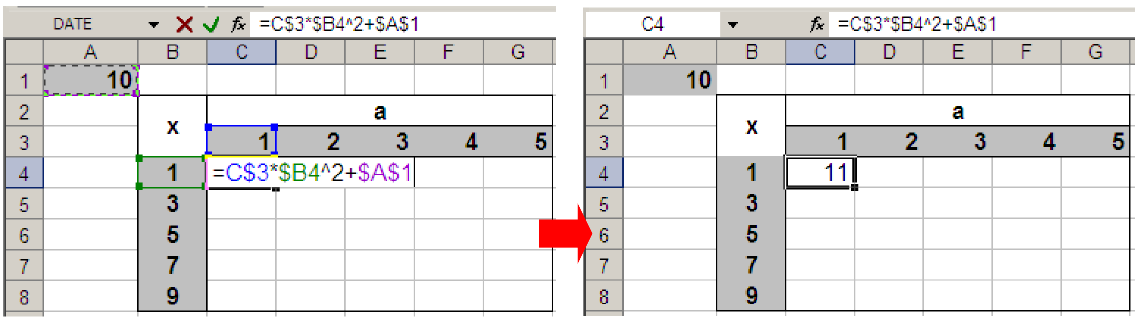


כאשר כתבנו את הנוסחה בתא A4 [איור משמאל] לאחר הקלדת [או לאחר הצבעה] על תא A1 אנו מקישים על מקש הפקודות [FUNCTIONS KEY] F4, הדבר גורם להוספת הסימן $ מימין לאות המציינת את העמודה וסימן $ מימין לספרה המציינת את השורה [כמובן שאפשר להוסיף את סימני ה- $ ע"י הקשתם בתהליך של **עריכה**], מרגע זה "קוּבַּע" התא A1 בנוסחה ובכל מקום שנעתיק אותה נקבל הפניה לתא A1. אנו רואים זאת באיור שמימין, לאחר שהעתקנו את תא A4 לתא A5 אנו רואים שהנוסחה בתא A5 כוללת את תא B5 [המוגדר כהפניה יחסית] ואת תא A1 [המוגדר כהפניה מוחלטת].

הפניה [כתובת] מעורבת

היא שילוב [בנוסחה] שמכיל הפניה מוחלטת לעמודה והפניה יחסית לשורה [=B3*$A1] או להיפך – הפניה מוחלטת לשורה והפניה יחסית לעמודה [=B3*A$1].

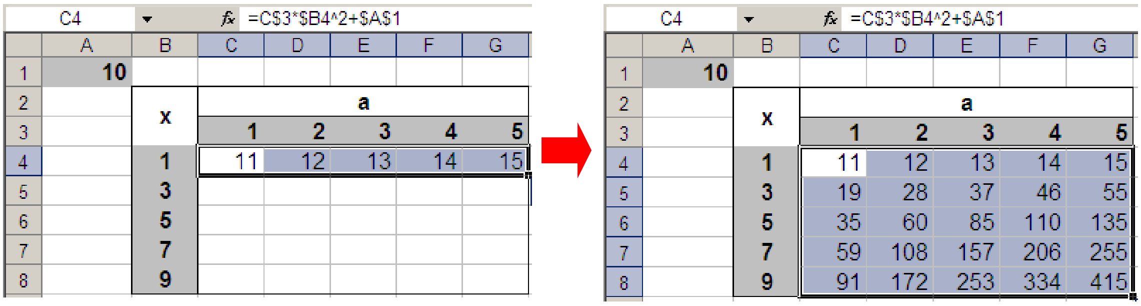
נדגים זאת בדוגמה הבאה: נתונה הנוסחה $Y = a \cdot x^2 + 10$, אנו רוצים לבנות טבלה באקסל אשר תיתן לנו את Y עבור ערכים שונים של a ו- x.



נבנה טבלה שבה בתאים C3:G3 נקליד את ערכי a הרצויים, בתאים B4:B8 נקליד את ערכי x הרצויים ובתא A1 נקליד את המספר 10 [האבר החופשי בנוסחה]. בתא C4 נכתוב את הנוסחה המתאימה, אנו רואים כי ערכי a יהיו עם הפניה מוחלטת לשורה 3 אך אם הפניה יחסית לעמודות, ערכי x יהיו עם הפניה מוחלטת לעמודה B אך עם הפניה יחסית לשורות. האבר החופשי יהיה עם הפניה מוחלטת לתא A1. את סימן ה- $ במקום הנכון

נקבל על ידי הקלקה חוזרת של מקש הפקודות F4, ההקלקה הראשונה מביאה להפניה מוחלטת [$A$1], הקלקה נוספת תביא להפניה מעורבת – עמודה יחסית ושורה מוחלטת [A$1], הקלקה נוספת תביא להפניה מעורבת – עמודה מוחלטת ושורה יחסית [$A1] והקלקה נוספת תביא להפניה יחסית הן בעמודה והן בשורה [A1].

נקיש עכשיו על הסימן √ ונקבל את התוצאה המתאימה בתא C4 [האיור הימני בעמוד קודם], עכשיו נגרור בעזרת ידית הגרירה את התא עד לתא G4 [כמו באיור שמשמאל] ואז נאחז שוב בידית הגרירה ונגרור את כל השורה עד לשורה 8 ונקבל את הטבלה כולה [האיור מימין].



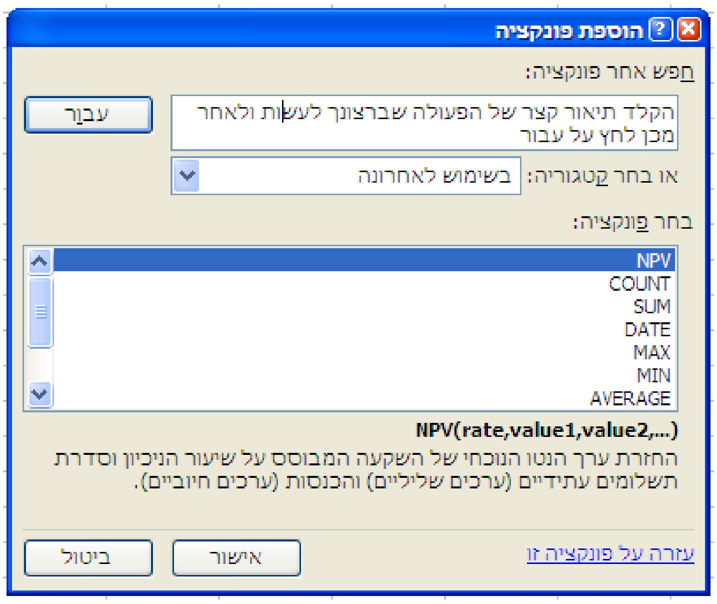
פרק 3 – הכרת אקסל, פונקציות מובנות

אקסל כולל מאות פונקציות מובנות. שימוש בפונקציה מובנית מהווה מעין "קיצור דרך" לביצוע פעולות שניתן לבצען באופן מפורט ומורכב יותר באמצעות אקסל.

בחירת הפונקציה מתוך ספרים הפונקציות נעשית בשתי דרכים:

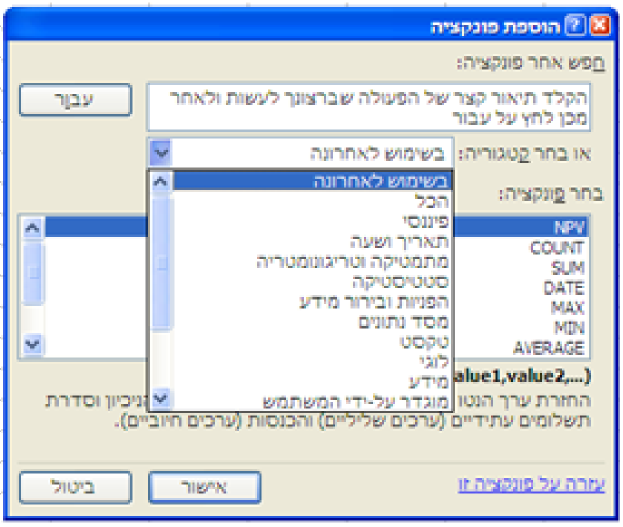
הקשה, בתפריט הראשי, על "נוסחאות" ובתת התפריט הנפתח הקשה נוספת על ***fx*** "הוספת פונקציה" או הקשה על ***fx*** שבשורה התחתונה של התפריט:

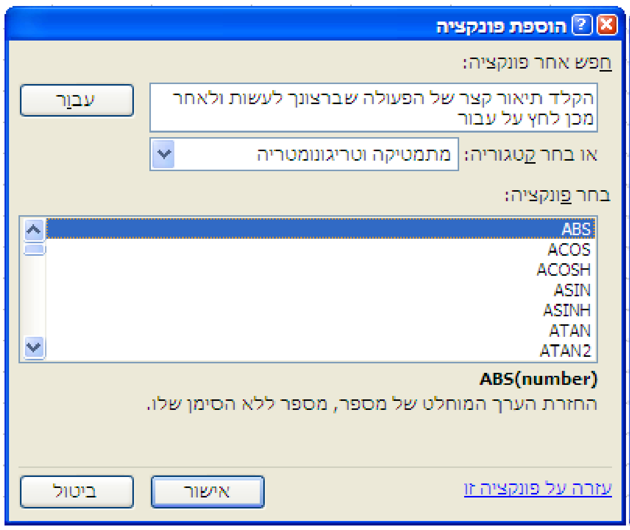
בשני המקרים נפתחת תיבת הדו שיח הבאה:



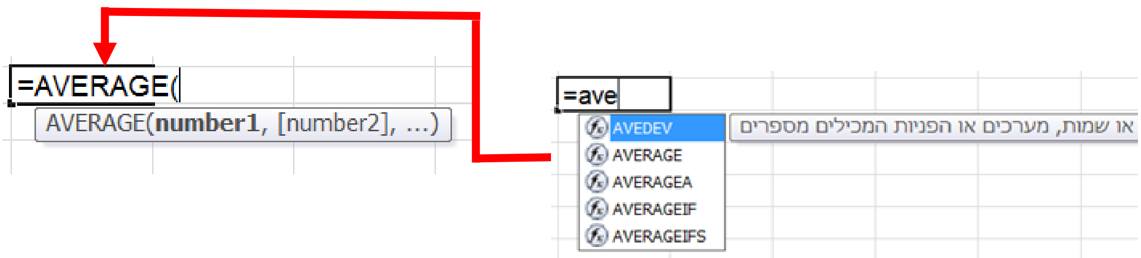
לבחירת הפונקציה ישנן 3 אפשרויות:

- הקלדת שם הפונקציה [לפחות 3 אותיות ראשונות שלה] בתיבה **הקלד תיאור קצר של הפעולה שברצונך לעשות** ולאחר מכן לחץ על עבור
- בחירת פונקציה מתוך אלו שהיו **בשימוש לאחרונה**, פונקציות אלו מוצגות בתיבה של **בחר פונקציה**
- פתיחה ע"י הקשה על החץ הנמצא משמאל ל- **או בחר קטגוריה**, נפתחת תיבה [הימנית בציור הבא] ובה נושאים שונים, נבחר נושא, למשל – מתמטיקה וטריגונומטריה ותיפתח תיבה נוספת [השמאלית בציור הבא] שבה נוכל לבחור את הפונקציה הרצויה לנו. כאשר אנו נמצאים על הפונקציה היא צבועה ותיאורה [מה היא עושה] מופיע בחלק התחתון של התיבה.





למעשה קיימת אפשרות נוספת לשימוש בפונקציות המובנות והיא על ידי הקלדתן באופן ישיר וזאת כאשר אנו מכירים כבר את הפונקציות מתוך שימוש קודם, רושמים בתא הפעיל את הסימן "=" ולאחריו 3 אותיות ראשונות של הפונקציה [לפעמים מספיקות 2] מתקבלות כל הפונקציות המובנות המתחילות באותן אותיות, מקישים הקשה כפולה על הפונקציה שאותה רוצים ומתקבל תפריט שאותו יש צורך למלא [ראה איור להלן] את הערכים עליהם יתבצע החישוב אפשר להקליד אולם בדרך כלל מסמנים אותם מתוך רשימה מוכנה.



נראה מספר דוגמאות:

המרת מעלות לרדיאנים 6.2831853 **=RADIANS(360)**	**חישוב e^{-3}** 0.0498 **=EXP(-3)**	**חישוב שטח מעגל שקוטרו 12** 113.10 **=PI()*12^2/4**
חישוב עצרת - 5! 120 **=FACT(5)**	**חישוב מספר ימי עבודה אפשריים בטווח תאריכים נתון** 256 **=NETWORKDAYS(M9,M10,M13:M18)**	
מציאת סינוס של זווית במעלות 0.5 **=SIN(RADIANS(30))**	**חישוב ערך נוכחי של השקעה** -105.7 **=NPV(0.15,-500,100,100,100,100,100,100)**	

פרק 4 – פונקציות סטטיסטיות מובנות

הפונקציות הסטטיסטיות המובנות מהוות חלק מהפונקציות הסטטיסטיות המובנות שתוארו לעיל. ככלל, פונקציה סטטיסטית יכולה לפעול על תאים המכילים ערכים מספריים בלבד [חלק מאוד קטן מהפונקציות פועל על תאים עם טקסט]. הערכים צריכים להיות ערכים "גולמיים" [הפונקציה אינה יודעת להתמודד עם "ערכים מקובצים"]. ברוב המקרים, הערכים יכולים להיות בשורה, בעמודה או בטווח של שורות ועמודות, אפשר שיהיו בטווח תאים ריקים, בחלק קטן מהמקרים [כמו למשל בחישובי רגרסיה הנתונים חייבים להיות בעמודות מתואמות [קרי, תמיד זוגות של ערכים] וללא תאים ריקים [חשוב לזכור שהמספר "0" הוא מספר לכל דבר ויובא בחשבון בחישוב, תא ריק הוא תא שבו לא רשום דבר או שבו רשום טקסט]. להלן ריכוז של חלק מהפונקציות היותר שכיחות שבשימוש, בהמשך יובאו דוגמאות:

הפונקציה	כתיבת הפונקציה	הסבר
AVERAGE	=AVERAGE(data range)	אחזור ממוצע חשבוני[1]
BINOMDIST	=BINOM.DIST(number, trial, probability, false)	אחזור של הסתברות התפלגות בינומית של איבר בודד
BINOMDIST	=BINOM.DIST(number, trial, probability, true)	אחזור של הסתברות התפלגות בינומית מצטברת
CONFIDENCE.N	=CONFIDENCE.NORM(alfa, std, size)	אחזור של רווח בר סמך לממוצע אוכלוסייה לפי התפלגות נורמלית
CONFIDENCE.T	=CONFIDENCE.NORM(alfa, std[2], size)	אחזור של רווח בר סמך לממוצע אוכלוסייה לפי התפלגות סטודנט
CORREL	=CORREL(data ranage1, data ranage2)	אחזור מקדם המתאם בין שתי סדרות נתונים תואמות
COUNT	=COUNT(data range)	מנייה/ספירה של מספר המספרים ברשימת ארגומנטים
COVAR	=COVARINCE.P(data ranage1, data ranage2)	אחזור השונות המשותפת של אוכלוסייה עבור שתי ערכות של נתונים מתואמים
COVAR	=COVARINCE.S(data ranage1, data ranage2)	אחזור השונות המשותפת של מדגם עבור שתי ערכות של נתונים מתואמים
FREQUENCY	=FREQUENCY(data range, bin range)	חישוב שכיחות המופע של ערכים בטווח ערכים מוגדר[3]
INTERCEPT	=INTERCEPT(data ranage1, data ranage2)	אחזור של ה"חותך" במשוואת רגרסיה ליניארית
MAX	=MAX(data ranage1, data ranage2)	אחזור של הערך הגדול ביותר של מערך נתונים

[1] אחזור=חישוב בחלק גדול מהפונקציות. האחזור מתבצע על ארגומנטים במערך נתונים המתווסף לפונקציה על ידי צביעה או על ידי הקלדה בתוך הפונקציה

[2] Standard Deviation

[3] הפונקציה היא סטטיסטית אולם היא מופעלת כ"נוסחת מערך" באופן שיוסבר בהמשך

הפונקציה	כתיבת הפונקציה	הסבר
MEDIAN	=MEDIAN(data ranage1, data ranage2)	אחזור החציון של מערך נתונים
MIN	=MIN(data ranage1, data ranage2)	אחזור של הערך הקטן ביותר של מערך נתונים
MODE	=MODE.SNGL(data range)	אחזור הערך השכיח של מערך נתונים
MODE	=MODE.MULT(data range)	אחזור הערכים השכיחים של מערך נתונים
NORMDIST	=NORM.DIST(x, mean, std, true)	אחזור השטח המצטבר שמתחת להתפלגות נורמלית מ- ∞ - עד x
NORMINV	=NORM.S.INV(probability)	אחזור Z לפי השטח המצטבר שמתחת להתפלגות נורמלית
NORMSDIST	=NORM.S.DIST(Z, true)	אחזור השטח המצטבר שמתחת להתפלגות נורמלית מ- ∞ - ועד Z נתון
PERCENTILE	=PERCENTILE.EXC(data range, k_th percentile)	אחזור x שהוא האחוזון/מאון ה-k-י של מערך נתונים[4]
POISSON	POISSON.DIST(x, mean ,false)	אחזור של הסתברות התפלגות פואסונית של איבר בודד
POISSON	POISSON.DIST(x, mean, true)	אחזור של הסתברות התפלגות פואסונית מצטברת
SLOPE	=SLOPE(data ranage1, data ranage2)	אחזור ה"השיפוע" במשוואת רגרסיה ליניארית
STANDARDIZE Returns a normalized value	=STANDARDIZE(x, mean, std)	אחזור ערך Z של x נתון בהתפלגות נורמלית
STDEVS	=STDEV.S(data range)	אחזור אומד לסטיית התקן של אוכלוסייה (בהתבסס על מדגם)
STDEVP	=STDEV.P(data range)	אחזור סטיית התקן של אוכלוסייה (בהתבסס על נתוני האוכלוסייה)
TINV	=T.INV(probability, df)	אחזור t לפי α (עבור זנב אחד)
TINV	=T.INV.2T(probability, df)	אחזור t לפי α (עבור 2 זנבות)
TTEST	=T.TEST(data range1,data range2,tails,type)	אחזור ההסתברות ששני מדגמים הגיעו משתי אוכלוסיות בסיס זהות בעלות אותו ממוצע
ZTEST	Z.TEST(data range, x, σ)	אחזור של ההסתברות תחת התפלגות נורמלית לערך נמוך מ- x [5]

[4] הרבעון הראשון והרבעון השלישי הם מקרים פרטיים של המאון ה- 25 והמאון ה- 75 בהתאמה
[5] הפונקציה זהה לפונקציה NORMDIST עם האופציה TRUE

פרק 5 – סטטיסטיקה תיאורית

הסטטיסטיקה התיאורית עוסקת בחישוב/הצגה/ארגון/סיכום/הצגה של מדדים הנוגעים לאוכלוסייה נבדקת לצורך תיאור תמציתי וקל לתפישה של נתוני האוכלוסייה הן לצורך קבלת הערכות או תמונה של אותה אוכלוסייה עצמה והן לצורך השוואת אותה אוכלוסייה עם אוכלוסיות אחרות דומות. בין המדדים נציין מדדים לנטייה מרכזית כמו ממוצע, שכיח וחציון; מדדים לפיזור כמו שונות, סטיית תקן, תחום [טווח]; מדדים למיקום יחסי כמו אחוזון [מאון], ציון תקן ועוד.

תרגיל מספר 1

חוקר פיתח מבחן המורכב מ- 40 שאלות סגורות [בשפת יום יום – שאלות אמריקאיות] למדידה ולתיקוף ההשכלה הכללית [מאורעות היסטוריים, מדע, אומנות ספרות וכו'], של אוכלוסייה. המבחן הועבר ל- 5 קבוצות [A, B, C, D ו- E] של מבוגרים, שנבחרו באקראי. בכל קבוצה 10 אנשים. תשובה נכונה מזכה בנקודה ולכן הציון המקסימלי לנבחן הוא 40. תוצאות המבחן מובאות בטבלה הבאה:

ציוני המבחן				
E	D	C	B	A
26	35	15	20	34
21	23	21	24	30
19	33	23	33	20
19	18	27	26	27
26	16	21	35	15
24	23	25	16	34
19	22	34	25	34
25	15	16	35	33
34	34	19	19	17
18	23	18	15	24

קבע, בעזרת פונקציות סטטיסטיות מובנות את המדדים/הנתונים הבאים: מספר הנבחנים, ממוצע, חציון, שכיח, אומד לסטיית התקן, סטיית התקן, ערך מינימלי וערך מקסימלי.

פתרון

חלק מהמדדים הנתונים הנדרשים נקבעים בדרך כלל ע"י ספירה [מספר הנבחנים], חלק ע"י סקירת הנתונים [למשל ערך מינימלי וערך מקסימלי] ואחרים על ידי חישוב [למשל ממוצע וסטיית התקן]. בעזרת הפונקציות הסטטיסטיות המובנות שבאקסל, ובהינתן שנתוני האוכלוסייה מאורגנים בקובץ אקסל, בעזרת כתיבת הפונקציה המתאימה מתקבל המדד/הנתון בהינף הקשה כפי שנראה בטבלה הבאה:

מספר הארגומנטים	50	=	=COUNT(B6:F15)
ממוצע	24.16	=	=AVERAGE(B6:F15)
שכיח	34	=	=MODE.SNGL(B6:F15)
חציון	23	=	=MEDIAN(B6:F15)
אומד לסטית התקן	6.61	=	=STD.S(B6:F15)
סטית התקן	6.54	=	=STD.P(B6:F15)
ערך מינ' בטווח	15	=	=MIN(B6:F15)
ערך מקס' בטווח	35	=	=MAX(B6:F15)

הפונקציה MODE

הפונקציה **MODE.SNGL** קובעת [מאחזרת] את ערך השכיח. לפונקציה זו יש בעיה מובנית ב- EXCEL [הידועה מאז ומתמיד] אותה נתאר כאן. לסדרת נתונים המובאת באיור הבא 2 שכיחים: 18 ו- 19. שימוש בפונקציה זו [במקרה שבו יש 2 שכיחים או יותר] מביא את השכיח בו נתקלת הפונקציה לראשונה בסקירת הנתונים. לשאלה מהי המשמעות של "להיתקל לראשונה?" לא נשיב כרגע [התשובה קלה לנתונים המאורגנים בעמודה אחת או שורה אחת אך לא לנתונים המאורגנים בטווח המכיל כמה עמודות וכמה שורות]. מכאן המסקנה היא פשוטה – אין לסמוך על חישוב השכיח באקסל עם הפונקציה הזאת ורצוי שלא להשתמש בה.

B13 | =MODE.SNGL(B2:B11)

	A	B	C	D	E	F	G
1							
2		16					
3		19					
4		18					
5		18					
6		18					
7		18					
8		19					
9		19					
10		16					
11		19					
12							
13		19					

במקום בפונקציה זו נשתמש בפונקציה **MODE.MULT** המאחזרת את כל השכיחים של נתונים במערך נתון. הפונקציה מחייבת שימוש ב"נוסחת מערך", לא נרחיב בנושא "נוסחת מערך" ונציג רק את טכניקת העבודה איתה.

תרגיל מספר 2

בהמשך לתרגיל מספר 1 נבדקה קבוצה נוספת של 50 אנשים שעברו את המבחן. תוצאות המבחן מובאות בטבלה שבהמשך. מהו השכיח/השכיחים של תוצאות המבחן?

פתרון

מערך הנתונים וחישובי השכיח/השכיחים מובאים בטבלה הבאה:

H2 {=MODE.MULT(B2:F11)}

	A	B	C	D	E	F	G	H
1		ציוני המבחן						השכיחים
2		15	18	22	25	33		15
3		15	19	23	25	34		19
4		15	19	23	26	34		23
5		15	19	23	26	34		34
6		16	19	23	26	34		24
7		16	20	24	27	37		#N/A
8		16	20	24	27	35		#N/A
9		17	20	24	30	35		#N/A
10		18	21	24	33	35		#N/A
11		18	21	25	33	36		#N/A
12		15						
13								
14								

=MODE.SNGL(B2:F11)

רואים שלציוני המבחן 5 שכיחים [אבל אם היינו משתמשים בפונקציית ה- MODE.SNGL היינו מקבלים שכיח אחד בלבד [שהוא הערך הראשון בטווח הנתונים שבו נתקלה הפונקציה. אגב, עכשיו, לאור נתוני השכיחים, אפשר לאדרג את הקורא ולשאול איך סוקרת הפונקציה את הנתונים?].

בכתיבת נוסחת המערך השלבים הם כאלו: יש לסמן/לצבוע את התאים בהם אנו מבקשים לקבל את התוצאות [במקרה שלנו H2:H11 ואז לרשום את הפונקציה [צריך להעריך כמה שכיחים יהיו, תמיד כדאי להוסיף ולהרחיב את מספר התאים המסומנים, כאשר תא יהיה לא ישים יופיע בתא N/A#]. עכשיו כאשר הנוסחה רשומה יש ללחוץ בו זמנית על המקשים CTRL+SHIFT+ENTER ומתקבלת התוצאה.

