מבוא לסטטיסטיקה והסתברות / ד"ר דוד סרג

בדיקות אי פרמטרית

בדיקות פרמטריות מבוצעות תחת הנחה של התפלגות נורמלית ועל משתנים כמותיים (סולמות יחסיים ומרווחיים). בשונה מאינדיקטורים לבדיקה בשונה התפלגות לא ידועה או עבור משתנים לא כמותיים (דירוגים ושמיים) משתמשים בשיטות אי-פרמטריות. לדוגמה ניתן לפתור בעיה גם בשיטה פרמטרית (דו-צדדית) וגם בשיטה אי-פרמטרית ונבחנים את הבעיה הבאה:

כדי לקבוע אם מטבע הוגן (סיכוי שווה לכל אחד מהצדדים), מטילים אותה 40 פעמים, מתקבלים 25 ראשים ו 15 כתבות. האם ברמת בטחון 95% היא הוגנת?

בשיטה פרמטרית כן בעיה של פרופורציה $A = z\sqrt{\frac{pq}{n}}$ כלומר בעיה דו-צדדית

$$A = 1.96\sqrt{\frac{0.5 \cdot 0.5}{40}} = 0.155$$

0.625

0.345 0.50 0.655

אנו קבלנו $\frac{25}{40} = 0.625$ ראשים, כלומר אנו בתחום השגיאה הסטטיסטי ונקבל את הטענה שהמטבע הוגן (H_0).

נקח בעיה אחרת: מטילים קובייה 120 פעמים ומקבלים תוצאות

	1	2	3	4	5	6
נצפה	15	22	25	18	20	20
צפוי	20	20	20	20	20	20

האם הקובייה הוגנת ברמת סה"ר (מובהקות 5%)

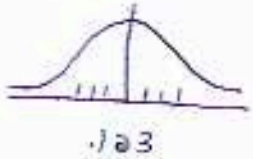
כאן אין באפשרותנו להשתמש בשיטה פרמטרית (פרופורציה), אבל נוכל להתיחס לפער בין הצפוי לנצפה. ברור שאם הפערים גדולים היא אינה הוגנת ואם קטנים היא "יותר הוגנת". נסכם את ריבועי הסטיות בין צפוי לנצפה ביחס לצפוי (מדוייק)

ערך מחושב $\chi^2_{calc} = \sum \frac{(\text{צפוי} - \text{נצפה})^2}{\text{צפוי}} = \frac{5^2}{20} + \frac{2^2}{20} + \frac{5^2}{20} + \frac{2^2}{20} + 0 = 2.9$

את הערך 2.9 נשווה לערך תיאורטי מטבלה שנקראת χ^2 המבוססת על המשפט הבא:

משפט: רבוע של משתנה נורמלי מתפלג לפי χ^2 עם דרגת חופש אחת.
סכום רבועים של n משתנים נורמליים בלתי תלויים מתפלג לפי χ^2 עם n דרגות חופש.

מדוע המשתנים שלנו הם נורמליים? $\frac{(\text{בפועל}-\text{בציפיה})^2}{\text{בפועל}}$



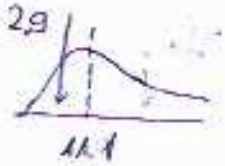
הערכים הנצפים מתפלגים סביב הצפוי כך שברוב המקרים יהיו קרובים לצפוי, וכל שנתרחק מהצפוי ההסתברות לכך תקטן וזה מצביע התנהגות נורמלית.

השמוש בטבלה הוא לפי $\chi^2(5, 0.95)$ הפרמטר הראשון הוא 5 דרגות חופש כי יש לנו בטבלה 5 מחלקות בלתי תלויות והשישית הוא המשלים שלהם ל 120 (לכן הוא תלוי). כמובן שכמו בטבלת סטודנט נתנו רק ערכים מקובלים כמו 0.975, 0.9, 0.95 וכו'

H_0: קרבה הוגנת

H_1: קרבה לא הוגנת

$\chi^2(5, 0.95) = 11.1 > 2.9$



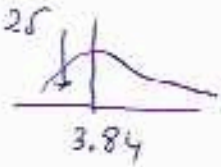
במילות הסטיות 2.9 קטנה מהערך התיאורטי, ולכן הקרבה עדיין הוגנת.

