

סטטיסטיקה יישומית



תוכן העניינים

1.	מבחנים אפרמטריים למדגם יחיד.....	1
2.	מבחנים אפרמטריים למדגמים מזווגים.....	6
3.	מבחנים אפרמטריים למדגמים בלתי תלויים.....	23
4.	מבחני חי בריבוע.....	30
5.	ניתוח שונות חד כיוונית.....	40
6.	מקדם המתאם (מדד קשר) הלינארי ומובהקותו.....	49
7.	רגרסיה פשוטה.....	72
8.	רגרסיה מרובה.....	84
9.	רגרסיה - שאלות ממבחנים.....	89

סטטיסטיקה יישומית

פרק 1 - מבחנים אפרמטריים למדגם יחיד

תוכן העניינים

1. מבחן הבינום 1

מבחנים אפרמטריים למדגם יחיד

מבחן הבינום – רקע

מבחן הבינום הינו מבחן סטטיסטי על הפרמטר p . הפרמטר p מייצג את פרופורציית ההצלחות באוכלוסייה כלומר, הסיכוי בניסוי בודד להצליח. ההשערות האפשריות על הפרמטר הן:

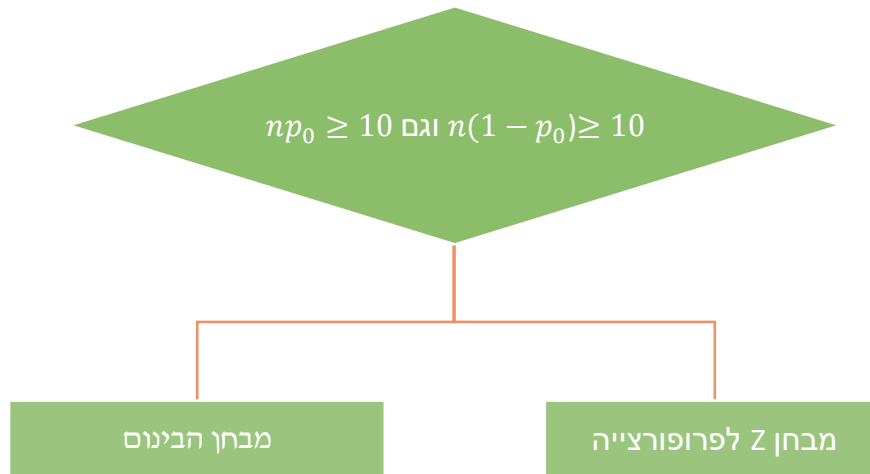
$H_0 : p = p_0$	$H_0 : p = p_0$	$H_0 : p = p_0$	השערת האפס:
$H_1 : p > p_0$	$H_1 : p < p_0$	$H_1 : p \neq p_0$	השערה אלטרנטיבית:

דוגמה:

במדינה אירופאית התנהל לפני 4 שנים משאל עם בדבר לגליזציה של הקנאביס. במשאל העם 40% מהאזרחים היו בעד לגליזציה של הקנאביס ובשל כך חוק הלגליזציה לא עבר באותה המדינה. במדגם שנעשה כיום בו השתתפו 15 אזרחים מהמדינה 10 ענו שהם תומכים בלגליזציה של הקנאביס במדינה. פרלמנטר מהמדינה חוקר האם כיום אחוז התומכים בלגליזציה של הקנאביס במדינה עלה יחסית לאחוז שהתקבל במשאל העם במדינה לפני 4 שנים. רשמו את השערות המחקר.

במבחן הבינום מבצעים מדגם אקראי בגודל n ומתבוננים במספר ההצלחות שהתקבלו במדגם, אותן נסמן ב- Y . בהנחה והתצפיות במדגם בלתי תלויות זו בזו אנו אומרים ש: $Y \sim B(n, p)$.

מבחן הבינום נכנס לקטגוריה של מבחנים אפרמטרים והוא בא כחלופה למבחן הפרמטרי על פרופורציה אחת כאשר התנאים לקירוב הנורמלי אינם מתקיימים.



דוגמה:

מהו המבחן הסטטיסטי המתאים? נמקו.

הטכניקה הנוחה ביותר למבחן הבינום היא לחשב את מובהקות התוצאה ולדחות את השערת האפס אם $\alpha \geq PV$. מובהקות התוצאה היא הסיכוי לתוצאות של המדגם וקיצוני יותר בהנחת השערת האפס. כזכור, פונקציית ההסתברות של ההתפלגות הבינומית היא:

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, \dots, n$$

כאשר התוחלת של ההתפלגות הבינומית היא: $E(Y) = np$. בעמוד הבא מצורפת טבלה של התפלגות בינומית מצטברת. זו טבלה שיכולה לעזור בתהליך החישוב.

דוגמה:

חשבו את מובהקות התוצאה. מה תהיה מסקנת המחקר ברמת מובהקות של 5%?

טבלת התפלגות בינומית מצטברת

n	x	p									
		0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.80	0.90
5	0	0.5905	0.3277	0.2373	0.1681	0.0778	0.0312	0.0102	0.0024	0.0003	0.0000
	1	0.9185	0.7373	0.6328	0.5282	0.3370	0.1875	0.0870	0.0308	0.0067	0.0005
	2	0.9914	0.9421	0.8965	0.8369	0.6826	0.5000	0.3174	0.1631	0.0579	0.0086
	3	0.9995	0.9933	0.9844	0.9692	0.9130	0.8125	0.6630	0.4718	0.2627	0.0815
	4	1.0000	0.9997	0.9990	0.9976	0.9898	0.9688	0.9222	0.8319	0.6723	0.4095
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0	0.3487	0.1074	0.0563	0.0282	0.0060	0.0010	0.0001	0.0000	0.0000	0.0000
	1	0.7361	0.3758	0.2440	0.1493	0.0464	0.0107	0.0017	0.0001	0.0000	0.0000
	2	0.9298	0.6778	0.5256	0.3828	0.1673	0.0547	0.0123	0.0016	0.0001	0.0000
	3	0.9872	0.8791	0.7759	0.6496	0.3823	0.1719	0.0548	0.0106	0.0009	0.0000
	4	0.9984	0.9672	0.9219	0.8497	0.6331	0.3770	0.1662	0.0474	0.0064	0.0002
	5	0.9999	0.9936	0.9803	0.9527	0.8338	0.6230	0.3669	0.1503	0.0328	0.0016
	6	1.0000	0.9991	0.9965	0.9894	0.9452	0.8281	0.6177	0.3504	0.1209	0.0128
	7	1.0000	0.9999	0.9996	0.9984	0.9877	0.9453	0.8327	0.6172	0.3222	0.0702
	8	1.0000	1.0000	1.0000	0.9999	0.9983	0.9893	0.9536	0.8507	0.6242	0.2639
	9	1.0000	1.0000	1.0000	1.0000	0.9999	0.9990	0.9940	0.9718	0.8926	0.6513
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
15	0	0.2059	0.0352	0.0134	0.0047	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.5490	0.1671	0.0802	0.0353	0.0052	0.0005	0.0000	0.0000	0.0000	0.0000
	2	0.8159	0.3980	0.2361	0.1268	0.0271	0.0037	0.0003	0.0000	0.0000	0.0000
	3	0.9444	0.6482	0.4613	0.2969	0.0905	0.0176	0.0019	0.0001	0.0000	0.0000
	4	0.9873	0.8358	0.6865	0.5155	0.2173	0.0592	0.0094	0.0007	0.0000	0.0000
	5	0.9978	0.9389	0.8516	0.7216	0.4032	0.1509	0.0338	0.0037	0.0001	0.0000
	6	0.9997	0.9819	0.9434	0.8689	0.6098	0.3036	0.0951	0.0152	0.0008	0.0000
	7	1.0000	0.9958	0.9827	0.9500	0.7869	0.5000	0.2131	0.0500	0.0042	0.0000
	8	1.0000	0.9992	0.9958	0.9848	0.9050	0.6964	0.3902	0.1311	0.0181	0.0003
	9	1.0000	0.9999	0.9992	0.9963	0.9662	0.8491	0.5968	0.2784	0.0611	0.0023
	10	1.0000	1.0000	0.9999	0.9993	0.9907	0.9408	0.7827	0.4845	0.1642	0.0127
	11	1.0000	1.0000	1.0000	0.9999	0.9981	0.9824	0.9095	0.7031	0.3518	0.0556
	12	1.0000	1.0000	1.0000	1.0000	0.9997	0.9963	0.9729	0.8732	0.6020	0.1841
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9948	0.9647	0.8329	0.4510
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9953	0.9648	0.7941
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	0	0.1216	0.0115	0.0032	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.3917	0.0692	0.0243	0.0076	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.6769	0.2061	0.0913	0.0355	0.0036	0.0002	0.0000	0.0000	0.0000	0.0000
	3	0.8670	0.4114	0.2252	0.1071	0.0160	0.0013	0.0001	0.0000	0.0000	0.0000
	4	0.9568	0.6296	0.4148	0.2375	0.0510	0.0059	0.0003	0.0000	0.0000	0.0000
	5	0.9887	0.8042	0.6172	0.4164	0.1256	0.0207	0.0016	0.0000	0.0000	0.0000
	6	0.9976	0.9133	0.7858	0.6080	0.2500	0.0577	0.0065	0.0003	0.0000	0.0000
	7	0.9996	0.9679	0.8982	0.7723	0.4159	0.1316	0.0210	0.0013	0.0000	0.0000
	8	0.9999	0.9900	0.9591	0.8867	0.5956	0.2517	0.0565	0.0051	0.0001	0.0000
	9	1.0000	0.9974	0.9861	0.9520	0.7553	0.4119	0.1275	0.0171	0.0006	0.0000
	10	1.0000	0.9994	0.9961	0.9829	0.8725	0.5881	0.2447	0.0480	0.0026	0.0000
	11	1.0000	0.9999	0.9991	0.9949	0.9435	0.7483	0.4044	0.1133	0.0100	0.0001
	12	1.0000	1.0000	0.9998	0.9987	0.9790	0.8684	0.5841	0.2277	0.0321	0.0004
	13	1.0000	1.0000	1.0000	0.9997	0.9935	0.9423	0.7500	0.3920	0.0867	0.0024
	14	1.0000	1.0000	1.0000	1.0000	0.9984	0.9793	0.8744	0.5836	0.1958	0.0113
	15	1.0000	1.0000	1.0000	1.0000	0.9997	0.9941	0.9490	0.7625	0.3704	0.0432
	16	1.0000	1.0000	1.0000	1.0000	1.0000	0.9987	0.9840	0.8929	0.5886	0.1330
	17	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9964	0.9645	0.7939	0.3231
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9924	0.9308	0.6083
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9992	0.9885	0.8784
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

שאלות

- (1) אחוז האנשים שאוכלים גלידות בחודשי החורף הינו 30%, קיים חשש בקרב מיצרי הגלידות כי השנה פחת אחוז אוכלי הגלידות בחודשי החורף. לשם כך נדגמו 12 אנשים אשר מתוכם 2 טענו שהם אוכלים גלידה בחודשי החורף.
- א. רשמו את השערות המוצגות בשאלה זו וציינו מהו המבחן הסטטיסטי המתאים.
- ב. מה תהיה המסקנה ברמת מובהקות של 10%?
- ג. מה תהיה המסקנה ברמת מובהקות גדולה יותר מ- 10%?
- (2) מטבע הוטל 15 פעמים במטרה לבדוק האם המטבע סימטרי.
- בסך הכול התקבלו בהטלות 3 פעמים עץ.
- א. בדקו ברמת מובהקות של 5% האם המטבע הוא סימטרי.
- ב. עבור אילו רמות מובהקות נוכל להסיק שהמטבע אינו סימטרי?
- (3) אחוז המובטלים במשק לפני 5 שנים היה 10%. מעוניינים לבדוק האם כיום אחוז המובטלים קטן לעומת זה שהיה לפני 5 שנים. במדגם של 20 אנשים התקבל מובטל אחד ויחיד.
- א. מהי רמת המובהקות הקטנה ביותר עבורה יוסק שכיום אחוז האבטלה נמוך מאשר לפני 5 שנים?
- ב. מה תהיה המסקנה אם רמת המובהקות תהיה 5%?
- (4) נניח ש 40% מהאוכלוסייה מכירה את המוצרים של חברת "רמקס". החברה מתכננת לצאת בקמפיין פרסום שמטרתו להעלות את המודעות לקיומם של המוצרים של החברה באוכלוסייה. בזמן הקמפיין נדגמו 20 אנשים אקראיים מתוכם 70% טענו שהם מכירים את המוצרים של חברת "רמקס".
- א. מהי מובהקות התוצאה?
- ב. כיצד מובהקות התוצאה הייתה משתנה אם במדגם 80% היו טוענים שהם מכירים את המוצרים של חברת "רמקס"?

- (5) חברה לתוספי מזון דיווחה שנטילת מולטי ויטמין מקטינה את הסיכוי לחלות במחלות חורף במהלך החורף. לפי משרד הבריאות 70% מהאוכלוסייה חולים במחלות חורף במהלך החורף. במחקר השתתפו 20 אנשים אשר נטלו במהלך שנה מולטי ויטמין. במהלך החורף נמצא ש 12 מתוכם חלו במחלות חורף במהלך החורף. את המחקר יש לבצע ברמת מובהקות של 5%. נסמן ב: PV את מובהקות התוצאה של המחקר.
- א. האם המשפט הבא נכון?
"אם כלל האוכלוסייה הייתה נוטלת מולטי ויטמין אז בסיכוי של 5% ניתן להגיד ש-70% מהאוכלוסייה הייתה חולה במחלת חורף".
- ב. האם המשפט הבא נכון?
"אם כלל האוכלוסייה הייתה נוטלת מולטי ויטמין אז בסיכוי של PV 70% מהאוכלוסייה הייתה חולה במחלת חורף".
- ג. חשבו את PV.

תשובות סופיות

- | | | |
|------------------------|-----------------------|------------------|
| (1) א. מבחן הבינום. | ב. לא נדחה את H_0 . | ג. לא ניתן לדעת. |
| | | |
| | | |
| | | |
| | | |
| (2) א. נדחה את H_0 . | ב. לפחות 0.0352 | |
| (3) א. 0.3917 | ב. לא נדחה את H_0 . | |
| (4) א. 0.0065 | ב. תקטן. | |
| (5) א. לא נכון. | ב. לא נכון. | ג. 0.2277. |

סטטיסטיקה יישומית

פרק 2 - מבחנים אפרמטרים למדגמים מזווגים

תוכן העניינים

1. מבחן הסימן 6
2. מבחן הסימן - על ידי שימוש בטבלה בינומית 9
3. מבחן הסימן - על ידי שימוש בקירוב הנורמלי 13
4. מבחן ווילקוקסון - על ידי שימוש בקירוב הנורמלי 17
5. תרגול בזיהוי מבחנים 20

מבחנים אפרמטרים למדגמים מזווגים

מבחן הסימן – רקע

מבחן הסימן הוא מבחן שמשמש בו כאשר לפנינו מדגם מזווג ולא ניתן להניח שהמשתנה הנחקר מתפלג נורמלית.

גם אם המשתנה הנחקר מתפלג נורמלית ניתן לבצע את מבחן הסימן אבל מבחן T למדגמים מזווגים יהיה מבחן עם עוצמה גבוהה יותר ולכן יש לבצע אותו. מבחן הסימן נחשב למבחן אפרמטרי – מבחנים אפרמטרים הינם כל המבחנים שאינם דורשים שהמשתנה הנחקר יתפלג נורמלית. מבחן הסימן נקרא כך כיוון שהוא דורש הגדרת סימן לכל תצפית:

(+) – אם מצב X גבוה ממצב Y.

(-) – אם מצב Y גבוה ממצב X.

(0) – אין הבדל בין המצבים.

במבחן הסימן נתעלם מהפרשים שהם 0, ולכן נסמן את מספר ההפרשים

האפקטיביים (השוניים מאפס) ב- n^* .

תחת השערת האפס נאמר שהסיכוי לקבל הפרש חיובי ($p+$) שווה לסיכוי לקבל הפרש שלילי ($p-$).

השערות המבחן:

$$\begin{array}{ll} H_0 : P_- = 0.5 & \text{או} & H_0 : P_+ = 0.5 \\ H_1 : P_- \neq, <, > 0.5 & & H_1 : P_+ \neq, <, > 0.5 \end{array}$$

נסמן ב $n(+)$ או ב S_+ את מספר התצפיות שקיבלו את הערך (+), ובאופן דומה:

נסמן ב $n(-)$ או ב S_- את מספר התצפיות שקיבלו את הערך (-).

ניתן לומר שבהנחת השערת האפס: $n(+), n(-) \sim B(n^*, 0.5)$.

נחשב את PV על סמך תוצאות המדגם בעזרת ההתפלגות הבינומית כך שאם

$PV \leq \alpha$ נדחה את השערת האפס.

במבחן הסימן אין התייחסות לגודל הפער בתצפיות אלא רק את כיוון ההבדל.

דוגמה (פתרון בהקלטה) :

קופות החולים טוענות כי רכישת תרופות שאינן דורשות מרשם רופא, הינן זולות יותר אצלן מאשר מברשות הפארם. דגמו 11 תרופות ובדקו את מחירן בבית המרקחת של קופות החולים וברשת הפארם. המחיר המוצג הינו עבור קפסולה בודדת: בדקו ברמת מובהקות של 5% באמצעות מבחן הסימן.

שם התרופה	קופת חולים	פארם
אדוויל	1.2	1.5
אקמול	2.6	2.6
אופטלגין	0.9	1.4
פוסטינור	3.5	3.2
סטרפסיל	1.1	1.4
נורפן	1.7	1.8
לורסטין	0.8	1.1
קולדקס	1.5	2
אלרגיז	2	2.8
נוסידקס	2	2.5
קורמיר	3	3.3

שאלות

- (1) רוצים לבדוק את הטענה שהציונים במבחן בסטטיסטיקה ב גבוהים מאשר בסטטיסטיקה א. נלקחו 10 סטודנטים שסיימו את סטטיסטיקה ב. עבור כל סטודנט נבדק מה הציון בסטטיסטיקה א ומה הציון בסטטיסטיקה ב. להלן התוצאות שהתקבלו:

א	62	74	68	94	82	67	65	84	78	80
ב	70	80	70	90	77	67	80	86	79	82

- א. בדקו ברמת מובהקות של 5% באמצעות מבחן הסימן.
 ב. כיצד התשובה לסעיף הקודם הייתה משתנה אם יוחלט לתת פקטור של 2 נקודות לכל הסטודנטים בשני המועדים?

- (2) מעוניינים לבדוק האם ההוצאות על "ג'אנק פוד" בקרב הסטודנטים רבות יותר בזמן הלימודים לעומת ימי החופשה.
 נדגמו 15 סטודנטים מקריים, אצל 13 ההוצאות בתקופת הלימודים היו גבוהות יותר מימי החופשה ואצל 2 נמוכות יותר.
 מה מסקנתך בר"מ של 0.05?

- (3) מעוניינים לבדוק האם סם מסוים משפיע על לחץ הדם. נלקחו 24 אנשים אשר נמדד להם לחץ הדם לאחר מכן ניתן להם הסם ושוב נמדדו להם את לחץ הדם. לחמישה אנשים לחץ הדם לא השתנה ל 15 אנשים לחץ הדם עלה וליתר לחץ הדם ירד אחרי לקיחת הסם. מה מסקנתכם ברמת מובהקות של 5%?

- (4) במדגם שנעשה על 15 משפחות השוו את רמת הביטחון העצמי של הבכור במשפחה לעומת הצעיר שבמשפחה. תוצאות המדגם הראו שאצל 7 משפחות רמת הביטחון העצמי של הבכור הייתה גבוהה יותר, אצל 3 משפחות רמת הביטחון העצמי של הצעיר הייתה גבוהה יותר ואצל 5 משפחות לא נמצא הבדל בין האחים מבחינת רמת הביטחון העצמי. טענת החוקר הייתה שבמשפחות לבכור ביטחון עצמי גבוה מזה של הצעיר במשפחה.
- א. מהי רמת המובהקות המינימלית עבורה יוחלט לקבל את טענת החוקר?
- ב. רמת הביטחון הוערכה על ידי פסיכולוג זוטר. פסיכולוג בכיר ביצע הערכה מחודשת וקבע שלמשפחה אחת במדגם הייתה הערכה שגויה: הפסיכולוג הזוטר קבע שלצעיר במשפחה יש ביטחון עצמי יותר גבוה למרות שלאח הבכור יש ביטחון עצמי יותר גבוה במשפחה הזו.
- מה יקרה לרמת המובהקות המינימלית שחושבה בשאלה הקודמת?

- (5) איזה מהטענות הבאות נכונות?

א. $n(+) + n(-) = n^*$

ב. $n(+) + n(-) = n$

ג. $n(+) = n(-)$

ד. $n(+) - n(-) = n^*$

תשובות סופיות

- (1) א. לא נדחה את H_0 . ב. לא תשתנה המסקנה.

- (2) נדחה את H_0 .

- (3) נדחה את H_0 .

- (4) א. 0.172. ב. תקטן.

- (5) א'.

מבחן הסימן (שימוש בטבלה של התפלגות בינומית) – רקע

מבחן הסימן הוא מבחן שמשמש שם בו כאשר לפנינו מדגם מזווג ולא ניתן להניח שהמשתנה הנחקר מתפלג נורמלית.

גם אם המשתנה הנחקר מתפלג נורמלית ניתן לבצע את מבחן הסימן אבל מבחן T למדגמים מזווגים יהיה מבחן עם עוצמה גבוהה יותר ולכן יש לבצע אותו. מבחן הסימן נחשב למבחן אפרמטרי - מבחנים אפרמטרים הינם כל המבחנים שאינם דורשים שהמשתנה הנחקר יתפלג נורמלית.

מבחן הסימן נקרא כך כיוון שהוא דורש הגדרת סימן לכל תצפית:

(+) – אם מצב X גבוה ממצב Y.

(-) – אם מצב Y גבוה ממצב X.

(0) – אין הבדל בין המצבים.

במבחן הסימן נתעלם מהפרשים שהם 0, ולכן נסמן את מספר ההפרשים

האפקטיביים (השוניים מאפס) ב- n^* .

תחת השערת האפס נאמר שהסיכוי לקבל הפרש חיובי ($p+$) שווה לסיכוי לקבל

הפרש שלילי ($p-$).