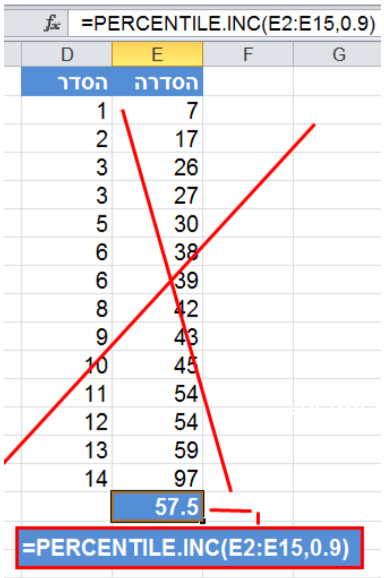
הפונקציה PERCENTILE

הפונקציה **PERCENTILE** קובעת/מחשבת [מאחזרת] את ערך המאון/האחוזון ה- K-י של סדרת נתונים. גם עם פונקציה זו [לפחות באחת מהפונקציות] יש בעיה הנובעת מאופן חישוב שונה ב- EXCEL [שונה מזה המתבצע בתוכנות סטטיסטיות אחרות ומזה הנלמד, אם כי במקרים מסוימים יש כן זהות בתשובות]. נתאר כאן בעזרת 2 דוגמאות את אופן השימוש:

תרגיל מספר 3

נתונה סדרת מספרים כבטבלה הבאה, נדרש לחשב את המאון ה- 90.

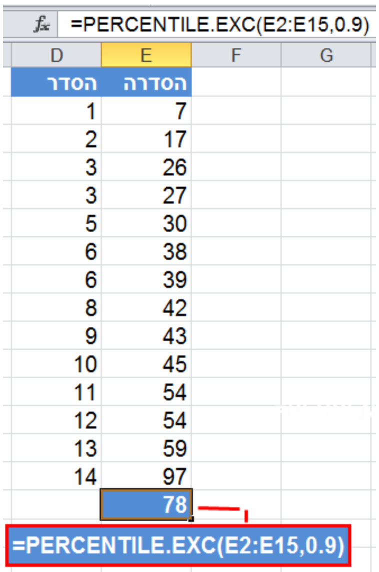
פתרון



=PERCENTILE.INC(E2:E15,0.9)

הסדר	הסדרה
1	7
2	17
3	26
3	27
5	30
6	38
6	39
8	42
9	43
10	45
11	54
12	54
13	59
14	97
	57.5

=PERCENTILE.INC(E2:E15,0.9)



=PERCENTILE.EXC(E2:E15,0.9)

הסדר	הסדרה
1	7
2	17
3	26
3	27
5	30
6	38
6	39
8	42
9	43
10	45
11	54
12	54
13	59
14	97
	78

=PERCENTILE.EXC(E2:E15,0.9)

נדגים את החישוב שהביא לתוצאה 78 [משמאל] שהוא דומה לחישובים מקובלים:

המיקום של המאון ה- 90 – $(14+1)\cdot\frac{90}{100}=13.5$ הערך המתאים למיקום/הסדר 13.5 נמצא בין הערך 59 לבין הערך 97 והוא מחושב כך $59+(97-59)\cdot(13.5-13)=78$ כלומר נעשית כאן אקסטרפולציה ליניארית כפי שמקובל עלינו לעשות בחישוב מאון לפי נתונים מקובצים.

החישוב שהביא [בצד ימין] לתוצאה של 57.5 הוא לא ברור לחלוטין ואין להשתמש בו.

הפונקציה FREQUENCY

בהמשך לתרגיל מספר 1, נרצה עכשיו לא רק להציג את הערכים ששכיחותם היא הגבוהה ביותר אלא את תדירות/שכיחות הופעתם של ערכים בתוך טווח מוגדר. זהו למעשה התחלת תהליך של **קיבוץ נתונים**. אפשר לבצע זאת [עם הפונקציות המובנות] בשתי דרכים. נתאר כאן את שתיהן:

I2 | {=FREQUENCY(B2:F11,H2:H6)}

	B	C	D	E	F	G	H	I	J
1	ציוני המבחן						BINS - תאים	שכיחות	
2	26	35	15	20	34		18	11	
3	18	23	18	15	24		23	15	
4	19	33	23	33	20		28	11	
5	19	18	27	26	27		33	4	
6	19	22	34	25	34		38	9	
7	21	23	21	24	30				
8	24	23	25	16	34				
9	25	15	16	35	33				
10	26	16	21	35	15				
11	34	34	19	19	17				
12	15		35						
13									
14	=MIN(B2:F11)		=MAX(B2:F11)		=FREQUENCY(B2:F11,H2:H6)				

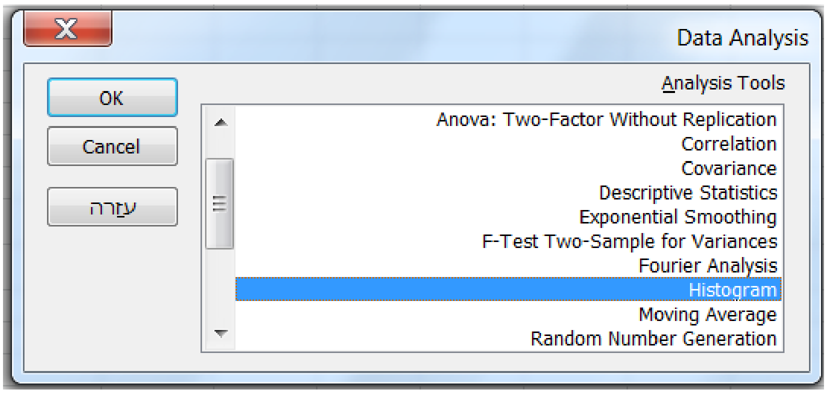
- שלב ראשון – לקבוע את הנתון שערכו הקטן ביותר והנתון שערכו הגדול ביותר.
- שלב שני – לקבוע את מספר התאים [BINS] וערך כל אחד מהם, אנו יודעים כי בטווח יש 50 נתונים, צריך לקבוע [אינטואיטיבית או לפי כללים שניתן למצוא בספרות] את מספר התאים.
- שלב שלישי – לרשום את ערכי התאים [בדוגמה – בטווח H2:H6]. ערך כל תא, בהתייחס לקיבוץ בתאים, הוא הגבול העליון המדומה של כל תא. נקבע את הגבול של התא הראשון H1 [18 באיור שלעיל] ואת הגבול של התא האחרון H6 [38 באיור שלעיל] הכלל הוא: L1≥Min [אם נבחר L1<MIM שכיחות התא תהיה 0] ו- L2≥Max [אם נבחר L2<Max המערכת תיצור תא נוסף בשם MORE ובו ייצברו כל הנתונים שערכם גדול מאותו L2].
- שלב רביעי – להימצא בתא I2 ולסמן את טווח התאים I2:I6 ולהקליד את הנוסחה =FREQUENCY(B2:F1,H1:H6). כיוון שהנוסחה היא נוסחת **מערך**, בסיום ההקלדה יש ללחוץ בוזמנית על המקשים CONTROL+SHIFT+ENTER ומתקבלת התוצאות [השכיחות] כפי שמופיע בדוגמה.

לצד התא 23 השכיחות היא 15, המשמעות היא שמספר הנתונים בטווח המוגדר שערכם שווה או קטן ל- 23 אך גדול מ- 18 הוא 15 ובאופן דומה ביתר התאים.

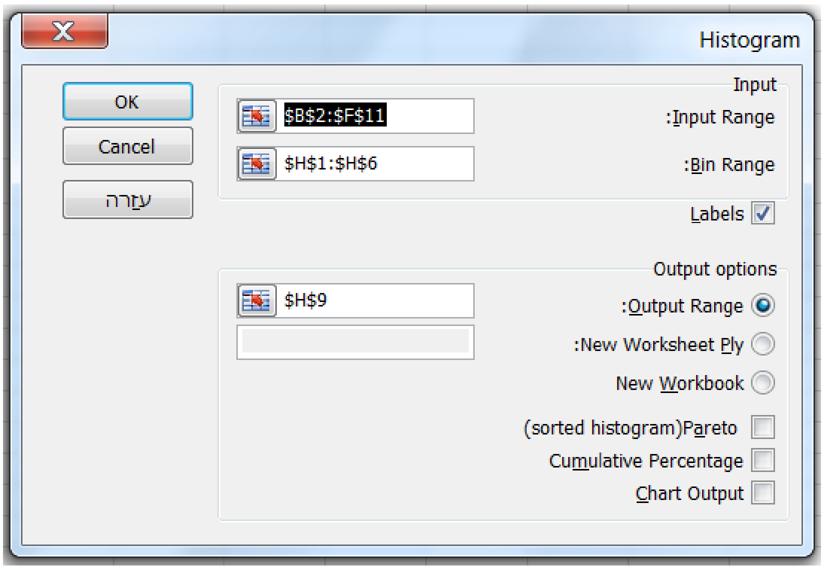
HISTOGRAM

האפשרות הנוספת לקביעת השכיחויות היא בעזרת אחד הכלים מתוך ה- **DATA ANALYSIS**. אם הכלי אינו מגודר יש צורך לשפעל [TO ACTIVATE] אותו בתהליך הבא: קובץ > אחרונים > אפשרויות > תוספות > Analysis ToolPak > נפתחת חלונית בה יש להוסיף √ בתיבה של Analysis ToolPak > אישור.

לאחר שהסתיים שלב השפעול, נפתח את הלשוניות > DATA ANALYSIS > Histogram > OK [כלומר להקיש בסוף על OK].



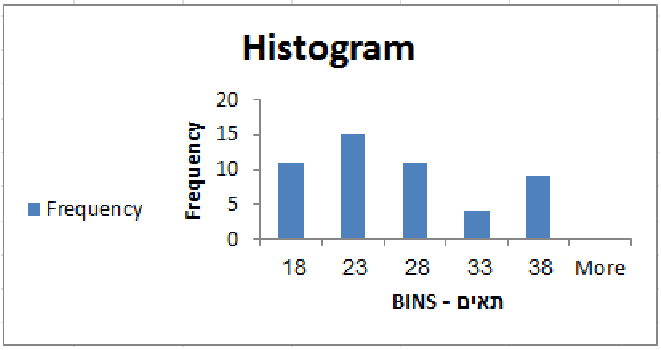
מתקבל המסך הבא לאחר שמולאו בו הנתונים:



יש לציין √ ליד Labels אם בטווח ה- BINS נכלל גם את הכותרת. מתקבלים הנתונים הבאים, הזהה כמובן לאלו שהתקבלו בעזרת הפונקציה FREQUENCY.

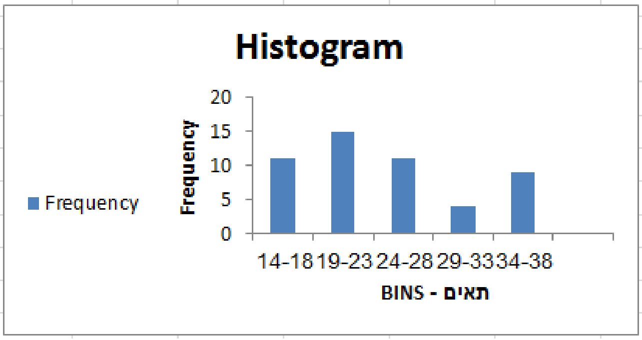
BINS - תאים	*Frequency*
18	11
23	15
28	11
33	4
38	9
More	0

אם במסך ה- HISTOGRAM שלעיל נציין גם √ ליד chart output נקבל את הגרף הבא, שהוא למעשה דיאגרמת מקלות, ושלאחר כמה עיבודים יהפוך להסטוגרמה:

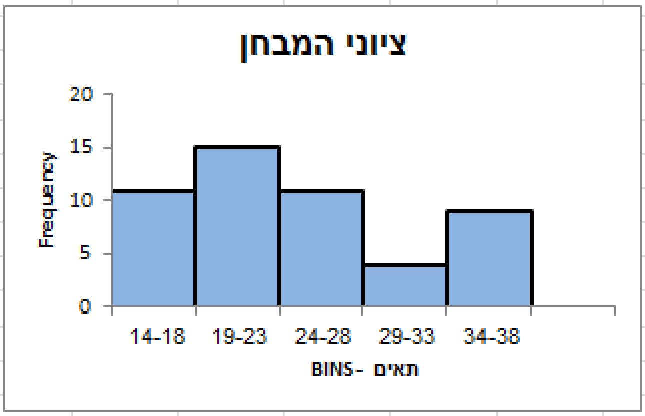


ניכנס ל"תאים-BINS" ושם במקום ערכי הגבול העליון [שהם מספרים] נרשום את הגבולות המדומים או האמיתיים של התאים כטקסט, יש שתי אפשרויות לרשום טקסט – לסמן את הטווח > עיצוב תאים > טקסט, או להימצא בכל תא ולפני הקלדה להקיש גרש ', נקבל את התאים והדיאגרמה הבאים:

BINS - תאים	Frequency
14-18	11
19-23	15
24-28	11
29-33	4
34-38	9



נקיש על אחד המלבנים ואז קליק ימני > עיצוב נקודת נתונים > רוחב מרווח > ללא מרווח. תתקבל הסטוגרמה שבה המלבנים מהווים מקשה אחת ללא גבולות ברורים, נעצב את אזור המלבנים, ואת הדיאגרמה מסביב ונקבל את ההסטוגרמה הבאה:



תרגיל מספר 4

בהמשך לתרגיל מספר 2 ולאחר שקיבצנו את הנתונים בתאים [כפי שבוצע לעיל], חשב את הממוצע, השונות, סטיית התקן, אומד לשונות ואומד לסטיית התקן לפי הנתונים המקובצים. אפשר לפתור את התרגיל בעזרת פעולות רגילות של האקסל ואפשר להיעזר בפונקציה **SUMPRODUCT.** הפונקציה מכפילה רכיבים מתאימים **במערכי נתונים** ומחזירה [מחשבת] את מכפלתם כפי שיתואר להלן:

(1) גבולות מדומים	(2) גבולות אמיתיים	(3) אמצע תא (x)	(4) שכיחות f(x)	(5) x*f(x)	(6) x^2*f(x)
14 - 18	13.5 - 18.5	16	11	176	2816
19 - 23	18.5 - 23.5	21	15	315	6615
24 - 28	23.5 - 28.5	26	11	286	7436
29 - 33	28.5 - 33.5	31	4	124	3844
34 - 38	33.5 - 38.5	36	9	324	11664
			50	1225	32375

=SUM(Q27:Q31)=A

=SUMPRODUCT(P27:P31,Q27:Q31)

=SUMPRODUCT(P27:P31,Q27:Q31)/=SUM(Q27:Q31)=B

הממוצע = 24.5

=SUMPRODUCT(P27:P31^2,Q27:Q31)=C

=C-A*B^2

השונות = 47.25

סטית התקן 6.87

=(C-A*B^2)/(A-1)

אומד לשונות = 48.21

אומד לסטית התקן = 6.94

הערות:

עמודות 1, 3 ו- 4 שבטבלת המקור הכרחיות לפתרון. עמודה 2 נדרשת כאשר יש צורך לחשב מאונים/אחוזונים, עשירונים ורבעונים.

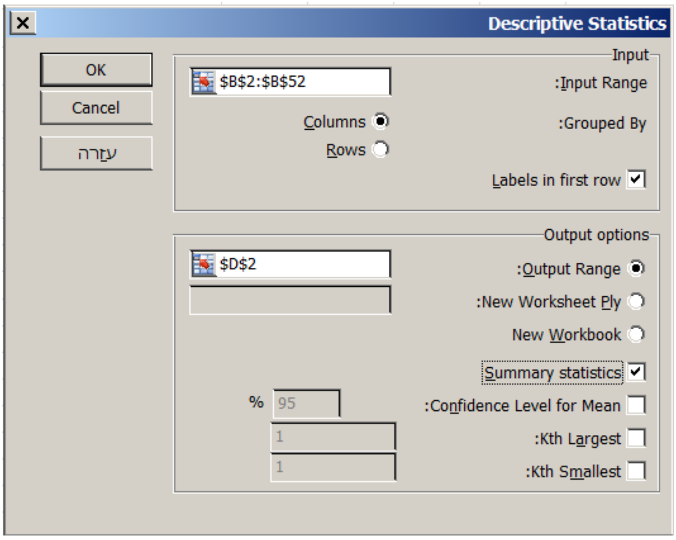
עמודות 5 ו- 6 שבטבלה הן עמודות הנדרשות לחישוב באופן ה"מסורתי" והן הובאו רק כדי לאזכרן. הפתרון כאן נעשה בעזרת הפונקציות המובנות SUM ו- SUMPRODUCT.

אם נשווה את הפרמטרים ממוצע וסטיית התקן שחושבו כאן אל מול הפרמטרים שחושבו קודם לכן על הנתונים הגולמיים, נגלה באופן לא מפתיע כי התוצאות קרובות אך לא זהות. קיבוץ בתאים מניח כי כל הנתונים [הגולמיים] שקובצו לתא מתפלגים באופן אחיד על פני התא או במילים אחרות שממוצע הנתונים הגולמיים שהוכנסו לתא שווה לממוצע התא ובפועל אין זה בהכרח כך.

הפונקציה DESCRIPTIVE STATISTICS

פונקציה זו היא כלי נוסף מתוך ה- **DATA ANALYSIS**. הפונקציה מאפשרת לקבל בדרך קלה ומהירה מדדים ופרמטרים סטטיסטיים של משתנה אחד או מספר משתנים.

ניכנס לפונקציה באופן דומה לכניסה לפונקציה HISTOGRAM שתוארה קודם לכן. נקיש על DATA ANALYSIS > descriptive statistics ומתקבל המסך הבא שבו יש למלא שדות, חלקן שדות חובה וחלקן שדות רשות. נמלא את השדות בנתונים המתייחסים לתרגיל מספר 1 [לאחר שנביא את הנתונים לעמודה אחת כיוון שפונקציה זאת, במתכונת המובאת, אינה יודעת לטפל בטווח נתונים שאינו עמודה]:



לאחר לחיצה על OK נקבל את המסך הבא:

ציוני המבחן	
Mean	24.16
Standard Error	0.9343
Median	23
Mode	34
Standard Deviation	6.6066
Sample Variance	43.647
Kurtosis	-1.1319
Skewness	0.3805
Range	20
Minimum	15
Maximum	35
Sum	1208
Count	50

שימו לב! המדדים המחושבים זהים למדדים שחושבו [על אותם נתונים] בתרגיל מספר 1. השכיח [MODE] הוא אמנם כמו שחושב בתרגיל מספר 1, אולם, במקרה אחר, שבו יש יותר משכיח אחד ייתכנו תשובות שונות. סטיית התקן המחושבת כאן היא "**אומד**" לסטיית התקן.

המדד Standard Error הוא סטיית התקן של ממוצעי מדגמים בגודל 50 והוא מחושב כך:

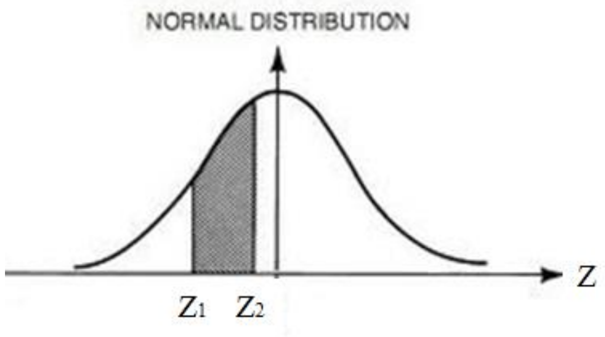
$$\text{Standard error} = \frac{\text{Standrd} \quad \text{Deviation}}{\sqrt{Count}} = \frac{\sigma}{\sqrt{n}}$$

המדדים Kurtosis ו- Skewness [צבועים בצהוב] מתייחסים לצורת העקום המאפיין את הנתונים והם לא יידונו בחוברת זו.

פרק 6 – התפלגויות

התפלגות נורמלית

ההתפלגות הנורמלית מאפיינת אוכלוסיות גדולות כמו גובה של אוכלוסייה, משקל יילודים, רמת [מנת] משכל, ערכי סוכר בדם. התפלגות זו נחשבת לחשובה שבהתפלגויות הסטטיסטיות משתי סיבות עיקריות, התפלגות זאת מאפיינת גם את התפלגות ממוצעי מדגמים הנלקחים מאוכלוסייה גדולה, גם אם אינה מאופיינת ע"י התפלגות נורמלית, ומאפשרת קיום למשפט הגבול המרכזי, כמו כן התפלגות זאת משמשת קירוב מצוין לקבוצה גדולה של התפלגויות סטטיסטיות, שחשיבותן המעשית רבה [ההתפלגות הבינומית וההתפלגות הפואסונית].



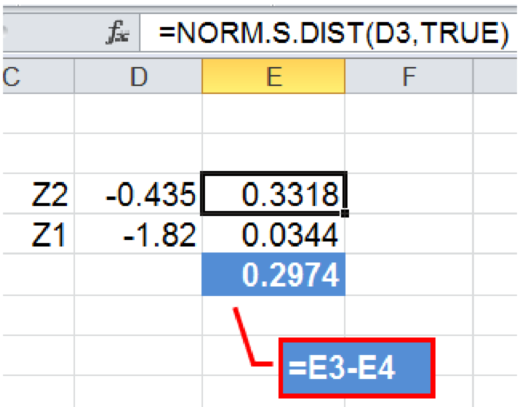
לצורך פתרון שאלה מסוימת אנו נדרשים לחשב את השטח, שמתחת לעקומה הנורמלית, הכלוא בין $Z_1 = -1.82$ לבין $Z_2 = -0.435$ [המתאימים לערכי X_1 ו- X_2 באוכלוסייה]. ברוב המקרים נמצא את השטח בעזרת טבלאות מוכנות. הטבלאות השכיחות הן כאלו הנותנות את השטח מ- $-\infty$ [מינוס אין סוף] ועד ל- Z חיובי נתון ועבור ערכי Z הנתונים עם 2 מקומות לאחר הנקודה. הפתרון מבוסס על עובדת הסימטריה [בתמונת ראי] של השטחים –

$$\Phi(Z=-0.435)-\Phi(Z=-1.82)=\Phi(Z=1.82)-\Phi(Z=0.435)$$

כדי להתקדם בחישוב עלינו לחשב בנוסף, בעזרת אקסטרפולציה את $\Phi(Z=0.435)$ שיהיה $(\Phi(Z=0.43)-\Phi(Z=0.44))/2=(0.1664+0.1700)/2=0.1682$, כמו כן נמצא בטבלה את $\Phi(Z=1.82)=0.4656$ ומכאן –

$$\Phi(Z=-0.435)-\Phi(Z=-1.82)=0.4656-0.1682=0.2974$$

נראה איך פותרים שאלה זאת בעזרת הפונקציה המתאימה ב- EXCEL –



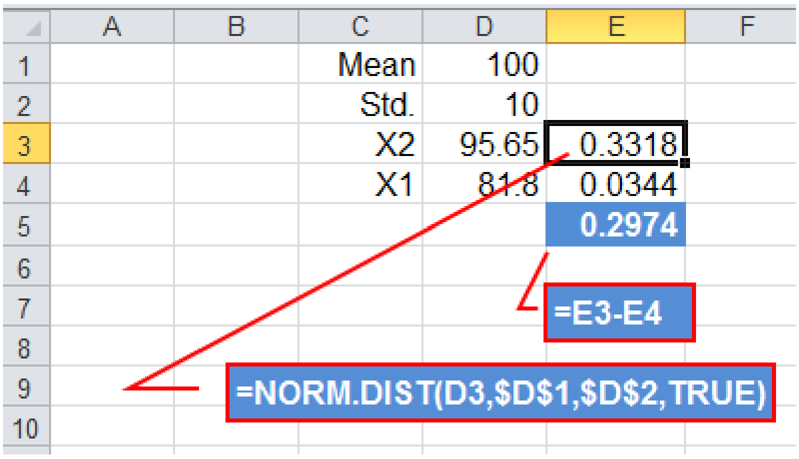
כאשר מקלידים את הפונקציה יש או לציין את ערך ה- Z או להפנות אל התא המכיל את ה- Z שעבורו רוצים את השטח וכן יש לבחור את TRUE הנותן את השטח המצטבר מ- -∞ ועד לאותו Z.

תרגיל מספר 5

ידוע כי באוכלוסייה מסוימת מנת המשכל [IQ] מתפלגת עם תוחלת של 100 וסטיית תקן 10. נדרש לחשב את שיעור האוכלוסייה שמנת המשכל שלהם נמצאת בטווח של 81.8 עד 95.65.

פתרון

הפעם [לאחר שהראנו כבר את המורכבות שבחישוב הידני] נביא רק את הפתרון בעזרת הפונקציה המתאימה שב- EXCEL – רוצים לקבל את השטח מ- -∞ ועד לאותו X. הפעם השתמשנו בפונקציה המחשבת [מאחזרת] את השטח לפי 3 פרמטרים: הממוצע, סטיית התקן וערך X נתון [למעשה בעזרת 3 הפרמטרים הללו מחושב ה- Z המתאים ולפיו השטח]. קיבענו בנוסחה את הממוצע וסטיית התקן כדי שבהעתקת הנוסחה עבור X1 תועתק הנוסחה הנכונה.