נפתור את בעיית המטבע בשיטה אי פרמטרית

	כאן	צד	
נצפה	25	15	
בפועל	20	20	

$\chi^2_{calc} = \frac{5^2}{20} + \frac{5^2}{20} = 2.5$

$\chi^2(1, 0.95) = 3.84$



לכן נקבל H_0 והמטבע הוגנת.

2 המקרים של המטבע והקרבה הם מקרים של התפלגות אחידה.

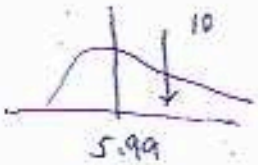
נפתור בעיה בלילית ולתר: תוקר טוען כי בכלאה של פרח אדום עם פרח לבן יתן 50% ורוד ו 25% לבן ואדום. נקח מידגם של 100 צמחים פרחים לבנים ואדומים ונתקבלו התוצאות הבאות:

הזם טרנת התוקר נכונה ברמת בטחון 95%

	אדום	לבן	ורוד
נצפה	35	15	60
בפועל	25	25	50

$$\chi^2_{calc} = \frac{10^2}{25} + \frac{10^2}{25} + \frac{10^2}{50} = 10$$

$$\chi^2(2, 0.95) = 5.99 < 10$$



H_0:

H_1: טענת החוקר

במלות פסיטיות גדולה ולכן הטענה איננה נכונה

אפשר לצפות כי H_1 מתיחס לרלבנטי שהתלבטה הנסיונית בצורה של התפלגות χ^2.

למעשה מבחן χ^2, מאפשר לבדוק לגבי התפלגות לא ידועה אם היא מתאימה להתפלגות שאינן חושבים שהיא כאן. תמיד H_0 מאשרת את ההתפלגות שאינן חושבים שהיא, ו H_1 שוללת אותה. נפתור עתה סידרת בדיקות להתפלגויות שונות.

<u>התפלגות פואסונית</u>: רוצים לבדוק אם התפלגות מספר העברות על נהגים חדשים מתפלג פואסונית עם ממוצע 2 עדות בשנה הראשונה. להלן טבלת נתונים

מס עבירות	0	1	2	3	4
מס נהגים נצפה	10	25	30	20	15
הסתברות	.135	.27			
מס נהגים צפוי	13.5	27			

האם ברמת בטחון 95% ההתפלגות פואסונית עם ממוצע 2?

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$P(0) = \frac{e^{-2}\cdot 2^0}{0!} = e^{-2} = 0.135$$

$$P(1) = \frac{e^{-2}\cdot 2^1}{1!} = 2e^{-2} = 0.27$$

$$P(2) =$$

להשלים את הטבלה, ולקבוע האם ההתפלגות פואסונית עם ממוצע 2.

$$\chi^2(4, 0.95) = 9.49$$

להשלמת הטבלה יש לשים לב להערות הבאות:

א. מספר הנהגים הצפוי הוא 100 מוכפלת בהסתברות המתושבת

ב. סך כל הנהגים בערכים הצפויים צריכים לא יגיע ל 100, שכן התפלגות פואסונית היא אינסופית בתיאוריה יכולה להיות יותר מ 4 עבירות, לכן המשבצת האחרונה צריכה להיות 4 ויותר עבירות הצפוי. כדי שתהיה התאמה בין צפוי לנצפה מחוננת כמות הנהגים יש להשלים את המשבצת האחרונה בצפוי ל 100

ג. מה יקרה אם השאלה תביע האם ההתפלגות פואסונית (ולא לגבי הממוצע) יש לחשב אותו מתוך הנתונים (ממוצע משוקלל

$$\bar{X} = \lambda = \frac{0\cdot 10 + 1\cdot 25 + 2\cdot 30 + 3\cdot 20 + 4\cdot 15}{100} = \frac{205}{100} = 2.05$$

ואז לחשב את הצפוי, לפי $\lambda = 2.05$. למעשה אין לחשב ממוצע עבור קבוצה פתוחה (הקבוצה +4 פתוחה), אלא שאפשר לומר כי 4 מייצג את הקבוצה שכן בהתפלגות פואסונית בכל שמתרחקים מן הממוצע ההסתברות קטנה ומספר עבורנו גדול הוא כנ"ל.