השערות המבחן:

$$\begin{array}{ll} H_0: P_- = 0.5 & \text{או} & H_0: P_+ = 0.5 \\ H_1: P_- \neq, <, > 0.5 & & H_1: P_+ \neq, <, > 0.5 \end{array}$$

נסמן ב $n(+)$ או ב S_+ את מספר התצפיות שקיבלו את הערך (+), ובאופן דומה:

נסמן ב $n(-)$ או ב S_- את מספר התצפיות שקיבלו את הערך (-).

ניתן לומר שבהנחת השערת האפס: $n(+), n(-) \sim B(n^*, 0.5)$.

נחשב את PV על סמך תוצאות המדגם בעזרת ההתפלגות הבינומית כך שאם

$PV \leq \alpha$ נדחה את השערת האפס.

במבחן הסימן אין התייחסות לגודל הפער בתצפיות אלא רק את כיוון ההבדל.

במקום להציב בפונקציית ההסתברות של ההתפלגות הבינומית.

נשתמש בטבלה של פונקציית ההסתברות המצטברת של ההתפלגות הבינומית עבור סיכוי של 0.5 להצלחה.

באמצעות הטבלה הבאה נוכל לחשב את PV.

טבלת הסתברות בינומית (מצטברת) עבור $p=0.5$

$\begin{matrix} X \\ n \end{matrix}$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	031	188	500	812	969	†										
6	016	109	344	656	891	984	†									
7	008	062	227	500	773	938	992	†								
8	004	035	145	363	637	855	965	996	†							
9	002	020	090	254	500	746	910	980	998	†						
10	001	011	055	172	377	623	828	945	989	999	†					
11		006	033	113	274	500	726	887	967	994	†	†				
12		003	019	073	194	387	613	806	927	981	997	†	†			
13		002	011	046	133	291	500	709	867	954	989	998	†	†		
14		001	006	029	090	212	395	605	788	910	971	991	999	†	†	
15			004	018	059	151	304	500	696	849	941	982	996	†	†	†
16			002	011	038	105	227	402	598	773	895	962	989	998	†	†
17			001	006	025	072	166	315	500	685	834	928	975	994	999	†
18			001	004	015	048	119	240	407	593	760	881	952	985	996	999
19				002	010	032	084	180	324	500	676	820	916	968	990	998
20				001	006	021	058	132	252	412	588	748	868	942	979	994
21				001	004	013	039	095	192	332	500	668	808	905	961	987
22					002	008	026	067	143	262	416	584	738	857	933	974
23					001	005	017	047	105	202	339	500	661	798	895	953
24					001	003	011	032	076	154	271	419	581	729	846	924
25						002	007	022	054	115	212	345	500	655	788	885

דוגמה: (פתרון בהקלטה)

קופות החולים טוענות כי רכישת תרופות שאינן דורשות מרשם רופא, הינן זולות יותר אצלן מאשר מברשתות הפארם. דגמו 11 תרופות ובדקו את מחירן בבית המרקחת של קופות החולים וברשת הפארם. המחיר המוצג הינו עבור קפסולה בודדת. בדקו ברמת מובהקות של 5% באמצעות מבחן הסימן.

שם התרופה	קופת חולים	פארם
אדוויל	1.2	1.5
אקמול	2.6	2.6
אופטלגין	0.9	1.4
פוסטינור	3.5	3.2
סטרפסיל	1.1	1.4
נורפן	1.7	1.8
לורסטין	0.8	1.1
קולדקס	1.5	2
אלרגיז	2	2.8
נוסידקס	2	2.5
קורמיר	3	3.3

שאלות

- (1) רוצים לבדוק את הטענה שהציונים במבחן בסטטיסטיקה ב גבוהים מאשר בסטטיסטיקה א. נלקחו 10 סטודנטים שסיימו את סטטיסטיקה ב. עבור כל סטודנט נבדק מה הציון בסטטיסטיקה א ומה הציון בסטטיסטיקה ב. להלן התוצאות שהתקבלו:

80	78	84	65	67	82	94	68	74	62	א
82	79	86	80	67	77	90	70	80	70	ב

- א. בדקו ברמת מובהקות של 5% באמצעות מבחן הסימן.
- ב. כיצד התשובה לסעיף הקודם הייתה משתנה אם יוחלט לתת פקטור של 2 נקודות לכל הסטודנטים בשני המועדים?
- (2) מעוניינים לבדוק האם ההוצאות על "ג'אנק פוד" בקרב הסטודנטים רבות יותר בזמן הלימודים לעומת ימי החופשה. נדגמו 15 סטודנטים מקריים, אצל 13 ההוצאות בתקופת הלימודים היו גבוהות יותר מימי החופשה ואצל 2 נמוכות יותר. מה מסקנתך בר"מ של 0.05?
- (3) מעוניינים לבדוק האם סם מסוים משפיע על לחץ הדם. נלקחו 24 אנשים אשר נמדד להם לחץ הדם לאחר מכן ניתן להם הסם ושוב מדדו להם את לחץ הדם. לחמישה אנשים לחץ הדם לא השתנה ל 15 אנשים לחץ הדם עלה וליתר לחץ הדם ירד אחרי לקיחת הסם. מה מסקנתכם ברמת מובהקות של 5%?
- (4) במדגם שנעשה על 15 משפחות השוו את רמת הביטחון העצמי של הבכור במשפחה לעומת הצעיר שבמשפחה. תוצאות המדגם הראו שאצל 7 משפחות רמת הביטחון העצמי של הבכור הייתה גבוהה יותר, אצל 3 משפחות רמת הביטחון העצמי של הצעיר הייתה גבוהה יותר ואצל 5 משפחות לא נמצא הבדל בין האחים מבחינת רמת הביטחון העצמי. טענת החוקר הייתה שבמשפחות לבכור ביטחון עצמי גבוה מזה של הצעיר במשפחה.
- א. מהי רמת המובהקות המינימלית עבורה יוחלט לקבל את טענת החוקר?
- ב. רמת הביטחון הוערכה על ידי פסיכולוג זוטר. פסיכולוג בכיר ביצע הערכה מחודשת וקבע שלמשפחה אחת במדגם הייתה הערכה שגויה: הפסיכולוג הזוטר קבע שלצעיר במשפחה יש ביטחון עצמי יותר גבוה למרות שלאח הבכור יש ביטחון עצמי יותר גבוה במשפחה הזו. מה יקרה לרמת המובהקות המינימלית שחושבה בשאלה הקודמת?

- (5) איזה מהטענות הבאות נכונות?

א. $n(+) + n(-) = n^*$

ב. $n(+) + n(-) = n$

ג. $n(+) = n(-)$

ד. $n(+) - n(-) = n^*$

תשובות סופיות

- (1) א. לא נדחה H_0 . ב. לא תשתנה המסקנה.
- (2) נדחה H_0 .
- (3) נדחה H_0 .
- (4) א. 0.172 ב. תקטן.
- (5) א'.

מבחן הסימן (שימוש בקירוב הנורמלי) – רקע

מבחן הסימן הוא מבחן שמשמש בו כאשר לפנינו מדגם מזווג ולא ניתן להניח שהמשתנה הנחקר מתפלג נורמלית.

גם אם המשתנה הנחקר מתפלג נורמלית ניתן לבצע את מבחן הסימן אבל מבחן T למדגמים מזווגים יהיה מבחן עם עוצמה גבוהה יותר ולכן יש לבצע אותו. מבחן הסימן נחשב למבחן אפרמטרי – מבחנים אפרמטרים הינם כל המבחנים שאינם דורשים שהמשתנה הנחקר יתפלג נורמלית. מבחן הסימן נקרא כך כיוון שהוא דורש הגדרת סימן לכל תצפית:

(+) – אם מצב X גבוה ממצב Y .

(-) – אם מצב Y גבוה ממצב X .

(0) – אין הבדל בין המצבים.

במבחן הסימן נתעלם מהפרשים שהם 0, ולכן נסמן את מספר ההפרשים

האפקטיביים (השוניים מאפס) ב- n^* .

תחת השערת האפס נאמר שהסיכוי לקבל הפרש חיובי ($p+$) שווה לסיכוי לקבל הפרש שלילי ($p-$).

השערות המבחן:

$$\begin{array}{ll} H_0 : P_- = 0.5 & \text{או} & H_0 : P_+ = 0.5 \\ H_1 : P_- \neq, <, > 0.5 & & H_1 : P_+ \neq, <, > 0.5 \end{array}$$

נסמן ב $n(+)$ או ב S_+ את מספר התצפיות שקיבלו את הערך (+), ובאופן דומה:

נסמן ב $n(-)$ או ב S_- את מספר התצפיות שקיבלו את הערך (-).

ניתן לומר שבהנחת השערת האפס: $n(+), n(-) \sim B(n^*, 0.5)$.

נחשב את PV על סמך תוצאות המדגם בעזרת ההתפלגות הבינומית כך שאם

$$PV \leq \alpha \quad \text{נדחה את השערת האפס.}$$

במבחן הסימן אין התייחסות לגודל הפער בתצפיות אלא רק את כיוון ההבדל. אם התנאים לקירוב נורמלי מתקיימים ניתן להמיר את ההתפלגות הבינומית להתפלגות נורמלית:

$$\text{אם: } n^* \cdot 0.5 \geq 5$$

$$\text{אז: } S_{\pm} \sim N\left(n^* \cdot \frac{1}{2}, n^* \cdot \frac{1}{4}\right)$$

דוגמה (פתרון בהקלטה) :

קופות החולים טוענות כי רכישת תרופות שאינן דורשות מרשם רופא, הינן זולות יותר אצלן מאשר מברשתות הפארם. דגמו 11 תרופות ובדקו את מחירן בבית המרקחת של קופות החולים וברשת הפארם. המחיר המוצג הינו עבור קפסולה בודדת. בדקו ברמת מובהקות של 5% באמצעות מבחן הסימן.

שם התרופה	קופת חולים	פארם
אדוויל	1.2	1.5
אקמול	2.6	2.6
אופטלגין	0.9	1.4
פוסטינור	3.5	3.2
סטרפסיל	1.1	1.4
נורפן	1.7	1.8
לורסטין	0.8	1.1
קולדקס	1.5	2
אלרגיז	2	2.8
נוסידקס	2	2.5
קורמיר	3	3.3

שאלות

- (1) רוצים לבדוק את הטענה שהציונים במבחן בסטטיסטיקה ב גבוהים מאשר בסטטיסטיקה א. נלקחו 10 סטודנטים שסיימו את סטטיסטיקה ב. עבור כל סטודנט נבדק מה הציון בסטטיסטיקה א ומה הציון בסטטיסטיקה ב. להלן התוצאות שהתקבלו:

80	78	84	65	67	82	94	68	74	62	א
82	79	86	80	67	77	90	70	80	70	ב

- א. בדקו ברמת מובהקות של 5% באמצעות מבחן הסימן.
 ב. כיצד התשובה לסעיף הקודם הייתה משתנה אם יוחלט לתת פקטור של 2 נקודות לכל הסטודנטים בשני המועדים?
- (2) מעוניינים לבדוק האם ההוצאות על "ג'אנק פוד" בקרב הסטודנטים רבות יותר בזמן הלימודים לעומת ימי החופשה. נדגמו 15 סטודנטים מקריים, אצל 13 ההוצאות בתקופת הלימודים היו גבוהות יותר מימי החופשה ואצל 2 נמוכות יותר. מה מסקנתך בר"מ של 0.05?
- (3) מעוניינים לבדוק האם סם מסוים משפיע על לחץ הדם. נלקחו 24 אנשים אשר נמדד להם לחץ הדם לאחר מכן ניתן להם הסם ושוב מדדו להם את לחץ הדם. לחמישה אנשים לחץ הדם לא השתנה ל 15 אנשים לחץ הדם עלה וליתר לחץ הדם ירד אחרי לקיחת הסם. מה מסקנתכם ברמת מובהקות של 5%?
- (4) במדגם שנעשה על 15 משפחות השוו את רמת הביטחון העצמי של הבכור במשפחה לעומת הצעיר שבמשפחה. תוצאות המדגם הראו שאצל 7 משפחות רמת הביטחון העצמי של הבכור הייתה גבוהה יותר, אצל 3 משפחות רמת הביטחון העצמי של הצעיר הייתה גבוהה יותר ואצל 5 משפחות לא נמצא הבדל בין האחים מבחינת רמת הביטחון העצמי. טענת החוקר הייתה שבמשפחות לבכור ביטחון עצמי גבוה מזה של הצעיר במשפחה.
 א. מהי רמת המובהקות המינימלית עבורה יוחלט לקבל את טענת החוקר?
 ב. רמת הביטחון הוערכה על ידי פסיכולוג זוטר. פסיכולוג בכיר ביצע הערכה מחודשת וקבע שלמשפחה אחת במדגם הייתה הערכה שגויה: הפסיכולוג הזוטר קבע שלצעיר במשפחה יש ביטחון עצמי יותר גבוה למרות שלאח הבכור יש ביטחון עצמי יותר גבוה במשפחה הזו. מה יקרה לרמת המובהקות המינימלית שחושבה בשאלה הקודמת?

- (5) איזה מהטענות הבאות נכונות?

א. $n(+) + n(-) = n^*$

ב. $n(+) + n(-) = n$

ג. $n(+) = n(-)$

ד. $n(+) - n(-) = n^*$

תשובות סופיות

- (1) א. לא נדחה H_0 . ב. לא תשתנה המסקנה.
- (2) נדחה H_0 .
- (3) נדחה H_0 .
- (4) א. 0.172 ב. תקטן.
- (5) א'.

מבחן ויילקוקסון למדגמים מזווגים (על ידי שימוש בקירוב הנורמלי) –

רקע

מתי נשתמש במבחן זה?

מבחן זה לא דורש הנחה של התפלגות נורמלית, אולם דורש ערכים מספריים המאפשרים חישוב הפרש בין ערכי X לערכי Y . מבחן זה הוא הגרסה הלא פרמטרית למבחן T למדגמים מזווג. נשתמש במבחן זה שיש משתנה כמותי שאינו מתפלג נורמלית או שיש משתנה מסולם סדר במדגם מזווג.

דוגמה (פתרון בהקלטה):

שני קונדיטורים מתחרים על מקום עבודה. נתנו לשניהם להכין 8 מאפים שונים כאשר כל אחד מהמאפים נאפה על ידי שניהם. בסופו של דבר בעל הקונדיטוריה נתן ציון לכל אחד מהאופים בעבור כל אחד מהמאפים. להלן הציונים שהתקבלו, ורוצים לבדוק את הטענה שאופה א טוב יותר מאופה ב.

אופה א	אופה ב
10	9
9	8
7	7
8	9
9	6
10	6
7	5
8	4

א. מהו המבחן הסטטיסטי המתאים?

ב. מהן השערות המחקר?

חישוב סטטיסטי המבחן:

1. נחשב את ההפרשים D_i לכל תצפית.
2. נוציא מהמדגם את כל התצפיות עם ההפרשים ששווים ל-0.
3. נדרג את ההפרשים הנותרים מהקטן אל הגדול בלי להתייחס לסימן ההפרש, כלומר מדרגים את הערכים המוחלטים של ההפרשים. הפרשים זהים מקבלים דרגה זהה שהיא הדרגה הממוצעת של המקומות שהם תופסים.
4. מסכמים את הדרגות של ההפרשים החיוביים ($R+$) ואת הדרגות של ההפרשים השליליים ($R-$).

דוגמה (פתרון בהקלטה) :

חשבו את W על סמך תוצאות המדגם.

אופה א	אופה ב
10	9
9	8
7	7
8	9
9	6
10	6
7	5
8	4

$$R_+ + R_- = \frac{n^*(n^*+1)}{2} : \text{מתקיים תמיד ש}$$

כמו כן, ניתן להגיד שהתוחלת והשונות של הסטטיסטיים הללו הם :

$$\sigma_{R_{\pm}}^2 = \frac{n^*(n^*+1)(2n^*+1)}{24} \quad \mu_{R_{\pm}} = \frac{n^*(n^*+1)}{4}$$

אם המדגם מספיק גדול, ניתן לבצע קירוב נורמלי לסטטיסטיים אלה באופן הבא :

$$Z_{\pm} = \frac{R_{\pm} - \mu_{R_{\pm}}}{\sqrt{\sigma_{R_{\pm}}^2}} \sim N(0,1) \quad R_{\pm} \sim N\left(\frac{n^*(n^*+1)}{4}, \frac{n^*(n^*+1)(2n^*+1)}{24}\right)$$

דוגמה (פתרון בהקלטה) :

א. מהי מובהקות התוצאה של מבחן זה?

ב. מה המסקנה ברמת מובהקות של 5%?

שאלות

- (1) נדגמו 8 לקוחות שקיבלו שירות ממוקד טלפוני. לקוחות אלה נתבקשו לתת הערכה על יעילות השירות ועל האדיבות שבשירות. הציונים ניתנו בסקאלה מ-1 (הערכה הנמוכה) עד 10 (הערכה הגבוהה ביותר). להלן התוצאות שהתקבלו:

5	7	5	2	3	4	8	7	הערכה על יעילות השירות	X
4	7	10	8	6	7	7	8	הערכה על אדיבות השירות	Y

בדקו ברמת מובהקות של 5% האם קיים הבדל בין הערכה על יעילות השירות להערכה על אדיבות השירות?

- (2) סטודנטים נתבקשו לתת חוות דעתם על רמת הקושי של הקורס (סקאלה של 1-5 כאשר 5=קשה ביותר) ועל רמת הקושי של הבחינות באותה סקאלה. הסטודנטים טוענים שהבחינה הייתה ברמה גבוהה יותר מהרמה של הקורס. להלן תוצאות המדגם:

4	5	1	2	3	4	2	3	4	1-קושי קורס
2	3	5	5	5	3	4	4	4	2-קושי בחינה

בדקו ברמת מובהקות של 5% את טענת הסטודנטים.

- (3) רוצים לבדוק את הטענה שהציונים במבחן בסטטיסטיקה ב גבוהים מאשר בסטטיסטיקה א. נלקחו 10 סטודנטים שסיימו את סטטיסטיקה ב. עבור כל סטודנט נבדק מה הציון בסטטיסטיקה א ומה הציון בסטטיסטיקה ב. להלן התוצאות שהתקבלו:

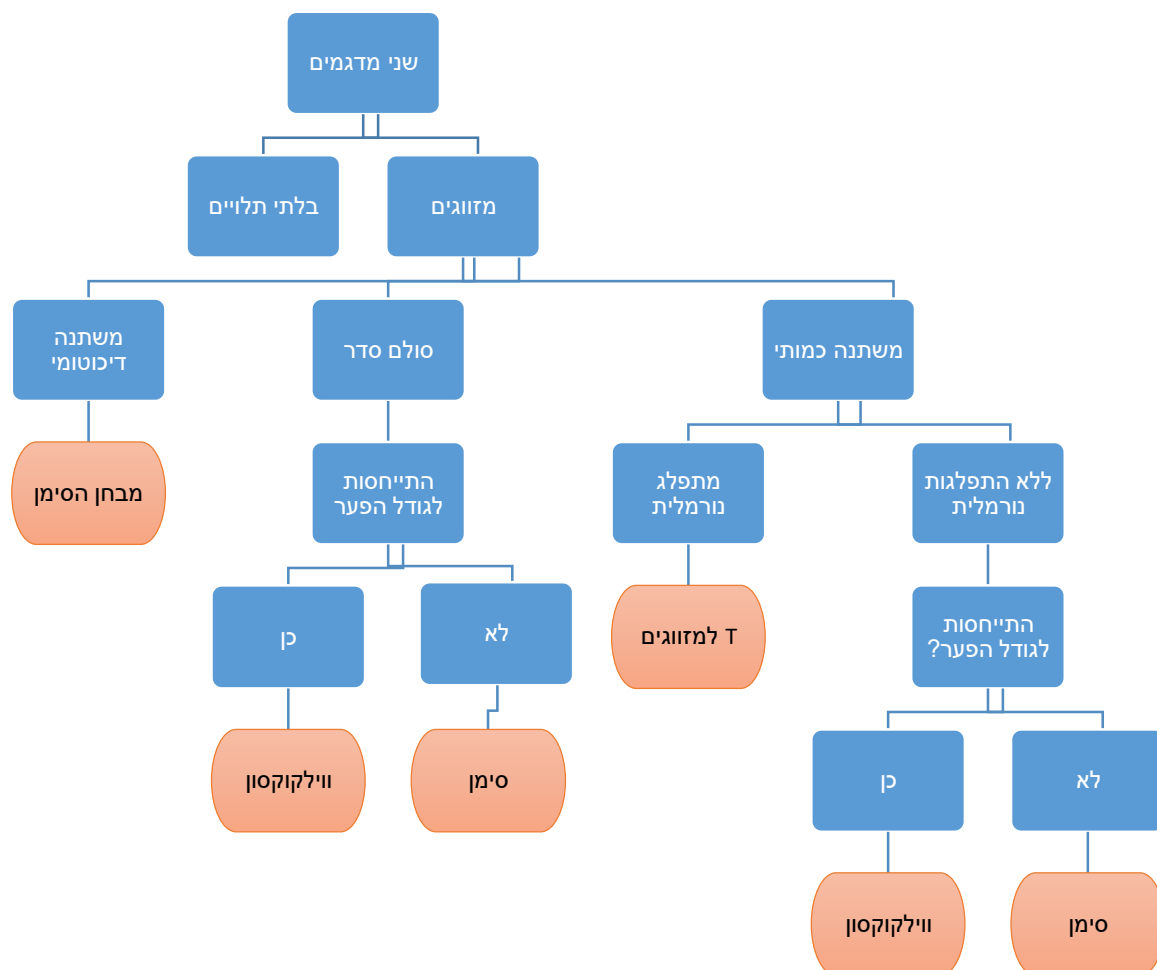
80	78	84	65	67	82	94	68	74	62	א
82	79	86	80	67	77	90	80	80	70	ב

- א. בדקו ברמת מובהקות של 5% באמצעות מבחן ווילקוקסון.
 ב. כיצד התשובה לסעיף הקודם הייתה משתנה אם יוחלט לתת פקטור של 2 נקודות לכל הסטודנטים בשני המועדים?
 ג. כיצד הייתה משתנה התשובה אם מסתבר שנפלה טעות ועבור הסטודנט הראשון ברשימה יש להחליף בנתונים את הציון של סטטיסטיקה ב עם סטטיסטיקה א?

תשובות סופיות

- (1) לא נדחה H_0 .
 (2) לא נדחה H_0 .
 (3) א. לא נדחה H_0 . ב. לא משתנה. ג. לא משתנה.

