תרגול 2 - ניתוח שונות חד כיווני, ניתוח שגיאות

מודל ניתוח שונות חד כיווני

נרצה לבחון השפעה של פקטור מסוים. מספר הרמות של הפקטור יכול להיות סופי או אינסופי.

המודל נחלק לשניים:

מודל פרמטרי – מספר הרמות של הפקטור סופי, כל הרמות נמדדות.

מודל אקראי – מספר רב של רמות, נדגמות רק חלק מהרמות. (בהמשך הסמסטר)

המודל הפרמטרי

עבור כל רמה של פקטור נדגום מספר תצפיות:

a - מספר הרמות של הגורם המסביר. (לעיתים מופיע גם ע"י הסימון k)

n_i - מספר התצפיות שנדגמו מרמה i של הגורם המסביר.

y_{ij} – התצפית ה-j ברמה ה-i.

$N=\sum_{i=1}^{a} n_i$ - סך התצפיות שנדגמו.

$\bar{y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ ממוצע התצפיות של רמה i (כשבכל רמה נלקחות n_i דגימות).

$\bar{y}=\frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}=\frac{1}{N}\sum_{i=1}^{a} n_i\bar{y}_i$ ממוצע כל התצפיות

המודל:

μ - תוחלת התצפיות

τ_i - הסטייה של רמה i מהתוחלת של כלל התצפיות

הסטיות מוגדרות כך ש: $\sum_{i=1}^{a} n_i\cdot\tau_i=0$, סכום הסטיות המשוקלל על כל הרמות חייב להתאפס.

$\varepsilon_{ij}\sim N(0,\sigma^2)$ - סטייה אקראית ("רעש") של התצפית ה-j ברמה i של הגורם המסביר.

$$\boxed{y_{ij}=\mu+\tau_i+\varepsilon_{ij} \quad \text{תצפית}}$$

המודל מתבסס על ההנחה כי התפלגות התצפיות בתוך כל רמה הינה נורמלית: $y_{ij}\sim N(\mu_i,\sigma^2)$

$\mu_i=\mu+\tau_i$ - התוחלת של רמה i מורכבת מתוחלת כללית ועוד Offset.

התפלגות ממוצע הסטיות האקראיות ברמה i: $\bar{\varepsilon}_i\sim N(0,\frac{\sigma^2}{n_i})$

התפלגות הממוצע הכולל של הסטיות האקראיות: $\bar{\varepsilon}\sim N(0,\frac{\sigma^2}{N})$

ממוצע רמה i: $\bar{y}_i=\frac{1}{n_i}\cdot\sum_{j=1}^{n_i}(\mu+\tau_i+\varepsilon_{ij})=\mu+\tau_i+\frac{1}{n_i}\cdot\sum_{j=1}^{n_i}\varepsilon_{ij}=\mu+\tau_i+\bar{\varepsilon}_i=\mu_i+\bar{\varepsilon}_i$

ממוצע כולל:

$$\bar{y}=\frac{1}{N}\sum_{i=1}^{a} n_i\bar{y}_i \underset{\bar{y}_i=\mu+\tau_i+\bar{\varepsilon}_i}{\Big\downarrow}= \frac{1}{N}\left[N\mu+\sum_{i=1}^{a} n_i\tau_i+\sum_{i=1}^{a} n_i\bar{\varepsilon}_i\right] \underset{\sum_{i=1}^{a} n_i\tau_i=0}{\Big\downarrow}= \mu+\frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n_i}\varepsilon_{ij}=\mu+\bar{\varepsilon}$$

נוסחאות לשונות

השונות הכוללת (נשווה כל תצפית לממוצע הכללי): $SS_{Tot}=\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2$

ניתן לפרק את השונות הכוללת לסכום הבא: $SS_{Tot}=\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2+\sum_{i=1}^{a}n_i(\bar{y}_i-\bar{y})^2$

כאשר: $SS_e=\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2$ סכום ריבועי השגיאות הנובעות מהרעש (השגיאה האקראית),

ו-$SS_{Tr}=\sum_{i=1}^{a}n_i(\bar{y}_i-\bar{y})^2$ סכום ריבועי ההפרשים הנובעים מרמת הפקטור (השונות בין הקבוצות)

ובסה"כ: $SS_{Tot}=SS_e+SS_{Tr}$

דרגות חופש:

$$df(SS_{Tot})=N-1 \qquad df(SS_{Tr})=a-1 \qquad df(SS_e)=a(n-1)=N-a$$

אמדים לשונות:(מנרמלים כל ביטוי לפי מס' דרגות החופש)

$$MS_{Tot}=\frac{SS_{Tot}}{N-1}$$

$$MS_{Tr}=\frac{SS_{Tt}}{a-1}$$

$$MS_e=\frac{SS_e}{N-a}$$

נוסחאות מקוצרות לחישוב השונות:

