תרגיל: קארל פרידריך גאוס

מה נלמד בתרגיל הזה?

איך עובדים עם נתונים בהתפלגות נורמלית: הפונקציות NORMDIST, NORMSDIST.
חישוב ציוני תקן: הפונקציה STANDARDIZE.
חישוב אחוזונים באמצעות הפונקציה PERCENTRANK.INC
איך להציג היסטוגרמה באמצעות Analysis Toolpak.

לפני התרגיל

למי שלא זוכר.ת - כדאי לחזור על החומר משיעורי סטטיסטיקה לגבי מהי התפלגות נורמלית ומה המשמעות של p-value.
לאחר מכן, צפו בסרטון הזה.

חישוב האחוזון של כל ציון

פתחו את הקובץ stats.xlsx בגיליון anxiety. יש כאן ציונים (פיקטיביים) של 200 ילדים בשאלון חרדה. הציונים בשאלון הם על סקאלה 0 עד 100. ככל שהציון גבוה יותר, רמת החרדה גבוהה יותר.

בהינתן רשימת הציונים המלאה, אפשר לחשב – עבור כל ציון – עד כמה הוא גבוה ביחס לציונים האחרים. מבחינה מספרית, נחשב עבור כל ציון את האחוזון שלו – כלומר, אחוז הציונים הנמוכים ממנו. למשל, האחוזון של הציון הכי נמוך הוא 0, והאחוזון של הציון הכי גבוה הוא 100%.
בתרגיל "תעביר את המלח בבקשה", כבר למדנו על פונקציה לחישוב אחוזונים: הפונקציה PERCENTILE מקבלת 2 ארגומנטים – רשימת ערכים, אחוזון – ומחזירה את הערך שנמצא באותו אחוזון. כאן, אנחנו רוצים לעשות את הפעולה ההפוכה: בהינתן ערך מסוים, לחשב את האחוזון שלו. הפונקציה שעושה את זה היא PERCENTRANK.INC.

1. חשבו את האחוזון של כל ציון ביחס לשאר הציונים ברשימה (התשובות: בסוף הקובץ).
2. הפונקציה מחזירה את האחוזונים בתור מספרים בטווח 0-1. השתמשו בעיצוב תאים כדי להציג את המספר בתור אחוז – מספר בין 0 ל-100, ברמת דיוק של ספרה אחת אחרי הנקודה.

אם שאלת המחקר שלנו היא "האם לאדם הזה יש רמת חרדה נמוכה מהקבוצה?", ה-p-value (חד זנבי) של התשובה הוא האחוזון עצמו. לדוגמה, אם אדם מסוים נמצא באחוזון ה-96, אז p-value עבור אותו אדם יהיה p=0.96: רמת החרדה שלו גבוהה יחסית, אז הסיכוי לקבל באופן מקרי רמת חרדה קיצונית יותר ביחס לשאלת המחקר (כלומר נמוכה יותר) הוא די גבוה.
אם שאלת המחקר שלנו היא "האם לאדם הזה יש רמת חרדה גבוהה מהקבוצה?", ה-p-value (חד זנבי) של התשובה הוא המספר שמשלים את האחוזון ל-100. לדוגמה, ה-p-value של האדם הנ"ל יהיה 4% (כלומר p=0.04): רמת החרדה שלו גבוהה יחסית, אז הסיכוי לקבל באופן מקרי רמת חרדה קיצונית יותר (גבוהה יותר) הוא די נמוך.

