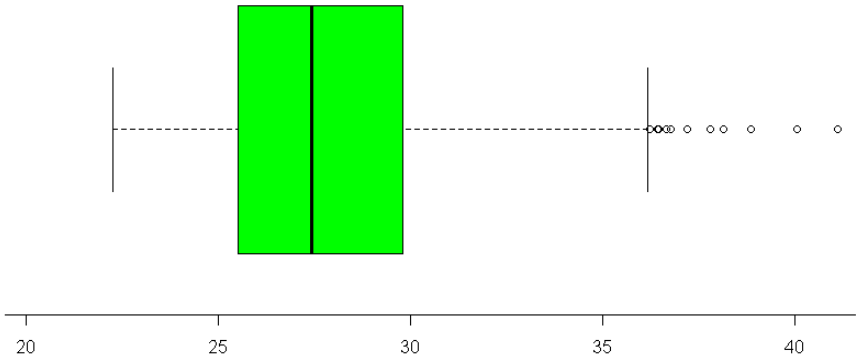
סטטיסטיקה לתלמידי הנדסת תעשיה וניהול

פתרון הבחינה - מועד ב

שאלות 1-2: השתמשו בדיאגרמת הקופסה שלפניכם לענות על שתי השאלות הבאות:
בדיאגרמה מוצגים נתוני BMI עבור גברים במדינה מסויימת:



1. מהי צורת ההתפלגות של ה-BMI באוכלוסייה זו?
   א. אסימטרית שמאלית
   ב. אסימטרית ימנית
   ג. סימטרית
   ד. לא ניתן לדעת, כיוון שישנן תצפיות חריגות
   ה. אף תשובה אינה נכונה

פתרון: התשובה הנכונה היא ב. ניתן לראות כי הזרוע הימנית של התרשים ארוכה יותר באופן ניכר מאורך הזרוע השמאלית, וכן יש תצפיות חריגות גבוהות בקצה הימני של הדיאגרמה. מכאן ניתן להסיק כי זוהי התפלגות אסימטרית ימנית.

2. השמנת יתר מוגדרת כמצב בו ה-BMI הוא 30 ומעלה. מהו בערך אחוז הגברים הסובלים מהשמנת יתר במדינה זו?
   א. לא ניתן לדעת, כיוון שישנן תצפיות חריגות
   ב. כ-50%
   ג. כ-25%
   ד. כ-33%
   ה. אף תשובה אינה נכונה

פתרון: התשובה הנכונה היא ג. הקצה הימני של הקופסה הוא ערכו של הרבעון העליון, וניתן לראות כי ערכו שווה בערך ל-30. מכאן שלכ-25% מהגברים יש ערך BMI של 30 לפחות, כלומר סובלים מהשמנת יתר.

3. איזה מהמשתנים הבאים - ניתן לדרג את ערכיו?
   א. משתנה בסולם רווח/אינטרוולי
   ב. משתנה בסולם סדר (אורדינלי)
   ג. משתנה בסולם שמי
   ד. משתנה בסולם יחס (מנה)

ה. א+ב+ג
ו. א+ב+ד
ז. ב+ג+ד

פתרון: התשובה הנכונה היא ו, כלומר א+ב+ד. סולם המדידה היחיד בן אין משמעות לסדר של הערכים האפשריים (ואף לא לערכים מספריים עצמם, אם יש כאלה) הוא סולם המדידה השמי.

4. איזה מההיגדים הבאים אינו נכון לגבי השאריות (כלומר שגיאות החיזוי) של קו הרגרסיה הלינארית?
   א. השאריות חייבות להיות בעלות ממוצע 0.
   ב. ההתפלגות של השאריות אינה בהכרח התפלגות נורמלית
   ג. סטיית התקן של השאריות בהכרח קטנה או שווה לסטיית התקן של Y
   ד. סטיית התקן של השאריות בהכרח קטנה או שווה לסטיית התקן של X

פתרון: התשובה היא ד.