תרגיל מספר 6

קבע את ערכי Z המתאימים לשטחים שמתחת לעקום התפלגות הנורמלית [הנמדדים מ- ∞-] שבטבלה הבאה:

0.005	0.01	0.02	0.025	0.05	0.10	0.90	0.95	0.975	0.98	0.99	0.995

פתרון

כמובן שנפתור זאת בעזרת הפונקציה המתאימה שב- EXCEL –

=NORM.S.INV(D2)

B	C	D	E	F	G	H	I	J	K	L	M	N
Φ	0.005	0.01	0.02	0.025	0.05	0.10	0.90	0.95	0.975	0.98	0.99	0.995
Z(Φ)	-2.5758	-2.3263	-2.0537	-1.9600	-1.6449	-1.2816	1.2816	1.6449	1.9600	2.0537	2.3263	2.5758

נראה לכם מוכר?

לסיום פרק זה נראה איך אנחנו יכולים ליצור בעצמנו טבלת שטח שמתחת להתפלגות נורמלית, נשים את הדגש על ערכי Z שליליים עם דיוק של 3 מקומות אחרי הנקודה.

C6 =NORM.S.DIST(($A6-C$4),TRUE)

	A	B	C	D	E	F	G	H	I	J	K
3		Φ									
4	Z	0	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
5	-1.99	0.0233	0.0232	0.0232	0.0231	0.0231	0.023	0.023	0.0229	0.0229	0.0228
6	-1.98	0.0239	0.0238	0.0237	0.0237	0.0236	0.0236	0.0235	0.0235	0.0234	0.0234
7	-1.97	0.0244	0.0244	0.0243	0.0242	0.0242	0.0241	0.0241	0.0240	0.0240	0.0239
8	-1.96	0.0250	0.0249	0.0249	0.0248	0.0248	0.0247	0.0246	0.0246	0.0245	0.0245
9	-1.95	0.0256	0.0255	0.0255	0.0254	0.0254	0.0253	0.0252	0.0252	0.0251	0.0251
10	-1.94	0.0262	0.0261	0.0261	0.0260	0.0259	0.0259	0.0258	0.0258	0.0257	0.0256
11	-1.93	0.0268	0.0267	0.0267	0.0266	0.0266	0.0265	0.0264	0.0264	0.0263	0.0263
12	-1.92	0.0274	0.0274	0.0273	0.0272	0.0272	0.0271	0.0271	0.0270	0.0269	0.0269
13	-1.91	0.0281	0.0280	0.0279	0.0279	0.0278	0.0277	0.0277	0.0276	0.0276	0.0275
14	-1.90	0.0287	0.0287	0.0286	0.0285	0.0285	0.0284	0.0283	0.0283	0.0282	0.0281
15	-1.89	0.0294	0.0293	0.0292	0.0292	0.0291	0.0290	0.0290	0.0289	0.0288	0.0288
16	-1.88	0.0301	0.0300	0.0299	0.0299	0.0298	0.0297	0.0296	0.0296	0.0295	0.0294
17	-1.87	0.0307	0.0307	0.0306	0.0305	0.0305	0.0304	0.0303	0.0303	0.0302	0.0301
18	-1.86	0.0314	0.0314	0.0313	0.0312	0.0312	0.0311	0.0310	0.0310	0.0309	0.0308

תרגיל מספר 7

גובהם של סטודנטים מתפלג נורמלית עם תוחלת של 176 ס"מ וסטיית תקן 6 ס"מ. מה ההסתברות:

[1] שגובהו של סטודנט שנבחר באקראי גדול מ- 180 ס"מ

[2] שבמדגם אקראי של 4 סטודנטים הגובה הממוצע יהיה גדול מ- 180 ס"מ?

[3] שבמדגם אקראי של 25 סטודנטים הגובה הממוצע יהיה גדול מ- 180 ס"מ?

[4] שבמדגם אקראי של 400 סטודנטים הגובה הממוצע יהיה גדול מ- 180 ס"מ?

לצורך פתרון השאלה נזכיר את משפט הגבול המרכזי:

אם מאוכלוסייה המתפלגת נורמלית בעלת תוחלת **μ** ושונות **σ^2** [ובצורה פורמלית – $X \sim N\left(\mu,\sigma^2\right)$] נלקח מדגם מקרי בגודל **n**, או שהאוכלוסייה מתפלגת בהתפלגות כלשהי עם תוחלת ושונות ידועים ובתנאי ש- **n** גדול מספיק, אזי ממוצע המדגם $\bar{X}$ מתפלג גם כן נורמלית כך:

$$\bar{X} \sim N\left(\mu,\frac{\sigma^2}{n}\right)$$

פתרון

הפתרון ה"מסורתי"/ה"רגיל":

$$P(X>180)=1-\Phi\left(\frac{180-176}{6}\right)=1-\Phi(0.667)=0.2525=25.25\% \quad [1]$$

$$P(\bar{x}>180)=1-\Phi\left(\frac{180-176}{6/\sqrt{4}}\right)=1-\Phi(1.333)=0.0912=9.12\% \quad [2]$$

באופן דומה אפשר לחשב גם את סעיף 3 ו- 4. אבל אפשר לנצל את היתרונות הרגילים של האקסל ועוד פונקציה מובנית אחת ולפתור את השאלה בקלות רבה כפי שנראה להלן:

	G	H	I	J	K	L	M	N	O
47	סעיף	μ	σ	n	X / $\bar{X}$	נדרש לחשב P	Z	Φ(Z)	P=1-Φ(Z)
48	1	176	6	1	180	P(X > 180)=	0.667	0.7475	0.2525
49	2	176	6	4	180	P($\bar{X}$ >180)=	1.333	0.9088	0.0912
50	3	176	6	25	180	P($\bar{X}$ >180)=	3.333	0.9996	0.0004
51	4	176	6	400	180	P($\bar{X}$ >180)=	13.33	1.0000	0.0000

=(K49-H49)/(I49/(J49)^0.5)

=NORM.S.DIST(M49,TRUE)

=1-N49

התפלגות t

התפלגות t של סטודנט [Student's t-distribution] או בפשטות התפלגות t היא התפלגות סימטרית פעמונית המתארת את הערכים הצפויים למדגם הנלקח מתוך אוכלוסייה המתפלגת נורמלית כאשר גודלו של המדגם קטן וסטיית התקן של האוכלוסייה אינה ידועה, או כאשר המדגם נלקח מתוך אוכלוסייה שאינה מתפלגת נורמלית. הצורה הכללית של התפלגות t דומה לזו של ההתפלגות הנורמלית וכאשר מספר הפרטים במדגם [גודל המדגם – n] גדול מאוד היא אף זהה לה [לצרכים מעשיים אם n>30 אפשר לראות את ההתפלגויות כדומות].

צורתה של התפלגות t משתנה בהתאם לדרגות החופש – d.f. [או בקיצור df]. כאשר df=n-1.

חישוב שטח מתחת לעקום התפלגות t בהינתן t מחושב/נתון:

	A	B	C	D	E
1	t	df	מס' "זנבות"	הסתברות	α
2	0.72	15	2	0.759	0.241
3	0.72	15	1	0.759	0.241
4	1.28	24	1	0.894	0.106
5	1.72	31	1	0.952	0.048
6	1.96	310	2	0.975	0.025
7	2.24	31	2	0.984	0.016
8	2.98	17	1	0.996	0.004
9	-1.72	31	1	0.048	0.952
10					
11	=T.DIST(A3,B3,C3)				

כמה הערות/הבהרות בהתייחס לפונקציה ולטבלה שלעיל:

- הפונקציה מאחזרת/מחשבת את ההסתברות מ- ∞- ועד לערך t נתון. כך למשל בדוגמה שבשורה 4 ההסתברות 0.894 מתייחסת לשטח שבין ∞- לבין t=1.28, לפיכך הערך α שהוא המשלים ל- 1 של ההסתברות מתייחס לשטח הזנב הימני
- הפונקציה מאחזרת/מחשבת את ההסתברות גם עבור ערכי t שליליים [בגרסאות אקסל קודמות לאקסל 2010 הדבר לא היה קיים]
- למרות שהפונקציה דורשת לציין את מספר הזנבות [C3 בדוגמה המודגשת] היא אינה מושפעת מנתון זה [ראה 2 הדוגמאות הראשונות בצבע ירקרק]
- יש להפעיל שיקול דעת בעת ניתוח התוצאות האם נדרש שימוש בערך t עבור זנב אחד [ואז האם הוא ימני או שמאלי] או עבור 2 זנבות

קביעת t בהינתן הסתברות ודרגות חופש:

E3 =T.INV.2T(B3,D3)

	A	B	C	D	E
1	הסתברות	α	α/2	df	t
2	0.995	0.010	0.005	15	2.9467
3	0.990	0.020	0.010	15	2.6025
4	0.980	0.040	0.020	24	2.1715
5	0.975	0.050	0.025	15	2.1314
6	0.950	0.100	0.050	4000	1.6452

=(1-A3)*2

=T.INV.2T(B3,D3)

כמה הערות/הבהרות בהתייחס לפונקציה ולטבלה שלעיל:

- הפונקציה מאחזרת/קובעת את ערך t תוך התייחסות לכך שמספר ה"זנבות" הוא 2. הערך המתאים להסתברות [probability] שיש למלא **בתוך הפונקציה** הוא ערך 2 הזנבות, כך לדוגמה בשורה 6 ההסתברות שיש למלא בפונקציה היא 0.1. הערך 0.95 מייצג את השטח מתחת לעקום מ- ∞- ועד לערך t נתון. בשורה זאת הערך t שווה ל- |1.6452| והוא משמש במבחני השערות חד צדדיים [כאשר המבחן הוא ימני הערך יהיה חיובי וכאשר המבחן שמאלי הערך יהיה שלילי] כשרמת המובהקות היא 5% ובמבחני השערות דו צדדיים כשרמת המובהקות היא 10%.
- חשוב להיות ערניים ולחשוב איזה ערך הסתברות יש להקליד לתוך הפונקציה כאשר נשתמש בפונקציה זאת במבחני השערות ובבניית רווחי סמך
- הפונקציה מאחזרת/קובעת את ערך t גם עבור ערך הסתברות [במובן של הנתון שבעמודה A] הקטן מ- 0.5 ומתקבלים ערכי t שליליים נכונים, אולם ערכי α [כפי שהובאו בטבלה שלעיל] יהיו חסרי משמעות ולכן יש לזכור להשתמש רק בערך הסתברות הגדול מ- 0.5
- שימו לב לשורה 6, מספר דרגות החופש הוא 4,000. כמובן שזו הגזמה והמטרה היחידה היא להראות כמה ערכי t דומים לערכים שהיו מתקבלים תוך שימוש בעקומה הנורמלית

בניית טבלת התפלגות t

D8 =T.INV.2T(D$4,$A8)

			למבחן חד-צידי			
$\alpha/2$	0.250	0.100	0.050	0.025	0.010	0.005
			למבחן דו-צידי			
α	0.500	0.200	0.100	0.050	0.020	0.010
הסתברות	0.750	0.900	0.950	0.975	0.990	0.995
דרגות חופש 1	1.0000	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
			1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
1600	0.6746	1.2821	1.6458	1.9614	2.3287	2.5789

=T.INV.2T(D$4,$A8)

תרגיל מספר 8

גובהם של שתילי גזר לאחר שבועיים מאז נשתלו מתפלג עם תוחלת של 12 ס"מ וסטיית תקן 3.2 ס"מ. מה ההסתברות:

[1] שבמדגם אקראי של 16 שתילים שנלקחו באקראי גובהם הממוצע של השתילים יהיה גדול מ- 13 ס"מ?

[2] שבמדגם אקראי של 1,600 שתילים שנלקחו באקראי גובהם הממוצע של השתילים יהיה גדול מ- 12.1 ס"מ?

פתרון

באופן עקרוני, כיוון שלא נכתב כי גובה שתילי הגזר מתפלג נורמלית – עלינו להניח כי ממוצע גובה השתילים במדגם בשני סעיפי השאלה מתפלג בהתפלגות t. אולם כיוון שבסעיף השני גודל המדגם גדול מאוד ניתן לצרכים מעשיים להניח שממוצע גובה השתילים במדגם זה מתפלג נורמלית.

הפתרון מובא להלן:

	A	B	C	D	E	F	G	H
1	**נתונים**							
2		μ = 12						
3		σ = 3.2						
4			**n**	**df**	$\bar{x}$	**t**	**הסתברות**	$P(\bar{x}' \geq \bar{x})$
5		**סעיף [1]**	16	15	13	1.2500	0.8848	**0.0576**
6		**סעיף [2]**	1600	1599	12.1	1.2496	0.8942	**0.0529**

=(E5-$B$2)/($B$3/C5^0.5)

=T.DIST(F5,D5,1)

=(1-G5)/2

בעקבות תרגיל זה, לא ניתן להתעלם מהתחושה שהיתרון היחיד שאנו רואים בשימוש בפונקציות המתאימות להתפלגות t הוא קבלת ערכים בצורה ממוחשבת במקום פשוט להסתכל בטבלה. נציין כי הטבלאות [שב- Hard copy] מוגבלות גם במספר דרגות החופש וגם בערכי ההסתברות ובכך יש יתרון לשימוש הממוחשב. כמו כן, בהמשך כאשר נדון בפונקציות מתוך Data Analysis נראה יתרונות נוספים.

התפלגות בינומית

חישובי הסתברות, הנוגעים להתפלגות בינומית, עשויים להיות מורכבים וכשהם נעשים באופן ידני או בעזרת מחשבונים הם מוגבלים [מעשית אך לא תיאורטית] לתחום מצומצם. נדגים את הדבר בעזרת הדוגמה הבאה:

תרגיל מספר 9

ידוע כי אוטובוסים של חברת "הדייקן והבטוח" יוצאים במועד המתוכנן מתחנת המוצא אך ב- 30% מהמקרים [**p**] מאחרים או מקדימים להגיע לתחנות הביניים. סוקר של משרד התחבורה הוצב באחת מתחנות הביניים והתבקש לבדוק את מספר האוטובוסים שהגיעו לתחנה שלא במועד [איחור או הקדמה של עד 5 דקות נחשבים כהגעה בזמן]. מה ההסתברות שמתוך 16 האוטובוסים [**n**] הראשונים שהגיעו לתחנה **יותר** מ- 6 [**x**] אוטובוסים אחרו או הקדימו?

פתרון

הפתרון ה"רגיל" נראה כך:

$$P(x>6)=1-\left[\binom{16}{0}\cdot 0.3^{0}\cdot 0.7^{16}+\binom{16}{1}\cdot 0.3^{1}\cdot 0.7^{15}+\binom{16}{2}\cdot 0.3^{2}\cdot 0.7^{14}+\ldots\ldots\binom{16}{6}\cdot 0.3^{6}\cdot 0.7^{9}\right]$$

באקסל ייראה הפתרון כך –

G2 =1-(BINOM.DIST(E4,E2,E3,TRUE))

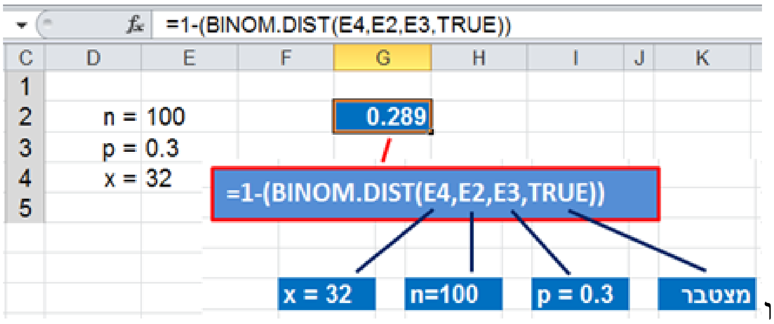
	C	D	E	F	G	H	I	J	K
1									
2		n =	16		0.175				
3		p =	0.3						
4		x =	6						

תרגיל מספר 10

בהמשך לתרגיל מספר 9, מה ההסתברות שמתוך 100 האוטובוסים הראשונים שהגיעו לתחנה **יותר** מ- 32 אוטובוסים אחרו או הקדימו?

פתרון

הפתרון ה"רגיל" מורכב יותר וכולל נוסחה עם הרבה אברים ובדרך כלל במצב זה אנו עוברים לפתרון המבוסס על שימוש בקרוב של ההתפלגות הנורמלית להתפלגות הבינומית. כפי שנראה הפתרון באקסל פשוט, קל ומדויק. הפתרון לתרגיל בעזרת הפונקציה המתאימה באקסל – לאחר שהקלדנו, בגיליון ה- EXCEL, את הנתונים ואת הנוסחה של התרגיל הקודם די שנחליף את ה- 16 ב- 100 ואת ה- 6 ב- 32 ונקבל את הפתרון לתרגיל.

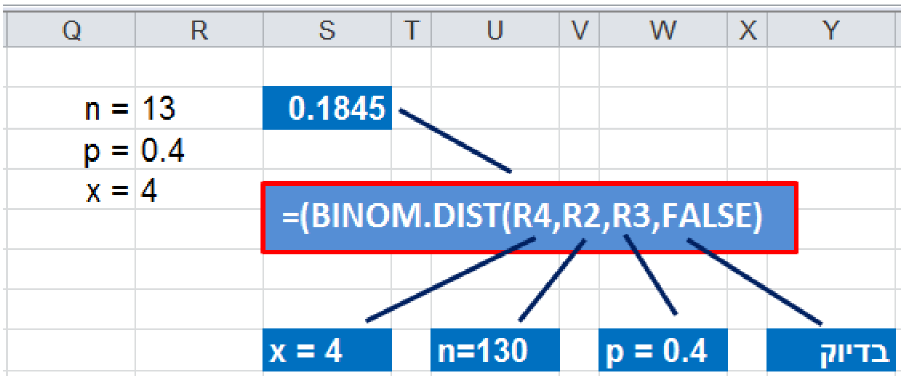


תרגיל מספר 11

באוסטרליה ל- 40% מהאוכלוסייה סוג דם O+. נלקח מדגם מקרי של 130 תושבים מהאוכלוסייה. מה ההסתברות שבמדגם יימצאו **בדיוק** 4 תושבים עם סוג דם O+?

פתרון

הפתרון דומה בעיקרון לזה של תרגיל מספר 9, אולם שם, לטובת התפלגות מצטברת [P(x≤6)] נעשה שימוש בפונקציה עם "**TRUE**" וכאן לטובת התפלגות בדידה [P(x=4)] נעשה שימוש בפונקציה עם "**FALSE**".



לסיום פרק זה נראה איך אנחנו יכולים ליצור בעצמנו טבלת הסתברות בינומית מצטברת – **TRUE** [אפשר גם ליצור טבלת הסתברות לערכים "בדיוק" – FALSE, אולם לא נעשה זאת משום שניתן לחלץ זאת בקלות מהטבלה המצטברת].

G6 =BINOM.DIST($A6,G$2,$H$1,TRUE)

P= 0.25

x \ n	10	11	12	13	14	15	16	17	18	19	20	25	30
0	0.0563	0.0422	0.0317	0.0238	0.0178	0.0134	0.0100	0.0075	0.0056	0.0042	0.0032	0.0008	0.0002
1	0.2440	0.1971	0.1584	0.1267	0.1010	0.0802	0.0635	0.0501	0.0395	0.0310	0.0243	0.0070	0.0020
2	0.5256	0.4552	0.3907	0.3326	0.2811	0.2361	0.1971	0.1637	0.1353	0.1113	0.0913	0.0321	0.0106
3	0.7759	0.7133	0.6488	0.5843	0.5213	0.4613	0.4050	0.3530	0.3057	0.2631	0.2252	0.0962	0.0374
4	0.9219	0.8854	0.8424	0.7940	0.7415	0.6865	0.6302	0.5739	0.5187	0.4654	0.4148	0.2137	0.0979
5	0.9803	0.9657	0.9456	0.9198	0.8883	0.8516	0.8103	0.7653	0.7175	0.6678	0.6172	0.3783	0.2026
6	0.9965	0.9924	0.9857	0.9757	0.9617	0.9434	0.9204	0.8929	0.8610	0.8251	0.7858	0.5611	0.3481
7	0.9996	0.9988	0.9972	0.9944	0.9897	0.9827	0.9729	0.9598	0.9431	0.9225	0.8982	0.7265	0.5143
8	1.0000	0.9999	0.9996	0.9990	0.9978	0.9958	0.9925	0.9876	0.9807	0.9713	0.9591	0.8506	0.6736
9		1.0000	1.0000	0.9999	0.9997	0.9992	0.9984	0.9969	0.9946	0.9911	0.9861	0.9287	0.8034
10				1.0000	1.0000	0.9999	0.9997	0.9994	0.9988	0.9977	0.9961	0.9703	0.8943
11						1.0000	1.0000	0.9999	0.9998	0.9995	0.9991	0.9893	0.9493
12								1.0000	1.0000	0.9999	0.9998	0.9966	0.9784
13										1.0000	1.0000	0.9991	0.9918
14												0.9998	0.9973

התפלגות היפרגאומטרית

חישובי הסתברות, הנוגעים להתפלגות היפרגאומטרית, עשויים להיות מורכבים וכשהם נעשים באופן ידני או בעזרת מחשבונים הם מוגבלים [מעשית אך לא תיאורטית] לתחום מצומצם. נדגים את הדבר בעזרת התרגיל הבא:

תרגיל מספר 12

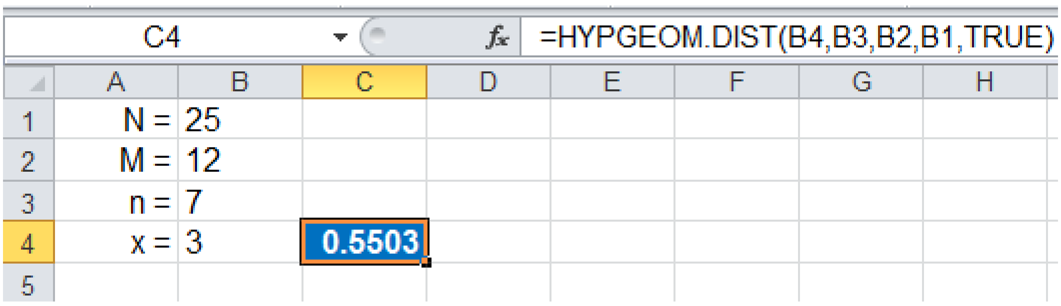
במשלוח של 25 צנצנות [**N**] המכילות [ע"פ הרשום על האריזה] כל אחת 200 טבליות של אומגה 3 [שמן דגים] 12 צנצנות [**M**] שבהן, בטעות, יש רק 175 טבליות. נלקח מדגם אקראי של 7 צנצנות [**n**]. מה ההסתברות שבמדגם לכל היותר 3 צנצנות [**x**] שבהן רק 175 טבליות?

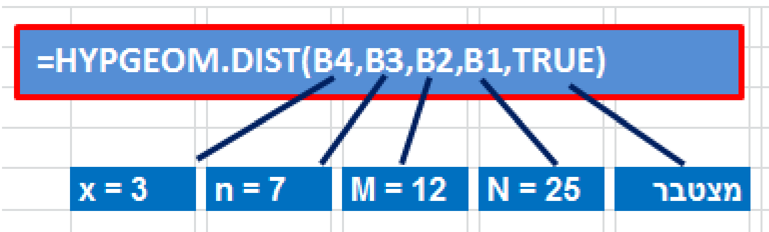
פתרון

הפתרון ה"רגיל" נראה כך:

$$P(x \le 3) = \left[\frac{\binom{12}{0}\cdot\binom{25-12}{7-0}}{\binom{25}{7}} + \frac{\binom{12}{1}\cdot\binom{25-12}{7-1}}{\binom{25}{7}} + \frac{\binom{12}{2}\cdot\binom{25-12}{7-2}}{\binom{25}{7}} + \frac{\binom{12}{3}\cdot\binom{25-12}{7-3}}{\binom{25}{7}} \right]$$

באקסל ייראה הפתרון כך:





תרגיל מספר 13

במשלוח של 250 צנצנות [**N**], המכילות [ע"פ הרשום על האריזה] כל אחת 200 טבליות של אומגה 3 [שמן דגים], 120 צנצנות [**M**] שבהן, בטעות, יש רק 175 טבליות. נלקח מדגם אקראי של 70 צנצנות [**n**]. מה ההסתברות שבמדגם לכל היותר 30 צנצנות [**x**] שבהן רק 175 טבליות?

פתרון

למעשה תרגיל זה דומה לתרגיל מספר 12 אולם כל אחד מהפרמטרים הוכפל פי 10. הפתרון ה"רגיל" כולל 31 אברים [x+1] אותם יש לחשב [וזה לא פשוט כלל] והוא נראה כך:

$$P(x \le 30)=\left[\frac{\binom{120}{0}\cdot\binom{250-120}{70-0}}{\binom{250}{70}}+\frac{\binom{120}{1}\cdot\binom{250-12}{70-1}}{\binom{250}{70}}+\ldots\ldots\ldots\ldots\frac{\binom{120}{30}\cdot\binom{250-120}{70-30}}{\binom{250}{70}}\right]$$

לאחר שהקלדנו, בגיליון ה- EXCEL, את הנתונים ואת הנוסחה של התרגיל הקודם די אם "נתלבש" על הפתרון ונחליף את ה- 25 ב- 250, את ה- 12 ב- 120, את ה- 7 ב- 70 ואת ה- 3 ב- 30 ונקבל את הפתרון לתרגיל:

C4 f_x =HYPGEOM.DIST(B4,B3,B2,B1,TRUE)

	A	B	C	D	E	F	G	H
1	N =	250						
2	M =	120						
3	n =	70						
4	x =	30	0.1911					
5								

התפלגות פואסונית

התפלגות פואסונית, הינה התפלגות של משתנה בדיד המאפיינת הסתברות של מופעים [מקרים] למרווח זמן או ליחידה אחרת, למשל מספר המכוניות המגיעות במשך שעה למכון לשטיפת מכוניות, מספר דליפות מים ל- 10 ק"מ של צינור מים ראשי ועוד.