הצגת התפלגות של הנתונים – היסטוגרמה

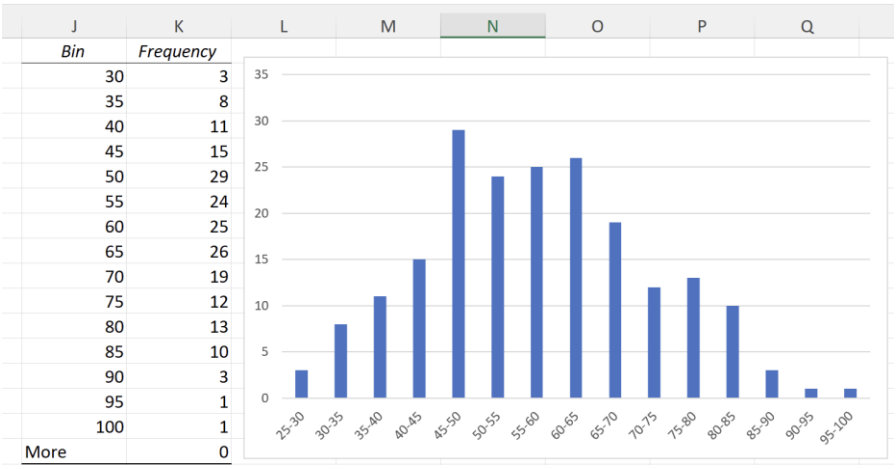
השימוש באחוזונים טוב כדי להשוות ציון ספציפי לרשימת ציונים נתונה. אבל במקרים רבים, אנחנו רוצים לבדוק אם הציון גבוה או נמוך כאשר לא נתונה לנו רשימת הציונים המלאה של כל האוכלוסיה הרלוונטית, אלא רק נתונים כלליים (ממוצע וסטיית תקן) לגבי התפלגות הציונים באוכלוסיה. המצב הכי נוח הוא כאשר התפלגות הציונים דומה לצורה של התפלגות נורמלית – זה מצב שמאפשר שימוש במגוון גדול של כלים סטטיסטיים כדי לנתח את הנתונים.

כל הזכויות שמורות © 2023 CC BY-NC-ND דרור דותן וריקרדו טרש, אוניברסיטת תל אביב (החוג ללקויות למידה וייעוץ חינוכי). מותר להשתמש בחומרי הקורס כמו שהם (ללא שינוי) למטרות לא-מסחריות בלבד תוך מתן קרדיט ליוצרים.

כדי לראות אם ציוני החרדה מתפלגים באופן שדומה להתפלגות נורמלית, נבדוק איך נראית ההתפלגות שלהם. נעשה זאת בדיוק כמו שריקרדו הראה בסרטון.

3. מה הציון הנמוך והגבוה ביותר?

4. הדרך להציג התפלגות של נתונים היא באמצעות היסטוגרמה: אנחנו מחלקים את הטווח הכולל בין הציון המינימלי למקסימלי למקטעים של טווחי ערכים באורכים שווים (לדוגמה, מקטעים באורך 5 בטווח בין 25 ל-100), וסופרים, עבור כל טווח ערכים, כמה ילדים הופיעו בו. ההיסטוגרמה של הנתונים שלנו נראית כך:



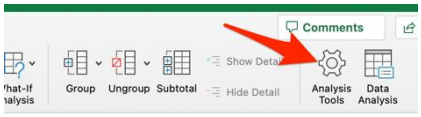
Bin	Frequency
30	3
35	8
40	11
45	15
50	29
55	24
60	25
65	26
70	19
75	12
80	13
85	10
90	3
95	1
100	1
More	0

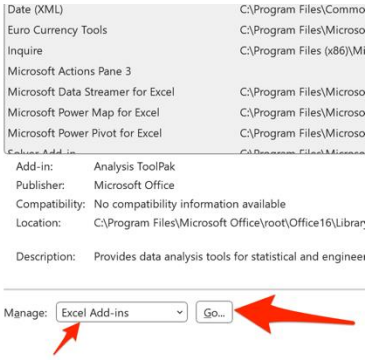
בעמודות J, K מופיעים נתוני ההיסטוגרמה: בעמודה J כתוב הגבול העליון של כל טווח ערכים, ובעמודה K כתוב כמה ילדים היו עם ציון חרדה באותו טווח ערכים. מימין, הנתונים מוצגים בתור תרשים עמודות: בציר X רשום טווח הערכים של כל עמודה, וציר Y הוא מספר הילדים. התרשים מציג את ההתפלגות: אם היינו מתקנים את המספרים על ציר Y כך שלא יציגו מספר ילדים אלא את האחוז מתוך כלל הילדים, התרשים היה משקף, בקירוב, את ההסתברות לקבל כל ערך.

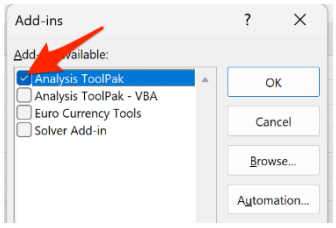
כדי ליצור את ההיסטוגרמה, נשתמש בכלי ניתוח נתונים של אקסל (Analysis tools).