ניתן להוכיח כי טענה א נכונה. ההוכחה דורשת ידע בסיסי בחדו"א ואלגברה. למעוניינים, ניתן לראות את ההוכחה בלינק זה.

טענה ב נכונה: המודל מניח כי ההתפלגות נורמלית, אך היא אינה חייבת להיות נורמלית. זו הנחה שצריך לבדוק. קיום ההנחה נדרש לצורך בדיקת השערות ובניית רווחי סמך אם גודל המדגם אינו מספיק גדול, אולם הנחה זו אינה דרושה כדי לאמוד את מקדמי הרגרסיה.

טענה ג נכונה: האינטואיציה היא שמודל הרגרסיה מסביר חלק מהשונות של Y, והשאריות מבטאות את חלק השונות של Y שאינה מוסברת. מאחר ושונות השאריות היא חלק השונות הלא מוסברת של Y, היא חייבת להיות קטנה מהשונות של Y. ניתן להוכיח טענה זו באופן אלגברי, על ידי שימוש בהגדרות בסיסות של אמדני המקדמים של משוואת הרגרסיה ושל השונות.

טענה ד אינה נכונה. לדוגמה, בקובץ נתוני המכוניות mtcars, במודל רגרסיה, השונות של המשתנה mpg היא 36.3, ובמודל רגרסיה בו המשנה המסביר הוא mpg והמשתנה המוסבר Y הוא qsec, שונות השאריות היא 2.63. האינטואיציה היא שהשונות של X אינה קשורה לשונות של Y, וכן השונות של השאריות נקבעת על פי מודל הרגרסיה והשונות של Y. אמנם מודל הרגרסיה כולל אינפורמציה שמקורה ב-X, אך גם אינפורמציה שמקורה ב-Y, כלומר השונות המוסברת לא תלויה רק ב-X. ניתן להראות באופן אלגברי, על ידי שימוש בהגדרות בסיסות של אמדני המקדמים של משוואת הרגרסיה ושל השונות, כי יש מצבים בהם השונות של X גדולה משונות השאריות ויש גם מצבים בהם השונות של X קטנה משונות השאריות

5. מה מהבאים אינו נכון לגבי המתאם (r) בין שני משתנים?
   א. המתאם יכול להיות קרוב ל -0 אם הנקודות יוצרות דפוס לא ליניארי.
   ב. במודל רגרסיה, r^2 מודד את חלק השונות ב-Y שמוסבר על ידי X.
   ג. אם ערכו של r קרוב ל-0 זה אומר שאין קשר בין X ל-Y.
   ד. המתאם לא ישתנה אם נוסיף קבוע לכל ערך של X

פתרון: התשובה היא ג.

r מודד את עצמת הקשר הלינארי בין המשתנים. ייתכן קשר בין המשתנים שאינו לינארי אך r יהיה קרוב לאפס. לדוגמה, כאשר ההתפלגות סימטרית ביחס לציר ה-Y, והנתונים, למשל, נמצאים על פרבולה (בנוסחה: $Y=X^2$). זוהי גם דוגמה לנכונות טענה א.

תשובה ב נכונה על פי הגדרת אחוז השונות המוסברת, וניתן להוכיח באופן אלגברי כי r^2 שווה לאחוז השונות המוסברת.

תשובה ד נכונה על פי הגדרות השונות המשותפת וסטיית התקן.

6. בנתונים בעלי אסימטריה שמאלית, אפשר לצפות ש:
   א. יותר מ -50% מהנתונים יהיו גדולים מהחציון.
   ב. הממוצע יהיה קטן מסטיית התקן.
   ג. הממוצע יהיה גדול מהחציון.
   ד. פחות מ-50% מהנתונים קטנים מהממוצע.
   ה. אף תשובה אינה נכונה

פתרון: התשובה היא ד.

בהתפלגות אסימטרית שמאלית הממוצע בדרך כלל קטן מהחציון, ולכן ניתן לצפות פחות מ-50% מהנתונים יהיו קטנים ממנו. ראו הסבר דומה לגבי התפלגות אסימטרית ימנית בפתרון של שאלה 4 במועד א.