גודל מדגם זהה בכל רמה (n)	**גדלי מדגם שונים בכל רמה (n_i)**
$SS_{Tr}=n\sum_{i=1}^{a}\bar{y}_i^{\,2}-N\bar{y}^2$	$SS_{Tr}=\sum_{i=1}^{a}n_i\bar{y}_i^{\,2}-N\bar{y}^2$
$SS_e=\sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2-n\sum_{i=1}^{a}\bar{y}_i^{\,2}$	$SS_e=\sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2-\sum_{i=1}^{a}n_i\bar{y}_i^{\,2}$
$SS_{Tot}=\sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2-N\bar{y}^2$	$SS_{Tot}=\sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2-N\bar{y}^2$

תוחלות אמדי השונות

$$E(SS_{Tr})=E[\sum_{i=1}^{a}n_i(\bar{y}_i-\bar{y})^2]=E[\sum_{i=1}^{a}n_i(\underbrace{\tau_i}_{a}+\underbrace{\bar{\varepsilon}_i-\bar{\varepsilon}}_{b})^2]=$$

$$\bar{y}_i=\mu+\tau_i+\bar{\varepsilon}_i \qquad \bar{y}=\mu+\bar{\varepsilon} \qquad (\ a\ +\ b\ )^2$$

$$E[\sum_{i=1}^{a}n_i\left(\tau_i^{\,2}+\underline{2\tau_i(\bar{\varepsilon}_i-\bar{\varepsilon})}_{*}-2\bar{\varepsilon}\bar{\varepsilon}_i+\bar{\varepsilon}_i^{\,2}+\bar{\varepsilon}^2\right)]=E[\sum_{i=1}^{a}n_i\left(\tau_i^{\,2}-\underline{2\bar{\varepsilon}\bar{\varepsilon}_i}_{**}+\bar{\varepsilon}_i^{\,2}+\bar{\varepsilon}^2\right)]$$

$$*\ 2\cdot E[\sum_{i=1}^{a}n_i\tau_i(\bar{\varepsilon}_i-\bar{\varepsilon})]=2\cdot[\sum_{i=1}^{a}n_i\tau_iE(\bar{\varepsilon}_i-\bar{\varepsilon})]=2\cdot\sum_{i=1}^{a}(n_i\tau_i\cdot 0)=0$$

$$**\ -2\sum_{i=1}^{a}n_i\bar{\varepsilon}\bar{\varepsilon}_i=-2\cdot\bar{\varepsilon}\sum_{i=1}^{a}n_i\bar{\varepsilon}_i=-2\cdot\bar{\varepsilon}\sum_{i=1}^{a}\sum_{j=1}^{n_i}\varepsilon_{ij}=-2\cdot\bar{\varepsilon}N\bar{\varepsilon}=-2N\bar{\varepsilon}^2$$

$$E(SS_{Tr})=\sum_{i=1}^{a}n_i\tau_i^2+\sum_{i=1}^{a}n_iE\left(\bar{\varepsilon}_i^{\,2}\right)-NE(\bar{\varepsilon}^2)$$

ידוע כי:

$$V(x)=E(x^2)-E^2(x) \quad\rightarrow\quad E(x^2)=V(x)+E^2(x)$$

$$E(\bar{\varepsilon}^2)=V(\bar{\varepsilon})+\underbrace{E^2(\bar{\varepsilon})}_{\text{תוחלת הרעש היא 0}} \quad\rightarrow\quad E(\bar{\varepsilon}^2)=V(\bar{\varepsilon})=\frac{\sigma^2}{N}$$

התוחלת של $E\left(\bar{\varepsilon}_i^{\,2}\right)$ תחושב באופן דומה ותהיה שווה ל- $\frac{\sigma^2}{n_i}$. לכן נקבל:

$$E(SS_{Tr})=\sum_{i=1}^{a}n_i\tau_i^2+\sum_{i=1}^{a}\frac{n_i\sigma^2}{n_i}-\frac{N\sigma^2}{N}=\sum_{i=1}^{a}n_i\tau_i^2+(a-1)\sigma^2$$

$$E(MS_{Tr})=\frac{E(SS_{Tr})}{a-1}=\frac{\sum_{i=1}^{a}n_i\tau_i^2}{a-1}+\sigma^2$$

