תשובות למועד X
מבחן בסטטיסטיקה להנדסת תוכנה

סמסטר אביב, 2020
הדס לפיד

א. בניסוי בדקו את רמות הגופרית בצמחים מגידול אורגני לעומת צמחים מגידול לא אורגני.
נתגלו רמות הגופרית הבאות לקילוגרם יבול (נתונים פיקטיביים)

גודל המדגם	שונות מדגמית	ממוצע המדגם (mg/kg)	
32	65000	2456	**אורגני**
48	68000	2870	**לא אורגני**

1. באיזה מבחן תשתמשו בכדי לבדוק האם רמות הגופרית בגידולים הלא אורגניים גבוהות מרמות הגופרית בגידולים האורגניים?

2. מהם התנאים שצריכים להתקיים בכדי שתוכלו להשתמש במבחן שהצעתם?

3. כיצד הייתם בודקים את קיום התנאים המקדימים למבחן?

4. בהנחה שהתנאים הדרושים מתקיימים, נסחו את ההשערה הסטטיסטית לבדיקת השאלה.

5. מה מספר דרגות החופש המתאים למבחן?

6. חשבו את האומדן הסטטיסטי של המבחן.

7. נתון כי $t_{0.975,df} = 1.991, t_{0.95,df} = 1.664$

נסחו את המסקנה, האם ניתן לקבל את הטענה כי רמות הגופרית בגידולים הלא אורגניים גבוהות מרמות הגופרית בגידולים האורגניים ברמת מובהקות של 5%? (השתמשו באומדן הקריטי המתאים לביסוס טענתכם)

תשובות לא':

1. מבחן t לא מצומד חד צדדי.

2. בכדי להשתמש במבחן זה יש להניח שוויון שונויות באוכלוסיות הנדגמות והתפלגות נורמליות משותפת של שני המדגמים.

3. בדיקת שוויון שונויות תתבצע על ידי מבחן לוין.

בדיקת נורמליות משותפת בודקים על ידי מבחן שפירו-ווילק על אוסף הנתונים של שני המדגמים יחדיו, המוסטים להתפלגות סביב ה-0 ע"י חיסור הממוצע המדגמי מכל אחת מהדגימות.

4. $H_0: \mu_{organic} = \mu_{non-organic}$

$H_A: \mu_{non-organic} > \mu_{organic}$

5. מספר דרגות החופש למבחן הוא 78 = 2-48+32

6. נחשב תחילה את סטיית התקן המשותפת של המדגמים:

$$S^2 = \frac{\sum_{i=1}^{2}(\boldsymbol{n_i} - \mathbf{1}) \cdot \boldsymbol{S_i^2}}{\sum_{i=1}^{2}(\boldsymbol{n_i} - \mathbf{1})} = \frac{31 \cdot 65000 + 47 \cdot 68000}{31 + 47} = 66807$$

$$S = \sqrt{66807} = 258.4718$$

ומכאן נחשב את האומדן הסטטיסטי:

$$t = \frac{|\overline{\boldsymbol{Y_1}} - \overline{\boldsymbol{Y_2}}|}{\boldsymbol{s} \cdot \sqrt{\left(\frac{\mathbf{1}}{\boldsymbol{n_1}} + \frac{\mathbf{1}}{\boldsymbol{n_2}}\right)}} = \frac{|2456 - 2870|}{258.472\sqrt{\left(\frac{1}{32} + \frac{1}{48}\right)}} = 7.02$$

7. ניתן לראות כי $t>t_{(0.95,df)}$ → 7.02>1.664

ולכן ניתן להסיק כי רמות הגופרית בגידולים הלא אורגניים גבוהות מרמות הגופרית בגידולים האורגניים ברמת בטחון של 95%.

ב. נתון כי סף רמת הגופרית המותר בצמח הוא 2800 (mg/kg).

באותו ניסוי, התפלגות הגידולים ביחס לסף רמות גופרית במדגם האורגני לעומת המדגם הלא אורגני היתה כדלקמן:

	אורגני	לא אורגני	סה"כ
מעל לסף <2800 (mg/kg)	2	12	14
מתחת לסף >2800 (mg/kg)	30	36	66
סה"כ	32	48	80

1. נסחו השערה סטטיסטית למציאת הבדל בין התפלגויות הגידולים האורגניים והלא אורגניים.

2. מה מספר דרגות החופש במבחן χ^2?

3. איזה תנאי צריך להתקיים בשביל לבצע מבחן זה?

4. בהנחה שהתנאי מתקיים, חשבו את האומדן הסטטיסטי לבדיקת ההשערות שניסחתם ב-1.

5. נתון: $\chi^2_{(df,0.975)} = 5.02, \quad \chi^2_{(df,0.95)} = 3.84, \quad \chi^2_{(df,0.90)} = 2.71$
הסיקו מסקנה סטטיסטית בנוגע להשערה שניסחתם ב-1 ברמת מובהקות של 5%. השתמשו באומדן הקריטי המתאים לביסוס ההשערה.

6. חשבו את ה-Odd-Ratio בהתבסס על נתוני המדגם.

7. מה ניתן להסיק ממדד זה לגבי ההבדלים בהתפלגות הרעילות בין הגידולים האורגניים והגידולים הלא אורגניים?

תשובות לשאלה ב:

1. $H_0: P^{\text{אורגני}}_{\text{מעל}} = P^{\text{לאאורגני}}_{\text{מעל}} = P^{expected}_{\text{מעל}}, P^{\text{אורגני}}_{\text{מתחת}} = P^{\text{לאאורגני}}_{\text{מתחת}} = P^{expected}_{\text{מתחת}}$

$H_A: otherwise$

או ורסיה חלקית של ההשערה (רק עבור קטגוריה אחת)

2. מספר דרגות החופש במבחן הוא: (2-1)*(2-1)=1

3. התנאי שצריך להתקיים על מנת לבצע את המבחן הוא שמספר התצפיות הצפויות ב-80% מהתאים יהיה גדול מ-5.