7. איזה מהבאים אינו מדד לפיזור?
   א. ממוצע הסטיות בריבוע מ- $\bar{x}$ (The mean of the squared deviations from $\bar{x}$)
   ב. ממוצע הסטיות המוחלטות מ- $\bar{x}$ (The mean of the absolute deviations from $\bar{x}$)
   ג. ממוצע הסטיות מ- $\bar{x}$ (The mean of the deviations from $\bar{x}$)
   ד. הטווח הבין רבעוני (interquartile range)
   ה. אף תשובה אינה נכונה

פתרון: התשובה היא ג. ממוצע הסטיות מהממוצע תמיד שווה לאפס. המדדים שמצוינים בסעיפים א, ב ו-ד הוזכרו בשיעור בנושא מדדי פיזור בחלק הסטטיסטיקה התיאורית של הקורס.

הנתונים הבאים מתייחסים לשאלות 8-9:
במשרד הבריאות חוששים שיותר מ-30% מבנות העשרה מעשנות כדי להישאר רזות, ומתכננים להתחיל תוכנית הסברה ארצית בנושא דימוי-גוף, אם אכן כך הדבר. החוקרים דגמו באקראי קבוצה של 1000 נערות בגילאי 12 עד 15, שלא עישנו. אחרי ארבע שנים הנערות נבדקו שוב, ו-310 אמרו שהן התחילו לעשן כדי לרזות. האם זוהי עדות לכך שיותר מ-30% מבנות העשרה מעשנות כדי להישאר רזות ?

8. ההשערה האלטרנטיבית במחקר זה היא:
   א. $p>0.31$
   ב. $p<0.31$
   ג. $p\leq 0.30$
   ד. $p>0.30$
   ה. אף תשובה אינה נכונה

פתרון: התשובה היא ד. לפי המתואר: במשרד הבריאות **חוששים שיותר מ-30% מבנות העשרה מעשנות כדי להישאר רזות**, ואת הטענה הזו רוצים לבדוק. הדרך לעשות זאת היא להציב אותה כהשערה אלטרנטיבית מול השערת האפס לפיה 30% (או אף פחות) מהנערות מעשנות כדי לרזות. נוכל לקבוע כי האחוז גדול מ-30 אם נדחה את השערת האפס.

9. מה יכולות להיות ההשלכות של טעות מסוג ראשון?

   א. שלא יתחילו בתוכנית ההסברה, בעוד שממדי התופעה בקרב נערות דורשים טיפול.

ב. שיתחילו בתוכנית ההסברה בקרב נערות, בעוד שממדי התופעה בקרב נערות לא דורשים טיפול
ג. שיתחילו בתוכנית ההסברה בקרב נערות בעוד שממדי התופעה אכן דורשים טיפול, אבל לא בדימוי הגוף אלא בהתמכרות לעישון
ד. שלא יתחילו בתוכנית ההסברה, כיוון שהראיות אינן חזקות מספיק
ה. אף תשובה אינה נכונה

פתרון: התשובה היא ב. טעות מסוג ראשון מתרחשת כאשר השערת האפס נכונה ולמרות זאת דוחים אותה. המשמעות של השערת האפס כאן היא ששיעור הנערות המעשנות כדי לרזות קטן או שווה ל-30%. במקרה כזה, אין צורך להתחיל בתכנית הסברה. דחיה מוטעית של השערת האפס תוביל לכך שבכל זאת יתחילו בתכנית כאשר אין צורך בכך.

ב-3 השאלות הבאות, ענו נכון / לא נכון:

10. רווח סמך ברמת סמך של 90% תמיד ארוך יותר מרווח סמך ברמת סמך של 80%

פתרון: הטענה נכונה. כאשר מדובר ברווחי סמך לתוחלות או פרופורציות, ניתן לראות כי ה-"זנבות" שנשארים בקצוות ההתפלגות הנורמלית (או התפלגות t) צרים יותר. זה התבטא גם בערכי z או t גבוהים יותר. למשל, ערך z עבור רווח סמך של 90% הוא 1.645, ועבור רווח סמך של 80% ערכו הוא 1.28.