לצרכים מעשיים ובעיקר לבקרת איכות סטטיסטית, משמשת ההתפלגות הפואסונית כקירוב להתפלגות הבינומית [לחישוב ההסתברות למציאת x חלקים פגומים במדגם.

ההסתברות הפואסונית מחושב על פי:

$$P(X \ge x)=1-\sum_{x=0}^{x-1}\frac{e^{-\lambda}\cdot\lambda^{x}}{x!} \qquad P(X \le x)=\sum_{x=0}^{x}\frac{e^{-\lambda}\cdot\lambda^{x}}{x!} \qquad P(X = x)=\frac{e^{-\lambda}\cdot\lambda^{x}}{x!}$$

חישובי הסתברות, הנוגעים להתפלגות פואסונית, עשויים להיות מורכבים וכשהם נעשים באופן ידני או בעזרת מחשבונים הם מוגבלים [מעשית אך לא תיאורטית] לתחום מצומצם. נדגים את הדבר בעזרת הדוגמה הבאה.

תרגיל מספר 14

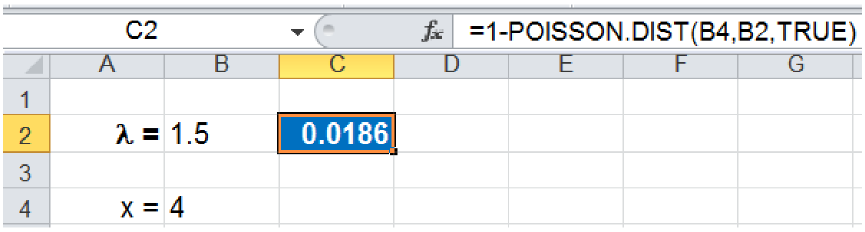
שיעור הנדבקים במחלת הנשיקה באוניברסיטאות ובמקומות עבודה הומי אדם הוא 1 ל-10,000 [במונחי הסתברות – 0.0001]. באוניברסיטה המונה 15,000 סטודנטים מה ההסתברות שיימצאו לפחות 5 סטודנטים שחלו במחלה זו?

פתרון

$$\lambda = 0.0001 \times 15{,}000 = 1.5$$

$$P(x \geq 5) = 1 - \left[\frac{e^{-1.5} \cdot 1.5^0}{0!} + \frac{e^{-1.5} \cdot 1.5^1}{1!} + \frac{e^{-1.5} \cdot 1.5^2}{2!} + \ldots \frac{e^{-1.5} \cdot 1.5^4}{4!} \right]$$

תוך שימוש בפונקציה המובנית באקסל ייראה הפתרון כך:



C2 | =1-POISSON.DIST(B4,B2,TRUE)

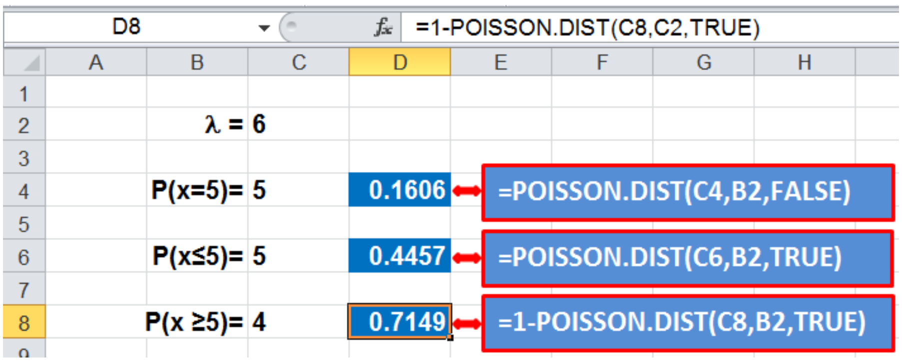
	A	B	C	D	E	F	G
1							
2	λ =	1.5	0.0186				
3							
4	x =	4					

תרגיל מספר 15

ממוצע כלי הרכב המגיעים למכון לשטיפת רכב במשך רבע שעה הוא 6. מה ההסתברות שבמשך רבע שעה יגיעו למכון השטיפה: [1] בדיוק 5 כלי רכב, [2] לא יותר מ- 5 כלי רכב, [3] יותר מ- 4 כלי רכב?

פתרון

כמו בתרגיל קודם – במקום לבצע חישובים מתמטיים, ניתן לקבל תשובות במהירות רבה תוך שימוש בפונקציה המובנית שב- EXCEL. התשובה תראה כך:



D8 | =1-POISSON.DIST(C8,C2,TRUE)

	A	B	C	D	E	F	G	H
1								
2		λ =	6					
3								
4		P(x=5)=	5	0.1606	=POISSON.DIST(C4,B2,FALSE)			
5								
6		P(x≤5)=	5	0.4457	=POISSON.DIST(C6,B2,TRUE)			
7								
8		P(x ≥5)=	4	0.7149	=1-POISSON.DIST(C8,B2,TRUE)			

תרגיל מספר 16

יש לשרטט עקום אפיון תכונות [O.C. CURVE] לתוכנית הדגימה הבאה: גודל המנה – N=1,000, גודל המדגם n=80, מספר הקבלה Ac=2, מספר הדחייה – Re=3 [במילים פשוטות – ממנות בנות 1,000 יש לקחת מדגם אקראי בגודל 80, אם במדגם יימצאו לכל היותר 2 פגומים – המנה תתקבל, בכל מקרה אחר – המנה תידחה].

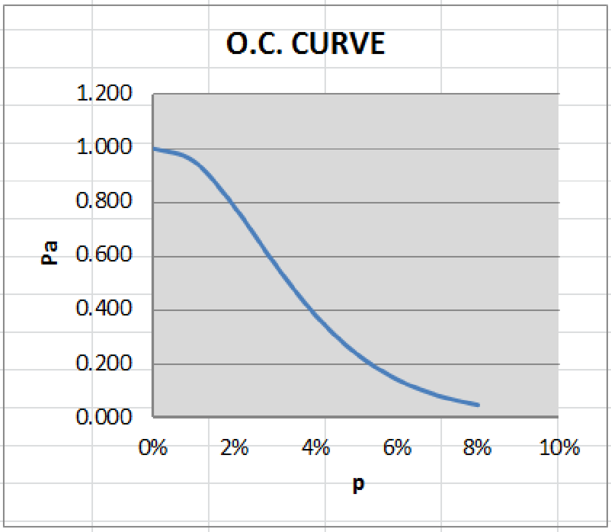
פתרון

עקום האפיון מאפשר לראות מהי הסתברות הקבלה [והדחייה] תחת תוכנית זו, של מנות ששיעור הפגומים [p] בהן הוא מ- 0% ועד שיעור מסוים. השיעור יחושב כך:

$$P(X < 2 \mid \lambda = p \cdot 80) = \sum_{x=0}^{x=2} \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

החישוב ב- EXCEL והגרף שהופק ממנו ייראו כך:

L	M	N
x=	2	
n=	80	
p	λ	Pa
0%	0	1.0000
1%	0.8	0.9526
2%	1.6	0.7834
3%	2.4	0.5697
4%	3.2	0.3799
5%	4	0.2381
6%	4.8	0.1425
7%	5.6	0.0824
8%	6.4	0.0463



=POISSON.DIST($M$2,M13,TRUE)

התפלגות מעריכית [אקספוננציאלית]

התפלגות מעריכית, הינה התפלגות שבה המשתנה המקרי הוא משך הזמן שיחלוף עד לקרות אירוע/התרחשות, למשל עד לקרות התקלה הראשונה במכשיר, עד שתתחולל רעידת אדמה, או עד שתפרוץ המלחמה הבאה, או עד שתגיע שיחת הטלפון הבאה.

הפונקציה המעריכית מיוצגת ע"י האיבר הראשון של פונקציית ההסתברות הפואסונית. השימוש העיקרי [והמעשי] של ההתפלגות המעריכית הוא באמצעות פונקציית ההתפלגות המצטברת שלה [עבור $x \ge 0$]:

$$F(x) = 1 - e^{-\lambda x}$$

תרגיל מספר 17

משך הזמן הממוצע להמתנה בתור לרכבת הרים ביורו דיסני הוא 22 דקות. מה ההסתברות שהילד עופר ימתין בתור לא יותר מ- 15 דקות? בין 15 דקות ל- 25 דקות? מה ההסתברות שלאחר המתנה של 15 דקות ימתין עופר עוד 15 דקות?

פתרון

אם נציין את הזמן הממוצע להמתנה כ- μ אזי יתקיים λ=1/μ.

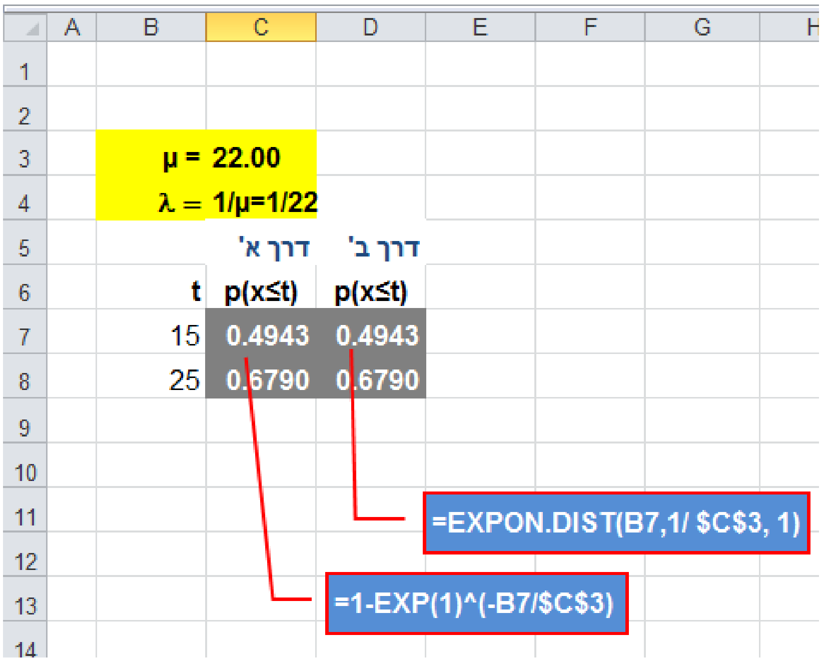
$$F(x \le 15) = 1 - e^{-\left(\frac{15}{22}\right)} = 0.4943$$

$$F(15 \le x \le 25) = \left[1 - e^{-\left(\frac{25}{22}\right)}\right] - \left[1 - e^{-\left(\frac{15}{22}\right)}\right] = 0.6790 - 0.4943 = 0.1847$$

התפלגות מעריכית מגודרת כ"התפלגות ללא זיכרון", המשמעות היא שלעובדה שעופר המתין בתור 15 דקות [ועדיין לא עלה על רכבת ההרים] אין כל השפעה על משך הזמן הנוסף שעליו להמתין. לכן, ההסתברות שלאחר המתנה של 15 דקות ימתין עופר עוד 15 דקות שווה גם כן 0.4943.

חישובי הסתברות, הנוגעים להתפלגות מעריכית, עשויים להיות, כפי שראינו לעיל, מורכבים וכשהם נעשים באופן ידני או בעזרת מחשבונים הם מוגבלים [מעשית אך לא תיאורטית] לתחום מצומצם.

הפתרון בעזרת פונקציות מובנות של אקסל אפשרי בשתי דרכים כפי שנראה להלן [דרך א' עוקבת למעשה אחר הפתרון שלעיל בעוד שדרך ב' מיישמת את הפונקציה הסטטיסטית המיועדת לנושא]:



שימושי ההתפלגות המעריכית בתורת האמינות

בתורת האמינות עושים שימוש בפונקציה זו כשאמינות מערכת מחושבת כך:

$$R_{(t)} = e^{-\lambda t} = e^{-t/m}$$

כאשר:

m = זמן ממוצע בין תקלות [Mena Time Between Failure – MTBF]

פונקציית אי אמינות המערכת מחושבת כך:

$$F_{(t)} = 1 - R_{(t)} = 1 - e^{-\lambda t}$$

כאשר λt קטן מאוד אזי בקרוב $e^{-\lambda t} = 1 - \lambda t$ [כאשר $0.01 \le \lambda t \le 0.1$ החישוב המקורב נכון עד ל- 2 מקומות לאחר הנקודה].

פרק 7 – סינון נתונים מתוך מסד נתונים

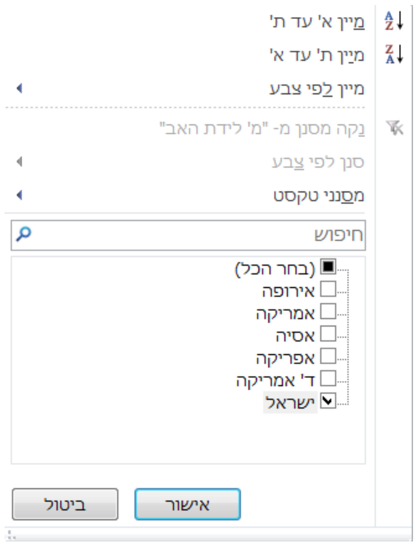
הנושא מוזכר קודם להצגת הפונקציות הסטטיסטיות שבהן מובנית [BUID IN] פונקציית התניה משום שבמקרים לא מעטים השילוב של סינון ופונקציה סטטיסטית מובנית "רגילה" מביאים לאותה תוצאה.

ללימוד הפונקציות נעשה שימוש בטבלה הבאה שהיא פלט [חלקי] מתוך יומן לידות של בית יולדות "דור העתיד":

פרטים אודות היולדת						פרטים אודות היילוד		
שם	גיל	לידות קודמות	מ' לידת האב	עליית משקל %	לחץ דם סיסטולי	מין	משקל בלידה	סוג דם
יערה קינן	24	0	אירופה	12	113	בת	2980	B
נינה אלפסי	32	2	אפריקה	15	86	בן	3520	A
דבורה כהן	19	0	אסיה	9	126	בן	3160	O
מיכל גולדשטיין	36	2	אמריקה	18	93	בת	3620	AB
דנה ארנון	34	4	ישראל	16	102	בת	3980	O
ברכה שיין	42	7	אירופה	22	85	בת	4040	A
סיגל לוי	39	5	אסיה	14	90	בן	2650	O
נילי בן דיין	18	0	ד' אמריקה	11	112	בת	3760	O
ג'ודי לאונד	44	2	אמריקה	15	98	בת	3250	A
מרינה בקשטיין	28	1	אירופה	8	105	בן	3300	AB
סמירה שיעב	45	3	ישראל	16	88	בן	3960	O
שולה גביאזדה	33	3	אירופה	25	135	בת	2840	B
לילך שוקרון	39	4	אסיה	18	89	בן	3750	B
קטיבה אלאסי	40	4	אפריקה	8	97	בן	2900	A
אסתר וקין	24	2	אפריקה	12	90	בן	3680	A
חן שילון	37	2	אמריקה	10	89	בת	4120	A
אסתר ברוך	40	1	ישראל	12	126	בת	3300	O
נטשה שימלה	22	1	אירופה	9	93	בת	3470	O
אפרת דוד	29	0	ישראל	6	128	בן	3170	AB
נדיה חורחה	41	2	ד' אמריקה	14	91	בת	3810	B
שלי בר זאב	30	0	ישראל	11	92	בת	3290	O

כדי להפעיל סינון יש לסמן את כל הטבלה, אור להימצא באחד התאים שלה, ואז: "בית" > "מיין וסנן" > סנן, מתקבל ריבוע קטן ובו משולש קטן הפוך בפינה הימנית התחתונה שבכותרת כל אחת מהעמודות – ריבוע זה מפעיל סינון, על כל הטבלה, על פי קריטריונים המובנים בתוך הסינון של העמודה.

אם נרצה לראות או לעבד רק את הנתונים של יולדות שמקום הולדת האב [של היילוד] הוא ישראל – נלחץ על הכפתור שבראש העמודה "מ' הולדת האב" ונקבל תיבה שבה מוצגים כל מקומות הלידה עם סימן √ בתוך ריבוע לידן, נסיר את הסימן עבור מקומות שאינם ישראל כפי שרואים באיור הבא:



נקיש אישור ונקבל את הטבלה המצומצמת הבאה:

פרטים אודות היולדת						פרטים אודות היילוד		
שם	גיל	לידות קודמות	מ' לידת האב	עליית משקל %	לחץ דם סיסטולי	מין	משקל בלידה	סוג דם
דנה ארנון	34	4	ישראל	16	102	בת	3980	O
סמירה שיעב	45	3	ישראל	16	88	בן	3960	O
אסתר ברוך	40	1	ישראל	12	126	בת	3300	O
אפרת דוד	29	0	ישראל	6	128	בן	3170	AB
שלי בר זאב	30	0	ישראל	11	92	בת	3290	O

על טבלה זאת נוכל לחשב פרמטרים סטטיסטיים או ללמוד נתונים אודות יולדות שמקום הולדת אב היילוד היא ישראל [שימו לב שבתוך הריבוע הקטן שבכותרת "מ' לידת האב מופיע משפך המראה שהסינון נעשה לפי עמודה זאת].

כתהליך קבוע, האחות הראשית מזמנת לשיחה קבוצתית, ביחד עם דיאטנית המחלקה, את כל היולדות שלחץ הדם הסיסטולי שלהן שווה או גדול מ- 95. לקבלת רשימת היולדות נלחץ על כפתור המיון של עמודת "לחץ דם סיסטולי" תיפתח תיבה ובה נקיש על מסנני מספרים > גדול ושווה ל- ואז בתיבה שתיפתח נקיש 95, תתקבל הטבלה הבאה:

פרטים אודות היולדת						פרטים אודות היילוד		
שם	גיל	לידות קודמות	מ' לידת האב	עליית משקל %	לחץ דם סיסטולי	מין	משקל בלידה	סוג דם
יערה קינן	24	0	אירופה	12	113	בת	2980	B
דבורה כהן	19	0	אסיה	9	126	בן	3160	O
דנה ארנון	34	4	ישראל	16	102	בת	3980	O
נילי בן דיין	18	0	ד' אמריקה	11	112	בת	3760	O
ג'ודי לאונד	44	2	אמריקה	15	98	בת	3250	A
מרינה בקשטיין	28	1	אירופה	8	105	בן	3300	AB
שולה גביאזדה	33	3	אירופה	25	135	בת	2840	B
קטיבה אלאסי	40	4	אפריקה	8	97	בן	2900	A
אסתר ברוך	40	1	ישראל	12	126	בת	3300	O
אפרת דוד	29	0	ישראל	6	128	בן	3170	AB

עכשיו נוכל לחשב, למשל את הממוצע וסטיית התקן של "משקל בלידה" של התינוקות שנולדו ליולדות שלחץ הדם הסיסטולי שלהן שווה או גבוה מ- 95 או כל פרמטר סטטיסטי אחר.

סינון של "גם וגם" [AND, ⋂]

אם רוצים לקבל רשימה, או לחשב פרמטר סטטיסטי כלשהו של יולדות שלחץ הדם שלהן הוא גם שווה או גבוה מ- 95 וגם שסוג הדם של היילוד שלהן הוא "O" נחזור אל הטבלה שלעיל ונבצע סינון נוסף לקבלת יילודים עם סוג דם "O". מתקבלת הטבלה הבאה:

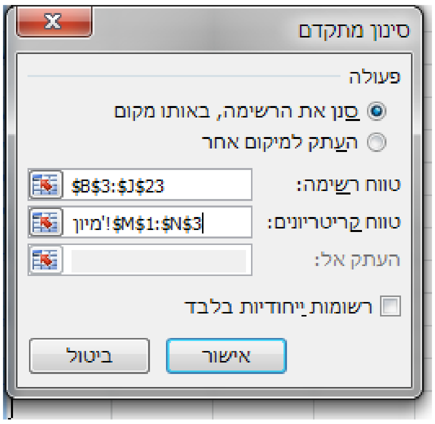
פרטים אודות היולדת						פרטים אודות היילוד		
שם	גיל	לידות קודמות	מ' לידה האם	עליית משקל %	לחץ דם סיסטולי	מין	משקל בלידה	סוג דם
דבורה כהן	19	0	אסיא	9	126	בן	3160	O
דנה ארנון	34	4	ישראל	16	102	בת	3980	O
נילי בן דייו	18	0	ד' אמריקה	11	112	בת	3760	O
אסתר ברוך	40	1	ישראל	12	126	בת	3300	O

שימו לב כי סימן המסנן מופיעה כרגע גם בעמודת "לחץ דם סיסטולי" וגם ב"משקל בלידה" שמשמעו – בוצע סינון של גם וגם על שתי עמודות אלו.

סינון של "או ו/או" [OR, ⋃]

קריטריונים	מ' לידה האם	לחץ דם סיסטולי
	ישראל	
		>95

אם רוצים לקבל רשימה, או לחשב פרמטר סטטיסטי כלשהו של יולדות שמקום הולדת האם הוא ישראל ו/או שלחץ הדם שלהן הוא שווה או גבוה מ- 95 נכין, לצד טבלת הנתונים המקורית, טבלת קריטריונים כמו באיור שמימין. יש להקפיד לרשום את כותרות העמודות עליהן יופעלו הקריטריונים בדיוק מוחלט כפי שהן מופיעות בטבלה הראשית [עדיף לעשות זאת בהעתקה]. לאחר פעולה זאת נפעיל ברצועת הכלים "נתונים" < "סנן" < "מתקדם". מתקבלת תיבה ובה יש למלא [בדקדקנות] את הפרטים הדרושים מילוי:



לפישוט ההדגמה נסמן "סמן את הרשימה, באותו מקום" [כדי שלא לפגוע בנתונים המקוריים רצוי להעתיק את הגיליון ועל הגיליון המועתק לבצע את הסינון], טווח הרשימה הוא טווח טבלת הנתונים המקורית עליה אנו עובדים הכוללת את כותרות העמודות כולן [המלל בשורה הראשונה של הטבלה הוא אינפורמציה בלבד ואין להכלילו בטווח], טווח הקריטריונים הוא הטווח המלא שבו רשומים הקריטריונים [לא חובה התא שבו מצוין "קריטריונים" שהוא למידע עבורנו].

מתקבלת הטבלה הבאה עליה נוכל לבצע חישובים סטטיסטיים:

פרטים אודות היולדת						פרטים אודות היילוד		
שם	**גיל**	**לידות קודמות**	**מ' לידת האב**	**עליית משקל %**	**לחץ דם סיסטולי**	**מין**	**משקל בלידה**	**סוג דם**
יערה קין	24	0	איחפה	12	113	בת	2980	B
דבורה כהן	19	0	אסיה	9	126	בן	3160	O
דנה ארנון	34	4	ישראל	16	102	בת	3980	O
נילי בן דיין	18	0	ד' אמריקה	11	112	בת	3760	O
ג'ודי לאונד	44	2	אמריקה	15	98	בת	3250	A
מרינה בקשטיין	28	1	איחפה	8	105	בן	3300	AB
סמירה שיעב	45	3	ישראל	16	88	בן	3960	O
שולה גביאזדה	33	3	איחפה	25	135	בת	2840	B
קטיבה אלאסי	40	4	אפריקה	8	97	בן	2900	A
אסתר בחך	40	1	ישראל	12	126	בת	3300	O
אפרת דוד	29	0	ישראל	6	128	בן	3170	AB
שלי בר זאב	30	0	ישראל	11	92	בת	3290	O

מ' לידת האב	**לחץ דם סיסטולי**
ישראל	
	>95

פרק 8 – פונקציית התניה (IF)

הנושא עולה בחוברת זו רק כהקדמה ליישום של פונקציות סטטיסטיות מובנות המשולבות עם פונקציות התניה.