זיהוי מבחנים סטטיסטיים – רקע



שאלות

(1) במטרה להשוות את רמת האפיה של שני קונדיטורים בחרו 9 מאפים שונים (קרואסון, בראוני וכדומה) ונתנו לכל אחד משני הקונדיטורים לאפות את 9 המאפים השונים. 18 המאפים שנאפו ניתנו למומחה שנתן ציון למאפים השונים. הציון שניתן הוא בין 1 ל-5 לפי ניסיונו וטעמו האישי של המומחה. מהו המבחן המתאים ביותר במקרה זה?

- א. מבחן ווילקוקסון.
- ב. מבחן הסימן.
- ג. מבחן T למזווגים.
- ד. מבחן למדגמים בלתי תלויים.

(2) שני מוסיקאים מפורסמים נתנו ציון בסולם של 1-10 לקולם של 8 מתמודדים בתוכנית ריאליטי ידועה. ציון 10 ניתן לקול שמצא חן ביותר בעיני המוסיקאי. מפיץ התוכנית רצה לבדוק האם יש הבדל בין המוסיקאים מבחינת הטעם. בטבלה הבאה נתונים הציונים של כל אחד מהמוסיקאים את שמונת המתמודדים:

8	7	6	5	4	3	2	1	
4	1	1	3	4	7	5	6	מוסיקאי א'
7	2	3	3	2	5	7	5	מוסיקאי ב'

מהו המבחן המתאים ביותר במקרה זה?

- א. מבחן ווילקוקסון.
- ב. מבחן הסימן.
- ג. מבחן T למזווגים.
- ד. מבחן למדגמים בלתי תלויים.

(3) במחקר בדקו לאנשים את רמת הסוכר בבוקר ואת רמת הסוכר בערב. מתוך 26 אנשים ל-3 רמת הסוכר הייתה זהה. ל-14 רמת הסוכר הייתה גבוהה יותר בשעות הערב. וליתר רמת הסוכר הייתה גבוהה יותר בשעות הבוקר. רוצים לבדוק ברמת מובהקות של 6% האם קיים הבדל בין רמת הסוכר בבוקר לרמת הסוכר בערב אצל האנשים. מהו המבחן המתאים ביותר במקרה זה?

- א. מבחן ווילקוקסון.
- ב. מבחן הסימן.
- ג. מבחן T למזווגים.
- ד. מבחן למדגמים בלתי תלויים.

- (4) חוקר מעוניין לבדוק את התפתחות היכולת לדחות סיפוקים מידיים בקרב ילדים. לשם כך, הוא משתמש במבחן לבדיקה של דחיית סיפוקים, ומעביר אותו בו זמנית ל-2 קבוצות גיל. מבחן זה מודד כמה זמן (בשניות) מסוגל הילד לדחות קבלה של תגמול מידי קטן על מנת לקבל תגמול גדול יותר בעתיד. התוצאות שמתקבלות הן הזמנים של הנחקרים בכל קבוצת גיל. מהו המבחן המתאים ביותר במקרה זה?
- מבחן ווילקוקסון.
 - מבחן הסימן.
 - מבחן T למזווגים.
 - מבחן למדגמים בלתי תלויים.
- (5) חברת משקאות יצאה בקמפיין שנוי במחלוקת. החברה מעוניינת לבדוק האם הקמפיין השפיע על הרגלי הצריכה. במחקר השתתפו נשאלים האם הם נהגו לרכוש את המשקה לפני הקמפיין והאם הם רכשו אותו לאחר הקמפיין. מהו המבחן המתאים ביותר במקרה זה?
- מבחן ווילקוקסון.
 - מבחן הסימן.
 - מבחן T למזווגים.
 - מבחן למדגמים בלתי תלויים.
- (6) מחקר התעניין בדפוסי שיחות הטלפון, שמנהל הפרט בעקבות פרידה מבן זוגו. במחקר השתתפו גברים ו-נשים (כולם נפרדו מבן זוגם). המשתתפים דיווחו על משך השיחות (בדקות; לפני ואחר הפרידה). שאלת המחקר בחנה האם פרידה מבן הזוג קשורה למשך השיחות (משך השיחות הוא משתנה שנהוג להתייחס אליו כמתפלג נורמלית). מהו המבחן המתאים ביותר במקרה זה?
- מבחן ווילקוקסון.
 - מבחן הסימן.
 - מבחן T למזווגים.
 - מבחן למדגמים בלתי תלויים.

תשובות סופיות

- א'.
- א'.
- ב'.
- ד'.
- ב'.
- ג'.

סטטיסטיקה יישומית

פרק 3 - מבחנים אפרמטריים למדגמים בלתי תלויים

תוכן העניינים

1. מבחן ווילקוקסון למדגמים בלתי תלויים..... 23

מבחן ווילקוקסון למדגמים בלתי תלויים – רקע

מבחן ווילקוקסון למדגמים בלתי תלויים נכנס לקטגוריות המבחנים האפרמטריים. מבחן זה רלבנטי כאשר רוצים להשוות בין שתי אוכלוסיות על סמך שני מדגמים בלתי תלויים. המשתנה התלוי הוא משתנה כמותי שאינו מתפלג נורמלית או משתנה מסולם סדר. מבחן זה הוא החלופה האפרמטרית למבחן הפרמטרי להשוואת תוחלות על סמך שני מדגמים בלתי תלויים.

דוגמה:

מחקר חינוכי מעוניין להשוות בין 2 שיטות חינוך. המחקר רוצה לבדוק האם קיים הבדל ברמת ביטחון העצמי של הילדים בשיטות החינוך השונות. נבחרו באקראי 5 ילדים שחונכו בשיטת A. כמו כן נדגמו באקראי 5 ילדים שחונכו בשיטת B. פסיכולוגים בחנו את 10 הילדים ונתנו ציון לביטחון העצמי בסקאלה של 1-20. מהן ההשערות ומהו המבחן הסטטיסטי המתאים?

שיטה A	שיטה B
16	14
17	14
20	19
10	9
18	8

כדי לבצע את המבחן יש לחשב על סמך תוצאות המדגם את סטטיסטי המבחן שנסמן באות U.

השלבים לחישוב סטטיסטי המבחן:

- נסדר את כלל התצפיות של המחקר בסדר עולה מהנמוך ביותר לגבוה ביותר אך יש לדעת כל תצפית מאיזה מדגם היא באה.
- נדרג את כלל התצפיות של המחקר (אם יש תצפיות עם ערכים זהים הדירוג שלהן יהיה ממוצע המקומות שהם תופסים)
- נחשב את W_1 - סכום הדירוגים של התצפיות השייכות למדגם 1,
- W_2 - סכום הדירוגים של התצפיות השייכות למדגם 2.
- נחשב את הגדלים הבאים:

$$U_1 = W_1 - \frac{n_1(n_1+1)}{2} \quad U_2 = W_2 - \frac{n_2(n_2+1)}{2}$$

- הסטטיסטי U: במבחן דו צדדי $U = \min(U_1, U_2)$ במבחן חד צדדי U יהיה ה- U_i , שאמור להיות יותר קטן לפי השערת המחקר.

כדי להגיע למסקנה יש שני סוגים של טבלאות סטטיסטיות.
טבלה מהסוג הראשון: עוזרת לנו לחשב את מובהקות התוצאה לאחר שחישבנו את
ה- U הסטטיסטי.

טבלה מהסוג השני שקובעת מראש את הערך הקריטי של U שנסמן ב U_c .

טבלה מהסוג הראשון	טבלה מהסוג השני
$n_1 - n_2 \leq 8$	n_1 או $n_2 \geq 9$
עוזרת לחשב על סמך תוצאות המדגם את P_v אם $P_v \leq \alpha$ דוחים את H_0 .	נותנת את הערך הקריטי U_c כל הכרעה: נדחה את H_0 אם $U \leq U_c$

דוגמה:

מה המסקנה ברמת מובהקות של 5%?

טבלאות למציאת מובהקות התוצאה במבחן ווילקוקסון למדגמים בלתי תלויים

 $n_2 = 3$

u	n_1		
	1	2	3
0	0.250	0.100	0.050
12	0.500	0.200	0.100
3	0.750	0.400	0.200
4		0.600	0.350
5			0.500
			0.650

 $n_2 = 4$

u	n_1			
	1	2	3	4
0	0.200	0.067	0.028	0.014
1	0.400	0.133	0.057	0.029
2	0.600	0.267	0.114	0.057
3		0.400	0.200	0.100
4		0.600	0.314	0.171
5			0.429	0.243
6			0.571	0.343
7				0.443
8				0.557

 $n_2 = 5$

u	n_1				
	1	2	3	4	5
0	0.167	0.047	0.018	0.008	0.004
1	0.333	0.095	0.036	0.016	0.008
2	0.500	0.190	0.071	0.032	0.016
3	0.667	0.286	0.125	0.056	0.028
4		0.429	0.196	0.095	0.048
5		0.571	0.286	0.143	0.075
6			0.393	0.206	0.111
7			0.500	0.278	0.155
8			0.607	0.365	0.210
9				0.452	0.274
10				0.548	0.345
11					0.421
12					0.500
13					0.579

$$n_2 = 6$$

u	n_1					
	1	2	3	4	5	6
0	0.143	0.036	0.012	0.005	0.002	0.001
1	0.286	0.071	0.024	0.010	0.004	0.002
2	0.428	0.143	0.048	0.019	0.009	0.004
3	0.571	0.214	0.083	0.033	0.015	0.008
4		0.321	0.131	0.057	0.026	0.013
5		0.429	0.190	0.086	0.041	0.021
6		0.571	0.274	0.129	0.063	0.032
7			0.357	0.176	0.089	0.047
8			0.452	0.238	0.123	0.066
9			0.548	0.305	0.165	0.090
10				0.381	0.214	0.120
11				0.457	0.268	0.155
12				0.545	0.331	0.197
13					0.396	0.242
14					0.465	0.294
15					0.535	0.350
16						0.409
17						0.469
18						0.531

$$n_2 = 7$$

u	n_1						
	1	2	3	4	5	6	7
0	0.125	0.028	0.008	0.003	0.001	0.001	0.000
1	0.250	0.056	0.017	0.006	0.003	0.001	0.001
2	0.375	0.111	0.033	0.012	0.005	0.002	0.001
3	0.500	0.167	0.058	0.021	0.009	0.004	0.002
4	0.625	0.250	0.092	0.036	0.015	0.007	0.003
5		0.333	0.133	0.055	0.024	0.011	0.006
6		0.444	0.192	0.082	0.037	0.017	0.009
7		0.556	0.258	0.115	0.053	0.026	0.013
8			0.333	0.158	0.074	0.037	0.019
9			0.417	0.206	0.101	0.051	0.027
10			0.500	0.264	0.134	0.069	0.036
11			0.583	0.324	0.172	0.090	0.049
12				0.394	0.216	0.117	0.064
13				0.464	0.265	0.147	0.082
14				0.538	0.319	0.183	0.104
15					0.378	0.223	0.130
16					0.438	0.267	0.159
17					0.500	0.314	0.191
18					0.562	0.365	0.228
19						0.418	0.267
20						0.473	0.310
21						0.527	0.355
22							0.402
23							0.451
24							0.500
25							0.549

$$n_2 = 8$$

u	n_1							
	1	2	3	4	5	6	7	8
0	0.111	0.022	0.006	0.002	0.001	0.000	0.000	0.000
1	0.222	0.044	0.012	0.004	0.002	0.001	0.000	0.000
2	0.333	0.089	0.024	0.008	0.003	0.001	0.001	0.000
3	0.444	0.133	0.042	0.014	0.005	0.002	0.001	0.001
4	0.556	0.200	0.067	0.024	0.009	0.004	0.002	0.001
5		0.267	0.097	0.036	0.015	0.006	0.003	0.001
6		0.356	0.139	0.055	0.023	0.010	0.005	0.002
7		0.444	0.188	0.077	0.033	0.015	0.007	0.003
8		0.556	0.248	0.107	0.047	0.021	0.010	0.005
9			0.315	0.141	0.064	0.030	0.014	0.007
10			0.387	0.184	0.085	0.041	0.020	0.010
11			0.461	0.230	0.111	0.054	0.027	0.014
12			0.539	0.285	0.142	0.071	0.036	0.019
13				0.341	0.177	0.091	0.047	0.025
14				0.404	0.217	0.114	0.060	0.032
15				0.467	0.262	0.141	0.076	0.041
16				0.533	0.311	0.172	0.095	0.052
17					0.362	0.207	0.116	0.065
18					0.416	0.245	0.140	0.080
19					0.472	0.286	0.168	0.097
20					0.528	0.331	0.198	0.117
21						0.377	0.232	0.139
22						0.426	0.268	0.164
23						0.475	0.306	0.101
24						0.525	0.347	0.221
25							0.389	0.253
26							0.433	0.287
27							0.478	0.323
28							0.522	0.360
29								0.399
30								0.439
31								0.480
32								0.520

טבלה למציאת U_c

ברמת מובהקות של 5% למבחן חד צדדי או ברמת מובהקות של 10% למבחן דו צדדי במבחן ווילקוקסון למדגמים בלתי תלויים.

n_1	n_2											
	9	10	11	12	13	14	15	16	17	18	19	20
1											0	0
2	1	1	1	2	2	2	3	3	3	4	4	4
3	3	4	5	5	6	7	7	8	9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100
16	42	48	54	60	65	71	77	83	89	95	101	107
17	45	51	57	64	70	77	83	89	96	102	109	115
18	48	55	61	68	75	82	88	95	102	109	116	123
19	51	58	65	72	80	87	94	101	109	116	123	130
20	54	62	69	77	84	92	100	107	115	123	130	138

שאלות

- (1) מעוניינים להשוות בין שתי קבוצות כדורסל. נלקחו 5 משחקים מקבוצה א' ושישה משחקים מקבוצה ב'. נבדק בכל משחק ועבור כל קבוצה מספר הנקודות שצברה במשחק. בדקו ברמת מובהקות של 5% האם קיים הבדל בין הקבוצות מבחינת הניקוד שצברה במשחק.

קבוצה א	קבוצה ב
68	82
82	74
78	82
94	64
87	67
	65

- (2) מעוניינים לבדוק האם קורס קיץ באנגלית משפר את יכולות האנגלית לתלמידי חטיבת ביניים. נלקחו 20 ילדים בגיל חטיבת הביניים ברמת אנגלית דומה. 12 מהם נשלחו לקורס קיץ והיתר לא. בסוף הקיץ כולם נבחנו במבחן באנגלית הציון הגבוה ביותר התקבל בקרב אחד שלא עשה את הקורס ושבעת הציונים הנמוכים ביותר היו גם בקרב תלמידים שלא עשו את הקורס. מה המסקנה ברמת מובהקות של 5%?
- (3) במחקר לבדיקת יעילות ויטמין C נבחרו 15 מתנדבים מבין עובדי המפעל. תשעה מהם נבחרו מקרית וקיבלו טיפול שוטף בוויטמין C, ואילו שאר המתנדבים (קבוצת הביקורת) קבלו גלולת סוכר. במשך שלוש שנות המחקר היו מספר ימי ההיעדרות בגלל ההצטננות:
קבוצת הטיפול: 1, 3, 9, 3, 4, 0, 8, 12, 16.
קבוצת הביקורת: 19, 7, 28, 13, 23, 12.
 בדקו ברמת מובהקות של 5% שמספר ימי המחלה במשך שלוש שנים מצטמצם ביותר מ-4 ימים עם לקיחת ויטמין C.

תשובות סופיות

- (1) לא נדחה את H_0 .
 (2) נדחה את H_0 .
 (3) לא נדחה את H_0 .

\\

סטטיסטיקה יישומית

פרק 4 - מבחני חי בריבוע

תוכן העניינים

- 1. מבחן טיב התאמה 30
- 2. מבחן לאי תלות 35

מבחן טיב התאמה – רקע

מבחן זה בא לבדוק האם אוכלוסייה מסוימת מתפלגת לפי התפלגות נתונה. המשתנה הנחקר מחולק למספר קטגוריות ויש לבדוק האם תוצאות המדגם תואמות להתפלגות הנתונה.

מבנה המבחן:

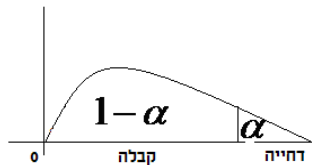
השערות:

המשתנה מתפלג לפי התפלגות מסוימת - H_0 .

אחרת - H_1 .

כלל הכרעה:

הערך הקריטי נקבע על סמך התפלגות חי בריבוע. התפלגות זו היא אסימטרית חיובית ותלויה בדרגות החופש. $d.f = K - 1$, כאשר K - מספר הקטגוריות.



הערך הקריטי הוא: $\chi^2_{1-\alpha, K-1}$, כלומר האחוזון ה- $1 - \alpha$ בהתפלגות חי בריבוע שדרגות החופש הן $K - 1$.

אם $\chi^2 > \chi^2_{1-\alpha, K-1}$, דוחים את השערת האפס.

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad \text{סטטיסטי המבחן:}$$

O_i - השכיחות שנצפתה במדגם בקטגוריה i .

p_i - הסתברות לקטגוריה i לפי השערת האפס.

$E_i = np_i$ - שכיחות צפויה במדגם לקטגוריה i בהנחת השערת האפס.

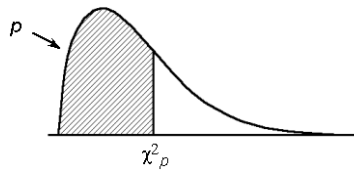
הערה:

תנאי כדי לבצע את המבחן הוא $E_i \geq 5$ לכל i . במידה ותנאי זה לא מתקיים יש אפשרות לאחד קטגוריות סמוכות עד שהתנאי יתקיים.

דוגמה (פתרון בהקלטה) :

במדינה מסוימת שלוש מפלגות. בפרלמנט הנוכחי התפלגות מספר המושבים היא 30% למפלגה A, 60% למפלגה B ו-10% למפלגה C. לקראת הבחירות המתוכננות בשבוע הבא נעשה סקר שכלל 300 אזרחים. בסקר התקבל ש-40% יצביעו למפלגה A, 50% למפלגה B ו-10% למפלגה C. האם תוצאות הסקר תואמות להתפלגות המושבים בפרלמנט הנוכחי? בדקו ברמת מובהקות של 5%.

טבלת התפלגות חי-בריבוע – ערכי החלוקה



df	p												
	.005	.01	.025	.05	.10	.25	.50	.75	.90	.95	.975	.99	.995
1	0.0 ⁴ 393	0.0 ³ 157	0.0 ³ 982	0.0 ² 393	0.0158	0.102	0.455	1.32	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	0.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	0.0717	0.115	0.216	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7

שאלות

- (1) במטרה לבדוק האם קובייה הוגנת, מטילים אותה 120 פעמים. התקבל 17 פעמים 1, 23 פעמים 2, 20 פעמים 3, 25 פעמים 4, 18 פעמים 5 ו-17 פעמים 6. מה מסקנתכם ברמת מובהקות של 5%?
- (2) מפעל מייצר סוכריות בצבעים כחול, אדום, ירוק וכתום. מעוניינים לבדוק שפרופורציית הסוכריות הכחולות גדולה פי 2 מכל צבע אחר. לצורך כך נדגמו באקראי 200 סוכריות והתקבל: 70 כחולות, 50 אדומות, 40 ירוקות והיתר כתומות. מה מסקנתכם ברמת מובהקות של 5%?
- (3) משרד החינוך טוען שבקרב השכירים במשק היחס בין השכירים בעלי השכלה נמוכה, תיכונית ואקדמאית הוא 1:2:1 בהתאמה. במדגם של 200 שכירים התקבלו 56 אנשים בעלי השכלה נמוכה, 105 בעלי השכלה תיכונית והיתר בעלי השכלה גבוהה.
- א. על סמך תוצאות המדגם, האם התפלגות ההשכלה היא כמו שמשדרד החינוך מפרסם? בדוק ברמת מובהקות של 5%.
- ב. בנו רווח סמך ברמת סמך של 95% לפרופורציית השכירים במשק בעלי השכלה אקדמאית.
- (4) 200 איש נתבקשו לבחור ספרה באקראי והנה התוצאות שהתקבלו:
- 18 איש בחרו בספרה 0, 24 איש בחרו בספרה 1, 17 איש בחרו בספרה 2, 19 איש בחרו בספרה 3, 20 איש בחרו בספרה 4, 18 איש בחרו בספרה 5, 22 איש בחרו בספרה 6 והיתר בחרו בספרות 7-9.
- א. על סמך התוצאות הללו האם בחירת הספרות אקראית? בדקו ברמת מובהקות של 2.5%.
- ב. תנו הערכה למובהקות התוצאה.
- ג. אם נגדיל את גודל המדגם פי 2 ונשמור על אותם יחסים של כמות האנשים במדגם שבחרו בספרות, כיצד הדבר ישפיע על ערכו של הסטטיסטי χ^2 ? מה תהיה המסקנה במקרה זה?
- (5) מעוניינים לבדוק האם קובייה היא הוגנת. הטילו את הקובייה פעמיים והתבוננו בסכום הוצאות. חזרו על התהליך 72 פעמים.
- להלן התוצאות שהתקבלו במדגם:
- מה המסקנה ברמת מובהקות של 5%?

מספר הטלות	סכום התוצאות
20	2-5
17	6-8
20	9-10
15	11-12

- (6) בפנס יש 4 סוללות. בבדיקה שנערכה ב-400 פנסים נמצאו סוללות פגומות לפי השכיחויות הבאות:

3 ומעלה	2	1	0	מספר הסוללות הפגומות
8	12	104	276	שכיחות

מעוניינים לבדוק על סמך תוצאות מדגם אלה האם הסיכוי לסוללה פגומה הוא 20%. מהי רמת המובהקות המינימלית עבורה נכריע שהסיכוי לסוללה פגומה אינו 20%?

- (7) מטילים מטבע עד שלראשונה מתקבל "ראש". חוזרים על התהליך 120 פעמים. נסמן ב- X את מספר ההטלות עד קבלת הראש. להלן התוצאות שהתקבלו:

x	1	2	3	4	5	6
מספר החזרות על התהליך	54	20	16	22	6	2

- א. בהנחה והמטבע הוגן, מהי ההתפלגות של X ?
- ב. בדקו האם המטבע הוגן, על סמך תוצאות המדגם ברמת מובהקות של 5%.

- (8) להלן השערות מחקר: $H_1: \text{else}$, $H_0: X \sim N(40, 2^2)$.

X	מתחת 36	36-40	40-44	מעל 44
מספר הדגימות	3A	50A	45A	2A

תוצאות המדגם הן:

מהו ערכו המקסימלי של A עבורו נקבל את H_0 ברמת מובהקות של 5%?