4. טבלת התדירויות הצפויות והאומדנים הסטטיסטים פר תא:

		אורגני	לא אורגני
מעל לסף	Expected counts	32*14/80=5.6	48*14/80=8.4
	cell χ^2	$(2-5.6)^2/5.6=2.314$	$(12-8.4)^2/8.4=1.543$
מתחת לסף	Expected counts	32*66/80=26.4	48*66/80= 39.6
	cell χ^2	$(30-26.4)^2/26.4=0.491$	$(36-39.6)^2/39.6=0.327$

$$\chi^2 = 2.314 + 1.543 + 0.491 + 0.327 = 4.675$$

5. 4.675>3.84 ולכן נדחה את השערת האפס ונאמר כי התפלגות הגידולים שונה בין הגידולים האורגנים והגידולים הלא אורגניים.

6. נחשב את טבלת ההסתברויות הנצפית:

	אורגני	לא אורגני
מעל לסף	2/32=0.0625	12/48=0.25
מתחת לסף	30/32=0.9375	36/48=0.75

כעת נחשב את ה-Odd-Ratio:

$$OR = \frac{0.0625 * 0.75}{0.25 * 0.9375} = 0.2$$

7. מכיוון שה-OR<0.5 ניתן להסיק כי אחוז חוצי הסף הרעיל בקרב הגידולים האורגנים נמוך משמעותית מאחוז חוצי הסף הרעיל בקרב הגידולים הלא אורגניים.

ג. בבדיקת הקשר בין הזמן שסטודנטים מבלים ברשתות החברתיות לבין ממוצע הציונים שלהם, סוכמו הנתונים באופן הבא:

זמן ברשתות החברתיות (שעות ביום)	גודל המדגם (n_i)	ממוצע ציונים	
		ממוצע ($\overline{Y_i}$)	סטיית התקן (s_i)
A: 1 or less	12	92.04	7.32
B: 2-3	68	85.14	9.48
C: 4 and more	23	78.43	9.33
All	103	84.45	

דווח גם כי הסכום המשוקלל של ריבועי ההפרשים בין ממוצע המדגמים לממוצע הכללי (ה- $MS_{between\ groups}$) היה 779.4, וכי השונות התוך קבוצתית (ה-SSE) היתה 8528.

1. נסחו השערת אפס והשערת אלטרנטיבית לבדיקת שוויון תוחלות הציונים בין קבוצות זמן החשיפה.

2. סכמו את הנתונים בטבלת ANOVA. לצורך כך, השלימו את המדדים החסרים (ה-$SS_{between}$, וה-MS_{error}) וחשבו את האומדן הסטטיסטי, F.

Source	df	SS	MS	F
Between groups	2		779.4	
Error	100	8528		

3. מה משמעות מדד ה-MS_{error}?

4. נתון כי F(0.95,df1=2,df2=100)= 3.087. נסחו את מסקנתכם בנוגע להשערה ברמת מובהקות של 5%, בהתבסס על האומדן הסטטיסטי שחישבתם בשאלה 2.

5. בהשוואות contrast בין תוחלות כל זוגות הקבוצות האפשריים, נמצאו רמות המובהקות הבאות.

Groups	**Pv**
C-A	0.000218
B-A	0.018919
C-B	0.006511

נתחו את ההבדלים לפי שיטת Binyamini-Hochberg (BH) עבור רמת מובהקות של 5% .

6. סכמו את הממצאים במילים.

7. איזה קריטריון אנו מגבילים לפי שיטת BH?

8. מה משמעות הקריטריון הזה?

9. לאיזה גודל הוא מוגבל כאשר מבצעים תיקון BH?

תשובות לשאלה ג:

1. $H_0: \mu_A = \mu_B = \mu_C$
   $H_A: otherwise$
2.

Source	df	SS	MS	F
Between groups	2	779.4*2=1559	779.4	779.4/85.3 = 9.14
Error	100	8528	8528/100=85.3	

3. MS_{error} משמש כאומדן לשונות הקבוצתית הממוצעת, הוא משמש אותנו במכנה בחישוב האומדן הסטטיסטי ב- F test. הוא אומד את השונות הממוצעת של כל המדגמים, תחת הנחת שוויון שונויות.

4. במבחן אנאליזת השונויות הזה 9.14>3.087, ז"א שהאומדן הסטטיסטי גדול מהאומדן הקריטי ולכן ניתן להסיק כי יש הבדלים בין תוחלות הציונים של הקבוצות השונות.

5.

Index	**Groups**	**Pv**	**BH Correction** $\frac{\alpha \cdot i}{m}$	**Hypothesis assertion**
1	C-A	0.000218	0.0167	*
2	C-B	0.006511	0.0333	*
3	B-A	0.018919	0.0500	*

כאשר i מתייחס לאינדקס ההשוואה בסדר עולה של ערכי ה-Pv שהתקבלו בהשוואות בין הקבוצות, ו-m מתייחס למספר ההשוואות.

6. לפי תיקון BH, נתגלו הבדלים משמעותיים סטטיסטית בין כל זוגות הקבוצות.

7. BH בא להגביל את ה-False Discovery Rate (FDR)

8. משמעו ה-FDR הוא אחוז ההשוואות בהן נטען H_A, בעוד H_0 הוא הנכון (אחוז ה-False positives) מתוך סך ההשוואות (m).

9. ע"י BH אנו מגבילים את ה-FDR לרמת המובהקות, α.