(א) ראשית, נוודא שהכלים האלה מופעלים באקסל שלנו (צריך לעשות את זה רק פעם אחת על כל מחשב) --

במחשבי Windows עם גירסה עדכנית של אקסל (צילום מסך ימני): בסרגל הראשון (File) נפתח את מסך ההגדרות של אקסל (options – בצד השמאלי תחתון של המסך). בהגדרות Add ins, נבחר להגדיר Excel add-ins ונלחץ על כפתור Go. במחשבי מק (צילום מסך שמאלי) / גירסאות אחרות של אקסל: בסרגל Data, נלחץ על כפתור Analysis tools:

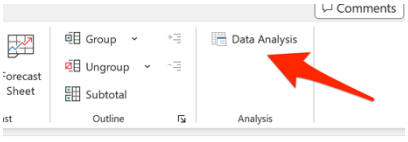






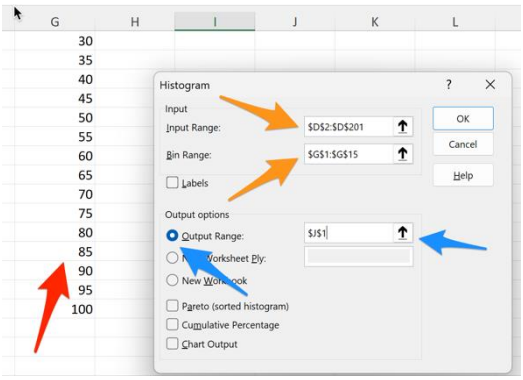
במסך שייפתח, נוודא שמסומן V על Analysis Toolpak:

(ב) במקום כלשהו על הגיליון שלנו, נכין את הגדרת המקטעים – הגבול העליון של כל מקטע. כאמור, אנחנו רוצים לחלק את טווח הציונים בין 25 ל-100 למקטעים באורך 5, אז הגבולות העליונים של המקטעים יהיו 30, 35, 40, 45 עד 100. אנחנו רשמנו את המספרים האלה בתאים G1:G15 (חץ אדום).



כל הזכויות שמורות © 2023 CC BY-NC-ND דרור דותן וריקרדו טרש, אוניברסיטת תל אביב (החוג ללקויות למידה וייעוץ חינוכי). מותר להשתמש בחומרי הקורס כמו שהם (ללא שינוי) למטרות לא-מסחריות בלבד תוך מתן קרדיט ליוצרים.

(ג) בסרגל Data, נלחץ על כפתור Data analysis, ונבחר באפשרות histogram. במסך שייפתח, נגדיר את טווח התאים בהם נמצאים ציוני החרדה (חץ כתום עליון) ואת טווח התאים בהם הגדרנו את הגבולות העליונים של המקטעים (חץ כתום תחתון). בנוסף, נגדיר איפה לשים את ההיסטוגרמה שמתקבלת (חץ כחול). במקרה זה ביקשנו לשים אותה באותו גיליון, אבל אפשר גם לשים אותה בגיליון/קובץ חדש.



התוצאה היא טבלת השכיחויות המקובצת שראינו בעמוד הקודם. בטבלה הזאת רשום כמה ציוני חרדה יש בכל אחד מטווחי הערכים שהגדרנו. כדי להציג את זה בתרשים, אפשר לסמן את הערכים בעמודה K וליצור תרשים עמודות (זו הדרך המקובלת להציג היסטוגרמה). שימו לב שבתרשים כאן גם הגדרנו שעל ציר X יוצג הגבול העליון של כל טווח.

5. בטבלת השכיחויות המקובצת (עמודה J) כתוב רק הגבול העליון של כל טווח ערכים. איך גרמנו לכך שבציר X של התרשים מוצג הטווח כולו, גם הגבול התחתון וגם העליון?

6. אנחנו רואים בתרשים שהנתונים מתפלגים פחות או יותר בהתאמה להתפלגות נורמלית. מהו הממוצע וסטיית התקן שלהם?

חישוב p-value בהשוואה להתפלגות נורמלית

כעת נחשב, עבור כל ציון, עד כמה הוא גבוה **ביחס להתפלגות**. זה מאד דומה לחישוב האחוזונים שעשינו בתחילת התרגיל, אבל הפעם לא נשווה את הציון לרשימת כל הציונים האחרים, אלא נשווה אותו אל ההתפלגות שלהם, תחת ההנחה שהציונים מתפלגים נורמלית עם ממוצע 57.5 וסטיית תקן 14.3. אנחנו כבר יודעים שההנחה הזאת פחות או יותר נכונה (רק "פחות או יותר", ולא "בדיוק", כי הציונים לא מתפלגים **בדיוק** בהתפלגות נורמלית אלא רק **בקירוב** נורמלית).