תשובות סופיות

- (1) לא נדחה H_0 .
- (2) לא נדחה H_0 .
- (3) א. לא נדחה H_0 . ב. $(0.14, 0.25)$.
- (4) א. לא נדחה H_0 . ב. בין 0.95 ל-0.975.
- ג. יגדל פי 2; המסקנה לא תשתנה.
- (5) נכריע שהקובייה אינה הוגנת.
- (6) 0.005.
- (7) א. $X \sim G(0.5)$. ב. נסיק שהמטבע לא הוגן.
- (8) 14.

מבחן חי בריבוע לאי תלות בין משתנים – רקע

מבחן לאי תלות מטרתו לבדוק האם קיים קשר בין שני משתנים. שני המשתנים שנבדקים צריכים להיות מחולקים למספר קטגוריות.

מבנה המבחן:

השערות:

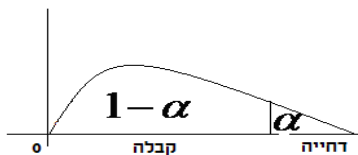
אין תלות בין המשתנים H_0 .

יש תלות בין המשתנים H_1 .

כלל הכרעה:

הערך הקריטי נקבע על סמך התפלגות חי בריבוע. התפלגות זו היא אסימטרית חיובית ותלויה בדרגות החופש $d.f = (r-1)(c-1)$. כאשר: r - מספר הקטגוריות של המשתנה שבשורות. c - מספר הקטגוריות של המשתנה שבעמודות.

הערך הקריטי הוא: $\chi^2_{1-\alpha, (r-1)(c-1)}$, כלומר האחוזון ה- $1-\alpha$ בהתפלגות חי בריבוע שדרגות החופש הן $(r-1)(c-1)$. אם $\chi^2 > \chi^2_{1-\alpha, (r-1)(c-1)}$ אז דוחים את השערת האפס.



$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

כאשר:

O_i - השכיחות נצפית במדגם בתא i .

E_i - שכיחות צפויה במדגם בתא i בהנחת השערת האפס.

$$E_i = \frac{f(x) \cdot f(y)}{n}$$

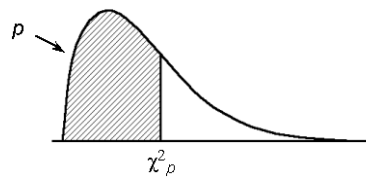
הערה:

תנאי כדי לבצע את המבחן הוא $E_i \geq 5$ לכל i . במידה ותנאי זה לא מתקיים יש אפשרות לאחד קטגוריות סמוכות עד שהתנאי יתקיים.
תנאי חלופי: אין E קטן מ-1 וגם אין ביותר מ 20% מהתאים E קטן מ-5.

דוגמה (הפתרון בהקלטה):

האם יש תלות בין המגדר לבין דעה מסוימת?
יש לבדוק ברמת מובהקות של 5% על סמך תוצאות הסקר:

המגדר / דעה	בעד	נגד	נמנע	סה"כ
גברים	50	40	10	
נשים	20	60	20	
סה"כ				

טבלת התפלגות חי-בריבוע – ערכי החלוקה χ^2_p 

df	p												
	.005	.01	.025	.05	.10	.25	.50	.75	.90	.95	.975	.99	.995
1	0.004393	0.00157	0.00982	0.02393	0.0158	0.102	0.455	1.32	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	0.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	0.0717	0.115	0.216	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7

שאלות

- (1) נבדקה התלות בין גודל הארגון לבין שביעות הרצון של העובדים. להלן התוצאות:

שביעות רצון גודל המפעל	נמוכה	בינונית	גבוהה	סה"כ
גדול	182	203	215	600
קטן	154	110	136	400
סה"כ	336	313	351	1000

מה המסקנה ברמת מובהקות של 2.5%?

- (2) מפעל עובד בשלוש משמרות. להלן מספר המוצרים הפגומים והתקינים בכל אחת מן המשמרות לפי מדגם שנעשה:

לילה	ערב	יום	
70	60	50	פגומים
800	700	600	תקינים

האם יש הבדל בין שיעורי הפגומים במשמרות השונות? הסיקו עבור רמת מובהקות $\alpha = 0.05$.

- (3) נדגמו 50 מוצרים ממפעל מסוים מתוך 30 מוצרים שיוצרו ביום 17 נבחרו לייצוא מתוך המוצרים שיוצרו בלילה 10 נבחרו לייצוא. האם יש קשר בין היות מוצר לייצוא למועד שבו הוא יוצר? בדקו ברמת בטחון של 95%.

- (4) במטרה לבדוק האם השתנו דפוסי ההצבעה למפלגות השונות בין שבוע שעבר לשבוע נלקחו שני סקרים אחד מהשבוע שעבר והאחר מהשבוע. להלן דפוסי ההצבעה שהתקבלו בסקרים אלה.

א. מהי רמת המובהקות המינמלית עבורה ניתן להחליט שהשתנו דפוסי ההצבעה משבוע שעבר לשבוע באופן מובהק?

ב. כיצד הייתה התשובה לסעיף א משתנה אם כל השכיחויות בטבלה של תוצאות המדגם היו מוכפלות פי 2?

ג. בנו רווח סמך לשיעור המצביעים למפלגה א השבוע ברמת סמך של 95%.

	מפלגה א	מפלגה ב	מפלגות אחרות	סה"כ
שבוע שעבר				
השבוע	143		253	550
סה"כ	243	314		1050

- (5) בחנות בגדים A בדקו את התפלגות הצבעים של הבגדים הנמכרים ביום מסוים. כמו כן בדקו את התפלגות הצבעים בחנות שכנה B:

מספר פריטים / צבע	שחור	לבן	אדום	כחול
חנות A	15	20	15	50
חנות B	60	20	10	20

- א. בדקו ברמת מובהקות של 5% האם התפלגות הצבעים בחנות A היא ביחס של 1:1:1:3 לטובת הכחול.
 ב. בדקו ברמת מובהקות של 2.5% האם קיים הבדל בין החנויות מבחינת התפלגות הצבעים של הפריטים הנמכרים.

- (6) סטודנט קיבל בבדיקת השערות ערך χ^2 (chi-square) השוו לאפס. הסטודנט הסיק כי לא קיימת תלות בין שני המשתנים שבדק, בכל רמת מובהקות. נכון / לא נכון? נמקו.

- (7) להלן טבלת O של שני משתנים שהתקבל במדגם כלשהו:

$f(x)$	Y_4	Y_3	Y_2	Y_1	
200					X_1
200					X_2
	160	120	60	60	$f(y)$

- מה צריכות להיות השכיחויות בתוך הטבלה כדי שמובהקות התוצאה (PV) תהיה 100%?

תשובות סופיות

- (1) נסיק שיש קשר בין גודל הארגון לשביעות הרצון של העובדים.
 (2) נסיק שאין הבדל מובהק בין שיעור הפגומים במשמרות השונות.
 (3) נסיק שאין קשר בין היות מוצא לייצוא למועד שבו הוא יוצר.
 (4) א. 10% ב. קטן ג. (0.223, 0.297)
 (5) א. נסיק שהתפלגות הצבעים בחנות היא כמו שמצוין.
 ב. נסיק שיש הבדל בין החנויות מבחינת התפלגות הצבעים.
 (6) נכון
 (7) להלן טבלה:

$f(x)$	Y_4	Y_3	Y_2	Y_1	
200	80	60	30	30	X_1
200	-8	60	30	30	X_2
400	160	120	60	60	$f(y)$

סטטיסטיקה יישומית

פרק 5 - ניתוח שונות חד כיוונית

תוכן העניינים

1. כללי 40

ניתוח שונות חד כיוונית

רקע תיאורטי

ניתוח שונות (חד כיוונית) הוא מבחן להשוואת תוחלות $(\mu_1, \dots, \mu_k)$ של k אוכלוסיות שונות. לכן, בניתוח שונות, השערות המחקר הן:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k \quad (\text{התוחלות של כל האוכלוסיות שוות})$$

$$H_1: \quad \text{אחרת} \quad (\text{לפחות שתיים מהתוחלות שונות})$$

ההנחות הדרושות לביצוע התהליך:

(2) בכל אוכלוסייה מתוך k האוכלוסיות ההתפלגות נורמלית.

(3) כל האוכלוסיות הן עם אותה שונות σ^2 .

(4) המדגמים בלתי תלויים זה בזה.

ישנו משתנה המבדיל בין הקבוצות השונות, הוא המשתנה הבלתי תלוי הנקרא גורם (factor). משתנה זה הוא קטגוריאלי עם k רמות (levels). כדי לבצע את התהליך יש לבצע מדגם מכל אוכלוסייה: נסמן ב- n_i את גודל המדגם בקבוצה i .

$$n = \sum_{i=1}^k n_i \quad \text{- מספר התצפיות סך הכול (בכל המדגמים).}$$

$\bar{X}_1$ - ממוצע המדגם הראשון, $\dots, \bar{X}_k$ - ממוצע המדגם ה- k .
 $\bar{X}$ - ממוצע כללי (של כל המדגמים).

$$SS_B = \sum_{i=1}^k n_i [\bar{X}_i - \bar{X}]^2 \quad \text{סכום ריבועים בין הקבוצות:}$$

$$SS_W = \sum_{i=1}^k n_i [n_i - 1] \cdot \hat{S}_i^2 \quad \text{סכום ריבועים בתוך הקבוצות:}$$

$$SS_T = \sum_{i=1}^k \sum_{j=1}^{n_i} [X_{ij} - \bar{X}]^2 \quad \text{סכום ריבועים כללי:}$$

$$SST = SSB + SSW$$

יש למלא את טבלת ניתוח השונות הבאה:

מקור השונות	סכום הריבועים SS	דרגות חופש df	ממוצע הריבועים MS	F
B - בין הקבוצות	SSB	$k-1$	$\frac{SSB}{k-1}$	$\frac{MSB}{MSW}$
W - בתוך הקבוצות	SSW	$n-k$	$\frac{SSW}{n-k}$	
T - סה"כ	SST	$n-1$		

$$F = \frac{\frac{SSB}{k-1}}{\frac{SSW}{n-k}} \sim F(k-1, n-k)$$

אזור דחיית $H_0 : 1-\alpha : F_{(k-1, n-k)}$

שאלות

- (1) מחקר מעוניין להשוות בין שלוש תרופות לשיכוך כאבים במטרה לבדוק האם קיים הבדל בין התרופות מבחינת הזמן בדקות שלוקח עד שהתרופה משפיעה. לצורך הבדיקה נלקחו 15 אנשים שסובלים מכאבי ראש. אנשים אלה חולקו באקראי לשלוש: קבוצה 1 קיבלה "אקמול" קבוצה 2 קיבלה "אופטלגין" קבוצה 3 קיבלה "נורופן". כל אדם במחקר מסר את מספר הדקות עד שהתרופה השפיעה עליו.
- א. מהו המשתנה התלוי ומהו המשתנה הבלתי תלוי במחקר?
מהו ה"גורם" וכמה רמות יש לו?
- ב. מהו המבחן הסטטיסטי המתאים כאן? רשמו את ההשערות.
- ג. מה הן ההנחות הדרושות כדי לבצע את המבחן הסטטיסטי שהצעת בסעיף הקודם?

- (2) בעיר מסוימת שלושה בתי ספר תיכון. ראש העיר התעניין לבדוק האם קיים הבדל בהצלחה של בתי הספר במקצוע מתמטיקה. לצורך כך הוא דגם מספר תלמידים שנבחנו במבחן הבגרות במתמטיקה ברמה של 3 יחידות בעירו ובדק עבור כל תלמיד מה ציון הבגרות שלו במתמטיקה. להלן הציונים שהתקבלו:

"הס"	"רבין"	"המתמיד"
85	98	78
83	62	65
74	55	70
85	80	90
75		56

- א. מהו המבחן הסטטיסטי המתאים?
רשמו את ההשערות ואת ההנחות של המבחן.
- ב. מהו גודל המדגם? מהו המשתנה הבלתי תלוי (factor) כמה רמות יש לו?
- ג. חשבו את הממוצע ואת סטיית התקן של הציונים בכל אחד מהמדגמים.
- ד. מלאו את טבלת ANOVA.
- ה. רשמו את כלל ההכרעה למבחן שהוצע בסעיף א ברמת מובהקות של 5%.
- ו. האם קיים הבדל בין בתי הספר בעיר מבחינת רמת הצלחת התלמידים במקצוע המתמטיקה? ענה על סמך הסעיפים הקודמים.

- (3) מעוניינים לבדוק האם יש הבדל בהשפעה של שיטות טפול שונות על לחץ הדם הסיסטולי (SBP) באוכלוסייה של קשישים. נבדקו 4 שיטות שונות. בטבלה המצורפת מרוכזים ממצאי המחקר.

השיטה	A	B	C	D
גודל המדגם	12	14	8	12
הממוצע	178	172	180	182
סטיית התקן	4	8	5	3

- א. רשמו את השערות המחקר וההנחות הדרושות כדי לבצע את המבחן המתאים.
- ב. מה מסקנת המחקר ברמת מובהקות של 5%?
- ג. האם יש צורך לבצע השוואות מרובות?

(4) שלושה אופים נתבקשו להכין עוגת שוקולד. לכל אופה בדקו את משך זמן ההכנה בדקות. כל אופה נדרש לאפות בכל יום 4 עוגות.

האם קיים הבדל בין האופים מבחינת תוחלת זמני ההכנה של העוגות?

בדקו ברמת מובהקות של 5%.

האופה	ניר	מוזס	שלום
סכום הזמנים	206	212	182
סכום ריבועי הזמנים	10644	11250	8982

(5) להלן טבלת ניתוח שונות חד כיוונית. במחקר בחנו 4 סוגי סוללות. רצו לבדוק האם לסוג הסוללה השפעה על תוחלת אורך החיים שלה. הפעילו את כל הסוללות על אותו מכשיר ובדקו את אורך החיים של כל סוללה בשעות. מה המסקנה ברמת מובהקות של 10%? רשמו את ההשערות וההנחות הדרושות.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	10.317	3	3.439	1.361	.279
Within Groups	60.648	24	2.527		
Total	70.964	27			

(6) להלן טבלת ANOVA בטבלה הושמטו חלקים. השלימו את החלקים בטבלה שהושמטו ומסומנים באותיות.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	357.450	ב	ג	ה	.000
Within Groups	א	17	ד		
Total	522.950	19			

7) חברת תרופות לקחה 15 אנשים ברמת בריאות דומה. החברה חילקה את האנשים ל שלוש קבוצות שוות בגודלן. לכל קבוצה ניתנה אותה תרופה במינון שונה (dosage). המינונים שניתנו הם: 10 מ"ג, 20 מ"ג ו-30 מ"ג. לאחר שעה מזמן לקיחת התרופה נבדק קצב פעימות הלב של כל אדם (pulse). הנתונים הוזנו לתוכנה סטטיסטית והתקבלו התוצאות הבאות:

ANOVA						pulse			
pulse						Tukey HSD ^a			
	Sum of Squares	df	Mean Square	F	Sig.	dosage	N	Subset for alpha = 0.05	
								1	2
Between Groups	414.400	2	207.200	19.733	.000	30.00	5	71.0000	
Within Groups	126.000	12	10.500			20.00	5		80.2000
Total	540.400	14				10.00	5		83.4000
						Sig.		1.000	.299
						Means for groups in homogeneous subsets are displayed.			
						a. Uses Harmonic Mean Sample Size = 5.000.			

Post Hoc Tests

Multiple Comparisons

		pulse Tukey HSD				
(I) dosage	(J) dosage	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
10.00	20.00	3.20000	2.04939	.299	-2.2675	8.6675
	30.00	12.40000 [*]	2.04939	.000	6.9325	17.8675
20.00	10.00	-3.20000	2.04939	.299	-8.6675	2.2675
	30.00	9.20000 [*]	2.04939	.002	3.7325	14.6675
30.00	10.00	-12.40000 [*]	2.04939	.000	-17.8675	-6.9325
	20.00	-9.20000 [*]	2.04939	.002	-14.6675	-3.7325

*. The mean difference is significant at the 0.05 level.

- בדקו ברמת מובהקות של 5% האם קיים הבדל בין המינונים השונים מבחינת תוחלת הדופק של האנשים? רשמו את ההשערות וההנחות הדרושות לצורך פתרון.
- הסבירו ללא חישוב כיצד הייתה משתנה התשובה לסעיף הקודם אם הינו מעלים את הדופק של כל התצפיות במחקר ב-2.
- האם יש צורך במחקר בהשוואת מרובות. נמקו!
- לטבלת ANOVA צורפו טבלאות של השוואות מרובות בשיטה הנקראת "טוקי". ברמת בטחון של 95% מה הם הממצאים לפי שיטה זו?

- (8) בעיר מסוימת רצו לבדוק האם קיים הבדל ברמה של התלמידים בין בתי הספר השונים בעיר. ביצעו מדגם מכל בית ספר ונתנו מבחן זהה לכל הנדגמים. לאחר מכן ריכזו את הנתונים בתוכנה סטטיסטית והפעילו ניתוח שונות. מצורפים הפלטים שהתקבלו. ענו על הסעיפים הבאים :
- כמה בתי ספר יש בעיר?
 - כמה תלמידים השתתפו בסך הכול במחקר?
 - האם קיים הבדל בין בתי הספר בעיר מבחינה רמת הציונים? בדקו ברמת מובהקות של 1%
 - בביטחון של 95% אילו בתי ספר שונים זה מזה ברמת התלמידים? נמקו והסבירו.

Oneway

ANOVA

grade

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7799.600	4	1949.900	13.586	.000
Within Groups	2870.400	20	143.520		
Total	10670.000	24			

Post Hoc Tests

Multiple Comparisons

grade

Scheffe

(I) school	(J) school	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	5.40000	7.57681	.971	-20.2543	31.0543
	3.00	36.80000*	7.57681	.003	11.1457	62.4543
	4.00	36.40000*	7.57681	.003	10.7457	62.0543
	5.00	-2.60000	7.57681	.998	-28.2543	23.0543
2.00	1.00	-5.40000	7.57681	.971	-31.0543	20.2543
	3.00	31.40000*	7.57681	.011	5.7457	57.0543
	4.00	31.00000*	7.57681	.013	5.3457	56.6543
	5.00	-8.00000	7.57681	.888	-33.6543	17.6543
3.00	1.00	-36.80000*	7.57681	.003	-62.4543	-11.1457
	2.00	-31.40000*	7.57681	.011	-57.0543	-5.7457
	4.00	-.40000	7.57681	1.000	-26.0543	25.2543
	5.00	-39.40000*	7.57681	.001	-65.0543	-13.7457
4.00	1.00	-36.40000*	7.57681	.003	-62.0543	-10.7457
	2.00	-31.00000*	7.57681	.013	-56.6543	-5.3457
	3.00	.40000	7.57681	1.000	-25.2543	26.0543
	5.00	-39.00000*	7.57681	.001	-64.6543	-13.3457
5.00	1.00	2.60000	7.57681	.998	-23.0543	28.2543
	2.00	8.00000	7.57681	.888	-17.6543	33.6543
	3.00	39.40000*	7.57681	.001	13.7457	65.0543
	4.00	39.00000*	7.57681	.001	13.3457	64.6543

*. The mean difference is significant at the 0.05 level.

Homogeneous Subsets

grade

Scheffe^a

school	N	Subset for alpha = 0.05	
		1	2
3.00	5	45.0000	
4.00	5	45.4000	
2.00	5		76.4000
1.00	5		81.8000
5.00	5		84.4000
Sig.		1.000	.888

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 5.000.

תשובות סופיות

(1) א. משתנה בלתי תלוי : סוג התרופה. ב. ניתוח שונות חד כיוונית

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : otherwise$$

משתנה תלוי : הזמן עד להשפעת התרופה בדקות.

ג. 1. מדגמים בלתי תלויים.

2. שוויון שונויות.

3. משתנים מתפלגים נורמלית.

(2) א. המבחן לניתוח שונות חד כיוונית.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : otherwise$$

הנחות :

1. מדגמים בלתי תלויים.

2. משתנים מתפלגים נורמלית.

3. שוויון שונויות.

ב. גודל המדגם : 14. משתנה ב"ת : בית הספר, בעל 3 רמות.

$$g. \bar{X} = 71.8, \hat{S} = 12.93, \bar{X} = 73.75, \hat{S} = 19.29, \bar{X} = 80.4, \hat{S} = 5.46$$

ד. להלן טבלה :

מקור השונות	SS	df	MS	F
B	200.6	2	100.3	
W	1904.75	11	173.2	
סה"כ	2105.35	13		0.58

$$h. F > 3.98$$

ו. נקבל את H_0 .

(3) א. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
 $H_1 : otherwise$
 ב. נדחה את H_0 .
 ג. כן.

הנחות :

1. מדגמים בלתי תלויים.

2. שוויון שונויות.

3. משתנים מתפלגים נורמלית.

(4) נקבל את H_0 : נכריע שאין הבדל מובהק בין האופים מבחינת תוחלת זמן הכנה.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad (5)$$

$$H_1 : otherwise$$

הנחות :

1. מדגמים בלתי תלויים.

2. שוויון שונות.

3. משתנים מתפלגים נורמלית.

נקבל את H_0 : לסוג סוללה אין השפעה של תוחלת החיים ברמת ביטחון של 10%.

(6) להלן טבלה :

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	357.450	2 ב	178.725 ג	18.36 ה	.000
Within Groups	165.5 א	17	9.735 ד		
Total	522.950	19			

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad (7)$$

$$H_1 : otherwise$$

הנחות :

1. מדגמים בלתי תלויים.

2. משתנים מתפלגים נורמלית.

3. שוויון שונות.

נדחה את H_0 : ברמת ביטחון של 5% קיים הבדל במינונים השונים מבחינת תוחלת הדופק.

$$\mu_{20} = \mu_{10} > \mu_{30} \quad (8)$$

$$25 \quad (8)$$

ג. נדחה את H_0 : יש לפחות שני בתי ספר בעיר עם תוחלת רמת ציונים שונה.

$$(\mu_3 = \mu_4) < (\mu_1 = \mu_2 = \mu_3) \quad (8)$$

סטטיסטיקה יישומית

פרק 6 - מקדם המתאם (מדד קשר) הלינארי ומובהקותו

תוכן העניינים

1. מקדם המתאם הלינארי (פירסון)..... 49
2. חישוב מקדם המתאם הלינארי (פירסון)..... 60
3. בדיקת השערות על מקדם המתאם הלינארי..... 65
4. בדיקת השערות על מקדם המתאם הלינארי באמצעות טבלה של ערכים קריטיים..... 69

מקדם המתאם (מדד קשר) הלינארי ומובהקותו

מדד הקשר הלינארי (פירסון) – מבוא

מעוניינים לבדוק עד כמה קיים קשר מסוג קשר לינארי (קו ישר) בין שני משתנים. שני המשתנים שאנו בודקים לגביהם קשר צריכים להיות משתנים כמותיים. מבחינת סולמות מדידה כל משתנה נחקר צריך להיות מסולם רווחים או מנה. בדרך כלל המשתנה המוצג כ- Y הוא המשתנה התלוי והמשתנה המוצג ב- X הוא המשתנה הבלתי תלוי. תיאור גרפי לנתונים נעשה על ידי דיאגרמת פיזור. בדיאגרמת פיזור אנחנו מסמנים כל תצפית בנקודה לפי שיעור ה- X ושיעור ה- Y שלה. דיאגרמת הפיזור נותנת אינדיקציה גרפית על הקשר בין שני המשתנים.