ב"פיצוי" על חישוב הפרמטר ממוצע מתוך המידע מקטינים את מספר דרגות החופש לכן χ^2 מבוטלה יהיה $\chi^2(4-1, 0.95)$

על כל פרמטר שמחוסר מהמידע יש להוריד דרגת חופש. למשל אם נרצה לבדוק התפלגות נורמלית ונחשב גם ממוצע וגם סטיית תקן, נצטרך להוריד 2 דרגות חופש.

<u>התפלגות בינומית:</u> חוקר טוען שהגיע "גולן" בארפרוקה אחוז הבנים הנולדות הוא 60%. כדי לבדוק את נכונות מידע בקרב 200 משפחות בנות 4 ילדים ונתקבלו התוצאות הבאות:

התפלגות בינומית

מס בנים	0	1	2	3	4
נצפה	20	33	90	39	18
הסתברות לצפוי	.0256	.1536			
	5.12	30.72			

$\underline{p=0.6}$ $\binom{n}{k} p^k (1-p)^{n-k}$

$P(0) = \binom{4}{0} \cdot 0.6^0\, 0.4^4 = 0.0256$

$P(1) = \binom{4}{1} \cdot 0.6^1\, 0.4^3 = 0.1536$

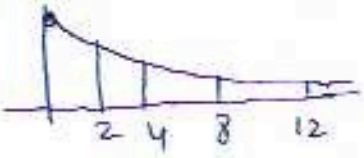
להשלים את הטבלה, לחשב $\chi^2(4, 0.95)$, χ^2_{calc} ולבדוק אם הטענה נכונה. שימו לב שכל הבדיקות נעשה ברמת בטחון 95%, אבל באדון ניתן לעשות כאלה ברמות בטחון שונות.

התפלגות מעריכית: להלן בדיקת אורך חיים של 200 מכשירי טלוויזיה בשנים.

אורך חיים	0-2	2-4	4-8	8-12
נצפה	70	50	45	35
שטח מצטבר		0.67	0.45	0.21
שטח התחום	0.33	0.22	0.24	0.21
צפוי	66	44	48	42

} תשובה דרך

בדוק את ההתפלגות היא מעריכית ברמת מובהקות 5% $\lambda = 5$



$e^{-\frac{8}{5}} = 0.21$

$e^{-\frac{4}{5}} = 0.45$

$e^{-\frac{2}{5}} = 0.67$

ההתפלגות המעריכית היא אינפלציה ולכן הקבוצה האחרונה היא בעלת התיישנות מ 8 ומעלה ולא רק מ 8-12 לפיכך נחשב את השטחים (רבוי פרק "הפצה") ונרבוע כל שטח ב 200 כדי למצוא מספר מכשירים בכל תחום

$$\chi^2_{calc} = \sum \frac{(\text{צפוי} - \text{נצפה})^2}{\text{צפוי}} = \frac{(70-66)^2}{66} + \frac{(50-44)^2}{44} + \frac{(45-48)^2}{48} + \frac{(35-42)^2}{42}$$

$$= 2.41$$

$\chi^2(3, 95\%) = 7.81$ טבלה

$2.41 < 7.81$

מקבלים את H_0

H_0: מעריכית 5 מאוזן

H_1: לא כן

אם הין שאלות האם ההתפלגות מעריכית ובמאוזן לא ניתן, יש לאמוד אותו מהנתונים ע"י: מאוזן משקלל כאשר אמצע הקבוצה מוכפל את הקבוצה.

$$\bar{X} = \frac{1\cdot 70 + 3\cdot 50 + 6\cdot 45 + 10\cdot 35}{200} = \frac{830}{200} = 4.15$$

ואנו מתחשבים מתדר שפי מאוזן 4.15 ובוקים.