באופן אינטואיטיבי, כדי לקבל רמת סמך יותר גבוהה צריך שהסיכוי כי רווח הסמך "יכסה" את הפרמטר יהיה יותר גבוה, ולרווח סמך יותר ארוך יותר קל לכסות את הפרמטר.

11. בכל התפלגות סימטרית, ציון התקן (z) של החציון תמיד שווה לאפס

פתרון: הטענה נכונה. ציון התקן של הממוצע שווה לאפס על פי הגדרת ציון התקן. בהתפלגות סימטרית החציון שווה לממוצע, ולכן גם ציון התקן של החציון יהיה שווה לציון התקן של הממוצע.

12. ברגרסיה, אם מקדם המתאם בין המשתנה המסביר והמשתנה המוסבר שלילי, אז אחוז השונות המוסברת הינו נמוך

פתרון: הטענה אינה נכונה. אחוז השונות המוסברת שווה לריבוע של מקדם המתאם. ככל שערכו של מקדם המתאם מתקרב לערך 1-, כך הריבוע שלו קרוב יותר ל-1.

13. לצורך מחקר על הישגי תלמידים לאחר הנהגת תכנית לימוד חדשה במתמטיקה שהונהגה בשנת 2018, נאספו עבור מדגם מייצג של 150 תלמידים נתונים עבור המשתנים הבאים:
    1. גיל התלמיד בשנים
    2. מין התלמיד
    3. תכנית הלימודים (ישנה/חדשה)
    4. מצב סוציו-אקונומי (נמוך/ביניים/גבוה)
    5. שיוך אתני של התלמיד (לבן, אפרו-אמריקאי, היספני, אסיאתי)
    6. ציון התלמיד במתמטיקה בסוף שנת 2017
    7. ציון התלמיד במתמטיקה בסוף שנת 2018

בשנת 2017 כל התלמידים למדו במסגרת תכנית הלימודים הישנה. ב-2018 חלקם למדו במסגרת התכנית הישנה וחלקם במסגרת התכנית החדשה. ניתן להניח כי המדגם מספיק גדול.

א. באיזה סולם מדידה נמדד כל משתנה?

ב. נתון כי 60 תלמידים היו ממצב סוציו-אקונומי נמוך, ו-50 ממצב סוציו-אקונומי גבוה. שרטטו סקיצה של דיאגרמת עמודות עבור נתון זה

ג. הסבירו באיזה שיטה סטטיסטית תשתמשו כדי לענות כל אחת משאלות המחקר הבאות:

1) האם תוחלת הציון במתמטיקה בסוף 2018 אצל תלמידים שלמדו בתכנית הלימודים החדשה גבוהה מתוחלת הציון אצל תלמידים שלמדו בתכנית הלימודים הישנה?

2) אם הציון של תלמיד גדול ב-2018 גדול ב-8 נקודות מציונו ב-2017, אז שינוי זה נחשב לשיפור בהבנת החומר במתמטיקה, ואם הציון בסוף 2018 קטן ב-8 נקודות מהציון ב-2017 זה נחשב להרעה בהבנת החומר (בכל מקרה אחר – אין שינוי ברמת הבנת החומר). האם יש קשר בין רמת הבנת החומר ובין מין התלמיד, וזאת רק עבור התלמידים שלמדו ב-2018 במסגרת תכנית הלימודים החדשה?