דוגמה (פתרון בהקלטה) :

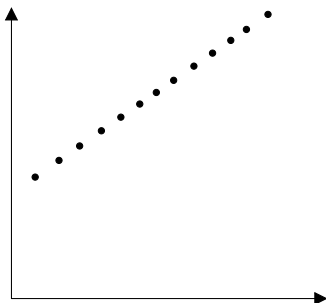
בבניין 8 דירות בדקו לכל דירה את מספר החדרים שלה וכמו כן את מספר הנפשות הגרות בדירה. להלן התוצאות שהתקבלו :

4	4	3	3	2	3	2	2	מספר חדרים בדירה
5	4	4	3	2	2	1	0	מספר הנפשות בדירה

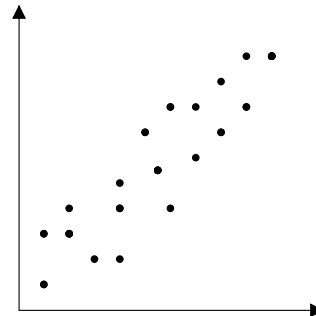
- (1) כמה תצפיות ישנן בדוגמה?
- (2) כמה משתנים ישנם בדוגמה, מי הם?
- (3) שרטטו לנתונים דיאגרמת פיזור.
- (4) מי המשתנה התלוי ומיהו המשתנה הבלתי תלוי?

דיאגרמות פיזור לקשר בין משתנים וניתוחם

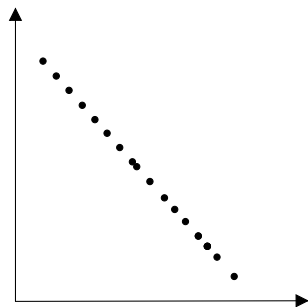
קשר לינארי חיובי מלא



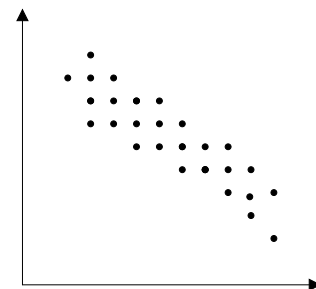
קשר לינארי חיובי חלקי



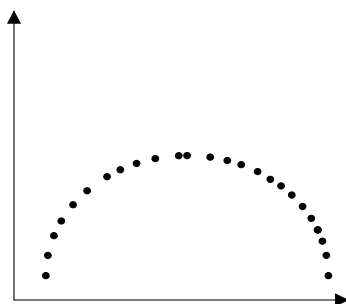
קשר לינארי שלילי מלא



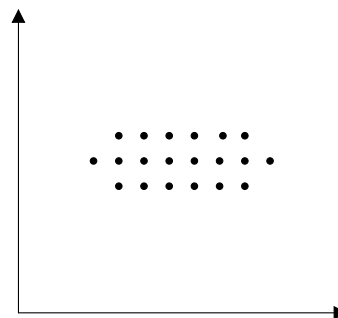
קשר לינארי שלילי חלקי



אין קשר לינארי



אין קשר



משמעות מקדם המתאם:

כדי לבדוק עד כמה קיים קשר לינארי בין שני המשתנים ישנו מדד קשר שנקרא גם מקדם המתאם הלינארי הידוע גם בשם מקדם המתאם של פירסון. מקדם מתאם זה מקבל ערכים בין 1 ל-1.

-1

0

1

מקדם מתאם 1-או 1 אומר שקיים קשר לינארי מלא בין המשתנים שניתן לבטאו על ידי נוסחה של קו ישר: $y = ax + b$.

מתאם חיובי מלא (מקדם מתאם 1):

קיים קשר לינארי מלא בו השיפוע a יהיה חיובי ואילו מתאם שלילי (מקדם מתאם-1) מלא אומר שקיים קשר לינארי מלא בו השיפוע a שלילי.

מתאם חיובי חלקי:

ככל שמשתנה אחד עולה לשני יש נטייה לעלות בערכו אבל לא קיימת נוסחה לינארית שמקשרת את X ל- Y באופן מוחלט ואילו מתאם שלילי חלקי אומר שככל שמשתנה אחד עולה לשני יש נטייה לרדת אבל לא קיימת נוסחה לינארית שמקשרת את X ל- Y באופן מוחלט. ככל שמקדם המתאם קרוב לאפס עוצמת הקשר יותר חלשה וככל שהמדד רחוק יותר מהאפס העוצמה יותר חזקה. לסיכום, מקדם המתאם בודק את עוצמת הקשר הלינארי, ואת כיוון הקשר.

מקדם המתאם הלינארי אינו מושפע מיחידות המדידה. כל שינוי ביחידות המדידה של המשתנים, לא ישנה את מקדם המתאם.

מדד הקשר הלינארי באוכלוסייה, שנקרא גם מקדם המתאם של פירסון או מדד הקשר של פירסון באוכלוסייה מסומן ב: ρ - פרמטר המאפיין את עוצמת הקשר הלינארי באוכלוסייה וכיוונו בין שני המשתנים הנחקרים. כאשר:

r - מדד הקשר הלינארי במדגם שמהווה אומד לפרמטר ρ .

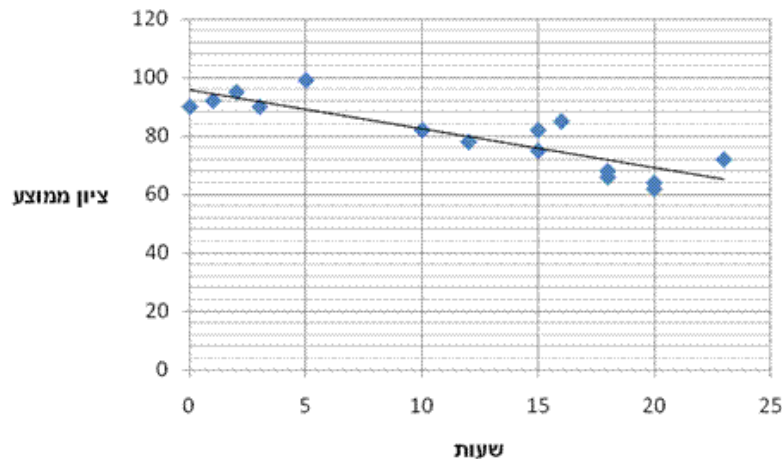
קיומו של מתאם בין שני משתנים אינו מצביע על סיבתיות בהכרח. למשל, אם נמצא מתאם חיובי בין כמות הסוכרזית שאדם אוכל לבין במשקל שלו אין זה אומר שהסיבה להשמנה היא הסוכרזית. מדד הקשר של פירסון הוא מדד קשר סימטרי, כלומר אם נחליף את X ב- Y התוצאה תהיה זהה.

דוגמה (פתרון בהקלטה):

- מה ניתן להגיד על מקדם המתאם של שני המשתנים על סמך דיאגרמת הפיזור ששרטטנו?
- אם היינו משנים את השרטוט כך שבציר האנכי היה המשתנה "מספר החדרים" ובציר האופקי היה "מספר הנפשות", האם הדבר היה משפיע על מדד הקשר של פירסון?

שאלות

- (1) חוקר רצה לאפיין את הקשר בין מספר השעות בשבוע שסטודנט מקדיש לבילויים לבין הציון הממוצע שלו בסוף הסמסטר. לשם כך הוא אסף נתונים של 15 סטודנטים ויצר דיאגרמת פיזור:



- א. מיהו המשתנה הבלתי תלוי?
 ב. מה ניתן לומר על כיוון הקשר בין מספר שעות הבילוי השבועיות לבין הציון הממוצע של הסמסטר? מה ניתן להגיד על עוצמת הקשר?
- (2) להלן טבלה המסכמת את מקדמי המתאם הלינארי בין ציוני מבחנים שונים שהתקבלו עבור תלמידים בכיתה מסוימת:

מתמטיקה	לשון	ספורט	
?	-0.7	?	ספורט
0.6	?	?	לשון
?	?	-0.1	מתמטיקה

- א. השלימו את מקדמי המתאם שמסומנים בסימן שאלה בטבלה.
 ב. בין אילו שני ציוני מקצועות שונים קיים מתאם בעל העוצמה החזקה ביותר?

- (3) במחקר נתבקשו לבדוק את הקשר בין מספר שעות התרגול של קורס לבין הציון הסופי שלו. להלן תוצאות מדגם שהתקבל:

שעות תרגול	ציון סופי
20	90
25	90
30	95
15	60
30	90
20	85
10	50

- א. מיהו המשתנה התלוי ומיהו המשתנה הבלתי תלוי בדוגמה זו?
 ב. שרטטו דיאגרמת פיזור לנתונים.
 ג. מה ניתן לומר על הקשר בין המשתנים במדגם?
 ד. מסתבר שבסופו של דבר נתנו פקטור של 5 נקודות לציון הסופי. כיצד הדבר היה משנה את מקדם המתאם של המדגם?

4) בתחנה המטאורולוגית רצו לבדוק את הקשר שבין הטמפרטורה במעלות צלזיוס לכמות המשקעים במ"מ. הם אספו נתונים על 10 ימים במהלך חודש ינואר. המתאם שהתקבל היה 0.8.

א. השלימו את המשפט:

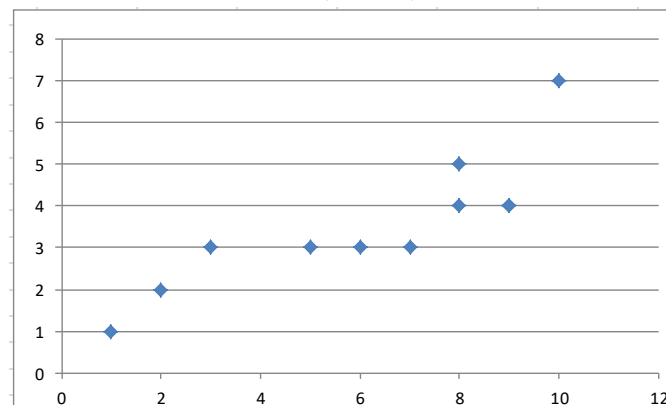
בחודש ינואר ככל שהטמפרטורה היומית נוטה לרדת, כך כמות המשקעים נוטה _____.

ב. הוחלט להעביר את הטמפרטורה למעלות פרנהייט על מנת שיוכלו

להשוות אותה לנתונים מארה"ב. נוסחת המעבר היא $F^0 = 32 + \frac{9}{5}C^0$.

כיצד הדבר ישפיע על מקדם המתאם בין הטמפרטורה במעלות פרנהייט לכמות המשקעים במ"מ?

5) להלן דיאגרמת פיזור המראה קשר בין שני משנים:



א. השלימו: ניתן לראות שהקשר הוא לינארי _____ (מלאו חלקי) כיוון הקשר הוא (חיובי/שלילי).

ב. השלימו: אם היינו מוסיפים תצפית שערך ה- X שלה הוא 4 וערך ה- Y שלה הוא 7, מקדם המתאם של פירסון היה _____ (גדלו קטן/לא משתנה).

שאלות רב ברירה (יש לבחור את התשובה הנכונה):

6) חוקר אקלים דגם כמה ימים בשנה ומדד את הטמפרטורה בטורונטו שבקנדה ואת הטמפרטורה בסידני שבאוסטרליה באותו היום. הוא חישב ומצא מקדם מתאם שלילי בין הטמפרטורה היומית בטורונטו לבין הטמפרטורה היומית בסידני. משמעות מקדם המתאם השלילי במדגם:

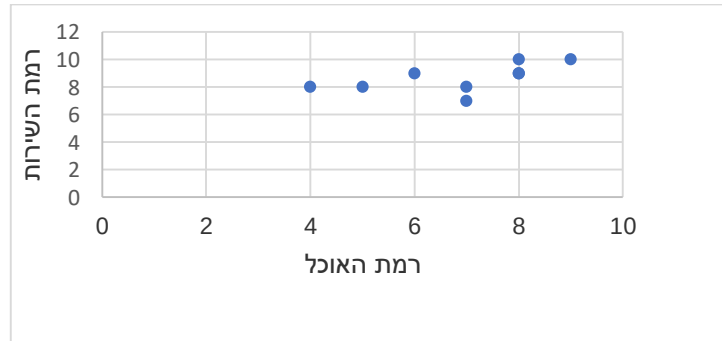
א. אין קשר בין הטמפרטורה בטורונטו לבין הטמפרטורה בסידני בימים שנדגמו.

ב. במדגם, רוב הטמפרטורות בטורונטו היו שליליות.

ג. ההפרש בין הטמפרטורה בטורונטו לבין הטמפרטורה באוסטרליה, במדגם זה, הוא שלילי.

ד. במדגם יש נטייה שהטמפרטורה יורדת בטורונטו לטמפרטורה לעלות בסידני.

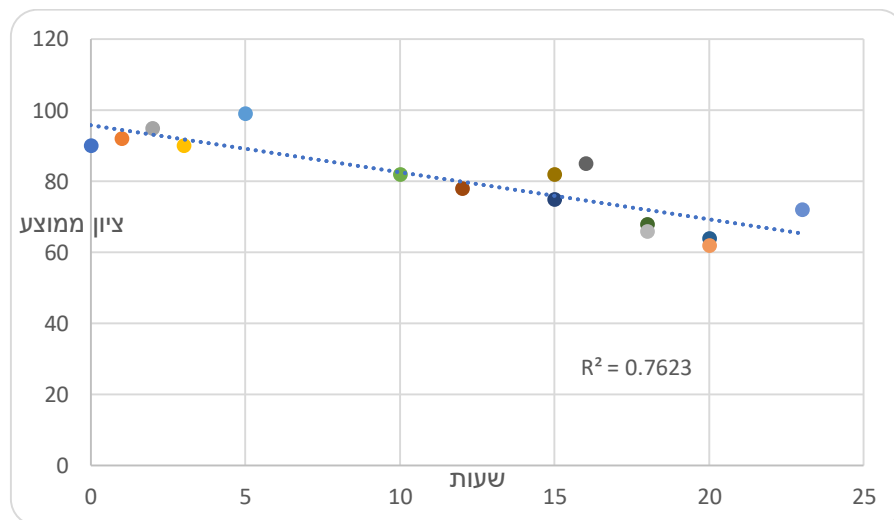
- 7) בסקר שביעות רצון שנערך בבית הקפה "פת לחם" התבקשו הלקוחות לדרג את מידת שביעות הרצון שלהם (בסולם 1-10) בשני נושאים: רמת האוכל ורמת השירות.



מה יהיה ערכו של מקדם המתאם (r)?

- א. $r = -0.3$
- ב. $r = 0$
- ג. $r = 1.125$
- ד. $r = 0.593$

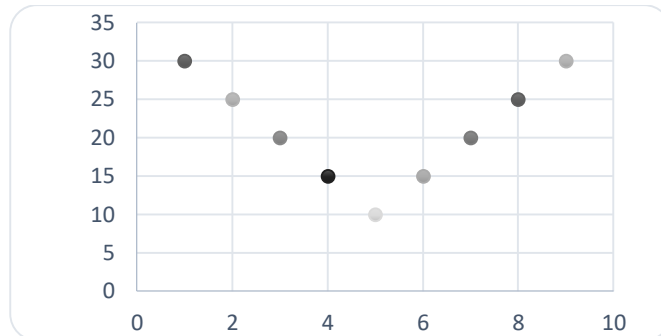
- 8) חוקר רצה לאפיין את הקשר בין מספר השעות בשבוע שסטודנט מקדיש לבילויים לבין הציון הממוצע שלו בסוף הסמסטר. לשם כך הוא אסף נתונים של 15 סטודנטים ויצר דיאגרמת פיזור.



מה ניתן לומר על כיוון הקשר במדגם בין מספר שעות הבילוי השבועיות לבין הציון הממוצע של הסמסטר?

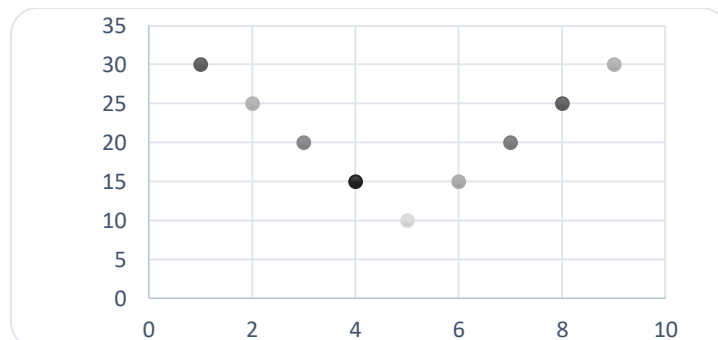
- א. ככל שמבליים יותר הציון נוטה לרדת.
- ב. אין קשר בין שעות הבילוי לציון.
- ג. ככל שמבליים פחות הציון נוטה לרדת.
- ד. ככל שהציון נוטה לרדת הסטודנט מבלה פחות.

9) התרשים הבא מתאר קשר בין שני משתנים, איזה מהמתאמים הבאים הוא המתאים ביותר לתיאור הקשר בין שני המשתנים?



- א. $r = 1$ היות ושני המשתנים יוצרים קוים ישרים.
- ב. $r = 2$ היות ויש שני קוים בעלי קשר מושלם.
- ג. $r = 0$ היות והקו יורד ואחר כך עולה באותו האופן.
- ד. $r = \pm 1$ היות ויש קו עולה וגם קו יורד.

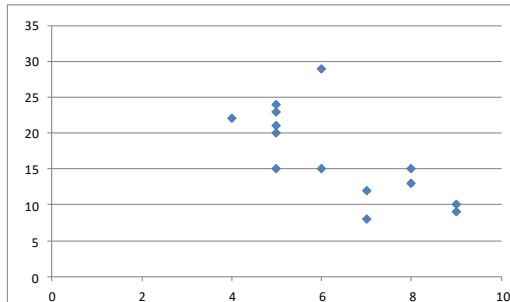
10) התרשים הבא מתאר דיאגרמת פיזור.



איזו טענה נכונה?

- א. בתרשים מוצג הקשר בין שני משתנים.
- ב. בתרשים מוצג הקשר בין 9 משתנים.
- ג. בתרשים מוצג הקשר בין 10 משתנים.
- ד. אין לדעת כמה משתנים מוצגים בתרשים.

בגרף הבא מתוארת דיאגרמת פיזור של שני משתנים:



X - (משתנה בלתי תלוי בציר האופקי)

Y - (משתנה תלוי).

במדגם התקבל $r^2 = 0.52$.

11) לאור הנתונים המופיעים בדיאגרמה, איזה מבין הערכים הבאים מתאים להיות התוצאה של r ?

א. -0.52

ב. 0.72

ג. -0.72

ד. 0.52

12) אם מקדם המתאם בין שני משתנים הוא 1, אזי:

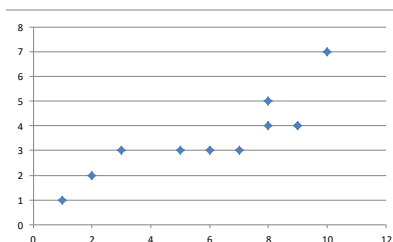
א. הערכים של המשתנים הם חיוביים.

ב. עבור כל תצפית ערך של משתנה אחד שווה לערך של המשתנה השני.

ג. הקשר הלינארי הוא בעוצמה חזקה.

ד. אף אחת מהתשובות לא בהכרח נכונה.

13) להלן דיאגרמת פיזור:



מה יהיה מקדם המתאם בין שני המשתנים?

א. 1

ב. 0.85

ג. 0.15

ד. 0

14) בבדיקת קשר בין שני משתנים התקבל: $r = -1$.

א. קיימת נוסחה לינארית הקושרת בין כל התצפיות.

ב. לא קיים קשר בין שני המשתנים.

ג. ככל שמשתנה אחד נוטה לרדת גם לשני יש נטייה לרדת.

ד. קיים קשר בין שני המשתנים, אך לא ניתן לדעת מאיזה סוג.

15) לפי הפתגם "רחוק מהעין, רחוק מהלב", יש קשר ____ בין קרבה פיזית לקרבה נפשית.

- א. חיובי
- ב. שלילי
- ג. אפסי
- ד. לא ניתן לדעת.

16) מבחן אמי"ר הינו מבחן מיון באנגלית של המרכז הארצי לבחינות והערכה. הציון המינימלי בבחינה הינו 150 והמקסימלי הינו 250. בקורס הכנה למבחן השתתפו 19 תלמידים. להלן הציונים שלהם על פי פלט שהתקבל:

	159
	170
	180
	185
	204
	224
	236
	212
	168
	189
	195
	163
	187
	206
	201
	223
	242
	203
	205
197.47	AVERAGE
536.25	VARPA

יש להוסיף עמודה נוספת לצד עמודת הציונים שתראה לכל תלמיד כמה נקודות חסרות לו כדי להשלים לציון המקסימלי בבחינה.

מה יהיה מקדם המתאם בין שתי העמודות (כלומר, מקדם המתאם בין הציון לבין הנקודות החסרות)?

- א. -1
- ב. 1
- ג. -0.5
- ד. 0.5

17) מקדם המתאם בין שטחי דירה למחיר שלהם חושב ונמצא 1.2. מה נובע מכך?

- א. ככל שהדירה גדולה יותר בשטחה כך היא יקרה יותר.
- ב. ככל שהדירה קטנה יותר בשטחה כך היא זולה יותר.
- ג. לא קיים קשר בין שטח הדירה למחיר הדירה.
- ד. מצב כזה שמתואר הנתונים לא אפשרי.

18) אם ניקח 10 אנשים ונרשום לכל אדם את הגובה במטר וכמו כן את הגובה בס"מ. מה יהיה מקדם המתאם בין גובה האדם במטר לגובה האדם בס"מ?