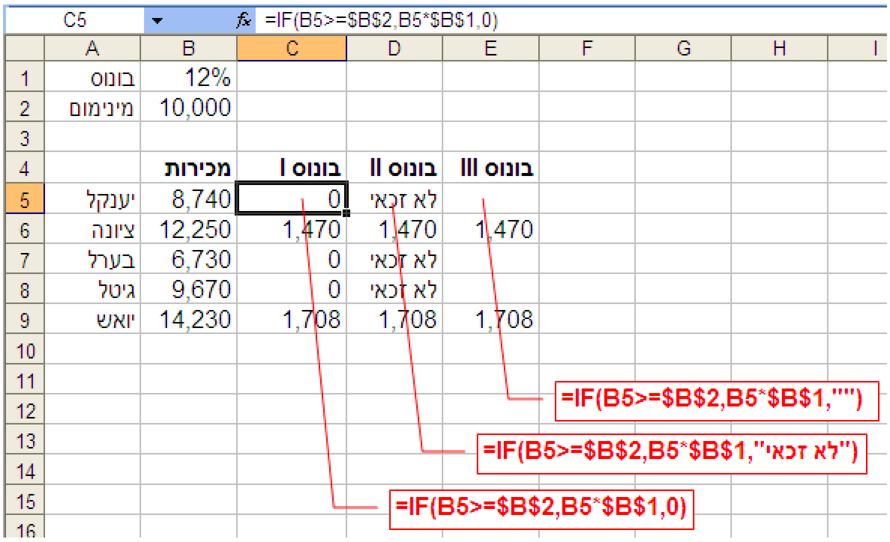
פונקציית התניה היא פונקציה המאפשרת לנו לבצע תִפלול [מניפולציה] על נתונים, בדרך כלל יהיו אלו נתונים בטבלה [מספרים וגם טקסט]. הפונקציה בוחנת תנאי לוגי [או מספר תנאים לוגיים בפונקציית התניה מורכבת], אם התנאי הלוגי מתקיים [TRUE] מתבצעת פעולה מסוימת המוגדרת בפונקציה ואם התנאי הלוגי אינו מתקיים [FALSE] מתבצעת פעולה מסוימת אחרת המוגדרת בפונקציה.

תחביר הפונקציה נראה כך [קרא משמאל לימין]:

=IF(התנאי הלוגי,עשה אם התנאי מתקיים,עשה אם התנאי אינו מתקיים)

הפעולה "עשה" שתתבצע כתוצאה מהחלטת התנאי [בין אם התשובה היא "התנאי מתקיים" ובין אם התשובה היא "התנאי אינו מתקיים", יכולה להיות חישובית, יכולה להיות ערך מספרי כלשהו או טקסט.

נדגים את השימוש בפונקציה IF בדוגמה הבאה המתייחסת לבחינת הזכאות לבונוס לעובדי מכירות כתלות בהיקף המכירות שלהם, ולחישוב הבונוס המגיע להם אם הם אכן זכאים לו. לנוחות [אך אין זו חובה] ולפישוט הנוסחה הוספנו בדוגמה בתא B1 את שיעור הבונוס באחוזים ובתא B2 את הסף התחתון לזכאות לבונוס [הטכניקה של הוצאת פרמטרים אל מחוץ לנוסחה מאפשרת מבחני רגישות מבלי לתקן את הנוסחאות עצמן. מומלץ לאמץ טכניקה זאת במקרים מתאימים].



	A	B	C	D	E
1	בונוס	12%			
2	מינימום	10,000			
3					
4		**מכירות**	**בונוס I**	**בונוס II**	**בונוס III**
5	יענקל	8,740	0	לא זכאי	
6	ציונה	12,250	1,470	1,470	1,470
7	בערל	6,730	0	לא זכאי	
8	גיטל	9,670	0	לא זכאי	
9	יואש	14,230	1,708	1,708	1,708

3 האפשרויות שמוצגות מדגימות לנו שה"עשה" יכול להיות חישוב [במקרה של TRUE], מספר, טקסט, או השארת תא ריק [במקרה של FALSE]. לגבי טקסט, הכלל הוא שיש להתחיל ולסיים את הטקסט במרכאות כפולות ["]. כאשר נקליד זוג מרכאות כפולות ברצף וללא טקסט ביניהן [""] – נקבל כתוצאה תא ריק.

פונקציית התניה (IF) מורכבת

נקראת גם פונקציה מקוננת [NESTING], היא פונקציה המאפשרת לנו לבצע תִפלול [מניפולציה] על נתונים, כאשר אנו נדרשים לבצע יותר מהתניה אחת על אותו משתנה, או כאשר אנו נדרשים להתניה מורכבת של שני משתנים.

בהמשך לדוגמה הקודמת, הוצעה שיטה למתן בונוס פרוגרסיבי - מי שמכירותיו 10,000 או יותר יקבל בונוס בשיעור של 12% בונוס ומי שמכירותיו מעל 13,600 יקבל בונוס בשיעור של 15% [במקום 12%]. יש כאן שתי התניות על אותו משתנה, הפתרון ייראה כך:

C5 | fx =IF(AND(B5>=$B$2,B5<$B$3),B5*$B$1,IF(B5>$B$3,B5*$C$1,"ללא בונוס"))

	A	B	C	D	E	F	G	H	I
1	בונוס	12%	15%						
2	מינימום	10,000							
3	מדרגה	13,600							
4		**מכירות**	**בונוס**						
5	יענקל	8,740	ללא בונוס						
6	ציונה	12,250	1,470						
7	בערל	6,730	ללא בונוס						
8	גיטל	9,670	ללא בונוס						
9	יואש	14,230	2,135						
10									
11	=IF(AND(B5>=$B$2,B5<$B$3),B5*$B$1,IF(B5>$B$3,B5*$C$1,"ללא בונוס"))								
12									

מנהל משאבי אנוש הציע לעודד מוכרים צעירים מתחת לגיל 30 ולתת להם בונוס בשיעור גדול יותר למשל 15%. הפעם יש לנו התניות על שני משתנים. הפתרון ייראה כך:

D5 | fx =IF(AND(C5>=$B$2,B5<30),C5*$B$3,IF(C5>=$B$2,C5*$B$1,0))

	A	B	C	D	E	F	G	H
1	בונוס	12%						
2	מינימום	10,000						
3	בונוס לצעירים	15%						
4		**גיל**	**מכירות**	**בונוס**				
5	יענקל	44	8,740	0				
6	ציונה	24	12,250	1,838				
7	בערל	38	6,730	0				
8	גיטל	56	9,670	0				
9	יואש	43	14,230	1,708				
10								
11	=IF(AND(C5>=$B$2,B5<30),C5*$B$3,IF(C5>=$B$2,C5*$B$1,0))							
12								

פרק 9 – פונקציות סטטיסטיות מובנות משולבות התניות

פונקציות התניה, כפי שראינו בפרק קודם, הן פונקציות המקיימות בדיקה לתנאי, או לתנאים, מוגדר/ים ועל פי העמידה/אי העמידה – מחושבת/מוצגת תוצאה. בפונקציות הסטטיסטיות המשולבות עם פונקציית התניה, מתבצעת הבדיקה [פעמים רבות הבדיקה היא בבחינת "סינון" הנתונים] ולאחריה מחושב פרמטר סטטיסטי על הנתונים שהם תוצאת הבדיקה. בדרך כלל הנתונים שהפונקציה עובדת עליהם הם טבלאיים.

הפונקציה AVERAGEIF

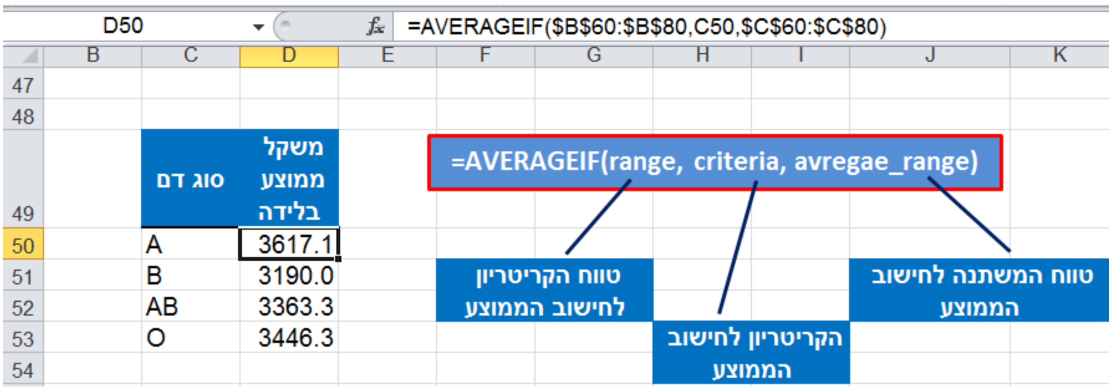
פונקציה זו ובדומה לה הפונקציות STDEV.S, SUMIF, COUNTIF מאפשרת לחשב פרמטר סטטיסטי תחת התניות שונות. נראה כמה דוגמאות:

תרגיל מספר 18

בהמשך לתרגילים הקודמים שהתבססו על מסד נתונים [חלקי] מתוך יומן לידות של בית יולדות "דור העתיד", נדרש לחשב את המשקל הממוצע של היילודים לפי קבוצות של סוג דם, כלומר מה משקלם הממוצע של יילודים שסוג דמם A, B, C ו- D.

פתרון

להלן הפתרון והנוסחאות הנוגעות:



הקריטריון הוגדר כאן ע"י הפניה לתאים C50:C53 [ע"י העתקה], אולם ניתן להקליד את הקריטריון לגוף הנוסחה כשהוא בא בין גרשיים, למשל "A", או ">=3000" אולם במקרה זה עם העתקת הנוסחה מועתק גם הקריטריון ה"מקורי" ואם נדרש צריך לתקנו ידנית.

הפונקציה SUMPRODUCT

הפונקציה מכפילה רכיבים מתאימים במערכים נתונים ומחזירה [מחשבת] את מכפלתם. ניתן לחשב בעזרתה באופן ישיר את הממוצע המשוקלל כאשר השכיחות נתונה כשכיחות יחסית, כאשר השכיחות אינה יחסית יש להוסיף חלוקה בסך השכיחויות.

נראה להלן 3 תרגילים המיישמים פונקציה זאת:

תרגיל מספר 19

לתפקיד ייצוגי בכיר שהתפנה במוסד ציבורי הציגו את מועמדותם 5 מעובדי המוסד. על מנת להגיע לבחינה הוגנת וטובה הוחלט לבחון את המועמדים לפי חמישה קריטריונים: וותק במוסד, השכלה פורמלית, שליטה בשפות זרות, הערכה תקופתית והשכלת בן/בת הזוג של המועמד. לכל קריטריון נקבעה שיטת קביעת הניקוד [שלא תפורט כאן] בו ומשקלו. המועמד העדיף הוא זה שהציון המשוקלל שלו הוא הגבוה ביותר.

פתרון

הטבלה הבאה מפרטת את ניקוד הקריטריונים לכל מועמד, את משקל הקריטריונים ואת הציון המשוקלל:

D102 =SUMPRODUCT(E102:I102,$E$101:$I$101)

קריטריון/משקל	וותק	השכלה	הערכה תקופתית	שפות זרות	השכלת בן הזוג	ציון משוקלל
	2	4	3	3	1	
מרינה בקשטיין	1	8	9	8	8	93.0
יניב בראל	8	5	8	7	8	89.6
אמין אגבריה	5	2	9	8	4	70.0
לאוניד גברניק	9	6	8	9	4	97.0
סמירה שיעב	2	4	9	6	6	70.4

=SUMPRODUCT(E102:I102,$E$101:$I$101)

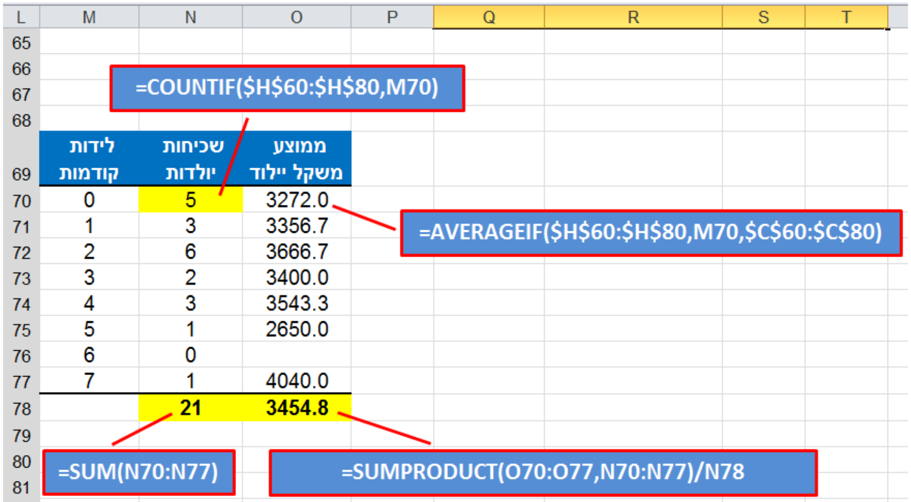
הציון המשוקלל הגבוה ביותר הוא של לאוניד גברניק והוא זה שנבחר לתפקיד.

הערה: הציון המשוקלל אינו בסולם של 1 עד 100 אבל לעניין הבחירה הדבר לא משנה כי חשוב סדר הציון המשוקלל. אם נרצה שהציון המשוקלל יהיה בסולם 1-100 צריך לשקלל את המשקלות או להביאן לשיעור מתוך 100. כיוון שסכום המשקלות הוא 13 אפשר לחשב ולקבוע כי משקל הוותק הוא 2/13x10=1.5846 או 15.3846% וכן הלאה.

תרגיל מספר 20

בהמשך לתרגילים קודמים , אנו רוצים לדעת מה המשקל הממוצע של יילודים לפי מספר לידות קודמות של היולדת, למשל – מה המשקל הממוצע של יילודים ליולדות שלהן 2 לידות קודמות. בהתאם לתוצאה זו מהו המשקל המשוקלל של כל היילודים?

פתרון



לעתים בנוסחאות נדרשת "התערבות" ידנית. בממוצע משקל יילוד [לא רואים כאן את הנוסחה כיוון שהיא כמו זו ששימשה בתרגיל קודם] אם מעתיקים את הנוסחה מתא O77 לטווח O71:O77 מתקבלת בתא O76 הוראה שגיאה "#DIV/0!" ואם לא נתערב ונמחק אותה הפונקציה SUMPRODUCT לא תעבוד.

בפלט המקורי מופיע בתא O76 הסימן #DIV/0! וזאת כיוון שבנוסחת הממוצע יש חלוקה ב- 0 [אין אף אישה שהיו לה 6 לידות קודמות]. כתוצאה מכך אותו סימן מופיע גם בתא O78, מחיקת הסימן המופיע בתא O76 תביא להופעת ערך בתא O78 המייצג את ממוצע הממוצעים שאפשר גם לקרוא לו הממוצע המשוקלל של כלל משקל כל היילודים.

תרגיל מספר 21

במשפטים בבית הדין הגבוה לצדק [בג"צ] צוות השופטים מונה 3 שופטים עד 7 שופטים. כתבת משפט של אחד מערוצי הטלוויזיה רצתה לאזכר, במסגרת כתבה רחבה שהתכוונה לערוך על הבג"צ, את **תוחלת** מספר השופטים במשפטי הבג"צ בחמש השנים האחרונות.

פתרון

הנתונים והחישוב מובאים בטבלה הבאה:

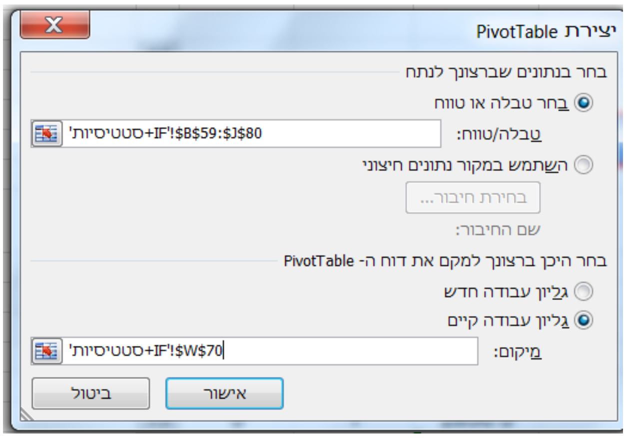
	C	D	E	F	G	H	I	J	K
118									
119		מספר שופטים	הסתברות						
120		X	P(X)	X*P(X)					
121		3	0.2	0.6		=D121*E121			
122		4	0.24	0.96					
123		5	0.28	1.4					
124		6	0.15	0.9					
125		7	0.13	0.91					
126				4.77		=SUM(F121:F125)			
127									
128				4.77		=SUMPRODUCT(D121:D125,E121:E125)			

לפי הנתונים שאספה הכתבת, תוחלת מספר השופטים במשפטי הבג"צ בחמש השנים האחרונות היא 4.77. אפשר כמובן להגיע לתוצאה זו ע"י החישוב בעמודה F וסכימתו בתא F126 ואפשר להגיע לזאת מיידית ע"י הפונקציה SUMPRODUCT כפי שהדבר התבצע בתא F128.

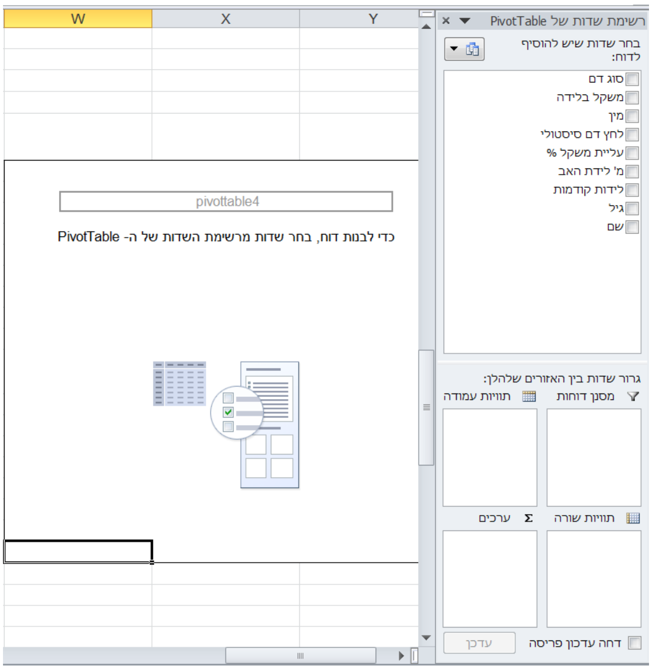
פרק 10 – טבלאות ציר

טבלת ציר [PIVOT TABLE] היא כלי טבלאי אינטראקטיבי המרכז נתוני טבלה, מספריים או טקסטואליים, ומאפשר לארגן אותם בחתכים שונים ולהפיק באופן פשוט ומיידי נתונים סטטיסטיים שונים.

תהליך בניית טבלת הציר מתחיל עם כניסה ללשונית "הוספה", לחיצה על " PIVOT TABLE" ואז נפתחת החלונית הבאה בה יש להגדיר 2 שדות: טווח הטבלה ומיקום הדוח. להגדרת טווח הטבלה מספיק להימצא באחד התאים של הטבלה והאקסל בוחר כברירת מחדל את הטווח [כולל כותרות העמודות] ומסמן את הטווח בקווים מקווקוים, חשוב מאוד לוודא כי אכן סומן הטווח הנכון ואם לא יש לתקנו. בהגדרת מקום היווצרות/הצגת הדוח יש 2 אפשרויות – או בגיליון חדש [ואז צריך לסמן את העיגול שלימינו] או בגיליון הקיים ואז גם צריך לסמן לימינו וגם להגדיר את הפינה השמאלית העליונה [בגיליון שטור A הוא הראשון משמאל] של הדוח.

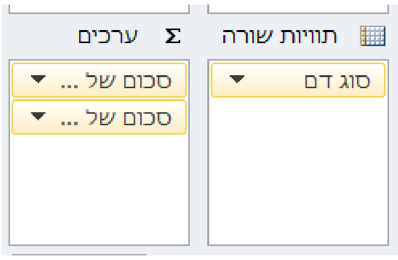


נקיש על "אישור" ונקבל את 2 טבלאות, האחת רשימה של שדות שיש להוסיף לדוח והשנייה תבנית של טבלת הציר [כשהיא ריקה בשלב זה].



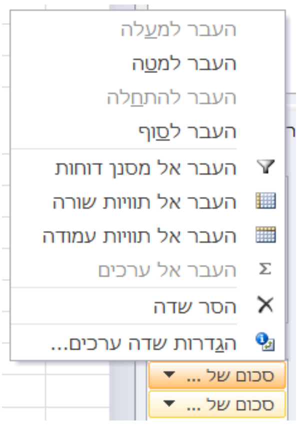
עכשיו אנו צריכים לייצר את הטבלה לפי סוג/מהות הנתונים הסטטיסטיים שאנו רוצים לקבל. נניח שאנו רוצים לבחון האם לסוגי הדם השונים יש השפעה על משקל היילוד ועל שיעור העלייה במשקל [של האם].

נגרור את "סוג הדם" לתיבת תוויות שורה ואת "משקל הלידה" ו"עלייה במשקל %" נגרור לתיבת ערכים. יתקבל מסך כבאיור משמאל:



מייד עם ביצוע גרירה לתיבות כאמור מתהווה הדוח כבטבלה הבאה:

תוויות שורה	סכום של משקל בלידה	סכום של עליית משקל %
A	25320	96
AB	10090	32
B	9570	55
O	27570	98
סכום כולל	**72550**	**281**



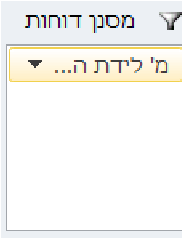
ברירת המחדל של טבלת ציר היא סכום העמודה. הנתון הסטטיסטי שיאפשר לראות האם לסוג הדם השפעה על משקל הלידה ועל העלייה במשקל [בעזרת מבחנים סטטיסטיים מתאימים] הוא ממוצע משקל הלידה. כדי לקבל את הממוצע נלחץ על "סכום של .." נפתחת חלונית בה נבחר ב"הגדרות של שדה ערכים" תיפתח שוב חלונית ובה נבחר באפשרות של "ממוצע". נחזור על פעולה זו עבור כל אחד מהערכים ונקבל את הטבלה הבאה [לאחר שעיצבנו אותה למספרים עם מקום אחד לאחר הנקודה]:

תוויות שורה	ממוצע של משקל בלידה	ממוצע של עליית משקל %
A	3617.1	13.7
AB	3363.3	10.7
B	3190.0	18.3
O	3446.3	12.3
סכום כולל	**3454.8**	**13.4**

הערה: הכיתוב "סכום כולל" המופיע בטבלה אינו במובן הרגיל של "סיכום" והוא מתייחס לכל סוג חישוב שבחרנו [ממוצע, סטיית תקן, ערך מינ', ערך מקס', ספירה וכו'] על כלל האוכלוסייה. בדוגמה שלנו – ממוצע של משקל בלידה של כל 21 היולדות הוא 3454.8 [תוכל לראות נתון זה בחישובים קודמים].

סינון דוחות

טבלת ציר מאפשרת לנו לקבל ממד נוסף לטבלה על ידי שימוש באפשרות של "סינון דוחות".



בהמשך לטבלה שבנינו נגרור את "מ' לידת האב" לתוך מסנן הדוחות ומייד נקבל את הטבלה הבאה:

מ' לידת האב	(הכל)	
תוויות שורה	**ממוצע של משקל בלידה**	**ממוצע של עליית משקל %**
A	3617.1	13.7
AB	3363.3	10.7
B	3190.0	18.3
O	3446.3	12.3
סכום כולל	**3454.8**	**13.4**

עכשיו נקיש על הכפתור שבפינה הימנית של התא "(הכל) מ' לידת האב" ונקבל את כל מקומות הלידה. אם נבחר ב"ישראל" נקבל את ממוצע משקל בלידה של ילודים ואת ממוצע של עליית משקל % בכל סוג דם שבטבלה לילודים שמקום לידת האב הוא בישראל. להלן הטבלה:

מ' לידת האב	ישראל	
תוויות שורה	**ממוצע של משקל בלידה**	**ממוצע של עליית משקל %**
AB	3170.0	6.0
O	3632.5	13.8
סכום כולל	**3540.0**	**12.2**

פרק 11 – מבחני השערות ורווח סמך

מבחן ורווחי סמך על תוחלת של אוכלוסייה μ כאשר השונות ידועה

תרגיל מספר 22

חוקר מתחום החינוך רצה לבחון האם רמת ה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם עולה/שונה/נמוכה מזו של ילדים הלומדים בחינוך הפורמלי. החוקר בחר מדגם אקראי של 12 ילדים בני 7 הלומדים בבית בעזרת ההורים ובחן את רמת ה- IQ שלהם. הנתונים מובאים להלן:

97, 112, 119, 84, 135, 96, 127, 103, 95, 107, 101, 98

בספרות המקצועית פורסם כי הנורמה של רמת ה- IQ אצל ילדים בני 7 מתפלגת נורמלית עם תוחלת של 100 וסטיית תקן השווה ל- 15. האם אפשר לקבוע, ברמת מובהקות של 5% כי רמת ה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם עולה/שונה מזו של ילדים הלומדים בחינוך הפורמלי?

פתרון

בטבלה הבאה בעמודה A מובאים תוצאות מדידת רמת ה- IQ של המדגם. לצד טבלת התוצאות טבלה אשר תאפשר לנו לקבוע האם אכן רמת ה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם עולה/שונה/נמוכה מזו של ילדים הלומדים בחינוך הפורמלי.