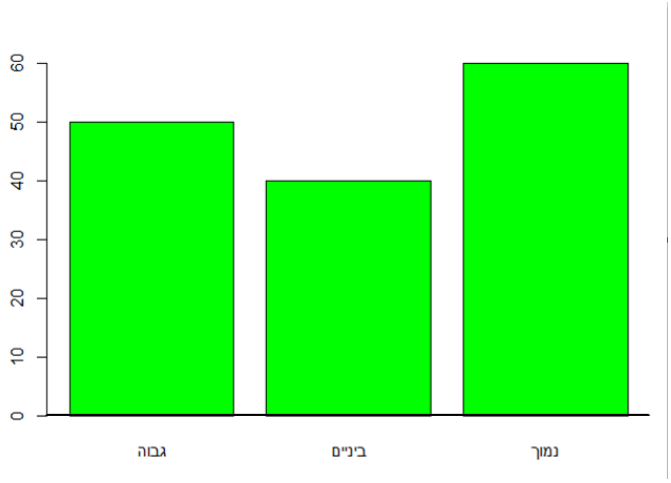
3) עבור כל התלמידים, האם היו הבדלים בין תוחלות הציונים ב-2017 עבור תלמידים מאוכלוסיות אתניות שונות?

.

פתרון:

א. מין התלמיד, שיוכו האתני, וכן תכנית הלימודים נמדדים בסולם מדידה שמי. גם אם נקודד את הערכים שלהם במספרים (למשל תכנית לימודים ישנה=1 ותכנית לימודים חדשה=2, לא תהיה משמעות לערכים המספריים.
המצב הסוציו-אקונומי נמדד בסולם סדר. תלמיד שמצבו הסוציו-אקונומי גבוה נמצא במצב יותר טוב מתלמיד שנמצא במעמד הביניים, וזה בתורו נמצא במצב יותר טוב מתלמיד ממצב סוציו-אקונומי נמוך. גיל התלמיד והציונים הם משתנים כמותיים. לכל ערכים אלה יש אפס מוחלט, כלומר הם נמדדים בסולם מנה.

ב. יש לשים לב כי אם יש 60 תלמידים במצב סוציו-אקונומי נמוך ו-50 במצב סוציו-אקונומי גבוה, הרי שיש גם 40 סטודנטים במצב הביניים. כמו כן, יש להקפיד כי סדר העמודות על ציר ה-X תואם את סדר הערכים של המשתנה (עקרונית לא חשוב אם מסדרים את העמודות מימין לשמאל או משמאל לימין, אך אם הסימון הוא בעברית מומלץ לסדר את העמודות מימין לשמאל). הסקיצה אמורה להיראות דומה לגרף הזה:



ג. המבחנים הסטטיסטיים המתאימים הם:

1) תלמידים מתחלקים לשתי קבוצות נפרדות על פי התכנית לפי למדו ב-2018. המבחן הסטטיסטי המתאים הוא מבחן להשוואת תוחלות בין שתי אוכלוסיות בלתי תלויות.

2) רמת הבנת החומר היא משתנה איכותי בסולם סדר: הרעה/אין שינוי/שיפור. תכנית הלימודים היא משתנה איכותי בסולם שמי. המבחן הסטטיסטי ההתאים הוא מבחן חי-בריבוע לבדיקת השערת אי-תלות, כאשר השערת האפס היא כי אין קשר בין המשתנים.

הנתון כי מעוניינים לבדוק את הקשר בין המשתנים רק עבור התלמידים שלמדו ב-2018 במסגרת תכנית הלימודים החדשה, מגדיר את האוכלוסייה לגביה נבדק הקשר, אך אינו רלוונטי למשתנים עבורם נבדק הקשר.

3) אנו רוצים להשוות בין התוחלות של ארבע אוכלוסיות בלתי תלויות, המוגדרות על ידי השיוך האתני של התלמידים. השיטה הסטטיסטית המתאימה למצב זה היא ניתוח שונות

14. בתהליך ייצור חדש, תפוקת החומר המיוצר בק"ג (yield) תלויה במשך הזמן (time) שבו התהליך מתבצע (בדקות) ובטמפרטורה הממוצעת (temp) במהלך התהליך (מעלות צלזיוס). כדי לבדוק את הקשר בין שלושת המשתנים, נערכו 50 ניסויים בהם התהליך בוצע על ידי אנשי מקצוע, כאשר בכל ניסוי הטמפרטורה ומשך הזמן נקבעו על ידי ה-"תחושה" של המבצע. טווח הטמפרטורות היה בין 55 ל-105 מעלות וטווח הזמן היה בין 2 ל-25 דקות.