- א. 1
- ב. 0
- ג. -1
- ד. לא ניתן לדעת.

- 19** נמצא מתאם חיובי בעוצמה גבוהה בין X – ציון בבגרות בלשון ל Y – ציון בבגרות במתמטיקה. אילו מהמשפטים הבאים נכון?
- א. ניתן לומר שאחת מהסיבות להבדלים שיש לסטודנטים במתמטיקה נובעים מההבדלים שיש להם בלשון.
- ב. קיימת נוסחה של קו ישר שקושרת בין ציון בבגרות במתמטיקה לציון בבגרות בלשון.
- ג. ללא יוצא מן הכלל, ניתן להגיד שכל תלמיד שמצליח יותר מתלמיד אחר בלשון גם יצליח יותר מאותו תלמיד במתמטיקה.
- ד. אף אחד מהטענות שהוצגו אינה בהכרח נכונה.

- 20** עבור סדרה של תצפיות מדדו את X ואת Y . נמצא שעבור כל התצפיות שהערך של Y ירד הערך של X בהכרח ירד ללא יוצא מן הכלל. מקדם המתאם של פירסון יהיה בהכרח :

- א. 1
- ב. -1
- ג. 0
- ד. אף אחת מהתשובות.

תשובות סופיות

- (1) א. שעות בילוי.
ב. הקשר חלקי, כיוון הקשר שלילי.
(2) א. להלן טבלה:
ב. ספורט ולשון.

מתמטיקה	לשון	ספורט	
0.1	-0.7	1	ספורט
0.6	1	-0.7	לשון
1	0.6	-0.1	מתמטיקה

- (3) א. ב"ת- מס' שעות התרגול, תלוי- ציון.
ג. קשר לינארי חיובי חלקי.
(4) א. לעלות.
(5) א. חלקי, חיובי.
ב. קטן.
ב. ראה גרף בפתרון וידאו.
ד. מקדם המתאם לא היה משתנה.
ב. לא ישפיע על מקדם המתאם.

(6) ד' (7) ד' (8) א' (9) ג' (10) א'

(11) ג' (12) ד' (13) ב' (14) א' (15) א'

(16) א' (17) ד' (18) א' (19) ד' (20) ד'

מדדי קשר – מדד הקשר הלינארי (פירסון) – רקע

המטרה היא לבדוק האם קיים קשר (קורלציה, מתאם) של קו ישר בין שני משתנים כמותיים. מבחינת סולמות המדידה קשר בין סולמות רווחים ומנה. בדרך כלל, X הוא המשתנה המסביר (הבלתי תלוי) ו- Y הוא המשתנה המוסבר (התלוי).

דוגמה:

נרצה להסביר כיצד השכלה של אדם הנמדדת בשנות לימוד X מסבירה את ההכנסה שלו Y . במקרה זה שנות ההשכלה זהו המשתנה המסביר (או הבלתי תלוי) ואנחנו מעוניינים לבדוק כיצד שינויים בשנות ההשכלה של אדם יכולים להסביר את השינויים שלו בהכנסה, ולכן רמת ההכנסה זהו המשתנה המוסבר התלוי במשתנה המסביר אותו.

שלב ראשון: נהוג לשרטט דיאגרמת פיזור. זו דיאגרמה שנותנת אינדיקציה ויזואלית על טיב הקשר בין שני המשתנים.

דוגמה:

מס' דירה	X	Y
1	3	2
2	2	2
3	4	3
4	3	3
5	5	4

בבניין של 5 דירות בדקו את הנתונים הבאים:
 X - מס' חדרים בדירה. Y - מס' נפשות הגרות בדירה.
 להלן התוצאות שהתקבלו:

נשרטט מנתונים אלה דיאגרמת פיזור (הדיאגרמה המלאה בסרטון). נתבונן בכמה מקרים של דיאגרמות פיזור וננתח אותן (הדיאגרמות המלאות בסרטון).

שלב שני: מחשבים את מקדם המתאם (מדד הקשר) שבודק עד כמה קיים קשר לינארי בין שני המשתנים. המדד (ניקרא גם מדד הקשר של פירסון) מכמת את מה שניראה בשלב הראשון רק בעין.
 המדד בודק את כיוון הקשר (חיובי או שלילי) ואת עוצמת הקשר (חלש עד חזק).
 מקדם מתאם זה מקבל ערכים בין -1 ל-1.
 מקדם מתאם -1 או 1 אומר שקיים קשר לינארי מוחלט ומלא בין המשתנים שניתן לבטאו על ידי הנוסחה: $y = bx + a$.

מתאם חיובי מלא (מקדם מתאם 1):

קיים קשר לינארי מלא בו השיפוע b יהיה חיובי ואילו מתאם שלילי מלא אומר שקיים קשר לינארי מלא בו השיפוע b שלילי (מקדם מתאם -1).

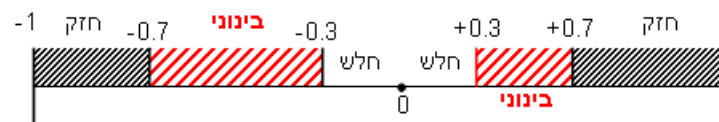
מתאם חיובי חלקי:

ככל שמשתנה אחד עולה לשני יש נטייה לעלות בערכו אבל לא קיימת נוסחה לינארית שמקשרת את X ל- Y באופן מוחלט.

מתאם שלילי חלקי:

ככל שמשתנה אחד עולה לשני יש נטייה לרדת אבל לא קיימת נוסחה לינארית שמקשרת את X ל- Y באופן מוחלט.

ככל שערך מקדם המתאם קרוב לאפס נאמר שעוצמת הקשר חלשה יותר וככל שמקדם המתאם רחוק מהאפס נאמר שעוצמת הקשר חזקה יותר:



מקדם המתאם יסומן באות r .

כדי לחשב את מקדם המתאם, יש לחשב את סטיות התקן של כל משתנה ואת השונות המשותפת.

$$COV(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n} = \frac{\sum xy}{n} - \bar{x} \cdot \bar{y} : \text{שונות משותפת}$$

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2 : \text{שונות של המשתנה } X$$

$$S_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n} = \frac{\sum_{i=1}^n y_i^2}{n} - \bar{y}^2 : \text{שונות המשתנה } Y$$

$$r_{xy} = \frac{COV(x, y)}{S_x \cdot S_y} : \text{מקדם המתאם הלינארי}$$

שאלות

- 1) להלן נתונים לגבי שישה תלמידים שנגשו למבחן. בדקו לגבי כל תלמיד את הציון שלו בסוף הקורס וכמו כן את מספר החיסורים שלו מהקורס.

מספר חיסורים	2	1	0	2	3	4
ציון	80	90	90	70	70	50

- א. שרטטו דיאגרמת פיזור לנתונים. מה ניתן להסיק מהדיאגרמה על טיב הקשר בין מספר החיסורים של תלמיד לציונו? מיהו המשתנה הבלתי תלוי ומיהו המשתנה התלוי?
- ב. חשבו את מדד הקשר של פירסון. האם התוצאה מתיישבת עם תשובתך לסעיף א'?
- ג. הסבירו, ללא חישוב, כיצד מקדם המתאם היה משתנה אם היה מתווסף תלמיד שהחסיר 4 פעמים וקיבל ציון 80?

X	Y
10	12
14	15
15	15
18	17
20	21

- 2) במחקר רפואי רצו לבדוק האם קיים קשר בין רמת ההורמון X בדם החולה לרמת ההורמון Y שלו. לצורך כך מדדו את רמת ההורמונים ההלו עבור חמישה חולים. להלן התוצאות שהתקבלו:
- א. מה הממוצע של כל רמת הורמון?
- ב. מהו מקדם המתאם בין ההורמונים? ומה משמעות התוצאה?

- 3) נסמן ב-X את ההכנסה של משפחה באלפי ₪. נסמן ב-Y את ההוצאות של משפחה באלפי ₪. נלקחו 20 משפחות והתקבלו התוצאות הבאות:

$$\sum_{i=1}^{20} Y_i = 200 \qquad \sum_{i=1}^{20} X_i = 240$$

$$\sum_{i=1}^{20} (Y_i - \bar{Y})^2 = 76 \qquad \sum_{i=1}^{20} (X_i - \bar{X})^2 = 76$$

$$\sum_{i=1}^{20} (X_i - \bar{X})(Y_i - \bar{Y}) = 60.8$$

- א. חשב את מדד הקשר הלינארי בין X ל-Y. מיהו המשתנה התלוי?
- ב. מה המשמעות של התוצאה שקיבלת בסעיף א'?

- (4) נסמן ב- X את ההכנסה של משפחה באלפי ₪. נסמן ב- Y את ההוצאות של משפחה באלפי ₪. נלקחו 20 משפחות והתקבלו התוצאות הבאות:

$$\sum_{i=1}^{20} Y_i = 200 \quad \sum_{i=1}^{20} X_i = 240$$

$$\sum_{i=1}^{20} Y_i^2 = 2080 \quad \sum_{i=1}^{20} X_i^2 = 2960$$

$$\sum_{i=1}^{20} X_i Y_i = 2464$$

חשבו את מדד הקשר הלינארי בין X ל- Y .

- (5) במוסד אקדמי ציון ההתאמה מחושב כך: מכפילים את הציון הממוצע בבגרות ב-3 ומפחיתים 2 נקודות. ידוע שעבור 40 מועמדים סטיית התקן של ממוצע הציון בבגרות הייתה 2.
מה מקדם המתאם בין ציון ההתאמה לציון הממוצע בבגרות שלהם?

- (6) להלן רשימת טענות, לגבי כל טענה קבעו נכון/לא נכון ונמקו.
- מתווך דירות המיר מחירי דירות מדולר לשקל. נניח שדולר אחד הוא 3.5 ₪. אם מתווך הדירות יחשב את מדד הקשר של פירסון בין מחיר הדירה בשקלים למחיר הדירה בדולרים הוא יקבל 1.
 - לסדרה של נתונים התקבל $\bar{X} = \bar{Y} = 6$, $S_x = S_y = 1$. לכן, מדד הקשר של פירסון יהיה 1.
 - אם השונות המשותפת של X ושל Y הינה 0 אז בהכרח גם מקדם המתאם של פירסון יהיה 0.

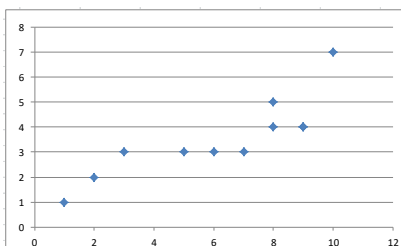
שאלות רב-ברירה:

- (7) נמצא שקיים מקדם מתאם שלילי בין הציון בעברית לציון בחשבון בבחינה לכן:
- הדבר מעיד שהציונים בכיתה היו שליליים.
 - ככל שהציון של תלמיד יורד בחשבון יש לו נטייה לרדת בעברית.
 - ככל שהציון של תלמיד עולה בחשבון יש לו נטייה לרדת בעברית.
 - אף אחת מהתשובות לא נכונה.

8) נלקחו 20 מוצרים ונבדק ביום מסוים המחיר שלהם בדולרים והמחיר שלהם בש"ח (באותו היום ערך הדולר היה-4.2ש"ח). מהו מקדם המתאם בין המחיר בדולר למחיר בש"ח?

- א. 1
 ב. 0
 ג. 4.2
 ד. לא ניתן לדעת.

9) להלן דיאגרמת פיזור: מה יהיה מקדם המתאם בין שני המשתנים?



- א. 1
 ב. 0.85
 ג. 0.15
 ד. 0

תשובות סופיות

- 1) א. משתנה תלוי: ציון, משתנה ב"ת: מס' חיסורים. ראה דיאגרמה בוידאו. ניתן להסיק שקיים קשר לינארי שלילי וחלקי בין מספר החיסורים לציון התלמיד.
 ב. -0.9325.
 ג. הקשר יישאר לינארי שלילי חלקי אך עוצמתו תחלש.
- 2) א. $\bar{y} = 16$, $\bar{x} = 15.4$ ב. $r_{xy} = 0.96$.
- 3) א. 0.8
 4) 0.8
 5) 1
 6) א. נכון. ב. לא נכון. ג. נכון.
 7) ג'.
 8) א'.
 9) ב'.

בדיקת השערות על מקדם המתאם הלינארי – רקע

מדד הקשר הלינארי באוכלוסייה, שנקרא גם מקדם המתאם של פירסון או מדד הקשר של פירסון באוכלוסייה מסומן ב: ρ - פרמטר המאפיין את עוצמת הקשר הלינארי וכיוונו בין שני המשתנים הנחקרים באוכלוסייה. כאשר: r - מדד הקשר הלינארי במדגם שמהווה אומדן לפרמטר ρ .

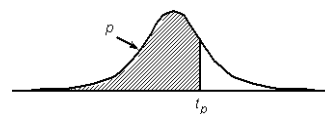
השערת האפס: תהיה שבאוכלוסייה לא קיים כלל קשר לינארי בין שני המשתנים $H_0: \rho = 0$.
ההנחה שעליה אנו מתבססים בתהליך היא ששני המשתנים הנחקרים מתפלגים דו נורמלית.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2)$$

סטטיסטי המבחן:

סטטיסטי זה מתפלג t עם $n-2$ דרגות חופש.

$H_0: \rho = 0$	$H_0: \rho = 0$	$H_0: \rho = 0$	השערת האפס:
$H_1: \rho > 0$	$H_1: \rho < 0$	$H_1: \rho \neq 0$	השערת המחקר:
$t \geq t_{1-\alpha}$	$t \leq -t_{1-\alpha}$	$t \geq t_{1-\alpha}$ ו $t \leq -t_{1-\alpha}$	כלל ההכרעה: אזור דחייה של השערת האפס

טבלת ערכים קריטיים של t - נספח: טבלת התפלגות T
 P 

דרגות חופש	0.75	0.90	0.95	0.975	0.99	0.995	0.9995
1	1.000	3.078	6.314	12.709	31.821	63.657	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	1.638	2.353	3.182	4.541	5.841	12.941
4	0.741	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	6.859
6	0.718	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	5.405
8	0.706	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.291

שאלות

(1) להלן נתונים על הוותק בעבודה (בשנים) ועל השכלה (בשנים) במדגם של 10 עובדים:

נחקר	1	2	3	4	5	6	7	8	9	10
X - וותק	13	18	2	8	16	9	5	28	17	24
Y - השכלה	12	14	17	8	11	12	13	8	12	15

מקדם המתאם חושב והתקבל: -0.31 .

- א. האם קיים מתאם בין וותק העובד להשכלתו? בדקו ברמת מובהקות של 5%?
 ב. אם הוותק של העובד היה נמדד בחודשים האם התשובה לסעיף א' הייתה משתנה?

(2) מחקר התעניין לבדוק את הקשר בין גיל נשים בהריון לרמת ההמוגלובין שלהן בדם בזמן הריון. נדגמו 7 נשים והתקבלו התוצאות הבאות:

נחקרת	1	2	3	4	5	6	7
המוגלובין	14.7	13.5	9.7	12	10.8	13	10.3
גיל	39	34	30	29	28	26	23

במדגם חושב מדד הקשר של פירסון להיות 0.7 .

- א. האם ניתן לומר שבמדגם אם אישה היא יותר מבוגרת אזי בהכרח יש לה יותר המוגלובין בדם?
 ב. האם ניתן לומר, ברמת מובהקות של 5%, שקיים מתאם בין גיל האישה שבהריון לבין רמת ההמוגלובין שלה בדם?

(3) בתחנה המטאורולוגית רצו לבדוק את הקשר שבין הטמפרטורה במעלות צלזיוס לכמות המשקעים במ"מ. הם אספו נתונים על 10 ימים במהלך חודש ינואר. המתאם שהתקבל היה -0.8 .

- א. בדקו ברמת מובהקות של 2.5% האם קיים קשר לינארי שלילי בחודש ינואר בין הטמפרטורה במעלות צלזיוס לבין המשקעים במעלות צלזיוס.
 ב. כיצד הייתה משתנה התשובה לסעיף א אם הינו מוסיפים עוד תצפיות למדגם?
 ג. על סמך טבלת T המצורפת עבור אילו רמות מובהקות ניתן להחליט שקיים קשר לינארי שלילי מובהק?

(4) מתווך דירות חישב את מקדם המתאם בין שטח דירה במרכז תל אביב לבין המחיר של הדירה עבור 17 דירות. מקדם המתאם שקיבל היה 0.6 .

- א. בדקו ברמת מובהקות של 5% האם ניתן להגיד שקיים קשר ישר עולה בין שטח הדירה לבין מחיר הדירה במרכז תל אביב?
 ב. מהי מובהקות התוצאה לבדיקת ההשערה שקיים קשר ישר עולה בין שטח הדירה לבין מחיר הדירה בתל אביב.

תשובות סופיות

- | | |
|---------------------------|---------------------------|
| ב. לא תשתנה. | (1) א. לא נדחה את H_0 . |
| ב. לא נדחה את H_0 . | (2) א. לא |
| ב. לא ניתן לדעת. | (3) א. נדחה את H_0 . |
| | ג. לפחות 0.005. |
| ב. $0.005 < P_v < 0.01$. | (4) א. נדחה את H_0 . |

בדיקת השערות על מקדם המתאם הלינארי (באמצעות טבלה של ערכים קריטיים) – רקע

מדד הקשר הלינארי באוכלוסייה, שנקרא גם מקדם המתאם של פירסון או מדד הקשר של פירסון באוכלוסייה מסומן ב: ρ - פרמטר המאפיין את עוצמת הקשר הלינארי וכיוונו בין שני המשתנים הנחקרים באוכלוסייה. כאשר:

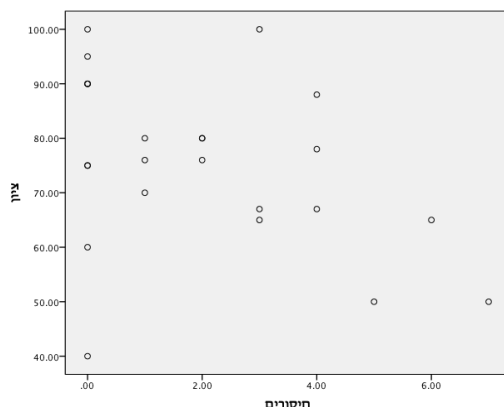
r - מדד הקשר הלינארי במדגם שמהווה אומדן לפרמטר ρ .

השערת האפס: תהיה שבאוכלוסייה לא קיים כלל קשר לינארי בין שני המשתנים: $H_0: \rho = 0$.
ההנחה שעליה אנו מתבססים בתהליך היא ששני המשתנים הנחקרים מתפלגים דו-נורמלית.
את מקדם המדגם הקריטי, שנסמן ב- r_c , נוציא מתוך טבלה של ערכים קריטיים שמצורפת בהמשך.

$H_0: \rho = 0$	$H_0: \rho = 0$	$H_0: \rho = 0$	השערת האפס:
$H_1: \rho > 0$	$H_1: \rho < 0$	$H_1: \rho \neq 0$	השערת המחקר:
$r \geq r_c$	$r \leq -r_c$	$r \geq r_c$ ו $r \leq -r_c$	כלל ההכרעה: אזור דחייה של השערת האפס

דוגמה (פתרון בהקלטה):

הדיקן ביקש לדגום סטודנטים כדי לבדוק את הקשר בין ציון הסטודנט בקורס למספר הפעמים שהוא החסיר שיעור בקורס.
דיאגרמת הפיזור שהתקבלה במדגם שבוצע:



מיהו המשתנה התלוי ומיהו המשתנה הבלתי תלוי במחקר?
מה ניתן לראות לגבי הקשר הלינארי בין המשתנים שהתקבל במדגם?

חושב האומדן למקדם המתאם הלינארי על סמך 24 הסטודנטים שנדגמו והתקבל: -0.389.

מה משמעות של מקדם המתאם שהתקבל במדגם?

האם ניתן להגיד ברמת מובהקות של 5% שקיים מתאם לינארי שלילי בין מספר החיסורים של הסטודנטים מהקורס לבין הציון של הסטודנטים בקורס?

טבלת ערכים קריטיים של מקדם המתאם הלינארי



0.0005	0.005	0.025	0.05	α / n
0.999	0.990	0.950	0.900	4
0.991	0.959	0.878	0.805	5
0.974	0.917	0.811	0.729	6
0.951	0.875	0.754	0.669	7
0.925	0.834	0.707	0.621	8
0.898	0.798	0.666	0.582	9
0.872	0.765	0.632	0.549	10
0.847	0.735	0.602	0.521	11
0.823	0.708	0.576	0.497	12
0.801	0.684	0.553	0.476	13
0.780	0.661	0.532	0.458	14
0.760	0.641	0.514	0.441	15
0.742	0.623	0.497	0.426	16
0.725	0.606	0.482	0.412	17
0.708	0.590	0.468	0.400	18
0.693	0.575	0.456	0.389	19
0.679	0.561	0.444	0.378	20
0.665	0.549	0.433	0.369	21
0.652	0.537	0.423	0.360	22
0.640	0.526	0.413	0.352	23
0.629	0.515	0.404	0.344	24
0.618	0.505	0.396	0.337	25
0.607	0.496	0.388	0.330	26
0.597	0.487	0.381	0.323	27
0.588	0.479	0.374	0.317	28
0.579	0.471	0.367	0.311	29
0.570	0.463	0.361	0.306	30
0.532	0.430	0.334	0.283	35

שאלות

- (1) להלן נתונים על הוותק בעבודה (בשנים) ועל השכלה (בשנים) במדגם של 10 עובדים:

נחקר	1	2	3	4	5	6	7	8	9	10
X - וותק	13	18	2	8	16	9	5	28	17	24
Y - השכלה	12	14	17	8	11	12	13	8	12	15

מקדם המתאם חושב והתקבל: -0.31 .

- א. האם קיים מתאם בין וותק העובד להשכלתו? בדקו ברמת מובהקות של 5%.
- ב. אם הוותק של העובד היה נמדד בחודשים האם התשובה לסעיף א' הייתה משתנה?

- (2) מחקר התעניין לבדוק את הקשר בין גיל נשים בהריון לרמת ההמוגלובין שלהן בדם בזמן הריון. נדגמו 7 נשים והתקבלו התוצאות הבאות:

נחקרת	1	2	3	4	5	6	7
המוגלובין	14.7	13.5	9.7	12	10.8	13	10.3
גיל	39	34	30	29	28	26	23

במדגם חושב מדד הקשר של פירסון להיות 0.7 .