הטבלה הבאה מרכזת את הנתונים והחישובים שנעשו[6]:

	A	B	C	D	E
1	רמת IQ במדגם	נתונים/חישובים	סימול	ערך	הפונקציה
2	97	תוחלת IQ באוכלוסיה	μ	100	
3	112	סטית תקן IQ באוכלוסיה	σ	15	
4	119	מספר פריטים במדגם	n	12	
5	84	רמת מובהקות	α	0.05	
6	135	שגיאת התקן באוכלוסיה	$\sigma_{\bar{x}}$	4.33	=D3/D4^0.5
7	96	ממוצע IQ במדגם	$\bar{X}$	106.17	=AVERAGE(A2:A13)
8	127	ערך Z למבחן חד צדדי	Zcritical	1.645	=NORM.S.INV(1-D5)
9	103	ערך Z למבחן דו צדדי	Z'critical	1.96	=NORM.S.INV(1-D5/2)
10	95	ערך קריטי עליון למבחן חד צדדי	K	**107.12**	=D2+D8*D6
11	107	ערך קריטי עליון למבחן דו צדדי	K_2	**108.49**	=D2+D9*D6
12	101	ערך קריטי תחתון למבחן דו צדדי	K_1	**91.51**	=D2-D9*D6
13	98				
14					מקרא -
15		נתוני התרגיל		חישובים ראשונים	חישובי ערכים קריטיים

6 בחוברת זו נקטנו בגישה לפיה קבלת החלטה על קבלה או דחייה של השערת האפס תעשה בדרך של חישוב ערך קריטי והשוואתו עם ערך המחושב. בחירה זאת נעשתה משום שהתוצאות יכולות לשמש מיידית לחישוב רווח בר סמך. גישה שנייה היא להשוות את ערך Z המחושב עם ערך קריטי של Z.

עכשיו נבדוק 2 השערות:

- האם רמת ה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם <u>עולה</u> על זו של ילדים הלומדים בחינוך הפורמלי?

השערת האפס היא שה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם שווה לזו של ילדים הלומדים בחינוך הפורמלי.

מבחינה פורמלית ייכתבו ההשערות כך: $H_1 = \bar{X} > \mu \quad H_O : \bar{X} = \mu$. השערת האפס תתקבל אם $\bar{X} < K$ כיוון שמתקיים $\bar{X} = 106.17 < K = 107.12$ אין לנו אלא להחליט שאי אפשר לדחות את השערת האפס והמשמעות המילולית היא: ה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם אינה עולה על זו של ילדים הלומדים בחינוך הפורמלי.

- האם רמת ה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם <u>שונה</u> מזו של ילדים הלומדים בחינוך הפורמלי?

מבחינה פורמלית ייכתבו ההשערות כך: $H_1 = \bar{X} \neq \mu \quad H_O : \bar{X} = \mu$. השערת האפס תתקבל אם $K_2 < \bar{X} < K_1$ כיוון שמתקיים $91.513 = K_2 < \bar{X} = 106.17 < K_1 = 108.49$ אין לנו אלא להחליט שאי אפשר לדחות את השערת האפס והמשמעות המילולית היא: ה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם אינה שונה מזה של ילדים הלומדים בחינוך הפורמלי.

מבחן ורווחי סמך על תוחלת של אוכלוסייה μ כאשר השונות לא ידועה

תרגיל מספר 23

נשתמש בנתוני תרגיל מספר 22 תוך ביצוע שינוי קל: חוקר מתחום החינוך רצה לבחון האם רמת ה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם עולה/שונה/נמוכה מזו של ילדים הלומדים בחינוך הפורמלי. החוקר בחר מדגם אקראי של 12 ילדים בני 7 הלומדים בבית בעזרת ההורים ובחן את רמת ה- IQ שלהם. הנתונים מובאים להלן:

97, 112, 119, 84, 135, 96, 127, 103, 95, 107, 101, 98

בספרות המקצועית פורסם כי הנורמה של רמת ה- IQ אצל ילדים בני 7 מתפלגת נורמלית עם תוחלת של 100 [שים לב! סטיית התקן של האוכלוסייה אינה ידועה].

פתרון

אפשר לפתור את התרגיל בעזרת טבלה דומה לזו שבה נעשה שימוש בתרגיל קודם [אלא שהפעם יהיה צורך לחשב את **אומד סטיית התקן** של המדגם ולהשתמש בו לחישוב K, K_1 ו- K_2]. אולם, התרגיל יכול להיפתר ביתר קלות תוך שימוש בפונקציה המובנית **DESCRIPTIVE STATISTICS** שהיא כלי נוסף מתוך ה- **DATA ANALYSIS**. [ראה בחומר הקודם]. לעניין מבחני השערות, חשוב להדגיש ולהזכיר כי בכלי זה הפלט מחושב תוך הנחה שהנתונים הם של מדגם [ולא אוכלוסייה], על כן מה שמחושב ומוצג כ- Standard Deviation הוא למעשה **אומד** לסטיית התקן ומה שמחושב ומוצג כ- Confidence Level הוא מחצית מאורך רווח הסמך המחושב בעזרת ערך ה- t המתאים לדרגות החופש ולרמת הביטחון אותה נקליד בפונקציה. נזכיר כי רמת ביטחון של 90% [צד שמאל בפלט שלהלן] מתאימה למבחן חד צדדי ברמת מובהקות של 5%. מחצית אורך רווח הסמך חושבה לפי $t_{11,0.95}$.

נראה להלן שני פלטים חלקיים [לנוחות ובהירות הושמטו פרמטרים לא רלוונטיים] שיעזרו לנו לפתור את התרגיל בקלות רבה:

תחשיבים לצורך מבחן חד צדדי	
Mean	106.17
Standard Error	4.2388
Standard Deviation	14.684
Count	12
Confidence Level(90.0%)	7.61

תחשיבים לצורך מבחן דו צדדי	
Mean	106.17
Standard Error	4.2388
Standard Deviation	14.684
Count	12
Confidence Level(95.0%)	9.33

הדבר שיש להקפיד, בעת השימוש בפונקציה, הוא קביעה נכונה של רמת הסמך [Confidence Level] אותה יש לרשום. במבחן החד צדדי, רמת המובהקות α=5% המתאימה למחצית רווח הסמך של 45% ולרמת סמך של 90% [לכן בחישוב, שנעשה על ידי הפונקציה, נלקח ערך t המתאים לזנב ימני של 5% עם 11 דרגות חופש].

באופן דומה במבחן הדו צדדי, ערך ה- t המתאים צריך להיות של מחצית רמת המובהקות α'=2.5% המתאימה למחצית רווח סמך של 47.5% ולרמת סמך של 95%, [לכן בחישוב שנעשה על ידי הפונקציה נלקח ערך t המתאים לזנב ימני של 2.5% עם 11 דרגות חופש].

עכשיו נבדוק 2 השערות:

- האם רמת ה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם **עולה** על זו של ילדים הלומדים בחינוך הפורמלי?

נחשב את K:

$$K = 100 + 7.61 = 107.61$$

מבחינה פורמלית ייכתבו ההשערות כך:

$$H_O : \bar{X} = \mu \quad H_1 = \bar{X} > \mu$$

השערת האפס תידחה אם $\bar{X} > K$ כיוון שמתקיים $\bar{X} = 106.17 < K = 107.61$ אין לנו אלא להחליט שאי אפשר לדחות את השערת האפס והמשמעות המילולית היא: ברמת מובהקות של 5% ה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם **אינה עולה** מזו של ילדים הלומדים בחינוך הפורמלי.

- האם רמת ה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם **שונה** מזו של ילדים הלומדים בחינוך הפורמלי?

נחשב את K1 ואת K2:

$$K_1 = 100 - 9.33 = 90.67 \text{ ו- } K_1 = 100 + 9.33 = 109.33$$

מבחינה פורמלית ייכתבו ההשערות כך:

$$H_O : \bar{X} = \mu \quad H_1 = \bar{X} \neq \mu$$

השערת האפס תידחה אם $K_2 < \bar{X} < K_1$

כיוון שמתקיים $109.33 = K_2 > \bar{X} = 106.17 > K_1 = 90.67$ אין לנו אלא להחליט שאי אפשר לדחות את השערת האפס והמשמעות המילולית היא: ברמת מובהקות של 5% ה- IQ אצל ילדים הלומדים בבית בעזרת הוריהם **אינה שונה** מזה של ילדים הלומדים בחינוך הפורמלי

מבחן ורווחי סמך על תוחלות של שתי אוכלוסיות כאשר השונויות לא ידועות והמדגמים בלתי תלויים

תרגיל מספר 24

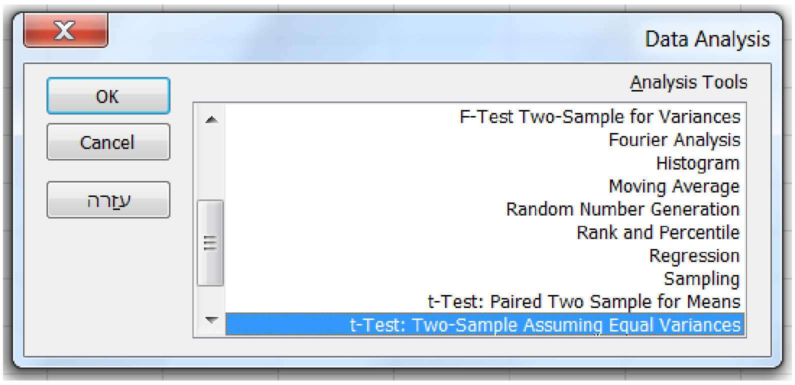
במטרה לבחון האם יש הבדל מובהק בכושר הגופני בין יחידות "מיוחדות" [להלן כוח א'] לבין יחידות שדה רגילות [להלן כוח ב'], בחר אגף כושר קרבי מדגם אקראי של חיילים מכל אחת מן היחידות וערך להם מבחן כושר זהה, על הליכון, בתנאי מעבדה. במבחן הכושר נמדדו מספר הקילומטרים שעבר כל חייל עד אשר הגיע הדופק שלו ל- 156 [אז גם הופסק המבחן לחייל]. תוצאות מבחן הכושר מובאות להלן:

כוח א'	כוח ב'
21	12
18	14
14	10
20	8
11	16
19	5
8	3
12	9
13	11
15	

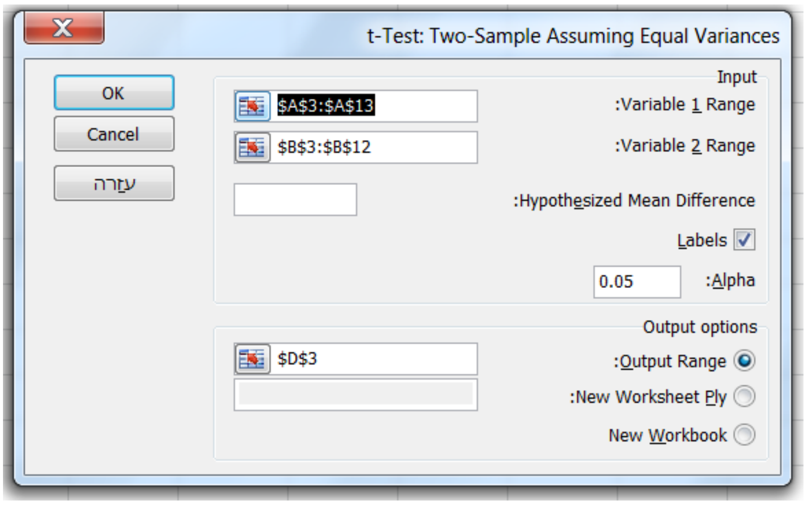
האם אפשר לקבוע ברמת מובהקות של 5% **שיש הבדל** בכושר הגופני של שתי היחידות?

פתרון

הניתוח נעשה בעזרת הפונקציה **"t-Test: Two-Sample Assuming Equal Variances"** אליה מגיעים במסלול הבא: נתונים > Data Analysis > t-Test: Two-Sample Assuming Equal Variances, מתקבל המסך הבא:



כאשר נקיש OK ויתקבל המסך הבא:



נמלא את הפרטים הנדרשים לפתרון. בהתייחס לרמת המובהקות, בשונה מהמקרה של מבחן ורווחי סמך על תוחלת של אוכלוסייה μ כאשר השונות לא ידועה שבו יש צורך להפעיל שיקול בשאלה איזה ערך נציב ב- Confidence Level כך שיתאים למבחן חד צדדי או דו צדדי, כאן התוצאה מתייחסת לשני המצבים. נקבל את הפלט הבא שבו תוצאות מבחן הכושר לצד התוצאות מובא ניתוח:

	A	B	C	D	E	F
1	כוח א'	כוח ב'		t-Test: Two-Sample Assuming Equal Variances		
2	21	12				
3	18	14			כוח א'	כוח ב'
4	14	10		Mean	15.1	9.7778
5	20	8		Variance	18.3222	16.9444
6	11	16		Observations	10	9
7	19	5		Pooled Variance	17.6739	
8	8	3		Hypothesized Mean Difference	0	
9	12	9		df	17	
10	13	11		t Stat	2.7553	
11	15			P(T<=t) one-tail	0.0068	
12				t Critical one-tail	1.7396	
13				P(T<=t) two-tail	0.0135	
14				t Critical two-tail	2.1098	

פיענוח [אינטרפרטציה] של הנתונים [תוך התייחסות לפרמטרים חדשים]:

Pooled Variance – אומד לשונות המשוקללת של 2 הקבוצות בהתאם לדרגות החופש [df] של כל אחת מהן לפי הנוסחה הבאה:

$$S^2_{pooled} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Hypothesized Mean Difference – ההשערה על גודל ההפרש בין ממוצעי הקבוצות. בתרגיל השערת האפס הייתה $H_0: \mu_1 - \mu_2 = 0$ אבל אפשר כמובן לבחון השערה על הפרש שונה מ- 0.

df – דרגות החופש של המבחן כולו המחושבות לפי $n_1 + n_2 - 2$.

t Stat – מוכר ומכונה במקרים רבים כ- tcalc. הערך המחושב של t ע"פ הנוסחאות שבפתיח הפרק. ערך זה יושווה לערכים הקריטיים של t כדי לקבוע האם לקבל או לדחות את השערת האפס.

P(T<=t) one-tail – היא אותה רמת מובהקות עבורה, עבור מבחן חד צדדי, שבה אי אפשר היה לדחות את השערת האפס. אם רמת המובהקות בתרגיל זה הייתה 0.68% או פחות [למשל 0.5%] – השערת האפס הייתה מתקבלת.

t Critical one-tail – הערך הקריטי [בערכים מוחלטים] של t למבחן חד צדדי. הערך נקבע ע"י רמת המובהקות α וע"י דרגות החופש df. בתרגיל זה $t_{critical}=t_{0.95,17} = 2.1098$. ערך זה משמש לכלל ההחלטה לגבי קבלת/דחיית השערת האפס – הכלל הוא קבל את השערת האפס אם **|t Stat|≤t Critical** אחרת דחה את השערת האפס.

P(T<=t) two-tail – הוא אותה רמת מובהקות [α] עבורה, עבור מבחן דו צדדי, שבה אי אפשר היה לדחות את השערת האפס. אם רמת המובהקות בתרגיל זה הייתה 1.35% או פחות [למשל 1%] – השערת האפס הייתה מתקבלת.

t Critical two-tail – הערך הקריטי [בערכים מוחלטים] של t למבחן דו צדדי. הערך נקבע ע"י רמת המובהקות α וע"י דרגות החופש df. בתרגיל זה $t_{critical}=t_{0.975,17}=1.7396$. ערך זה משמש לכלל ההחלטה לגבי קבלת/דחיית השערת האפס – הכלל הוא קבל את השערת האפס אם **t Stat≤t Critical** אחרת דחה את השערת האפס.

בתשובה לשאלה שנשאלה " האם אפשר לקבוע ברמת מובהקות של 5% **שיש הבדל** בכושר הגופני של שתי היחידות?" התשובה היא **השערת האפס תידחה** ואפשר להסביר זאת [לפי הפלט] בשתי צורות:

כפי שצוין לעיל, הכלל הוא קבל את השערת האפס אם **|t Stat|≤t Critical** אחרת דחה את השערת האפס. כיון שמתקיים: $t_{stat.} = 2.7553 > t_{critical} = 2.1098$ אין לנו אלא להחליט שאפשר לדחות את השערת האפס והמשמעות המילולית היא: ברמת מובהקות של 5% **יש הבדל** בכושר הגופני של שתי היחידות. יתרה מכך, אפשר לקבוע כי הכושר הגופני של כוח א' גבוה יותר מזה של כוח ב'.

הדרך השנייה היא בעזרת תוצאת **P(T<=t) two-tail** השווה ל- 0.0135 [1.35%], כאשר רמת המובהקות הנדרשת שווה או קטנה מערך זה – אי אפשר היה לדחות את השערת האפס. רמת המובהקות בתרגיל זה היא 5% ולפיכך אי אפשר לקבל את השערת האפס.

מבחן ורווחי סמך על הפרש תוחלות של שתי אוכלוסיות נורמליות כאשר המדגמים מזווגים

תרגיל מספר 25

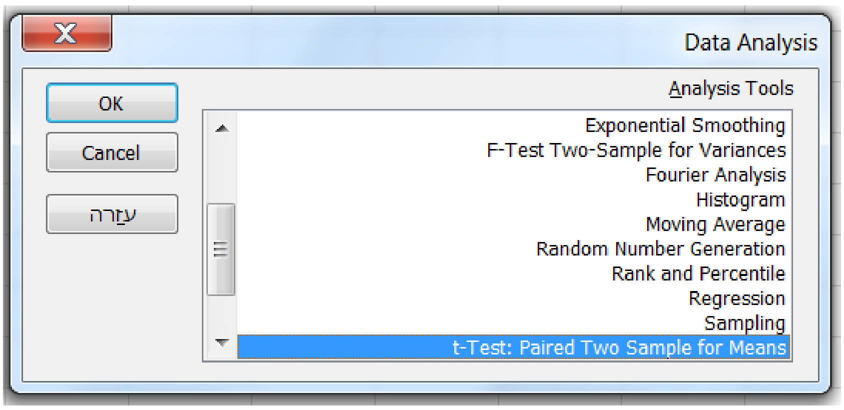
מדריך ליוגה חזר מהשתלמות ארוכה בנושא "יישומי יוגה לשיפור תפקוד קוגניטיבי". המדריך טען שפיתח סדרת תרגילים המיועדת לתלמידים בגילאים 12-18, אותם יש לבצע שעתיים קודם למועד המבחן, המשפרים את ביצועי המבחן [לכל התלמידים – גם כאלו שהישגיהם טובים]. לבדיקת טענת המדריך נלקח מדגם אקראי של תלמידי כיתה י' מבית הספר "לומדים ונהנים" אשר נבחנו יומיים קודם לכן במקצוע היסטוריה בת זמננו והם נבחנו שוב באותו נושא במבחן שנבדק ונקבע לגביו שהוא בעל רמת קושי מקבילה למבחן הקודם כאשר שעתיים קודם לתחילת המבחן הם סדרת תרגילי יוגה. תוצאות שני המבחנים מובאות להלן:

ציון אחרי	ציון לפני
69	65
81	78
68	56
79	76
59	47
90	88
74	67
77	74
65	61
63	56

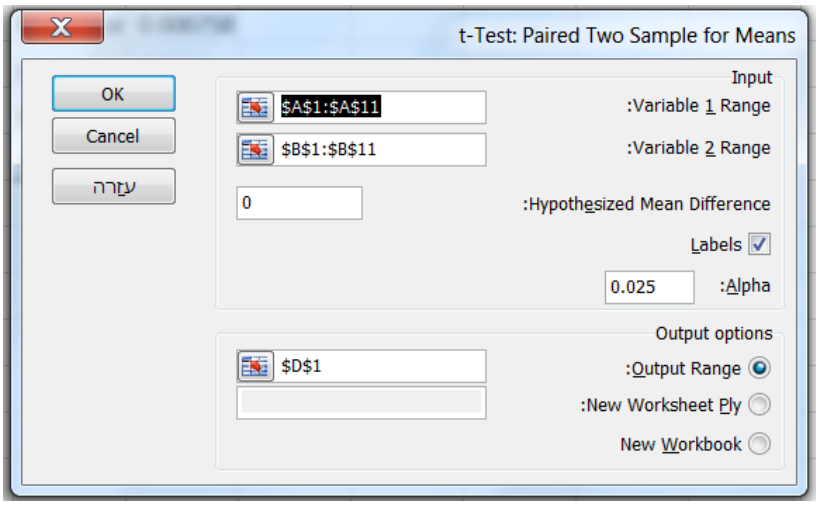
האם אפשר לקבוע ברמת מובהקות של 2.5% אפשר לקבל את טענת מדריך היוגה?

פתרון

הניתוח נעשה בעזרת הפונקציה **"t-Test: Paired Two Sample for Means"** אליה מגיעים במסלול הבא: נתונים > Data Analysis >t-Test: Paired Two Sample for Means , מתקבל המסך הבא:



כאשר נקיש OK ויתקבל המסך הבא:



נמלא את הפרטים הנדרשים לפתרון. בהתייחס לרמת המובהקות [Alpha] יש להפעיל שיקול דעת באשר לערך רמת המובהקות אם וכאשר עורכים מבחן דו צדדי כאשר השאלה היא "האם יש הבדל" או כאשר עורכים מבחן חד צדדי כאשר השאלה, כמו בדוגמה שבתרגיל זה, היא "האם חל שיפור". נקבל את הפלט הבא שבו תוצאות מבחן הכושר לצד התוצאות מובא ניתוחן:

	A	B	C	D	E	F
1	ציון לפני	ציון אחרי		**t-Test: Paired Two Sample for Means**		
2	65	69				
3	78	81			*ציון לפני*	*ציון אחרי*
4	56	68		Mean	66.8	72.5
5	76	79		Variance	152.6	89.4
6	47	59		Observations	10	10
7	88	90		Pearson Correlation	0.98	
8	67	74		Hypothesized Mean Difference	0	
9	74	77		df	9	
10	61	65		t Stat	-4.854	
11	56	63		P(T<=t) one-tail	0.000	
12				t Critical one-tail	2.262	
13				P(T<=t) two-tail	0.001	
14				t Critical two-tail	2.685	

פיענוח [אינטרפרטציה] של הנתונים [תוך התייחסות לפרמטרים חדשים]:

Mean, Variance – ממוצע וסטיית התקן של תוצאות המבחן הראשון "ציון לפני" וממוצע וסטיית התקן תוצאות המבחן השני "ציון אחרי" שלפני התחלתו בצעו הנבחנים תרגילי יוגה. חשוב לציין כי הפונקציה לא "מתרשמת" מהכיתוב "לפני" ו"אחרי" ובחישוביו היא יוצרת עמודת הפרשים [D] השווה ל- $X_A - X_B$ לכן בעת ההגעה למסקנות נצטרך להפעיל שיקול דעת ובמקרה זה השאלה תתהפך ותהיה האם אפשר לקבוע כי הציון

לפני הוא נמוך יותר. אפשרות נוספת [אם אנחנו רוצים להתאים את החישוב ל"ראש שלנו"] הוא לרשום את תוצאות הציון אחרי בעמודה A ואת תוצאות הציון לפני בעמודה B.

Hypothesized Mean Difference – ההשערה על גודל ממוצע ההפרשים של n תצפיות מזווגות – $\overline{D}$. בתרגיל השערת האפס הייתה $H_0 : \mu_D = 0$ אבל אפשר כמובן לבחון השערה על הפרש שונה מ- 0.

Pearson Correlation – מקדם המתאם של פירסון [הנושא ייסקר בהמשך החוברת] המבטא את עוצמת הקשר [במקרה זה לא את הכיוון] בערכים מוחלטים. הערך 0.98 מצביע על מקדם מתאם [קשר] גבוה ביותר בין הציון "לפני" לבין הציון "אחרי" אבל אין בזה דבר כדי להצביע על נכונות השערת האפס.

df – דרגות החופש של המבחן המחושבות לפי $n-1$ כאשר n מתייחס למספר הכולל של זוגות התצפיות.

t Stat – מוכר ומכונה במקרים רבים כ- tcalc. הערך המחושב של t ע"פ הנוסחאות שבפתיח הפרק. ערך זה יושווה לערכים הקריטיים של t כדי לקבוע האם לקבל או לדחות את השערת האפס.

P(T<=t) one-tail – הוא אותה רמת מובהקות עבורה, עבור מבחן חד צדדי, שבה אי אפשר היה לדחות את השערת האפס. אם רמת המובהקות בתרגיל זה הייתה 0.05% או פחות [למשל 0.01%] – השערת האפס הייתה מתקבלת.

t Critical one-tail – הערך הקריטי [בערכים מוחלטים] של t למבחן חד צדדי. הערך נקבע ע"י רמת המובהקות α וע"י דרגות החופש df. בתרגיל זה 4.854- = $t_{critical}=t_{0.975,9}$.
ערך זה משמש לכלל ההחלטה לגבי קבלת/דחיית השערת האפס – הכלל הוא קבל את השערת האפס אם **t Critical**≥|**t Stat**| אחרת דחה את השערת האפס.

P(T<=t) two-tail – היא אותה רמת מובהקות עבורה, עבור מבחן דו צדדי, שבה אי אפשר היה לדחות את השערת האפס. אם רמת המובהקות בתרגיל זה הייתה 0.09% או פחות [למשל 0.05%] – השערת האפס הייתה מתקבלת.

t Critical two-tail – הערך הקריטי [בערכים מוחלטים] של t למבחן דו צדדי. הערך נקבע ע"י רמת המובהקות α וע"י דרגות החופש df. בתרגיל זה $t_{critical}=t_{0.9875,17}=2.262$.
ערך זה משמש לכלל ההחלטה לגבי קבלת/דחיית השערת האפס – הכלל הוא קבל את השערת האפס אם **t Critical**≥|**t Stat**| אחרת דחה את השערת האפס.