בסופו של דבר נותחו הנתונים במודל רגרסיה לינארית. להלן תוצאות חלקיות מהמודל:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.363					
R Square						
Adjusted R Square	0.095					
Standard Error	44.058					
Observations	50					
ANOVA						
	df	*SS*	*MS*	*F*	*gnificance F*	
Regression	2	13828.2	6914.1	3.6	0.03628	
Residual	47	91230.1	1941.1			
Total	49					
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	120.1	62.3	1.9	0.05987	-5.2	245.4
temp	1.6	0.7	2.2	0.03523	0.1	3.1
time	-1.0	1.2	-0.8		-3.4	1.4

א. מהי משוואת הרגרסיה?
ב. מהו אחוז השונות המוסברת על ידי הרגרסיה?
ג. מהי סטיית התקן של השאריות?
ד. האם ה-p-value עבור המקדם של time: קטן מ-0.05, גדול מ-0.05, שווה בדיוק ל-0.05 או שלא ניתן לדעת?
ה. האם מקדם המתאם בין yield ו-temp הינו חיובי, שלילי, שווה ל-0 או שלא ניתן לדעת?
ו. בניסוי בו הטמפרטורה הייתה 84 מעלות שארך 8.6 דקות, כמות החומר שיוצרה הייתה 210 ק"ג. מהו ערך השארית עבור תצפית זו?
ז. מה הניבוי לתפוקה כאשר הטמפרטורה היא 60 מעלות ומשך הזמן הוא 20 דקות?
ח. כיצד ישתנה הניבוי לכמות התפוקה אם בניסוי מסויים הטמפרטורה תועלה ב-10 מעלות?

פתרון

הערה כללית: פתרון שאלה זו אינו כולל הסברים מקיפים. כל השאלות הן שאלות סטנדרטיות שנדונו בשיעורים (קישור למצגת השיעורים) ובתרגול (קישור למצגת פתרון התרגיל). הקלטות השיעורים זמינות ביוטיוב.

א. המשוואה היא $\text{Yield} = 120.1 + 1.6 \cdot \text{temp} - 1.0 \cdot \text{time}$

ב. אחוז השונות המוסברת מחושב על יד חלוקת הערך של SS בשורת Regression בטבלת ניתוח השונות (ANOVA) בחלוקת הערך של SS בשורת ה-Total. ערך ה-Total אמנם אינו נתון, אך ניתן לחשב אותו על ידי סיכום שני הערכים שנתונים בעמודה. לכן החישוב הוא:

$$SS_{Total} = 13828.2 + 91230.1 = 105058.3$$
$$R^2 = \frac{13828.2}{105058.3} = 0.1316$$

ג. סטיית התקן של השאריות היא הערך Standard Error בטבלת Regression Statistics והיא שווה ל-44.058.

ד. רווח הסמך (ברמת סמך של 95%) למקדם של time מכיל את האפס, ולכן לא דוחים את השערת האפס כי המקדם הוא 0. לכן ה- p-value גדול מ-0.05.

ה. סימנו של מקדם המתאם זהה לסימן של מקדם הרגרסיה. מקדם הרגרסיה של temp הוא חיובי ולכן גם מקדם המתאם בין yield ו-temp הינו חיובי,

ו. הניבוי לתפוקה הוא $120.1 + 1.6 \cdot 84 - 1.0 \cdot 8.6 = 245.9$ ק"ג. השארית שווה לערך האמיתי פחות הניבוי, כלומר השארית שווה ל- $210 - 245.9 = -35.9$ ק"ג.

ז. באופן דומה לסעיף הקודם, הניבוי הוא $120.1 + 1.6 \cdot 60 - 1.0 \cdot 20 = 196.1$ ק"ג.

ח. הניבוי ישתנה ב- $1.6 \cdot 10 = 16$ ק"ג.