מבחן בסטטיסטיקה להנדסת תוכנה

פתרון למועד X

סמסטר חורף, 2021

עמית שטקל

שאלה א (40 נק')

בסקר הצבעה לבחירות הקרובות ביקשו הסוקרים לבדוק האם קיימת תלות בין רמת ההשכלה של הבוחרים להעדפות הפוליטיות שלהם. בפרט, רצו לבדוק האם אחוזי ההצבעה מתפלגים באופן שונה בקרב קבוצות ההשכלה.

הנתונים שנמדדו סוכמו בטבלת השכיחויות הבאה.

כאשר X מתאר את רמת השכלה (1: יסודי, 2: תיכון, 3: אקדמאי)

ו-Y מתאר את המפלגה הפוליטית (A: מפלגה א', B: מפלגה ב', C: מפלגה ג').

X	Y		
Observed	A	B	C
1	100	200	300
2	100	100	100
3	80	10	10

1. איזה מבחן מתאים לבדיקת ההשערה?

מבחן χ2-Goodness-of-Fit Test

2. אילו תנאים מקדימים צריכים להתקיים על מנת שנוכל להשתמש במבחן זה?

לפחות ב 80% מהתאים מספר התצפיות **הצפוי** גדול מ-5

3. בהנחה שהתנאים מתקיימים, נסחו השערה סטטיסטית מתאימה לבדיקת ההשערה.

נניח ש-i מייצג את קבוצת ההשכלה *i=1,2,3*

$H_0: P_A^i = P_A^{expected}, P_B^i = P_B^{expected}, P_C^i = P_C^{expected}$

$H_1: otherwise$

4. מה מספר דרגות חופש המתאים למבחן?

df = (3-1)*(3-1)=4

5. חשבו את האומדן הסטטיסטי המתאים למדגם

נחשב את השכיחויות הצפויות לפי המשוואה: $\widehat{m}_{ij} = \frac{\sum n_i * \sum n_j}{N}$

X	Y			Totals
Observed	A	B	C	
1	100	200	300	600
2	100	100	100	300
3	80	10	10	100
Totals	280	310	410	1000

expected	A	B	C
1	168=600*280/1000	186	246
2	84	93	123
3	28	31	41

נחשב את האומדנים הסטטיסטיים התאיים (פר קבוצה/מפלגה)

X^2	A	B	C
1	27.52=(100-168)^2/168	1.05	11.85
2	3.05	0.53	4.30
3	96.57	14.23	23.44

נסכום את האומדנים התאיים לחישוב האומדן הסטטיסטי הכולל:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} t(x,y) = \frac{(n_{ij}-\widehat{m}_{ij})^2}{\widehat{m}_{ij}} = 182.54$$

6. נתונים הערכים הקריטיים הבאים:

$$\chi^2_{(df,0.975)} = 11.14, \quad \chi^2_{(df,0.95)} = 9.49, \quad \chi^2_{(df,0.90)} = 7.779$$

נסחו מסקנה סטטיסטית לגבי ההשערה הנבדקת ברמת מובהקות של 5% בהתבסס על האומדן שחישבתם בשאלה 5. (השתמשו באומדן הקריטי המתאים לביסוס טענתכם, לא לשכוח לכתוב מסקנה מילולית).

מכיוון ש- 182.54 > 9.49 = X^2(0.95, 4) אנו דוחים את השערת האפס וטוענים כי לא מתקיים שוויון התפלגויות בין קבוצות ההשכלה השונות. או במילים אחרות, קיימות העדפות שונות למפלגות שונות בין קבוצות ההשכלה השונות.

שאלה ב (40 נק)

במחקר בדקו את הקשר בין הרגלי עישון לבין ריכוז הרטינול בפלסמת הדם.

לצורך העניין חילקו את הנבדקים לארבע קבוצות: כאלו שלא עישנו מעולם, גמולים מעישון, מעשנים קל ומעשנים כבד. רמות הרטינול בדם של ארבע הקבוצות מסוכמות בטבלה הבאה:

הרגלי עישון		n	ממוצע רטינול בפלסמת הדם (ng/ml)	סטיית התקן של רמת הפלסמה בדם (ng/ml)
A	לא עישנו מעולם	25	811.4	274.5
B	גמולים מעישון	30	619.5	244.5
C	מעשנים קל	13	525.6	143.7
D	מעשנים כבד	12	404.3	221.2
סה"כ		80	631.9	

1. באיזה מבחן נשתמש על מנת לבדוק האם קיים שוני בין רמות הרטינול בדם של הקבוצות השונות?

נסחו השערה סטטיסטית לבדיקת ההשערה.