בתשובה לשאלה שנשאלה בתרגיל השערת האפס תידחה בשל העובדה שמתקיים:

t Stat = 4.8541 ≥ t Critical =2.262

אכן, ברמת מובהקות של 2.5% אפשר לקבל את טענת מדריך היוגה לפיה ביצוע תרגילי יוגה [מסוימים] שעתיים קודם לתחילת מבחן תביא לשיפור בציוני הנבחנים.

פרק 12 – רגרסיה ליניארית בשיטת הריבועים הפחותים ומתאם

תרגיל מספר 26

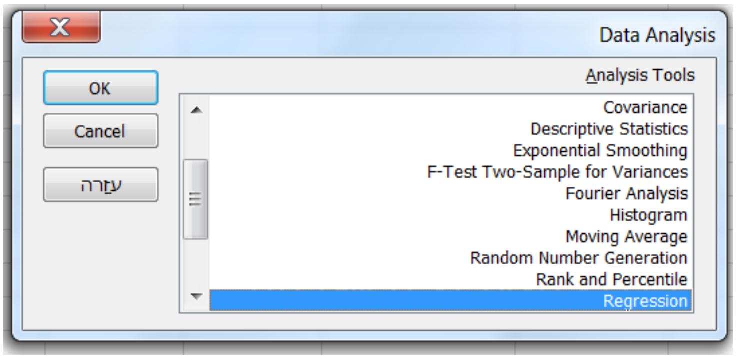
רשת פיצות בארה"ב תכננה פתיחת סניף חדש של הרשת. מנהלי הפיתוח של הרשת חשבו כי מיקום הסניף החדש בקרבת מתחם מכללה הוא רעיון טוב. הנהלת הרשת העריכה כי מספר הסטודנטים במכללה קשור/משפיע על היקף המכירות השנתי של סניף מסעדה הממוקם סמוך למתחם המכללה. כדי לבדוק את הקשר בין מספר הסטודנטים במכללה [X] לבין היקף המכירות השנתי של מסעדה הממוקמת ליד קמפוס המכללה נלקח מדגם אקראי של 10 מסעדות הממוקמות בסמיכות למתחמי מכללות ונבדק היקף המכירות השנתי שלהן [Y]. תוצאות המדגם מובאות להלן:

מסעדה מספר	מספר סטודנטים במכללה באלפים	מכירות שנתיות באלפי $
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

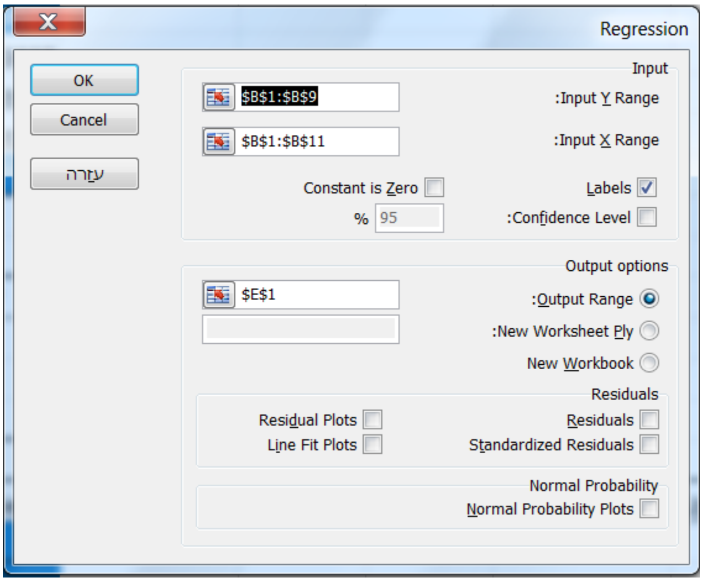
קבע את משוואת הרגרסיה הליניארית אשר תאפשר לנבא את היקף המכירות השנתי של מסעדה שתיפתח בסמוך למתחם מכללה. קבע את מקדם המתאם שבין מספר הסטודנטים במכללה לבין המכירות השנתיות. האם לאור התוצאות כדאי לרשת להקים סניף חדש ליד מתחם מכללה?

פתרון

הניתוח נעשה בעזרת הפונקציה **"Regression"** אליה מגיעים במסלול הבא: נתונים < Data Analysis <Regression, מתקבל המסך הבא:



כאשר נקיש OK ויתקבל המסך הבא:



נמלא את הפרטים הנדרשים לפתרון. חשוב להקדיש מחשבה לשאלה איזה משתנה הוא המסביר/המשפיע – X ואיזה משתנה הוא המשתנה המושפע/הנאמד – Y, לעניין מקדם המתאם אין לכך חשיבות אולם לעניין משוואת הרגרסיה הדבר חשוב. לאחר הקשה על OK נקבל את הפלט הבא שבו הודגשו הנתונים החשובים לנו [ברמה הנלמדת] לצד התוצאות מובא ניתוח:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	**0.9501**							
R Square	0.9027							
Adjusted R Square	0.8906							
Standard Error	13.8293							
Observations	**10**							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	14200	14200	74.248	2.54887E-05			
Residual	8	1530	191.25					
Total	9	15730						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	**60**	9.226	6.503	0.00019	38.725	81.275	38.725	81.275
מספר סטודנטים במכללה באלפא	**5**	0.580	8.617	0.00003	3.662	6.338	3.662	6.338

פיענוח [אינטרפרטציה] של חלק מהנתונים:

Multiple R – הוא מקדם המתאם [Correlation], בערכו המוחלט, בין המשתנה X [המסביר] לבין המשתנה Y [הנאמד]. השימוש במילה **Multiple** נובע מהעובדה שפונקציה דומה באקסל יודעת לחשב את משוואת הרגרסיה ואת מקדם המתאם כאשר יש 2 משתנים מסבירים.

R Square – או R^2 הוא מקדם המתאם בריבוע. ערך זה מייצג את שיעור ההשתנות של המשתנה Y המוסבר ע"י המשתנה X.

Adjusted R Square – הוא הערכה טובה יותר לשיעור ההשתנות של המשתנה Y המוסבר ע"י המשתנה [הערכה המביאה בחשבון גם את גודל המדגם].

ANOVA [Analysis of Variance] – קבוצת נתונים/פרמטרים המתייחסת לניתוח שונות. בנתונים אלו נתייחס בעיקר לנתון ***Significance F***, לא נעמיק על דרך חישובו אולם ככול שערכו נמוך מ- 0.05 כך אפשר לקבוע שמודל הרגרסיה כולו אמין ומובהק.

Coefficients – קבוצת נתונים/פרמטרים המתייחסת למקדמי משוואת הרגרסיה הליניארית ובדיקת השערות עליהם [בדיקת "חוזקם" ו"מהימנותם"]. גם כאן נתייחס בעיקר לנתונים של ***P-value***, לא נעמיק על דרך חישובם אולם ככול שערך כל אחד מהם נמוך מ- 0.05 כך אפשר לקבוע שהחותך [**Intercept**] או השיפוע [**SLOPE**] מובהקים.

Intercept – או "החותך", הערך a במשוואת הרגרסיה $\hat{Y} = a + bX$ שמשמעו נקודת חיתוך קו הרגרסיה את ציר ה- Y [במילים אחרות – אומד של Y עבור X=0].

מספר סטודנטים במכללה באלפים – או השיפוע [**SLOPE**], הערך b במשוואת הרגרסיה.

לאור הנתונים, משוואת הרגרסיה היא:

$$\hat{Y} = 60 + 5X$$

מקדם המתאם r = 0.9501, יש לשים לב לכך שאת כיוונו של מקדם המתאם [שלילי/חיובי] הוא ככיוונו של השיפוע [b].

באשר לשאלה האם לאור התוצאות כדאי לרשת להקים סניף חדש ליד מתחם מכללה? על פניו נראה כי אכן כדאי לרשת להקים סעיף חדש ליד מתחם מכללה. יחד עם זאת ולמרות מקדם המתאם הגבוה במיוחד והעובדה שניתוח מובהקות מקדמי המדגם מצביע על היותם מובהקים, **עדיין הקשר בין "סיבה" ל"מסובב" אינו ברור חד משמעית, ייתכן שיש משתנה נוסף, משתנה מתערב, או משתנה מסביר שלא הובא בחשבון**, למשל מספר המסעדות הקיימות כבר בקרבת מתחם המכללה.

אם רוצים לקבל את משוואת הרגרסיה ואת מקדם המתאם בלבד, אפשר לעשות זאת בעזרת הפקודות הבדידות הבאות:

H15

	B	C	D	E	F	G
1	מספר סטודנטים במכללה באלפים	מכירות שנתיות באלפי $				
2	2	58	0.950122955 =	=CORREL(B1:B11,C1:C11)		
3	6	105				
4	8	88	60 =	=INTERCEPT(C1:C11,B1:B11)		
5	8	118				
6	12	117	5 =	=SLOPE(C1:C11,B1:B11)		
7	16	137				
8	20	157				
9	20	169				
10	22	149				
11	26	202				

תרגיל מספר 27

בהמשך לתרגיל 26 ולאחר שאנשי הכספים של הרשת ראו את התוצאות, הוחלט לחזור ל- 10 המתחמים שנבדקו, לבדוק את מספר בתי האוכל שבמתחם ולהוסיפם כמשתנה נוסף ברגרסיה ולחשב משוואת רגרסיה דו משתנית. מה תהינה תוצאות הבדיקה [מקדם המתאם, משוואת הרגרסיה ומשמעם] ביחס להחלטה על פתיחת סניף חדש?

פתרון

תהליכי העבודה יהיו זהים לאלו שנעשו בתרגיל 26 אלא שבשדה "Input X Range" יש להוסיף את הטווח הכולל גם את X1 וגם את X2.

להלן פלט [הפעם חלקי] הכולל את נתוני המקור ואת התוצאות:

מעסדה מספר	מספר סטודנטים במכללה באלפים - X1	מספר בתי אוכל במתחם - X2	מכירות שנתיות באלפי $ - Y
1	2	4	58
2	6	4	105
3	8	7	88
4	8	5	118
5	12	5	117
6	16	6	137
7	20	5	157
8	20	4	169
9	22	4	149
10	26	4	202

SUMMARY OUTPUT

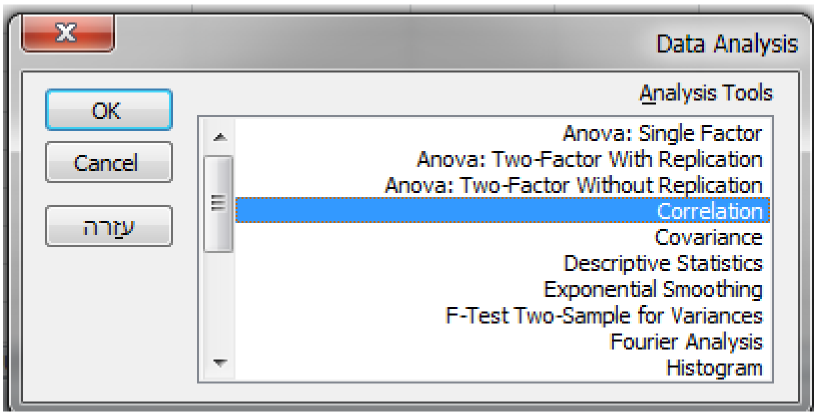
Regression Statistics	
Multiple R	0.9534
R Square	0.9090
Adjusted R Square	0.8830
Standard Error	14.3010
Observations	10
	Coefficients
Intercept	77.03
מספר סטודנטים במכללה באלפים - X1	4.91
מספר בתי אוכל במתחם - X2	-3.28

לאור הנתונים, משוואת הרגרסיה היא:

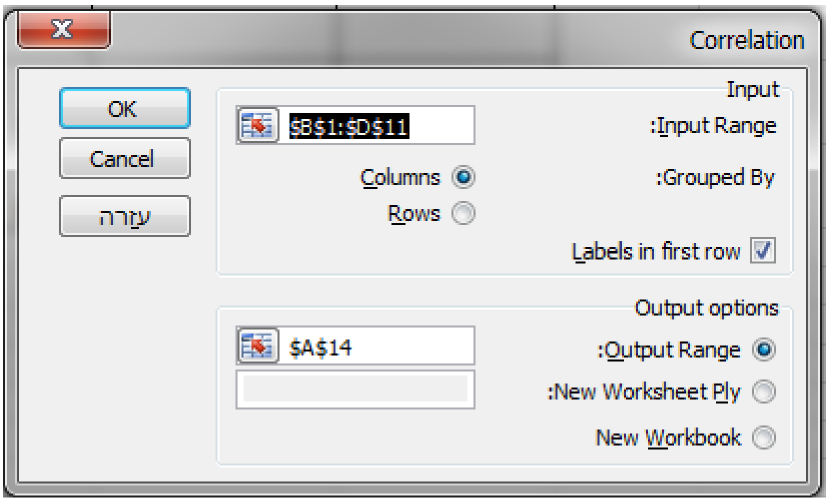
$$\hat{Y} = 77.03 + 4.91X_1 - 3.28X_2$$

מקדם המתאם המרובה, המושפע משני המשתנים המסבירים/המנבאים, שווה בערכו המוחלט ל- r=0.9534. יש לשים לב לכך שאי אפשר לקבוע [מהפלט] את כיוונו הקשר המשולב [שלילי/חיובי] אלא אם נחשב את ערכו של $\hat{Y}$ עבור זוגות של ערכי X1 ו- X2.
אולם בעזרת **מטריצת מקדם מתאם** שאת חישובה נראה להלן – אפשר לראות את מקדם המתאם, הן בגודלו והן בכיוונו, בין כל שני משתנים:

הניתוח נעשה בעזרת הפונקציה **"Correlation"** אליה מגיעים במסלול הבא: נתונים > Data Analysis >Correlation, מתקבל המסך הבא:



כאשר נקיש OK ויתקבל המסך הבא שבו נמלא את השדות כפי שנראה להלן:



נקיש OK ונקבל את הפלט הבא [הפלט מביא רק את מטריצת מקדמי המתאם, כאן בחרנו להביא גם את נתוני המקור]:

	A	B	C	D
1	מסעדה מספר	מסר סטודנטים במכללה באלפים - X1	מספר בתי אוכל במתחם - X2	מכירות שנתיות באלפי $ - Y
2	1	2	4	58
3	2	6	4	105
4	3	8	7	88
5	4	8	5	118
6	5	12	5	117
7	6	16	6	137
8	7	20	5	157
9	8	20	4	169
10	9	22	4	149
11	10	26	4	202
12				
13				
14		מסר סטודנטים במכללה באלפים - X1	מספר בתי אוכל במתחם - X2	מכירות שנתיות באלפי $ - Y
15	מסר סטודנטים במכללה באלפים - X1	1		
16	מספר בתי אוכל במתחם - X2	-0.217	1	
17	מכירות שנתיות באלפי $ - Y	0.950	-0.283	1

אפשר לראות כי מקדם המתאם שבין מספר הסטודנטים במכללה לבין מכירות שנתיות באלפי $ הוא גבוה מאוד וחיובי [0.950] בעוד שמקדם המתאם שבין מספר בתי האוכל במדגם לבין מכירות שנתיות באלפי $ הוא נמוך מאוד ושלילי [-0.283].

תרגיל מספר 28

במחלקה האפדימולוגית ערכו מחקר אודות סטריליזציה של מכשירים המשמשים בחדר טיפולים של מרפאות גסטרולוגיה. במחקר בדקו את השפעת חשיפה, במשכי זמן שונים, של 3 סוגי חיידקים לקרני גאמה [המשמשות לסטריליזציה]. לצורך הבדיקה נלקחו 3 טסיות – בטסית אחת 840 חיידקים מסוג א', בטסית שנייה כ- 840 חיידקים מסוג ב' ובטסית שלישית – כ- 840 חיידקים מסוג ג'. הטסיות נחשפו לקרני גאמה לפרק זמן של 10 שניות ובתום כל שנייה נספרו מספר החיידקים שהושמדו בה. הנתונים מובאים בטבלה הבאה:

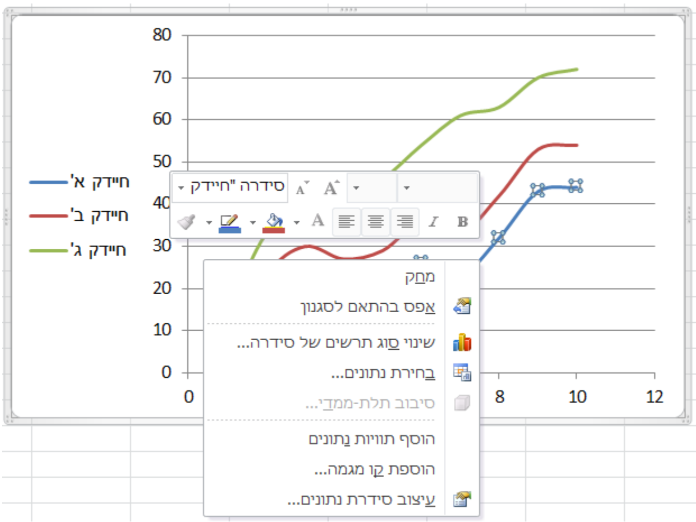
	A	B	C	D
1	זמן החשיפה	חיידק א'	חיידק ב'	חיידק ג'
2	1	0	4	14
3	2	13	23	33
4	3	20	30	40
5	4	17	27	42
6	5	19	29	46
7	6	26	36	54
8	7	23	33	61
9	8	32	42	63
10	9	43	53	70
11	10	44	54	72

מהי משוואת הרגרסיה הליניארית לחישוב אומדן של החיידקים שמתו [בכל אחד מסוגי החיידקים] כתלות במשך החשיפה בשניות לקרני גאמה [המספרים בטבלה מציינים את מספר החיידקים שמתו באותה שנייה ולא את המספר המצטבר של חיידקים שמתו].

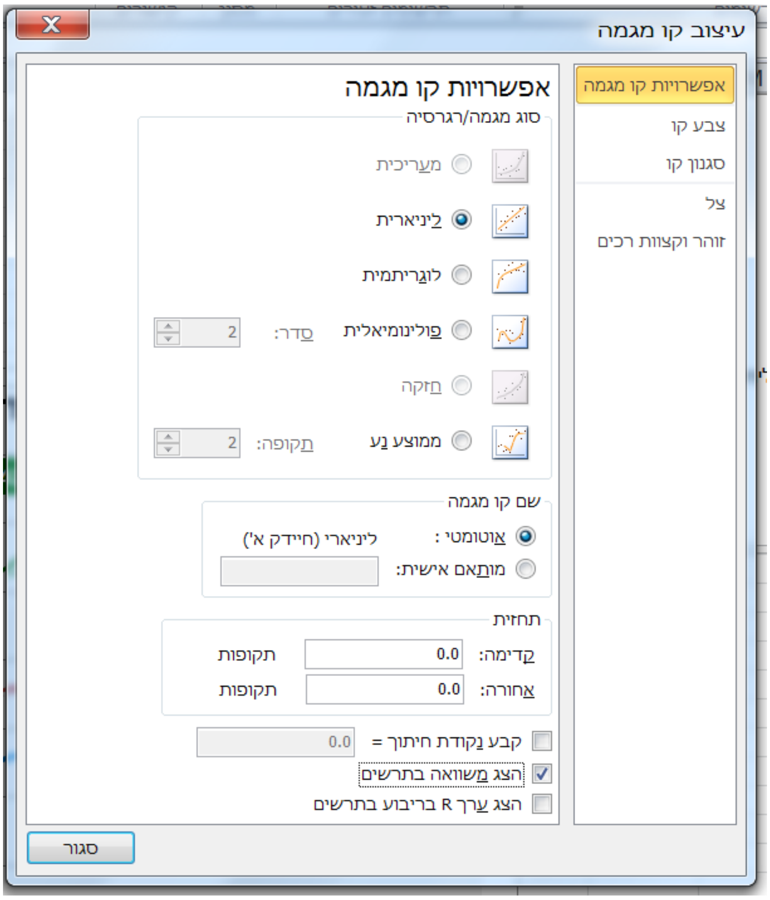
ניתן לפתור את התרגיל בדרכים שהוצגו עד כה. כאן תובא דרך נוספת – בעזרת גרף [לא נלמד עד כה הפקת גרפים וההנחה היא שלקורא יש ידע בסיסי בהכנת גרפים].

נסמן על ידי הקלקה את הטבלה לרבות הכותרות [A1:D11], אח"כ נלחץ על הוספה > פיזור > פיזור עם קווים חלקים [אפשר גם "קווים חלקים עם סימנים" כי תמיד אפשר להסיר את הסימנים] ונקבל מיד את התרשים [התרשים לא יהיה מושלם – לעתים יהיה צורך להפוך את כיוון ציר ה- X ובכל מקרה יהיה צורך לתת כותרות לצירים ולגרף בכלל].

נימצא על אחד מ- 3 הגרפים ונקליק עליו ייפתח המסך הבא:

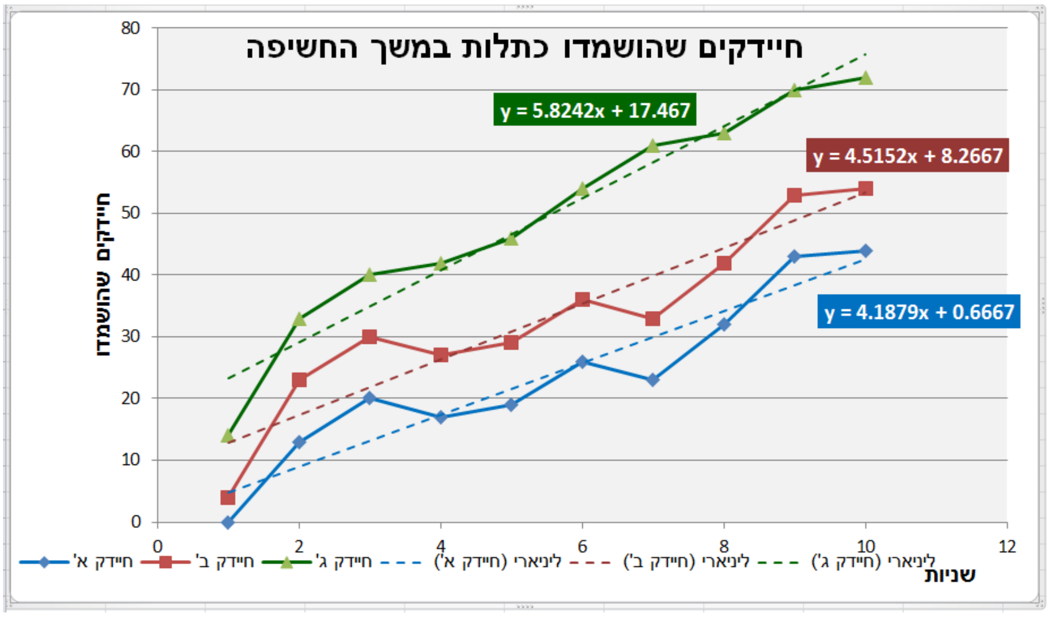


נקיש על "הוספת קו מגמה" וייפתח המסמך הבא:



כברירת מחדל יופיע סימון על אפשרויות קו מגמה – **ליניארית** [אם לא – נסמן זאת] וגם קו מגמה **אוטומטי** [במצב זה רוכש קו המגמה את השם של המשתנה, יש אפשרות לבחור שם אחר ע"י הקשה על **מותאם אישית**]. עכשיו נוסיף √ בריבוע של **הצג משוואה בתרשים** ונקיש – סגור. נחזור על אותה פעולה לגבי 2 הגרפים הנוספים, אפשר גם להציג את הערך של R^2 בתרשים ע"י הוספת הסימן √ בריבוע של **הצג ערך R בריבוע בתרשים**.

נקבל את התרשים הבא [לאחר כמה עריכות לקבלת תצוגה נאה יותר]:



לתשומת לבך! כאשר משנים/מתקנים את נתוני המקור לפיהם נבנה הגרף – הגרף מתעדכן באופן אוטומטי אולם לא כן המשוואה. תופעה זו נכונה לגבי כל הפונקציות המובנות [בשונה מתחשיבים "מסורתיים" המקושרים ביניהם ושינוי בנתונים משנה מייד את התוצאות].

פרק 13 – מבחן חי בריבוע [$\chi 2$] לבדיקת אי תלות

האקסל אינו נותן פתרון [לפחות כיום] מלא בעזרת פונקציה מובנית או באמצעות כלי ב- Data Analysis למבחן חי בריבוע. יחד עם זאת וכפי שנראה בתרגיל הבא ניתן להסתייע בכמה מכלי האקסל לפתרון ובעיקר כאשר הנתונים לבדיקת אי התלות הם גולמיים [כלומר טרם עובדו לטבלה].