- א. האם ניתן לומר שבמדגם אם אישה היא יותר מבוגרת אזי היא בהכרח יש לה יותר המוגלובין בדם?
- ב. האם ניתן לומר, ברמת מובהקות של 5%, שהמתאם בין גיל האישה שבהריון לבין רמת ההמוגלובין שלה בדם הוא חיובי?

- (3) בתחנה המטאורולוגית רצו לבדוק את הקשר שבין הטמפרטורה במעלות צלזיוס לכמות המשקעים במ"מ. הם אספו נתונים על 10 ימים במהלך חודש ינואר. המתאם שהתקבל היה -0.8 .

- א. בדוק ברמת מובהקות של 2.5% האם קיים קשר לינארי שלילי בחודש ינואר בין הטמפרטורה במעלות צלזיוס לבין המשקעים במ"מ?
- ב. כיצד הייתה משתנה התשובה לסעיף א אם הינו מוסיפים עוד תצפיות למדגם?

תשובות סופיות

- (1) א. לא נדחה את H_0 .
ב. לא תשתנה.
- (2) א. לא.
ב. נדחה את H_0 .
- (3) א. נדחה את H_0 .
ב. לא ניתן לדעת.

סטטיסטיקה יישומית

פרק 7 - רגרסיה פשוטה

תוכן העניינים

1. כללי 72

הגרסיה פשוטה:

רקע:

הגרסיה ליניארית פשוטה מסתמכת על המתאם הליניארי בין המשתנה התלוי (המנובא) לב"ת (המנבא).

$$r = \frac{\text{cov}(x, y)}{S_x \cdot S_y} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)} \cdot \sqrt{\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)}} = \frac{S_{XY}}{\sqrt{S_{XX}} \cdot \sqrt{S_{YY}}} : \text{מקדם המתאם}$$

המודל באוכלוסייה: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

כאשר:

β_0 הוא החותך.

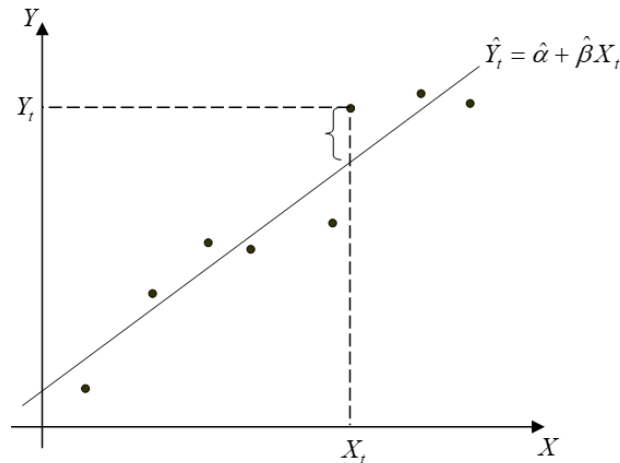
β_1 הוא שיפוע.

ε_i הינו גורם הטעות מסביב לקו הליניארי.

המודל הנאמד (על סמך מדגם): $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

לסיכום:

1. במודל $y_i = \alpha + \beta x_i + \varepsilon_i$, α ו- β הם מספרים קבועים אך לא ידועים. אנו יכולים להעריך אותם ולקבל אומדים (תהליך קבלת האומדנים נקרא אמידה).
2. $\hat{\alpha}$ הוא האומד ל- α . $\hat{\beta}$ הוא האומד ל- β .
3. אומדי ריבועים פחותים (אר"פ) הם אומדים שחושבו בשיטת הריבועים הפחותים. אומדי הריבועים הפחותים מסומנים בד"כ ע"י 'כובעי' - $\hat{\beta}, \hat{\alpha}$.
4. בעוד α ו- β הם קבועים, $\hat{\alpha}$ ו- $\hat{\beta}$ הם משתנים מקריים. מדוע? מפני שבכל מדגם מתקבלים $\hat{\alpha}$ ו- $\hat{\beta}$ אחרים.
5. את α ו- β אי אפשר לדעת, ולכן אי אפשר לדעת מהו הקו האמיתי, וכן אי אפשר לדעת את ε .
6. אפשר לדעת את e , שהיא הסטייה מקו הרגרסיה. נגדיר זאת באופן הבא:
 - עבור X_i , הערך הצפוי של המשתנה המוסבר ($\hat{Y}_i$) המתקבל לפי הרגרסיה הוא: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$.
 - הסטייה של התצפית (Y_i) מהערך הצפוי לפי הרגרסיה ($\hat{Y}_i$) היא: $e_i = Y_i - \hat{Y}_i$.



האומדים של הרגרסיה $(\hat{\alpha}, \hat{\beta})$:

שיטת האמידה של α ושל β נקראת שיטת הריבועים הפחותים
Ordinary Least Squares (OLS).

השאלה הנשאלת בשיטת אמידה זו היא:

איזה $\hat{\alpha}$ ו- $\hat{\beta}$ יביאו למינימום את סכום ריבועי טעויות האמידה.

$$\min_{\hat{\alpha}, \hat{\beta}} \sum e_t^2 = \min_{\hat{\alpha}, \hat{\beta}} \sum (y_t - \hat{y}_t)^2 = \min_{\hat{\alpha}, \hat{\beta}} \sum [y_t - (\hat{\alpha} + \hat{\beta}x_t)]^2 = ?$$

ובתרגום מתמטי:

מתוך גזירת הפונקציה הזו מתקבלים האומדים $\hat{\alpha}$ ו- $\hat{\beta}$: $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{S_{XY}}{S_{XX}} = \frac{COV(X, Y)}{V(X)} = r \frac{S_Y}{S_X}$$

מבחני המובהקות:

$$\begin{aligned} H_0: \beta &= 0 \\ H_1: \beta &\neq 0 \end{aligned}$$

השערות:

בגרסיה פשוטה בה יש לנו רק מנבא אחד: ניתן לבצע מבחן F למובהקות משוואת הרגרסיה או מבחן T למובהקות מקדם הרגרסיה (הביטא).

משמעות דחיית השערת האפס: משוואת הרגרסיה מובהקת, מקדם הרגרסיה מובהק, הקשר בין X ל- Y מובהק.

ולחיפך – אם השערת האפס לא נדחת: אין הוכחה לקשר בין המשתנים X ו- Y , משוואת הרגרסיה איננה מובהקת וכך גם מקדם הרגרסיה.

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{(1-r^2)SST}{n-2}$$

אמידת שונות הטעויות:

מבחן F:

מבחן זה נעשה על מנת לבדוק האם משוואת הרגרסיה מובהקת.

המבחן מתבסס על פירוק סכום הריבועים: $SST = SSR + SSE$

$$S_Y^2 = r^2 S_Y^2 + (1-r^2) S_Y^2$$

טבלת ניתוח שונות (טבלת ANOVA):

מקור	סכום ריבועים SS	דרגות חופש $d.f.$	ממוצע סכום ריבועים $MS = \frac{SS}{d.f.}$	F
מודל הרגרסיה	SSR	1	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
שאריות	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
סה"כ	SST	$n-1$		

כלל הכרעה:

אם: $F_{st} > F_c \alpha(1, n-2)$ נדחה את השערת האפס.

מבחן t:

מבחן זה נעשה על מנת לבדוק האם מקדם הרגרסיה מובהק.

$$t_{st} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{s.e.}(\hat{\beta}_1)} \sim t_{c(n-2)} \quad \text{סטטיסטי המבחן}$$

$$t_{stt} = \frac{r^2 \sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{אם השערת האפס מתייחסת ל-} \beta_0 = \beta_0 \text{ (בדר"כ):}$$

כלל הכרעה :

השערה דו צדדית $H_1 : \beta_1 \neq \beta_{1,0}$	השערה חד צדדית שמאלית $H_1 : \beta_1 < \beta_{1,0}$	השערה חד צדדית ימנית $H_1 : \beta_1 > \beta_{1,0}$	
$t_{\text{statistic}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{s.e.}(\hat{\beta}_1)} = \frac{r^2 \sqrt{n-2}}{\sqrt{1-r^2}}$ $\text{s.e.}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{SXX}}$			סטטיסטי המבחן
$ t_{\text{statistic}} \geq t_{n-2, 1-\alpha/2}$	$t_{\text{statistic}} \leq -t_{n-2, 1-\alpha}$	$t_{\text{statistic}} \geq t_{n-2, 1-\alpha}$	אזור דחייה
$2 * P(t_{n-2} > t_{\text{statistic}})$	$P(t_{n-2} > t_{\text{statistic}})$	$P(t_{n-2} > t_{\text{statistic}})$	P-VALUE

• שימו לב כי במודל של רגרסיה ליניארית פשוטה ערך ה- t סטטיסטי

שהתקבל שווה בדיוק לשורש של ערך F המחושב : $t = \sqrt{F}$
 $Pvalue = Pvalue$

רווח סמך לאמידת β :

$1 - \alpha = (\text{גבול תחתון} \leq \beta \leq \text{גבול עליון}) \cdot p$

$$\hat{\beta}_1 \pm t_{n-2, 1-\frac{\alpha}{2}} \cdot \text{s.e.}(\hat{\beta}_1)$$

מדד טיב ההתאמה R^2 :

מדד שנותן את פרופורציית השונות המוסברת. כמה מהשונות של Y מוסברת על ידי השונות של X :

$$0 \leq R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \leq 1 : (X \text{ מסביר את כל השונות של } Y)$$

מהשונות של Y).

נרצה פרופורציית שונות מוסברת קרובה ככל האפשר ל-1.

אחוז השונות המוסברת : $R^2 \cdot 100$.

רבי"ס שמטרתו לאמוד את תוחלת ערכי המשתנה התלוי (μ_0) עבור ערך מסוים של המשתנה ה"ב" (x_0). במילים אחרות אנו מתבקשים לאמוד את הניבוי באוכלוסייה עבור ערך מסוים של X .

האומד הנקודתי (הסטטיסטי) סביבו בנוי הרבי"ס הוא הניבוי במדגם עבור אותו

$$\text{ה-} X : \hat{\mu}_0 = \hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0$$

$$\text{נוסחת הרב"ס : } \hat{\mu}_0 \pm t_{n-2} \left(\frac{\alpha}{2} \right) \cdot \sqrt{MSRES \cdot \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SSX} \right)}$$

טעות התקן/גודל הרב"ס מושפעים מ-4 גורמים :

1. $MSRES$ - האומד לשונות הטעויות. ככל שגדל, טעות התקן/הרב"ס גדלים ולהפך.
2. n - גודל המדגם. ככל שגדל, טעות התקן/הרב"ס קטנים ולהפך.
3. SSX - מונה השונות של X (קשור לתופעת קיצוץ תחום). ככל שגדל, טעות התקן/הרב"ס קטנים ולהפך.
4. $(x_0 - \bar{x})$ - הסטייה של ערך X המסוים מהממוצע של X . ככל שגדלה טעות התקן/הרב"ס גדלים ולהפך.

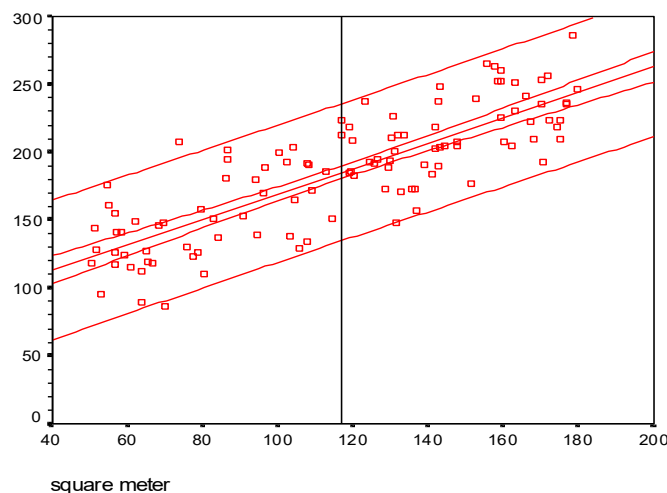
רב"ס לערכי Y עבור ערך מסוים של X :

רב"ס שמטרתו לאמוד את כל טווח ערכי Y (y_0) עבור ערך X מסוים (x_0).

$$\text{נוסחת הרב"ס : } \hat{\mu}_0 \pm t_{n-2} \left(\frac{\alpha}{2} \right) \cdot \sqrt{MSRES \cdot \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SSX} \right)}$$

ניתן לראות כי גם רב"ס זה בנוי סביב האומדן הנקודתי לתוחלת ערכי Y עבור ערך ה- X המסוים ($\hat{\mu}_0$).

ההבדל בין רב"ס לערכי Y לבין הרב"ס לתוחלת ערכי Y בא לידי ביטוי בטעות התקן. ניתן לראות כי טעות התקן של הרב"ס לערכי Y גדולה יותר מטעות התקן של הרב"ס לתוחלת ערכי Y . כאשר כל יתר הפרמטרים נשארים קבועים רב"ס זה יהיה רחב יותר מן הרב"ס לתוחלת. התרשים הבא מתאר רב"ס לתוחלת ולערך המשתנה התלוי וממחיש זאת בבירור :



שאלות:

קו הרגרסיה:

(1) מתווך דירות בתל אביב רצה לבדוק איך משפיע גודלה של דירה על המחיר שבו

היא נמכרת. הוא הניח 2 הנחות מקדימות:

1. רק גודל הדירה משפיע על מחיר הדירה באופן שיטתי. כל שאר הדברים המשפיעים על מחיר הדירה הם אקראיים ולא ניתנים לחיזוי.

2. ההשפעה של גודל הדירה על מחיר הדירה היא ליניארית.

גודל הדירה הינו X ומחיר הדירה הינו Y . מודל המתווך: $y_i = \alpha + \beta x_i + \varepsilon$.

המתווך אסף נתונים על 6 דירות, שנמכרו בחודש האחרון באותו אזור:

מספר הדירה	גודל הדירה במ"ר	מחיר הדירה באלפי דולרים
1	$X_1 = 70$	$Y_1 = 190$
2	$X_2 = 70$	$Y_2 = 210$
3	$X_3 = 80$	$Y_3 = 250$
4	$X_4 = 100$	$Y_4 = 290$
5	$X_5 = 120$	$Y_5 = 360$
6	$X_6 = 120$	$Y_6 = 380$

א. מקדם המתאם בין גודל הדירה למחיר הדירה. מה משמעותו?

ב. קו הרגרסיה לניבוי מחיר הדירה באמצעות גודל הדירה ופרשו את משמעות המקדמים.

ג. המחיר החזוי על פי קו הרגרסיה של דירה בגודל 100 מ"ר.

מבחן F:

(2) בצעו מבחן F לבדיקת הקשר בין גודל הדירה למחירה ברמת מובהקות של 1%.

מבחן t:

(3) בהמשך לדוגמא הנ"ל:

א. בצעו מבחן t למובהקות מקדם הרגרסיה ברמת מובהקות של 1%.

ב. בדקו את הטענה כי עליה במ"ר אחד תעלה את מחיר הדירה ביותר מ-\$2000.

ג. מהו ה-pvalue של מובהקות הקשר בין גודל הדירה למחירה. מה משמעותו?

קשר בין מבחן F למבחן t:

4) חשבו את סטטיסטי המבחן F על סמך סטטיסטי המבחן t שקיבלתם. מה ה-pvalue של מבחן F?

5) חשבו רב"ס לאמידת מקדם הרגרסיה ברמת סמך של 0.99. השוו עם תוצאות מבחן t.

6) חשבו את אחוז השונות המוסברת של מחיר הדירה על ידי גודלה.

רווח בר סמך לתוחלת:

7) השאלה מבוססת על נתוני דוגמא מס' 2 (ראו סרטון) והפלטים הבאים:

Case Summaries

	N	Mean	Std. Deviation	Minimum	Maximum
SIZE square meter	112	116.740	39.13942	50.46	179.76
PRICE thousands \$	112	185.0664	44.45345	86.20	286.56

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)					.000	60.979	91.015
SIZE square meter	.062	.005	.823	15.173	.000		

a. Dependent Variable: PRICE thousands \$

Descriptive Statistics

	Mean	Std. Deviation	N
SIZE square meter	116.740	39.139	112
PRICE thousands \$	185.066	44.453	112
PRE_1 Unstandardized Predicted Value	185.066	36.568	112
RES_1 Unstandardized Residual	.000	25.277	112

חשב רב"ס ברמת סמך של 95% לתוחלת מחיר הדירה כאשר שטח הדירה הוא 100 מ"ר.

רווח בר סמך לערכי נעלם:

- (8) חשב רב"ס ברמת ביטחון של 95% למחיר הדירה עבור שטח דירה של 100 מ"ר. מה ההבדל בין רב"ס זה לרב"ס הקודם?

תרגול מסכם:

- (9) בפיצוציית "שלמה המלך" חושדים כי מספר הלקוחות המבקרים בפיצוצייה תלוי במחיר המכירה של הבירה במקום. לשם בדיקת הנושא ערכו ניסוי בו בכל שבוע שינו את מחיר הבירה במקום ומנו את מספר הלקוחות שהגיעו במשך אותו שבוע. משך הניסוי 7 שבועות עוקבים. להלן נתוני הניסוי:

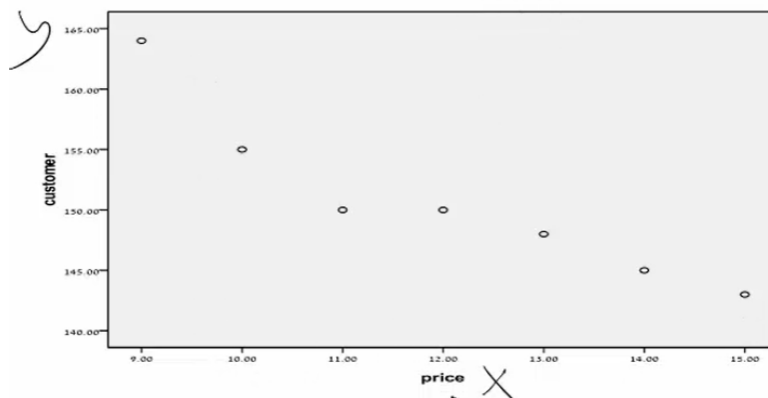
שבוע 7	שבוע 6	שבוע 5	שבוע 4	שבוע 3	שבוע 2	שבוע 1	
9	10	11	12	13	14	15	מחיר הבירה
164	155	150	150	148	145	143	כמות הלקוחות

- אמדו את מודל הרגרסיה ע"י חישוב מקדמי הרגרסיה.
- חשבו את מקדם המתאם r_{xy} .
- אמדו את השונות של שאריות המודל.
- בצעו בדיקה גראפית של אקראיות השאריות.
- חשבו את אחוז השונות המוסברת. מה משמעותה?
- בצעו חיזוי לכמות הלקוחות אם מחיר הבירה יהיה 16 ₪. האם להערכתכם ניתן להסתמך על חיזוי זה?
- בצעו מבחן F לבדיקה האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות.
- בצעו מבחן t לבדיקה האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות המבקרים בפיצוצייה ברמת מובהקות 5%. השוו את התוצאות.
- אמדו את מקדם הרגרסיה ברמת סמך של 0.95. השוו את התוצאה עם הסעיף הקודם.
- כתבו דו"ח קצר על הממצאים.

קריאת פלטים של SPSS:

10) על סמך הנתונים של השאלה הקודמת התקבלו הפלטים הבאים:

דיאגרמת הפיזור (scatter plot):



סטטיסטיקה תיאורית (descriptive statistics):

Descriptive Statistics			
	Mean	Std. Deviation	N
customer	150.7143	7.01699	7
Price	12.0000	2.16025	7

פלט מקדם המתאם (correlations):

Correlations			
		customer	Price
Pearson Correlation	customer	1.000	-.935
	price	-.935	1.000
Sig. (1-tailed)	customer	.	.001
	price	.001	.
N	customer	7	7
	price	7	7

פלט model summary :

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.935 ^a	.873	.848	2.73470

a. Predictors: (Constant), price

b. Dependent Variable: customer

פלט ניתוח שונות (ANOVA) :

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	258.036	1	258.036	34.503	.002 ^a
	Residual	37.393	5	7.479		
	Total	295.429	6			

a. Predictors: (Constant), price

b. Dependent Variable: customer

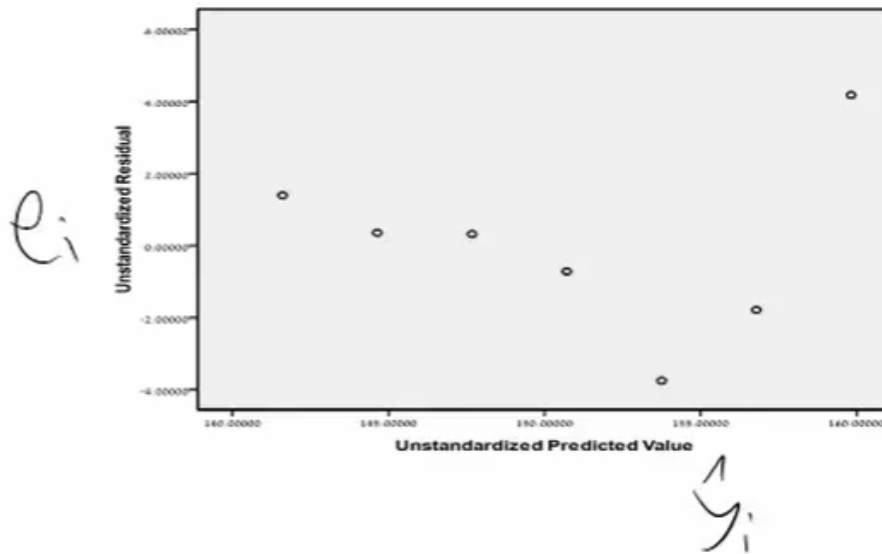
פלט מקדמי הרגרסיה (coeffitients) :

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	187.143	6.287		29.765	.000
	Price	-3.036	.517	-.935	-5.874	.002

a. Dependent Variable: customer

גרף ניתוח שאריות:



על סמך הפלטים הנתונים :

- א. מהו מודל הרגרסיה שנאמד?
- ב. מהו מקדם המתאם r_{xy} ?
- ג. מהי השונות של שארית המודל?
- ד. האם נמצא דפוס מיוחד בשאריות?
- ה. מהו אחוז השונות המוסברת?
- ו. על פי מבחן F : האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות המבקרים בפיצוצייה ברמת מובהקות 5%?
- ז. על פי מבחן t : האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות המבקרים בפיצוצייה ברמת מובהקות 5%? השוו את התוצאות.
- ח. מה ה-pvalue של המבחנים הסטטיסטיים? מה משמעותו?
- ט. בדקו האם קיים קשר חיובי מובהק בין המשתנים ברמת מובהקות 5%?