הפונקציה שמחשבת את המיקום של ציון מסוים על ההתפלגות, כלומר ה"אחוזון" שלו, היא NORMDIST.
הפונקציה מקבלת 4 ארגומנטים:

א. הציון אותו נבדוק
ב. הממוצע של ההתפלגות הנורמלית
ג. סטיית התקן של ההתפלגות הנורמלית
ד. איזה סוג של ערך אנחנו רוצים שהפונקציה תחזיר. אם נרשום פה 1 (או TRUE), הפונקציה תחזיר את ה"אחוזון" של הציון ביחס להתפלגות – כלומר, מה אחוז הציונים הנמוכים ממנו – זה מה שאנחנו רוצים במקרה הזה.
אם נרשום 0, הפונקציה תחזיר מספר שמשקף את הסיכוי לקבל את אותו ציון בדיוק. מתמטית, זה לא בדיוק הסיכוי לקבל את הציון: כיוון שההתפלגות הנורמלית רציפה, היא מניחה שמבחינה עקרונית ציון החרדה יכול להיות כל מספר שהוא, כלומר ייתכנו אינסוף ציונים אפשריים והסיכוי לקבל בדיוק ציון מסוים הוא אפסי. אבל הערך שהפונקציה תחזיר כן נמצא ביחס ישר לקבל את הציון הספציפי.

7. השתמשו בפונקציה NORMDIST כדי לחשב את המיקום של כל ציון ביחס להתפלגות נורמלית עם ממוצע 57.5 וסטיית תקן 14.3.

8. אם כבר עשית את התרגיל "אקסלאמי", או יודע.ת איך לכתוב פונקציה בתור ארגומנט של פונקציה אחרת: תקן.י את הנוסחה כך שהממוצע וסטיית התקן לא יופיעו בה בתור מספרים, אלא יחושבו מתוך רשימת הציונים שמופיעה בקובץ.

9. תקנו את הנוסחה כך שתציג את ערך p-value חד זנבי: את הסיכוי לקבל באופן מקרי ציון חרדה גבוה יותר מהציון של אותו ילד.

כל הזכויות שמורות © 2023 CC BY-NC-ND דרור דותן וריקרדו טרש, אוניברסיטת תל אביב (החוג ללקויות למידה וייעוץ חינוכי). מותר להשתמש בחומרי הקורס כמו שהם (ללא שינוי) למטרות לא-מסחריות בלבד תוך מתן קרדיט ליוצרים.

ציוני תקן

יש עוד דרך לתאר את המיקום של ערך מסוים (ציון חרדה) ביחס להתפלגות: ציון תקן. נהוג לסמן אותו בעזרת האות z.

המשמעות של "ציון תקן" זה להמיר את הציון המקורי לסקאלה בה z=0 מייצג את הממוצע, ו-z=1 מייצג את הציון שנמצא סטיית תקן אחת מעל הממוצע. לדוגמה, אם ממוצע ההתפלגות הנורמלית הוא 57.5 וסטיית תקן היא 14.3, אז ציון תקן z=0 תואם את הציון הגולמי 57.5, ציון תקן z=1 תואם את הציון הגולמי 71.8, וציון תקן z=2 תואם את הציון הגולמי 86.1.

בהתפלגות נורמלית, הנוסחה לחישוב ציון תקן, בהינתן הציון הגולמי score, היא: $z = \frac{score - average}{standard\ deviation}$

10. כתבו נוסחה לחישוב ציון תקן של ציון חרדה מסוים.

11. בצעו את אותו חישוב באמצעות הפונקציה STANDARDIZE, שמקבלת 3 ארגומנטים: הציון הגולמי, הממוצע, וסטיית התקן.

ראינו קודם את הפונקציה NORMDIST. יש גירסה נוספת של הפונקציה הזאת, NORMSDIST, שלא מקבלת כארגומנט את הציון הגולמי אלא את ציון התקן. אין צורך לתת לה את הממוצע וסטיית התקן כארגומנטים נוספים, כי הפרמטרים האלה כבר ידועים (ממוצע ציוני התקן הוא תמיד 0 וסטיית התקן שלהם תמיד 1).

חישובי התפלגות נורמלית בכיוון ההפוך: מאחוזון לציון גולמי

ראינו איך לחשב, בהינתן ציון מסוים של משתנה שמתפלג נורמלית, את המיקום שלו בהתפלגות (ההסתברות לקבל ציון נמוך ממנו). הפונקציה NORMINV עושה את החישוב ההפוך: בהינתן המיקום בהתפלגות (ההסתברות), היא מחזירה את הציון. כמובן, גם הפונקציה הזאת צריכה לקבל, בתור שני ארגומנטים נוספים, את הממוצע וסטיית התקן של ההתפלגות.