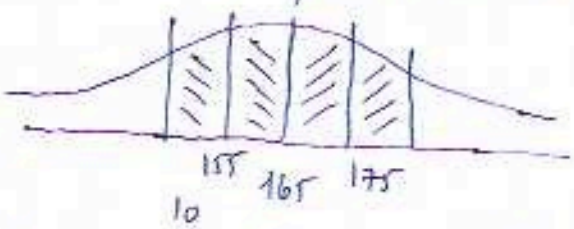
התפלגות נורמלית: יש לאמוד ע"י 2 פרמטרים, תוחלת וסטיית תקן.

תרגיל: כדי לבדוק את הטענה שהתפלגות הגבהים באוכלוסיה מסוימת מתפלגת נורמלית עם מאוזן 165 וסטיית תקן 10 נלקח ממנה מדגם מקרי של 1000 ונתקבלו התוצאות הבאות, מצאו

מעל 195	185-195	175-185	165-175	155-165	145-155	135-145	פחות מ 135	גובה
1	16	130	335	340	150	25	3	שכיחות נצפית
1.3	21.5	135.9		341.3	135.9	21.5	1.3	שכיחות צפויה

שימו לב שהשכיחות הצפויה סימטרית מאחר שהיא מבוססת על התפלגות נורמלית

במובן שחושב בדפוס מבוסס על חשבי הסתברות מתוך "הפסמון" שבין הממוצע 165 וסטיית התקן 10



הסתברות שלא כל תחום מוכפלת במובן ב 1000. נניח מובהקות 10%

$$\chi^2_{calc} = \sum \frac{(\text{נצפה} - \text{תיאורטי})^2}{\text{תיאורטי}}$$, טבלה $\chi^2(7, 90\%) = 12.9$

ומתוך החישוב ניתן לקבוד אם תתקבל H_0 או H_1

בחישוב נרמלנו ממוצע 165 סטית תקן 10 H_0: לעומת H_1:

$N(165, 10)$

אם הממוצע או סטיית התקן לא נתונים יש לחשב אותם מהמדגם ולהוריד דרגת חופש נוספת ואם שניהם לא נתונים מורידים עוד 2 דרגות חופש.

תרגיל נוסף: תקלה המדינה בוחן 400 מלוכלך ותוכנן כך שהבעיות הרבעי של הגיע תהיה הגדולה בפולה מאשר הרבעונים האחרים. לפלן ההגלה בפועל

רבעון	I	II	III	IV
הגלה בפועל	220	250	180	350
הגלה צפויה (לפי תכנון)	200	200	200	400

האם התקלה מוגבר לפי המתוכנן ברמת בטחון 90% יש לחשב את הבדל:

$$\chi^2_{calc} = \frac{20^2}{200} + \frac{50^2}{200} + \frac{20^2}{200} + \frac{50^2}{400}$$

$$= 2 + 12.5 + \cdots$$

$\chi^2(3, 90\%)$ טבלה $= 6.25$

ברור ש $\chi^2_{calc} > \chi^2$ טבלה

לכן לא הייתה עמידה בתקלה לפי המתוכנן.

בדיקת תלות/אי תלות

עד עכשיו פיתחנו למעשה בשיטות אי פרמטרית בדיקת של טוב התאמה, כאשר הנתונים הנסיוניים מתאימים להתפלגות מסוימת.

ניתן בטכניקה זו למבחן אם קיימת תלות בין משתנים, למשל, כאשר חיסון לשפעת מקטין את התחלואה?

תרגיל: כדי לבדוק אם קיימת תלות בין מתן חיסון לתחלואה נלקח המדגם הבא: 200 שחוסנו מהם חלו 50, 300 לא חוסנו מהם חלו 150.

	חולה	בריא	סה"כ
מחוסנים	50 (60)	150 (140)	200
לא מחוסנים	100 (90)	200 (210)	300
סה"כ	150	350	500

האם ברמת הסתברות 95% יש תלות בין החיסון לתחלואה.