וכאשר $\forall i=1...k \quad n_i=n$:

$$E(MS_{Tr})=\frac{E(SS_{Tr})}{a-1}=n\frac{\sum_{i=1}^{a}\tau_i^2}{a-1}+\sigma^2$$

ובאותו האופן ניתן לחשב עבור השונות בתוך הקבוצות:

$$E(SS_e)=(N-a)\sigma^2$$

$$E(MS_e)=\frac{E(SS_e)}{(N-a)}=\sigma^2$$

מבחן ההשערות:

השערת האפס היא כי כל התוחלות שווה, כלומר $\tau_i=0, \forall i$, ולכן: $E(MS_{Tr})=E(MS_e)=\sigma^2$.

$$\begin{cases} H_o:\tau_1=\tau_2=\cdots=0 \\ H_1: else \end{cases} \Rightarrow \begin{cases} H_o: E(MS_{Tr})/E(MS_e)=1 \\ H_1: E(MS_{Tr})/E(MS_e)>1 \end{cases}$$

$$F_{emp}=\left.\frac{MS_{Tr}}{MS_e}\right|_{H_0} \sim F(k-1,N-k)$$

$$F_{cr}=F_\alpha(k-1,N-k)$$

$$\boxed{\text{אם } F_{emp}>F_{cr} \text{ נדחה את } H_0}$$

לוח ניתוח שונות

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_{emp}
Treatment	SS_{Tr}	a-1	$MS_{Tr}=\frac{SS_{Tr}}{a-1}$	$\frac{MS_{TR}}{MS_e}$
Error	SS_e	N-a	$MS_e=\frac{SS_e}{N-a}$	
Total	SS_{Tot}	N-1	$MS_{Tot}=\frac{SS_{Tot}}{N-1}$	

אמדים ורב"סים

אמד לתוחלת: $\hat{\mu}=\bar{y}$

רב"ס לתוחלת: $\mu \in \bar{y}\pm t_{\frac{\alpha}{2}}(N-a)\sqrt{\frac{MS_e}{N}}$

אמד לסטיית רמה i: $\hat{\tau}_i=\bar{y}_i-\bar{y}$

רב"ס לסטיית רמה i: $\hat{\tau}_i \in (\bar{y}_i-\bar{y})\pm t_{\frac{\alpha}{2}}(N-a)\sqrt{MS_e(\frac{1}{n_i}-\frac{1}{N})}$

תוחלת לרמה: $\hat{\mu}_i=\bar{y}_i$

רב"ס לתוחלת של רמה: $\mu_i \in \bar{y}_i\pm t_{\frac{\alpha}{2}}(N-a)\sqrt{\frac{MS_e}{n_i}}$

אמד לשונות הרעש: $\hat{\sigma}^2=MS_e$

רב"ס לשונות הרעש: $\frac{(N-a)MS_e}{\chi^2_{\frac{\alpha}{2}}(N-a)}\le\sigma^2\le\frac{(N-a)MS_e}{\chi^2_{1-\frac{\alpha}{2}}(N-a)}$

תזכורת – p-value

באופן כללי, משמעות ערך ה-p-value היא הסיכוי לטעות מסוג I – הסיכוי לקבל תוצאה קיצונית לפחות כפי שקיבלנו במבחן, בהנתן ש-H_0 נכונה. במקרה של ניתוח שונות, המשמעות היא הסיכוי לקבל הבדל מסויים בין הרמות השונות של הפקטור, למרות שבמציאות (שאותה אנחנו לא יודעים) אין כלל הבדל. מכאן, שאנו מעוניינים ב- p-value קטן ככל האפשר. בהינתן p-value (לדוגמא מפלט של r), צריך רק להשוות אותו לרמת המובהקות הנדרשת (לרוב 5% או 1%). אם ה-p-value שקיבלנו קטן מרמת המובהקות הדרושה – נדחה את H_0.

תרגיל 1

במפעל ישנן 4 מכונות שאמורות להיות זהות. המכונות מייצרות שבבי סיליקון במנות של 1000 יח'. בתום הייצור בודקים את כל היחידות וסופרים כמה פגומים היו. מכל מכונה נלקחו 5 דגימות.

מכונה \ תצפית	1	2	3	4	5
1	56	55	62	59	60
2	64	61	50	55	56
3	45	46	45	39	43
4	42	39	45	43	41

א. בדוק האם המכונות זהות בתוחלתן במובהקות של 1%

ב. הערך את כל הפרמטרים של המודל ובנה עבורם רב"ס במובהקות של 5%

פתרון:

א.