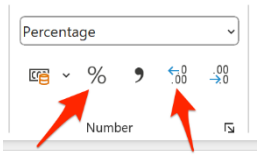
12. כתבו נוסחה שמחשבת מהו הציון החרדה שבדיוק 93% מהציונים נמוכים ממנו (תחת הנחת התפלגות נורמלית עם הממוצע וסטיית התקן שכבר חישבנו).

13. איך יצרנו את הנתונים הפיקטיביים בקובץ stats.xlsx? בשביל ליצור נתון פיקטיבי אחד, הגרלנו באופן אקראי את המיקום בהתפלגות (מספר בין 0 ל-1), והשתמשנו בפונקציה NORMINV כדי לתרגם אותו לציון חרדה. כדי לקבל קובץ עם 200 ציונים, חזרנו על הנוסחה הזאת 200 פעמים. קבענו ממוצע 60 וסטיית תקן 15 (וכמו שראינו בתרגיל הזה, התפלגות הציונים המקריים שקיבלנו באמת היתה עם ממוצע וסטיית תקן מאד קרובים למספרים האלה).

פתחו קובץ חדש ומלאו אותו ב-200 ציוני חרדה מקריים, בדיוק באותה צורה.

כל הזכויות שמורות © 2023 CC BY-NC-ND דרור דותן וריקרדו טרש, אוניברסיטת תל אביב (החוג ללקויות למידה וייעוץ חינוכי). מותר להשתמש בחומרי הקורס כמו שהם (ללא שינוי) למטרות לא-מסחריות בלבד תוך מתן קרדיט ליוצרים.

התשובות לשאלות בקובץ זה

1. הנוסחה לשורה 2: PERCENTRANK.INC($D2:$D201,D2)
2. 

3. טווח הציונים הוא 25.2 עד 95.5.

כדי למצוא את זה, אפשר לסדר את הגיליון לפי עמודת הציון. לחליפין אפשר להשתמש בנוסחאות לחישוב הציון הנמוך/הגבוה ביותר: MAX(D2:D201), MIN(D2:D201)

5. בעמודה I כתבנו נוסחה שמציגה, לכל טווח ערכים, את הטווח המלא. למשל הנוסחה ב-I2 היא:

```
=(J2-5) & "-" & J2
```

בתרשים, הגדרנו ששמות הקטגוריות (ציר X) יילקחו מעמודה I.

6. הממוצע 57.5, סטיית התקן 14.3.
7. אם הציון רשום בתא D2, הנוסחה היא: =NORMDIST(D2, 57.5, 14.3, 1)
8. =NORMDIST(D2, AVERAGE($D$2:$D$201), STDEV.S($D$2:$D$201), 1)
9. מחשבים את המשלים ל-1. אם הנוסחה בשאלה 6 (או 7) רשומה בתא E2, נכתוב =1-E2

לחליפין אפשר לכתוב בבת אחת: =1 - NORMDIST(D2, 57.5, 14.3, 1)

10. נכתוב בתא D202 את הממוצע: AVERAGE(D2:D201)

בתא D203 את סטיית התקן: STDEV.S(D2:D201)

ציון התקן של השורה השניה הוא (D2-$D$202)/$D$203

לחליפין אפשר לכתוב הכל בנוסחה אחת:

```
=(D2 - AVERAGE($D$2:$D$201)) / STDEV.S($D$2:$D$201)
```

11. =STANDARDIZE(D2, $D$202, $D$203)
12. =NORMINV(0.93, 57.5, 14.3)
13. בקובץ החדש, בתא A1 נגריל מספר מקרי בין 0 ל-1: =RAND()

בתא B1 נכתוב נוסחה שהופכת את המספר הזה לציון בהתפלגות הנורמלית: =NORMINV(A1, 60, 15)

נעתיק את שתי הנוסחאות האלה לשורות 2-200.

כל הזכויות שמורות © 2023 CC BY-NC-ND דרור דותן וריקרדו טרש, אוניברסיטת תל אביב (החוג ללקויות למידה וייעוץ חינוכי). מותר להשתמש בחומרי הקורס כמו שהם (ללא שינוי) למטרות לא-מסחריות בלבד תוך מתן קרדיט ליוצרים.