תשובות סופיות:

- (1) א. $r = 0.987$. ב. $\hat{Y}_t = -27.32 + 3.29X_t$. ג. 301.68 אלף דולר.
- (2) $F_{st} = 21.198$, יש עדות לקשר מובהק.
- (3) א. $t_{stt} = 4.604$. ב. יש עדות לכך. ג. $pvalue < 0.001$.
- (4) $t_{stt}^2 = 147$, $pvalue < 0.001$.
- (5) $p(2.061 \leq \beta \leq 4.519) = 0.99$.
- (6) 97.4% .
- (7) $p(163.889 \leq \mu_{100} \leq 174.24) = 0.95$.
- (8) $p(119.036 \leq Y_{100} \leq 219.09) = 0.95$, רחב יותר.
- (9) א. $\hat{y}_i = 187.143 - 3.0357x_i$. ב. $r = -0.93457$. ג. $\hat{\sigma}^2 = 7.4785$.
 ד. ראו סרטון. ה. $R^2 = 0.873$. ו. $\hat{y} = 138.5714$, כן.
 ז. $F_{stt} = 34.5 > F_c 0.05(1,5) = 6.6$, יש עדות לכך. ח. $t_{st} = -5.87395$.
 ט. $p(-4.36 \leq \beta \leq -1.709) = 0.95$. י. ראו סרטון.
- (10) א. $\hat{y}_i = 187.143 - 3.036x_i$. ב. $r_{xy} = 0.935$. ג. $MSE = 7.479$.
 ד. לא. ה. $R^2 = 0.874$. ו. $F = 34.503$.
 ז. $t = -5.874$. ח. $pvalue = 0.002$. ט. ראו סרטון.

סטטיסטיקה יישומית

פרק 8 - רגרסיה מרובה

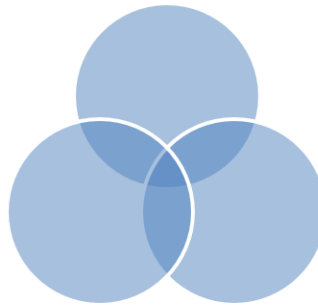
תוכן העניינים

1. כללי 84

הגרסיה מרובה:

רקע:

ניבוי המשתנה התלוי באמצעות יותר ממשתנה ב"ת אחד.
 המודל באוכלוסייה: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$.
 מקדמי מודל הרגרסיה המרובה:
 α = חותך אחד שמשמעותו: הציון המנובא כאשר כל המשתנים הב"ת = 0.
 $\beta_1 \dots \beta_j$ = מקדמי השיפוע. מס' הבטות = למספר המשתנים הב"ת במודל.
 משמעות מקדם השיפוע β_j : ההשפעה הייחודית של המשתנה הב"ת המסוים לניבוי המשתנה התלוי, בניכוי השפעתם של כל יתר המשתנים הב"ת האחרים המצויים במשוואת הרגרסיה.



אמידת מודל הרגרסיה המרובה:

ברגרסיה מרובה, כמו ברגרסיה פשוטה, שיטת האמידה הטובה ביותר היא שיטת הריבועים הפחותים. כלומר, נרצה להביא את סכום הטעויות בניבוי למינימום.
 מפתרון פונקציית הריבועים הפחותים נקבל את אומדי הרגרסיה: $\hat{\alpha}, \hat{\beta}_1 \dots \hat{\beta}_j$.

מבחני מובהקות:

1. מבחן F למובהקות הרגרסיה:
 בדיקה האם קיים קשר ליניארי בין המשתנה התלוי Y לבין לפחות אחד מהמשתנים המסבירים.

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

ההשערות הן: $H_1: \text{Not } H_0 = \text{at least one of the } \beta\text{'s is not } 0$

טבלת ניתוח שונות (ANOVA):

מקור	סכום ריבועים SS	דרגות חופש $d.f.$	ממוצע סכום ריבועים $MS = \frac{SS}{d.f.}$	$F_{st} \sim F_{k,n-k-1}$
מודל הגרסיה	SSR	K	$MSR = \frac{SSR}{K}$	$F_{st} = \frac{MSR}{MSE}$
שאריות	SSE	$n-k-1$	$MSE = \frac{SSE}{(n-k-1)}$	
סה"כ	TSS	$n-1$		

סטטיסטי המבחן: $F_{st} = \frac{MSR}{MSE}$

כלל הכרעה: נדחה את H_0 אם: $F_{st} \geq F_{k,n-k-1}^{1-\alpha}$

חישוב סכומי הריבועים:

$$TSS = \sum_{i=1}^n y_i^2 - n\bar{y}^2$$

$$SSR = R^2 \cdot TSS$$

$$SSE = (1 - R^2)TSS$$

פרופורציית השונות המוסברת R^2 :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

בגרסיה מרובה אומד זה לפרופורציית השונות המוסברת הוא בעייתי שכן הוא מושפע ממספר המשתנים הב"ת במודל. אומד זה יכול רק לגדול בהוספת משתנים ב"ת למודל ולכן לא ייתן לנו אינדיקציה האם כדאי היה לי להוסיף אותם למודל או לא.

האומד המתוקן לפרופורציית השונות המוסברת $AdjR^2$:

$$\bar{R}^2 = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

בניגוד ל- R^2 לוקח בחשבון את מספר המשתנים הב"ת במודל. יכול שלא לגדול ואף לקטון בהוספת משתנה ב"ת שלא תורם תרומה משמעותית לניבוי.

2. מבחן t למובהקות משתנה ב"ת יחיד :

$$\begin{aligned} H_0 : \beta_j &= 0 \\ H_1 : & \text{else} \end{aligned}$$

השערות :

סטטיסטי המבחן וכלל הכרעת השערת האפס :

$$|t_{\hat{\beta}_j}| = \left| \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}} \right| > t_{\left(T-k-1; 1-\frac{\alpha}{2}\right)}$$

$$\hat{\beta}_j \pm t_{n-k-1, 1-\frac{\alpha}{2}} \cdot s.e.(\hat{\beta}_j) : \beta_j \text{ לא מידת ה-}$$

שאלות:

- 1) לצורך בדיקת ההשערה שקיים קשר בין מספר המוניות בעיר באר שבע (y) לבין מספר התושבים בעיר באלפים (x_1) ומספר הרכבים הפרטיים באלפים (x_2). הוחלט לבנות מודל רגרסיה מהצורה: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, על סמך הנתונים הבאים: $\sum y_i = 1673$, $\sum y_i^2 = 338,657$, $MSE = 119.789$.
- א. ע"ס הנתונים הנ"ל, השלימו את טבלת ניתוח השונות הבאה.
- איזו השערה ניתן לבדוק באמצעותה? כתוב את ההשערה ובחן אותה.

SOURCE	SS	DF	MS	F
Regression				
Error				
Total		8		

- ב. חשבו את מדד טיב ההתאמה. הסבר את משמעותו.
- ג. נתונה טבלת המקדמים (החלקית) הבאה :

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-511.727	114.9476				
X 1	9.208785		3.732167			
X 2	-8.79921	4.420456				

- i. רשמו את האומדן למשוואת הרגרסיה ופרשו את מקדמיה.
- ii. בחנו את ההשערה כי קיים קשר בין מספר הרכבים הפרטיים לבין מספר המוניות ברמת מובהקות של 5%.
- iii. בנו רווח סמך למקדם של מספר התושבים בעיר

- iv. ענה ללא חישוב (על סמך הסעיפים הקודמים) – האם קיים קשר בין מספר התושבים לבין מספר המוניות ברמת מובהקות 5%?
- v. מהי תחזית מס' המוניות בבאר שבע עבור 100,000 תושבים ו-52,000 מכוניות פרטיות?
- vi. האם ניתן לסמוך על תחזית זאת?

תרגול מסכם:

- (2) מעוניינים למצוא קשר בין מחיר הדירה (ב-\$) לבין ארבעה משתנים מסבירים: (1) שטח הדירה ו-(2) גודל שטח האמבטיה (ב-Sqft) וכן (3) מרחק הדירה מהים ו-(4) מהאוניברסיטה (במיילים). לשם כך נדגמו מספר דירות והריצו רגרסיה אשר בה המשתנה המוסבר הוא מחיר הדירה. להלן פלט הרגרסיה שהתקבל:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.952 ^a			

a. Predictors: (Constant), Sea_Dist, Apartment, Bath

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression					.000 ^a
	Residual					
	Total	1940484.615	25			

a. Predictors: (Constant), Univ_Dist, Bath, Sea_Dist, Apartment

b. Dependent Variable: Price

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-265.514	146.673		-1.810	.085
	Apartment		.449	.722	6.572	
	Bath	4.256		.297	2.687	.014
	Sea_Dist	-32.114	11.090	-.223		.009
	Univ_Dist	11.746	9.439	.095	1.244	.227

a. Dependent Variable: Price

ענו על הסעיפים הבאים :

- א. מלאו את התאים החסרים בטבלה (אם לא ניתן למלא את כל התאים החסרים באופן מלא נמקו באופן מפורש מדוע לא ניתן).
- ב. כתבו את האומדן למשוואת מחיר הדירה בצורה מפורשת על סמך הפלט הנ"ל. פרשו את מקדמי הרגרסיה.
- ג. בדקו האם ארבעת הגורמים ביחד אכן מסבירים את מחיר הדירה. הסבירו את המסקנה שהגעתם אליה. השתמשו ברמת מובהקות 5%.
- ד. הסבירו מהו ערך ה-Pvalue ומה ניתן להסיק ממנו לגבי המשתנים המסבירים?
- ה. בנו רווח סמך למקדם גודל שטח האמבטיה. השתמשו ברמת מובהקות של 2%.
- ו. ברמת מובהקות של 5% יש לבדוק האם המרחק מהאוניברסיטה אכן משפיע על מחיר הדירה.
- ז. האם במודל הרגרסיה הנוכחי ניתן לוותר על גורם המרחק מהים? השתמשו ברמת מובהקות 1%.
- ח. בדקו את ההשערה כי קיים קשר חיובי בין גודל הדירה למחירה ברמת מובהקות של 5%.

תשובות סופיות:

(1) א. השערה: $H_0 : \beta_1 = \beta_2 = 0$
 $H_1 : \text{at least one of the } \beta\text{'s is not } 0$

SOURCE	SS	DF	MS	F
Regression	26945.784	2	13472.892	112.414
Error	718.733	6	119.789	
Total	27664.517	8		

- ב. 97.4% .i. $\hat{y}_i = -511.727 + 9.208x_{1i} - 8.799x_{2i}$
- ii. אין עדות לכך. .iii. $p(3.17 \leq \beta_1 \leq 15.24) = 0.95$
- iv. כן. .v. 321 .vi. כן.
- (2) א. ראו סרטון. ב. $\hat{y}_i = -256.514 + 2.95x_{1i} + 4.256x_{2i} - 32.114x_{3i} + 11.746x_{4i}$
- ג. לפחות אחד מהמשתנים הב"ת שונה מאפס באוכלוסייה.
- ד. ראו סרטון. ה. $p(1.016 \leq \beta_2 \leq 7.496) = 0.98$
- ו. לא. ז. לא. ח. יש עדות לכך.

סטטיסטיקה יישומית

פרק 9 - רגרסיה - שאלות ממבחנים

תוכן העניינים

89	1. מבחן 1
93	2. מבחן 2

מבחן 1:

שאלות:

1) להלן תוצאות הרצת רגרסיה של Y בתלות ב-X עבור 10 תצפיות (חלק מהנתונים הושמטו בכוונה מהפלט, אך ניתנים לחישוב על ידך).

$$\sum (X_i - \bar{X})^2 = 1475.6 \quad \text{נתון כי:}$$

מקור	סכום ריבועים SS
רגרסיה	SSR = 2148.6
שאריות	SSE = ?
סה"כ	SST = ?

משתנה	מקדם	טעות תקן	ערך סטטיסטי (מתוקנן)	p-value מובהקות
	b_i	S_{b_i}	t	
קבוע (חותך)	-24.7	11.3	?	?
X	1.20	?	10.5	

א. מהו SST?

i. לא ניתן לקבוע.

ii. 1994.42

iii. 2304.1

iv. 1629.78

ב. האם הרגרסיה מובהקת? בדוק לפי p value.

i. הרגרסיה מובהקת.

ii. הרגרסיה אינה מובהקת.

(2) לפניך פלט רגרסיה פשוטה (ממנו הושמטו נתונים שבאפשרותך להשלים), המתאר את ציון המבחן כפונקציה של מספר התרגילים שהגיש הסטודנט במהלך הסמסטר, ידוע כי כל הנחות המודל תקפות.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.842105598
R Square	0.709141839
Adjusted R Square	0.688366256
Standard Error	5.315523758
Observations	16

ANOVA

	Df	SS	MS	F	Significance F
Regression	1	964.4329004	964.4329004	34.13343	4.28E-05
Residual	14	395.5670996	28.25479283		
Total	15	1360			

	Coefficients	Standard Error	t Stat	P-value
Intercept	50.99134199	4.318918368	11.80650747	1.15E-08
מספר התרגילים	4.086580087	0.699471573	5.842381939	4.28E-05

א. ע"פ הנתונים, אחוז השונות של ציוני המבחן המוסברת ע"י מספר התרגילים שהגיש הסטודנט, היא _____. אם נוסף משתנים נוספים, אחוז השונות המוסברת _____, ו-Adjusted R _____.

i. 84%, יגדל, לא ניתן לקבוע ללא נתונים נוספים.

ii. 84%, יקטן, יגדל.

iii. 70.9%, יגדל, לא ניתן לקבוע ללא נתונים נוספים.

iv. 70.9%, יגדל, יקטן.

v. 68.8%, יגדל, יגדל.

vi. 68.8%, יקטן, יקטן.

ב. מהו הרב"ס של שיפוע הרגרסיה β_1 ? (בר"מ של 1%).

i. $2.58 < \beta < 5.58$

ii. $2 < \beta < 6.17$

iii. $1.74 < \beta < 6.88$

iv. $2.86 < \beta < 5.3$

במטרה לנבא בצורה טובה יותר את הצלחת הסטודנטים בבחינה, החליט החוקר להוסיף 2 משתנים נוספים לניתוח הרגרסיה. מספר השיעורים בהם נכח הסטודנט, ומספר השעות שלמד לבחינה. לפניכם הפלט החסר:

SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.880163577			
R Square	0.774687922			
Adjusted R Square	0.718359902			
Standard Error	5.053253294			
Observations	16			
ANOVA				
	df	SS	MS	F
Regression	3	?	351.1918579	13.75315
Residual	12	306.4244263	25.53536886	
Total	15	?		

	Coefficients	Standard Error	t Stat	P-value
Intercept	37.08959571	8.726136945	4.250402663	0.001127
מספר השיעורים	2.610000755	1.429904588	1.82529714	0.092932
מספר התרגילים	3.198068014	1.239162836	2.58082951	0.024061
שעות לימוד לבחינה	-0.108373802	1.021202927	-0.106123669	0.917238

ג. בדוק את ההשערה כי הרגרסיה מובהקת לכל אחד מהמשתנים המסבירים בר"מ של 1%.

- i. הרגרסיה אינה מובהקת לכל המשתנים שנבדקו.
- ii. לא ניתן לקבוע מהנתונים האם הרגרסיה מובהקת.
- iii. הרגרסיה מובהקת למשתנה מספר התרגילים, אך אינה מובהקת למשתנים מספר השיעורים ומספר שעות הלימוד לבחינה.
- iv. הרגרסיה מובהקת למשתנה מספר התרגילים ומספר השיעורים, אך אינה מובהקת למשתנה שעות הלימוד לבחינה.

ד. מהו SSR של הרגרסיה המרובה?

- i. $SSR = 964.4$
- ii. $SSR = 1053.57$
- iii. $SSR = 694.57$
- iv. $SSR = 853.57$
- v. לא ניתן לחשב את SSR מהנתונים שהתקבלו.

תשובות סופיות:

- (1) א. iii. ב. i.
(2) א. iii. ב. ii. ג. i. ד. ii.

מבחן 2:

שאלות:

- (1) הועלתה השערה שהוצאות האחזקה של מערכת לעיבוד נתונים קשורות למספר שעות השימוש השבועיות במערכת. להלן תוצאות חלקיות של פלט EXCEL של ניתוח רגרסיה בין Y הוצאות אחזקה שנתיות (במאות \$) ו-X מספר שעות השימוש השבועיות.

SUMMARY OUTPUT

Regression Statistics					
Multiple R					
R Square					
Adjusted R Square					
Standard Error					
Observations 10					
ANOVA					
	df	SS	MS	F	Significance F
Regression		860.051			0.00012
Residual					
Total		1004.525			
	Coefficients	Standard Error	t Stat	P-value	
Intercept	10.528	3.745		0.023	
שעות שימוש	0.953		6.901		

א. לאור התוצאות ה-p-value בבדיקת ההשערה: $H_0: \beta_1 \leq 0.5$.
 $H_1: \beta_1 > 0.5$

- i. $P_v < 0.005$
- ii. $0.005 < P_v < 0.01$
- iii. $0.01 < p_v < 0.025$
- iv. $0.025 < p_v < 0.05$
- v. $P_v > 0.05$

ב. אם היינו מריצים רגרסיה בה Y מספר שעות השימוש השבועיות ואילו המשתנה המסביר X , הוצאות אחזקה שנתיות (במאות \$), אזי השיפוע של קו הרגרסיה יהיה:

- i. 0.953
- ii. 1.049
- iii. 10.528
- iv. 0.095
- v. 0.898

2) הועלתה השערה שמספר התקלות ברכב Y קשורה לגיל הנהג X . לשם כך נלקח מדגם של 10 נהגים.

כמו כן חושבו הסכומים הבאים:

$$\sum X_i^2 = 14,227, \sum X_i = 363, \sum Y_i = 13, \sum Y_i^2 = 29, \sum X_i Y_i = 366$$

$$\hat{Y} = 4.96 - 0.1X_i \quad \text{על ידי:}$$

א. ערכו של מקדם המתאם הליניארי בין מספר התקלות לבין גיל הנהג הוא:

- i. -10.59
- ii. -0.9395
- iii. 0.8826
- iv. -0.1

החוקר לא היה מרוצה מעוצמת הקשר ולכן החליט להוסיף לרגרסיה את המשתנים הבאים: מספר הק"מ שהמכונית נסעה (באלפי ק"מ) וסוג הרכב. במדגם נכללו 2 סוגי רכבים: A ו-B כאשר סוג B קודד כערך 0 וסוג רכב A קודד כערך 1. להלן פלט הרגרסיה המרובה. שימו לב כי חלק מהערכים בפלט חסרים:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	
R Square	
Adjusted R Square	
Standard Error	0.204047
Observations	10

ANOVA				
	df	SS	MS	F
Regression	3			0.00002
Residual	6	0.249811		
Total	9			

	Coefficients	Standard Error	t Stat	P-value
Intercept	1.407943	0.71094		
X1 גיל הנהג	-0.03094	0.014611		
X2 סוג הרכב	0.239574	0.152994		
X3 ק"מ באלפים	0.027666	0.005329		

- ב. להלן מספר טענות לגבי מובהקות הרגרסיה ומובהקות המשתנה סוג הרכב ברמת מובהקות 0.1 :
- הרגרסיה מובהקת והמשתנה סוג הרכב מובהק ברמת מובהקות 0.1.
 - הרגרסיה מובהקת, אך המשתנה סוג הרכב אינו מובהק ברמת מובהקות 0.1.
 - הרגרסיה אינה מובהקת, אך המשתנה סוג הרכב מובהק ברמת מובהקות 0.1.
 - הרגרסיה והמשתנה סוג הרכב אינם מובהקים ברמת מובהקות 0.1.
- ג. SST בפלט הרגרסיה המרובה שווה ל :
- 12.1
 - 16.0
 - 20.0
 - אין מספיק נתונים לחשבו.
- ד. לאור התוצאות, רב"ס למקדם המשתנה מספר הקילומטרים, בר"מ 5% :
- (0.015, 0.04)
 - (0.017, 0.038)
 - (0.02, 0.035)
 - אין מספיק נתונים לחשבו.

ה. אם היינו מקודדים את סוג רכב B כערך 1 וסוג רכב A קודד בערך 0 משוואת הרגרסיה הייתה:

i. נשאר ללא שינוי.

$$\hat{Y} = 1.4079 - 0.0309X_{1i} - 0.2395X_{2i} + 0.0276X_{3i} \quad \text{ii.}$$

$$\hat{Y} = 1.6475 - 0.0309X_{1i} - 0.2395X_{2i} + 0.0276X_{3i} \quad \text{iii.}$$

iv. לא ניתן לדעת ללא הרצה מחדש.

ו. החוקר רצה להוסיף משתנה מסביר נוסף X_4 מספר השנים שחלפו מאז קבלת רישיון הנהיגה. להלן פלט הרגרסיה (חלק מהנתונים חסרים) עם 4 המשתנים המסבירים:

	Coefficients	Standard Error	t Stat	P-value
Intercept	1.5	0.71		
X1 גיל הנהג	5.1	14.1		
X2 סוג הרכב	0.25	0.2		
X3 ק"מ באלפים	0.02	0.008		
מס' השנים שחלפו מאז	-5.13094			
X4 קבלת רישיון		0.8		

על סמך נתונים אלו:

- חשש סביר למולטיקוליניאריות.
- התחזית הנקודתית של מספר התקלות תהיה דומה לזו של הרגרסיה (הקודמת) עם 3 המשתנים המסבירים.
- קיימת קורלציה גבוהה בין חלק מהמשתנים המסבירים.
- כל התשובות נכונות.

3 במפעל מסוים הורץ מודל של רגרסיה ליניארית פשוטה על 8 עובדים כאשר Y תפוקת העובד ו- X גיל העובד. נמצא שהחלק המוסבר על ידי הרגרסיה הוא 74. הטעויות שהתקבלו מופיעות בחלקן בטבלה שלהלן:

e_8	e_7	e_6	e_5	e_4	e_3	e_2	e_1
0	2	-2	2	?	2	-3	0

אחת מהטענות שלהלן נכונה:

- מקדם ההסבר בין X ל- Y הוא 0.74.
- לא ניתן לחשב את מקדם המתאם המרובה.
- $SSR = 18$.
- $SSE = 74$.
- אף אחת מהטענות איננה נכונה.

תשובות סופיות:

- (1) א. ii. ב. v.
(2) א. ii. ב. i. ג. i. ד. i. ה. ii. ו. iv.
(3) א'.