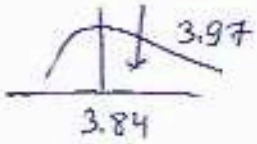
אם אין קשר בין החיסון לתחלואה אזי חייב יחס החולים ב 2 הקבוצות להיות כמו יחס הבריאים. כיוון שכך מדגם אינו כמודן סטיה מסוימת נחפש אם הסטיה בתוצאה בשאלה אי לא. ובכן יש לחשב את הכפלי עבור כל משבצת. נחשב אותו עבור המשבצת השמאלית העליונה (מחוסן, בריא) $140 = \frac{350 \cdot 200}{500}$. בהסתכלות שמתוך 350 בריאים 2/5 מחוסנים. כלומר 140 אנשים. טכניקת החישוב נעשה בשיטת "הכלבה" במובנה תמיד יעמוד סה"כ המודגם (מחוסנים ולא מחוסנים) ובמונה תבוא "הכלבה" = מכפלה של סה"כ הבריאים עם סה"כ המחוסנים. נרשום את הכפלי בטבלה בקבה המשבצת המתאימה, אי שנבנה טבלה נפרדת לכפלי. במהרה שלנו יש רק דרגת חופש אחת כי ברגע שמלאנו משבצת אחת כל היתר מתמלאות בהתאם לסה"כ הערכים. ניתן לבדוק בקלות אי חישוב הכפלי בכל משבצת בנפרד ולראות שהם כמומים מתאימים (בדוק זאת!)

$$\chi^2_{calc} = \frac{(150-140)^2}{140} + \frac{10^2}{60} + \frac{10^2}{210} + \frac{10^2}{90} = 3.97$$

H_0: אין תלות

H_1: יש תלות.

$$\chi^2(1, 0.95) = 3.84$$



אנו נקבל את H_1 כלומר יש תלות. עצמת התלות מחושבת לפי נוסחת קרמר

$$V = \sqrt{\frac{\chi^2_{calc}}{n \cdot \min(c-1, r-1)}} = \sqrt{\frac{3.97}{500 \cdot 1}} = 0.09$$

r = מספר שורות

c = מספר עמודות

$\min(c-1, r-1)$

הנמוך מבין $c-1$ ל $r-1$

ניתן להוכיח שרמת התלות מקבלת ערכים $0 \le V \le 1$

כאשר $V=1$ משמעותו תלות מלאה, $V=0$ אין תלות כלל. במקרה שלנו יש תלות חלשה ואמנם במבחן הסטיות מהבדלי קטנה למדי, 3.97 לעומת 3.84

תרגיל: בנה מטריצה שבה $V=1$, ומטריצה שבה $V=0$

	בריא	חולה	
מחוסן			200
לא מחוסן			300
			500

מה מספר דרגות החופש עבור מטריצה מסדר 9×10?

תשובה: 72 דרגות חופש.

מה מספר דרגות החופש במטריצה 3×2? תשובה 2.

מספר דרגות החופש עבור מטריצה $r\times c$ היא $(r-1)*(c-1)$

תרגיל נוסף: להלן טבלת מתגיסים ליחידות בצבא לפי מבצרים.

מבצר	חרדי	דרוזי	פריפריה	סה"כ
נח"ל	22 / 60	33 / 10	/ 40	110
משמר הגבול	/ 10	42 / 70	/ 60	140
גולני	30 / 10	/ 40	/ 100	150
	80	120	200	400

$$(\text{חרדי, נח"ל}) = \frac{80\cdot110}{400} = 22$$

$$(\text{דרוזי, נח"ל}) = \frac{120\cdot110}{400} = 33$$

$$(\text{חרדי, גולני}) = \frac{80\cdot150}{400} = 30$$

$$(\text{דרוזי, משמר הגבול}) = \frac{120\cdot140}{400} = 42$$

השלם את הטבלה

ומדוע הסכום יש תלות בין המבצר ליחידה בצבא.