מבחן אנאליזת שונויות (Analysis of variance)

$H_0: \mu_A = \mu_B = \mu_C = \mu_D$

$H_A: otherwise$

2. מהן ההנחות שצריכות להתקיים על מנת שנוכל לבצע מבחן זה?

- שוויון שונויות בין קבוצות המדגם
- נורמליות של התפלגות המדד המשותפת לכל הקבוצות

3. מה מספר דרגות החופש לאומדן הסטטיסטי במבחן?

df1 = k-1 = 4-1 = 3

df2 = n-k = 80-4 = 76

4. הדגימו את חישוב ה-MSB. שימו לב כי ממוצע הרטינול הכללי בדם נתון בטבלה המופיעה למעלה.

$$MSB =$$

$$\frac{25*(811.4-631.9)^2+30*(619.5-631.9)^2+13*(525.6-631.9)^2+12*(404.3-631.9)^2}{3}$$

$$=526212$$

5. מה מייצג ה-MSB? (במשפט אחד, כולל פרוש הקיצור)

ה-MSB הוא ה-Mean Squared Error Between Groups, מייצג את השונות הבין קבוצתית הממוצעת, או את מרחק ממוצעי הקבוצות מהממוצע הכללי.

6. נתון כי ה- **SSE** שווה ל- 4328269, וה-**SSB** שווה ל- 1578354.

חשבו את האומדן הסטטיסטי F (מספיקה ספרת דיוק אחת אחרי הנקודה)

$$MSB=\frac{1578354}{4-1}=526118,\qquad MSE=\frac{4328269}{80-4}=56951$$

$$F(3,76)=\frac{MSB}{MSE}=\frac{526118}{56951}=9.2$$

7. אם היינו מוציאים את הקבוצה "לא עישנו מעולם" מי מהאומדנים היה משתנה יותר ה-MSE או ה-MSB? הסברו מבלי לחשב מחדש את טבלת ה-ANOVA.

ה-MSB היה משתנה יותר, כי אנו מניחים שוויון שונויות, ולכן ה-MSE לא אמור להיות מושפע דרמטית משינוי גודל המדגם.

לעומתו, ה-MSB משקף את השונות הבין קבוצתית, ולקבוצה זאת היו רמות רטינול הגבוהות ביותר, ולכן סביר שהפרש הממוצעים מהממוצע הכללי ישתנה, הוא ה-MSB.

8. באיזה כוון ישתנה המדד המושפע מהוצאת קבוצת "לא עשנו מעולם"?

ה-MSB ירד, מכיוון שזאת הקבוצה שהממוצע שלה הוא הרחוק ביותר מהממוצע הכללי.

9. נתונים F סטטיסטי הבאים

F(0.95,df1,df2)=2.725

F(0.975,df1,df2)= 3.293

F(0.05,df1,df2)= 0.117

האם ניתן לדחות את השערת האפס? נסחו את המסקנה הסטטיסטית בהתבסס על האומדן הקריטי המתאים לרמת מובהקות 0.05.

ניתן להסיק כי קיים שוני בין רמות הרטינול של הקבוצות השונות ברמת מובהקות של 0.05 מכיוון ש

$F_{3,76}=9.2>F_{0.95,3,76}=2.725$

שאלה ג (20 נק')

במדגם הרגלי העישון שצויין בשאלה הקודמת רצו להשוות בין זוגות של קבוצות הניסוי השונות באמצעות התפלגות studentized (התפלגות Tukey).

נתון כי $q^*(p,df1,df2) = 3.7148$. כמו כן, נתון $MSE = 56951$.

1. מהן מספר דרגות החופש לאומדן הסטטיסטי הקריטי q^* במקרה זה?

df1=k=4, df2=n-k=76

2. מהו ערך p המתאים לבדיקה ברמת מובהקות α=0.05?

0.95

3. פרטו את חישוב האומדן הסטטיסטי המתאים לבחינת השאלה.

$$q_{B-C} = \frac{|619.5 - 525.6|}{\sqrt{0.5 * 56951 * \left(\frac{1}{30} + \frac{1}{13}\right)}} = 1.6756$$

4. האם קיים הבדל בין רמות הרטינול בדם של קבוצת ה"גמולים מעישון" לקבוצת ה"מעשנים קל"?

הסיקו מסקנה סטטיסטית התואמת ל-q_{B-C} המחושב

$1.6756<3.7148 \rightarrow q_{B-C}<q_{0.95,4,76}$

ניתן להסיק כי אין הבדל משמעותי סטטיסטית בין רמות הרטינול הממוצעות בדם של קבוצת ה"גמולים מעישון" לקבוצת ה"מעשנים קל".