תרגיל מספר 29

התרחבות תרבות שתיית הבירה בישראל והקמתן של מבשלות בוטיק, הביאו את עורכי הירחון "דרך האוכל והשתייה" [החריפה] לערוך סקר שיענה על השאלה: האם יש תלות בין העדפת סוגי בירה לבין מקום המגורים. נקבעו 3 סוגי בירה: קלה, רגילה וכהה [בירה כהה אינה מתייחסת ל"בירה שחורה" המוכרת מהעבר]. נקבעו 3 אזורי מגורים: חדרה-גדרה [להלן "מרכז"], דרומה מגדרה [להלן "דרום"] וצפונה מחדרה [להלן "צפון"]. נבחר מדגם אקראי ומייצג של 150 בוגרים, נשים וגברים, הרשאים ע"פ החוק לרכוש ולשתות בירה והם נשאלו למקום מגוריהם ולבירה המועדפת עליהם. תוצאות המדגם מובאות בטבלה הבאה [הטבלה המקורית היא טבלה המשך שבה כעל הנתונים רשומים ב- 3 העמודות הראשונות]:

בירה מועדפת	אזור מגורים	נשאל מס'	בירה מועדפת	אזור מגורים	נשאל מס'	בירה מועדפת	אזור מגורים	נשאל מס'	בירה מועדפת	אזור מגורים	נשאל מס'	בירה מועדפת	אזור מגורים	נשאל מס'
כהה	מרכז	121	רגילה	מרכז	91	רגילה	דרום	61	כהה	מרכז	31	קלה	צפון	1
כהה	מרכז	122	רגילה	מרכז	92	רגילה	מרכז	62	קלה	מרכז	32	קלה	צפון	2
כהה	מרכז	123	קלה	דרום	93	רגילה	מרכז	63	קלה	מרכז	33	רגילה	צפון	3
רגילה	דרום	124	קלה	צפון	94	כהה	דרום	64	רגילה	צפון	34	קלה	דרום	4
רגילה	דרום	125	כהה	מרכז	95	קלה	צפון	65	רגילה	צפון	35	רגילה	מרכז	5
רגילה	מרכז	126	קלה	צפון	96	קלה	צפון	66	רגילה	דרום	36	רגילה	מרכז	6
כהה	מרכז	127	קלה	צפון	97	קלה	צפון	67	רגילה	דרום	37	כהה	דרום	7
כהה	מרכז	128	רגילה	צפון	98	רגילה	צפון	68	רגילה	מרכז	38	כהה	צפון	8
כהה	מרכז	129	רגילה	דרום	99	רגילה	דרום	69	כהה	מרכז	39	כהה	מרכז	9
כהה	צפון	130	רגילה	מרכז	100	רגילה	מרכז	70	כהה	מרכז	40	כהה	מרכז	10
קלה	צפון	131	רגילה	מרכז	101	רגילה	מרכז	71	קלה	מרכז	41	קלה	מרכז	11
כהה	מרכז	132	כהה	דרום	102	קלה	דרום	72	רגילה	צפון	42	קלה	מרכז	12
קלה	מרכז	133	קלה	צפון	103	קלה	צפון	73	כהה	צפון	43	רגילה	צפון	13
קלה	מרכז	134	כהה	מרכז	104	כהה	מרכז	74	רגילה	דרום	44	כהה	צפון	14
רגילה	צפון	135	כהה	מרכז	105	כהה	מרכז	75	קלה	צפון	45	רגילה	דרום	15
כהה	צפון	136	קלה	מרכז	106	קלה	מרכז	76	כהה	צפון	46	קלה	דרום	16
רגילה	דרום	137	קלה	מרכז	107	קלה	מרכז	77	רגילה	צפון	47	רגילה	דרום	17
רגילה	דרום	138	רגילה	צפון	108	רגילה	צפון	78	רגילה	דרום	48	רגילה	מרכז	18
רגילה	מרכז	139	כהה	צפון	109	רגילה	צפון	79	רגילה	מרכז	49	כהה	מרכז	19
כהה	מרכז	140	רגילה	דרום	110	רגילה	דרום	80	רגילה	מרכז	50	כהה	מרכז	20
כהה	מרכז	141	כהה	דרום	111	כהה	דרום	81	כהה	דרום	51	כהה	מרכז	21
קלה	צפון	142	קלה	דרום	112	קלה	דרום	82	קלה	צפון	52	כהה	מרכז	22
קלה	צפון	143	רגילה	מרכז	113	רגילה	מרכז	83	כהה	מרכז	53	קלה	דרום	23
רגילה	צפון	144	כהה	מרכז	114	כהה	מרכז	84	כהה	מרכז	54	רגילה	דרום	24
רגילה	דרום	145	כהה	מרכז	115	כהה	מרכז	85	קלה	מרכז	55	רגילה	מרכז	25
רגילה	מרכז	146	רגילה	דרום	116	כהה	דרום	86	קלה	מרכז	56	כהה	מרכז	26
רגילה	מרכז	147	רגילה	מרכז	117	כהה	צפון	87	רגילה	צפון	57	כהה	מרכז	27
כהה	דרום	148	כהה	מרכז	118	כהה	מרכז	88	רגילה	צפון	58	כהה	מרכז	28
קלה	צפון	149	כהה	מרכז	119	כהה	מרכז	89	רגילה	דרום	59	כהה	צפון	29
כהה	מרכז	150	כהה	מרכז	120	כהה	דרום	90	כהה	דרום	60	קלה	צפון	30

והשאלה היא: האם ברמת מובהקות של 5% אפשר לקבל את הטענה **שאין** תלות בין מקום המגורים לבין העדפת סוג בירה?

פתרון

נבנה טבלת ציר, כפי שהוסבר קודם לכן בחוברת, שתרכז את תוצאות הנבחנים לפי אזור מגורים וסוג בירה מועדפת ונקבל את הטבלה הבאה:

ספירה של בירה מועדפת	תוויות עמודה			
תוויות שורה	דרום	מרכז	צפון	סכום כולל
כהה	12	38	3	53
קלה	3	13	20	36
רגילה	21	22	18	61
סכום כולל	**36**	**73**	**41**	**150**

מבנה הטבלה הוא המבנה שבו באופן ה"מסורתי" אנו פותרים שאלות הנוגעות למבחני חי-בריבוע. עכשיו צריך ל"עבד" את נתוני הטבלה:

ערכים נצפים Oi

בירה מועדפת	אזור מגורים			סכום כולל
	דרום	מרכז	צפון	
כהה	12	38	3	**53**
קלה	3	13	20	**36**
רגילה	21	22	18	**61**
סכום כולל	**36**	**73**	**41**	**150**

ערכים צפויים Ei

בירה מועדפת	אזור מגורים			סכום כולל
	דרום	מרכז	צפון	
כהה	12.7	25.8	14.5	**53**
קלה	8.64	17.5	9.84	**36**
רגילה	14.6	29.7	16.7	**61**
סכום כולל	**36**	**73**	**41**	**150**

=$U44*R$47/$U$47

=SUM(R52:T52)

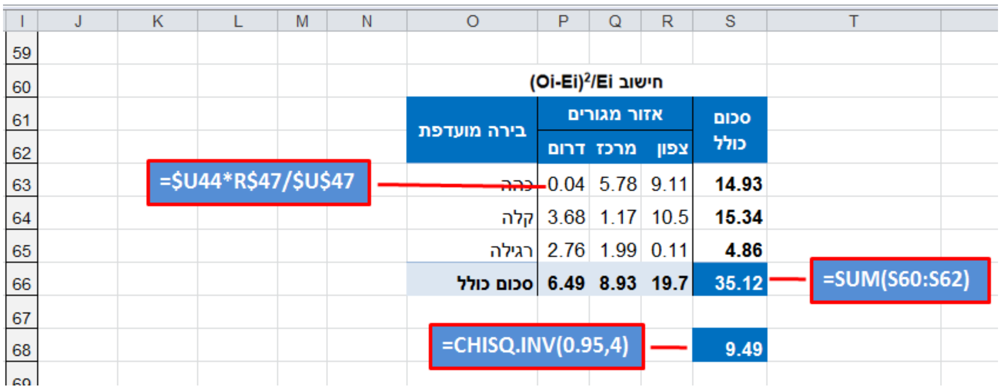
=CHISQ.TEST(P44:R46,P52:R54)

P= 0.0000004

הפונקציה =CHISQ.TEST(P44:R46,P52:R54) עושה את החישוב [שאותו אנו עושים בפתרון המסורתי כפי שנראה בהמשך] של $\chi^2_{calc.} = (O_i - E_i)^2 / E_i$ וקובעת את הערך P שעבורו ועבור כל ערך הגדול ממנו נדחה את השערת האפס.

כיוון ש- P=0.0000004<0.05 – נדחה את השערת האפס ונקבע כי אכן יש תלות בין אזור המגורים לבין העדפת סוג בירה.

הפתרון ה"מסורתי" ייראה כך [גם אותו נבצע באקסל בהמשך לחישוב הקודם]:



חישוב $(Oi-Ei)^2/Ei$

בירה מועדפת	אזור מגורים: צפון	מרכז	דרום	סכום כולל
כהה	9.11	5.78	0.04	14.93
קלה	10.5	1.17	3.68	15.34
רגילה	0.11	1.99	2.76	4.86
סכום כולל	19.7	8.93	6.49	35.12
				9.49

כיוון שמתקיים $\chi^2_{calc.} = 35.12 > \chi^2_{critc.} = 9.49$ - נדחה את השערת האפס ונקבע כי אכן יש תלות בין אזור מגורים לבין העדפת סוג בירה.

פרק 14 – ניתוח שונות חד כיווני – Analysis of Variance [ANVOVA]

על מנת לבדוק את טיבן של מספר שיטות שונות של גידול ירקות [השקיה למשל, שיטות לימוד, טיפול טרמי למוצרי מתכת וכו' [במובן של השפעתן על פרמטר מסוים כמו למשל יבול בק"ג ליחידת שטח, ציון במבחן, חוזק קריעה במונחים של ק"ג למ"מ מרובע], נבחנות מספר קבוצות של פריטים מהאוכלוסייה הנוגעת כשעל כל קבוצה מופעל שיטה/תהליך שונה שאת השפעתו רוצים לבדוק כאשר יתר תנאי הקבוצה דומים ביותר [לא נכתוב זהים משום שזהות מוחלטת בלתי אפשרית]. בתום הבדיקה בודקים, בכל קבוצה, את הפרמטר שהיה אמור להיות מושפע מהשיטה שהופעלה עליו בכל אחת מן הקבוצות.

ניסוח ההשערות:

השערת האפס (H_0) מניחה שאין הבדל בין ה"שיטות" שהופעלו וההשערה האלטרנטיבית (H_1) מניחה שיש הבדל בין ה"שיטות".

פתרון לשאלה שכזאת כאשר מספר השיטות הוא 2 בדיוק ייעשה בעזרת מבחן t להפרש תוחלות של שתי אוכלוסיות, אולם כאשר מספר השיטות הוא 3 או יותר חייבים להשתמש בניתוח שונות חד כיווני.

תרגיל מספר 30

במכון לחקר היבולים, נבחנה השפעה של 3 שיטות השקיה שונות על יבול התירס. לצורך הבחינה נבחרו באקראי 12 חלקות של תירס ששטח כל אחת מהן 1,000 מ"ר. כל אחת משיטות השקיה יושמה ב- 4 חלקות שנבחרו באקראי[7] [פרט לשיטות ההשקיה, כל החלקות קבלו טיפול זהה]. בסוף הניסוי התקבלו התוצאות הבאות של היבול [בק"ג ל- 1,000 מ"ר].

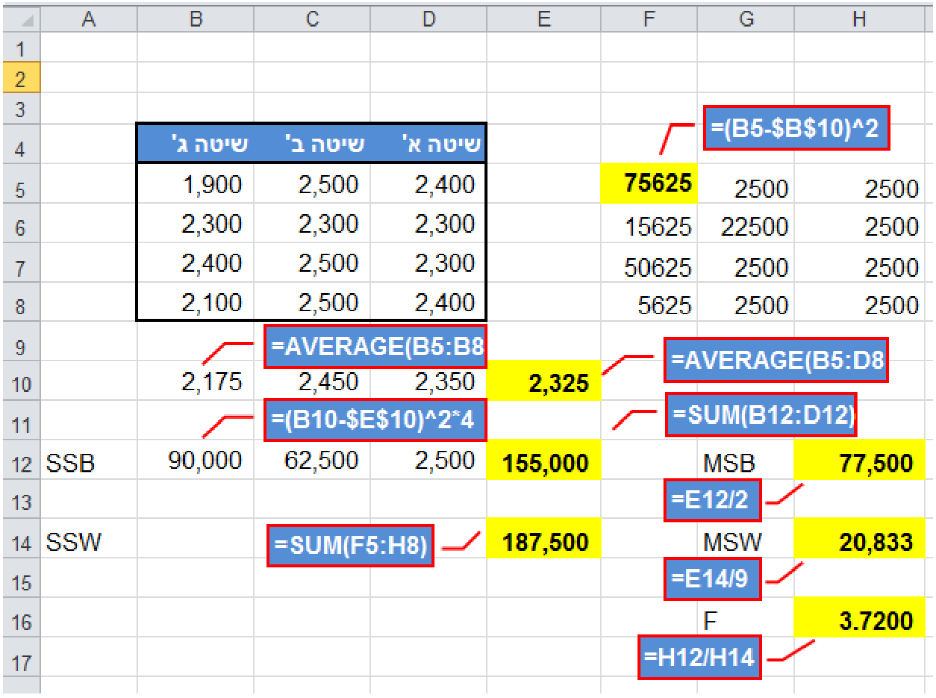
שיטה א'	שיטה ב'	שיטה ג'
2,400	2,500	1,900
2,300	2,300	2,300
2,300	2,500	2,400
2,400	2,500	2,100

האם אפשר לקבוע, ברמת מובהקות של 5% כי לשיטות ההשקיה השונות אין השפעה על תוחלת יבול התירס? באיזו רמת מובהקות ניתן לקבוע כי אין הבדל בין תוחלת היבול בשיטות ההשקיה השונות?

[7] בתרגיל זה כל שיטה יושמה ב- 4 חלקות אך אין זה תנאי. למעשה אפשר ששיטה א' תיושם ב- 5 חלקות, שיטה ב' ב- 4 חלקות ושיטה ג' ב- 3 חלקות.

פתרון

אפשר לערוך את החישובים הנדרשים תוך שימוש בכלי האקסל ה"רגילים" כלהלן:



	A	B	C	D	E	F	G	H
1								
2								
3								
4		שיטה ג'	שיטה ב'	שיטה א'				
5		1,900	2,500	2,400		75625	2500	2500
6		2,300	2,300	2,300		15625	22500	2500
7		2,400	2,500	2,300		50625	2500	2500
8		2,100	2,500	2,400		5625	2500	2500
9								
10		2,175	2,450	2,350	2,325			
11								
12	SSB	90,000	62,500	2,500	155,000		MSB	77,500
13								
14	SSW				187,500		MSW	20,833
15								
16							F	3.7200
17								

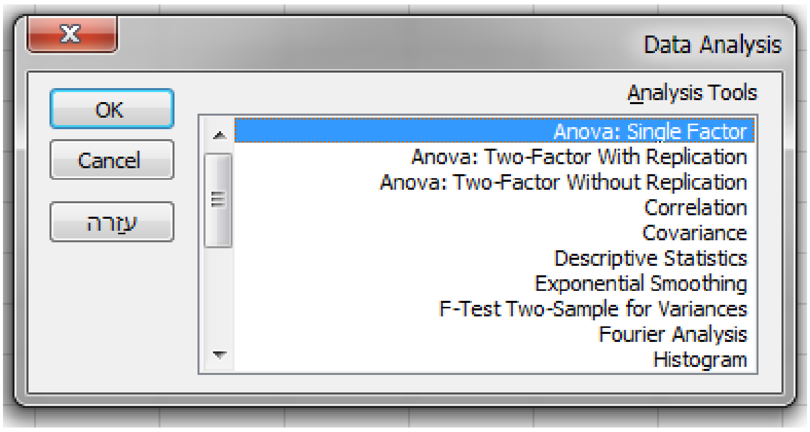
אם נסכם את החישובים לטבלה המקובלת נקבל:

מקור	סכום ריבועים	דרגות חופש - d.f	ממוצע ריבועים	F יחס
בין	SSB=155,000	k-1=3-1=2	$MSB = \frac{SSB}{k-1} = 77{,}500$	$F = \frac{MSB}{MSW} = 3.7200$
בתוך	SSW=187,500	n-k=12-3=9	$MSW = \frac{SSW}{n-k} = 20{,}833$	

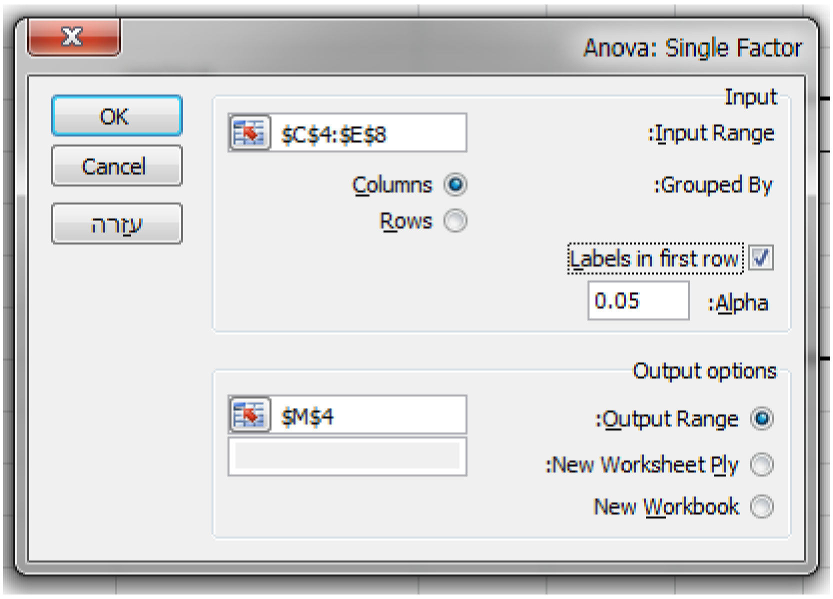
נדחה את Ho אם $F_{מחושב} > F_{critc.} = F_{\alpha,k-1,n-k}$. כיוון ש- $F_{מחושב}=3.72$ ו- $F_{critic.}=F_{0.05,2,9}=4.2565$ [אותו מצאנו בטבלה המתאימה] – **ניתן לקבוע כי אין הבדלים בין היבולים בחלקות השונות.**

לעניין השאלה השנייה על רמת מובהקות שבה ניתן לקבוע כי אין הבדל בין תוחלת היבול בשיטות ההשקיה השונות נשיב בהמשך.

לאקסל פונקציה סטטיסטית מובנית המאפשרת לנו למצוא את הפתרון בקלות ובמהירות. הניתוח נעשה בעזרת הפונקציה **"Anova"** אליה מגיעים במסלול הבא: נתונים < Data Analysis >Anova: single factor, מתקבל המסך הבא:



כאשר נקיש OK ויתקבל המסך הבא בו נמלא את הפרטים הנדרשים לפתרון. במסגרת ה- Input Range נכללים גם כותרות העמודות, חשוב להקפיד לציין כי הנתונים מובאים בעמודות וכי בשורה הראשונה נכללות כותרות העמודות [טקסט].



לאחר הקשה על OK נקבל את הפלט הבא שבו הודגשו הנתונים החשובים לנו.

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
שיטה ג'	4	8700	2175	49167		
שיטה ב'	4	9800	2450	10000		
שיטה א'	4	9400	2350	3333.3		
ANOVA						
Source of Variation	***SS***	***df***	***MS***	***F***	***P-value***	***F crit***
Between Groups	155000	2	77500	**3.72**	**0.0665**	**4.2565**
Within Groups	187500	9	20833			
Total	342500	11				

אנו רואים כי התקבל אותו ערך של $F_{מחושב}$ ושל F_{critc}. [כפי שחושבו בחישוב ה"מסורתי" ולכן המסקנה תהיה אותה מסקנה, היינו – לא ניתן לדחות את השערת האפס ובמילים אחרות: לשיטות ההשקיה השונות אין השפעה על היבול.

ועכשיו נתייחס לשאלה: באיזו רמת מובהקות ניתן לקבוע כי אין הבדל בין תוחלת היבול בשיטות ההשקיה השונות?

P-value=0.0665 נותן את התשובה לשאלה. עבור כל רמת מובהקות הגבוהה מ-6.65% [כדי למנוע ספק: עבור כל **$\alpha \le 0.0665$**] לא ניתן יהיה לדחות את השערת האפס ובמילים אחרות – אין הבדל בין תוחלת היבול בשיטות ההשקיה השונות. עבור כל רמת מובהקות הקטנה מ- 6.65% [כדי למנוע ספק: עבור כל **$\alpha \ge 0.0665$** למשל עבור **$\alpha = 0.1$**] נדחה את השערת האפס ונקבע כי יש הבדל בין תוחלת היבול בשיטות ההשקיה השונות.

הוצאת
הוד-עמי
ספרי מחשבים

ת.ד. 6108 הרצליה 46160 **הזמנות באתר**
info@hod-ami.co.il www.**hod-ami**.co.il

קטלוג ספטמבר 2015

	כולל	עמ'	מחיר*
אינטרנט - מפתחי אתרים/גרפיקה			
אמא, אבא - בניתי אתר באינטרנט (HTML)		256	29
הגדל את הכנסות העסק שלך באמצעות **פרסום בגוגל Google AdWords**		300	249
מבוא לתכנות בסביבת אינטרנט – הכל כלול בספר אחד HTML + JS + ASP		192	9
מבוא לתכנות בסביבת אינטרנט - מבוא ו- **HTML** - חלק 1 מתוך 3 - מהד' 3		320	59
מבוא לתכנות בסביבת אינטרנט - **JavaScript** - חלק 2 מתוך 3		240	49
מבוא לתכנות בסביבת אינטרנט - **ASP** - חלק 3 מתוך 3 - מהד' 3		192	49
HTML5 המדריך לבניית אתרים ולמערכות WEB, הדור הבא – מהד' 2 חדש		384	159
The **Java** Tutorial סדנת לימוד	CD	768	179
JavaScript סדנת לימוד		586	159
מדריך **ASP.NET MVC 4**		514	199
ASP.NET 3.5 סדנת לימוד בשפות #C ו-VB		824	99
תכנות			
Code Complete – מדריך מעשי לפיתוח תוכנה		288	139
לחפש באגים, מדריך מעשי לבודק תוכנה, מהד' 3		350	169
Visual C# 3.0 סדנת לימוד		656	99
ללמוד **C** - מהד' 3		480	139
שפת **אסמבלי** למחשב האישי, מהד' 2		314	89
יסודות התכנות ב-**VBA** לתוכנת **Excel**, מהד' 4		152	95
PC - חומרה, תוכנה ורשתות			
מדריך Hacking ואבטחת מידע, מהד' 2		428	169
מדריך **חומרה ותוכנה** לטכנאי PC - מהד' 5	CD	752	189
Windows 7/8 נושאים מתקדמים		140	49
מדריך **רשתות** לטכנאי PC ולמנהלי רשת - מהד' 4		608	219
Windows			
Windows 8.1 מדריך למשתמש חדש		544	149
Windows 8 מדריך למשתמש		438	19
Windows 7 צעד-אחר-צעד		272	19

* מחיר מומלץ לצרכן כולל מע"מ

היכנס לאתר להתעדכן בספרים החדשים ומבצעים

תוכן עניינים ופרקים לדוגמה www.**hod-ami**.co.il

	כולל	עמ'	מחיר*
גרפיקה			
Flash – ספר הדרכה ותרגילים		122	64
אינדיזיין – ספר הדרכה ותרגילים		132	72
Illusatrator – ספר הדרכה ותרגילים		120	64
Photoshop צעד אחר צעד (צבע מלא, למתחילים), מהד' 3		200	159
מדריך לתוכנת העיצוב והאנימציה **3ds max** (2 כרכים)	CD	1400	289
OFFICE			
יישומי סטטיסטיקה בגיליון אלקטרוני Excel חדש		86	69
טבלאות ציר – ניתוח נתונים חכם		116	87
יסודות התכנות ב-**VBA** לתוכנת **Excel**, מהד' 4		152	95
Word 2010 צעד אחר צעד		160	69
Excel 2010 סדנת לימוד		344	129
PowerPoint 2010 צעד אחר צעד		202	69
Outlook 2010 צעד אחר צעד		190	69
Access 2010 סדנת לימוד		360	179
עוד ספרים בגרסאות קודמות (2007 ו-2003) ניתן למצוא באתר הוד-עמי			
ניהול, כלכלה ושונות			
לחפש באגים, מדריך מעשי לבודק תוכנה, מהד' 3		350	169
ניהול ממוקד לעשות יותר עם מה שיש (כריכה קשה) - מהד' 4		358	133
לי זה עולה יותר (תמחיר) (כריכה קשה) - מהד' 3		368	129
מערכות מידע			
Oracle SQL יכולות מתקדמות		626	249
SAS (**Statistical Analysis System**) – ספר לימוד		256	169
בסיסי נתונים ושפת **SQL** – עקרונות ועיצוב		648	149
ניתוח מערכות מידע כולל מתודולוגיית ה- **UML**		818	229
המדריך העברי השלם **UML**		346	329
ספרים דיגיטליים			
לרכישת ספרים לצפייה בפורמט PDF היכנס לאתר לקטגוריה "ספרים דיגיטליים"			
קבצי תרגול לספרים			
קבצי תרגול לספרים שונים תמצא באתר בקטגוריה "קבצי תרגול לספרים"			

* מחיר מומלץ לצרכן כולל מע"מ. קטלוג 9/2015

היכנס לאתר להתעדכן בספרים החדשים ומבצעים

תוכן עניינים ופרקים לדוגמה www.**hod-ami**.co.il