מכונה \ תצפית	1	2	3	4	5	$\bar{y}_i$ ממוצע רמה
1	56	55	62	59	60	58.4
2	64	61	50	55	56	57.2
3	45	46	45	39	43	43.6
4	42	39	45	43	41	42

במקרה זה מס' הדגימות בכל רמה זהה, ולכן הממוצע הכולל הוא ממוצע ממוצעי הרמות. $\bar{y} = 50.3$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_{emp}
Treatment	1135	4-1=3	378.33	29.79
Error	203.2	20-4=16	12.7	
Total	1338.2	20-1=19		

$$F_{cr} = F_{0.01}(3,16) = 5.292$$

$F_{emp} > F_{cr}$ ולכן ברמת מובהקות של 1% ניתן לדחות את ההשערה כי המכונות זהות בתוחלתן.

ב. אמד לתוחלת: $\hat{\mu} = \bar{y} = 50.3$

רב"ס לתוחלת: $\mu \in \bar{y}_{\cdot\cdot} \pm t_{\frac{\alpha}{2}}(N-k)\sqrt{\frac{MS_e}{N}} = 50.3 \pm t_{0.025}(16)\sqrt{\frac{12.7}{20}} = 50.3 \pm 1.7$

רמה	אמד לתוחלת	רב"ס לתוחלת	אמד לסטיית הרמה	רב"ס לסטיית הרמה
1	58.4	(56, 61.8)	8.1	(5.2, 11)
2	57.2	(53.8, 60.6)	6.9	(4, 9.8)
3	43.6	(40.2, 47)	-6.7	(-9.6, -3.8)
4	42	(38.6, 45.4)	-8.3	(-11.2, -5.4)

אמד לשונות הרעש: $\hat{\sigma}^2 = MS_e = 12.7$, רב"ס לשונות הרעש: $7.04 \le \sigma^2 \le 29.42$

תרגיל 2 (חלק משאלה ממבחן תשס"ט, מועד א')

שאלה 1 (40 נק')

בבחינה שנערכה בבית-הספר לכלכלה חולקו הנבחנים לשלוש כיתות. שתי כיתות נבחנים במבנה ביה"ס לכלכלה וכיתה נוספת בבניין דן-דוד. בבניין דן-דוד נערכו ביום הבחינה שיפוצים. לאחר הבחינה טענו הסטודנטים שנבחנו בבניין דן-דוד כי הם נאלצו להיבחן בתנאים לא הוגנים ולכן ראויים הם לפקטור.

טבלת הציונים:

דן דוד	כלכלה 2	כלכלה 1
82	86	89
76	80	85
76	78	82
73	78	82
69	76	79
68	70	75

כדי לחזק את טענתם הגישו הסטודנטים למרצה הקורס את פלט ניתוח השונות הבא:

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	192	2	96	3.73057	0.04841	3.68232
Within Groups	386	15	25.7333			
Total	578	17				

א. מהי המסקנה הנובעת מפלט ניתוח השונות?

ב. תוצאה זאת אינה מאמתת את טענת הסטודנטים, מדוע?

ג. איזה מבחן נוסף יש לבצע על מנת לאמת את הטענה? בצעו את המבחן ברמת מובהקות 0.05, מהי מסקנתכם?

פתרון

א. המסקנה הנובעת מפלט השונות היא כי ברמת מובהקות 0.05, לא ניתן לקבוע כי תוחלת הציונים בשלוש הכיתות זהה.

ב. הפלט מראה כי אכן יש הבדל בין הכיתות אך מן הפלט לא ניתן להסיק כי אכן תוחלת ציון של הכיתה שנבחנה בדן-דוד נמוכה יותר מתוחלות שתי הכיתות האחרות.

ג. נוכל לבצע ניתוח שונות בין שתי הכתות שנבחנו בכלכלה:

F crit	P-value	F	MS	df	SS	ANOVA
F crit	P-value	F	MS	df	SS	Source of Variation
4.964603	0.197617	1.904762	48	1	48	Between Groups
			25.2	10	252	Within Groups
				11	300	Total

הפלט מראה כי אין הבדל בין תוחלת הציונים של הכתות שנבחנו בכלכלה ברמת מובהקות של 5%.

כעת נוכל לאחד את הציונים של כתות כלכלה ולבצע ניתוח שונות בין ציונים אלו לציונים בכתת דן דוד. להלן הנתונים לאחר סידור מחדש:

כתת דן דוד	כתות כלכלה
82	89
76	85
76	82
73	82
69	79
68	75
	86
	80
	78
	78
	76
	70

להלן פלט ניתוח השונות:

						ANOVA
F crit	*P-value*	*F*	*MS*	*df*	*SS*	*Source of Variation*
4.543077	0.01045	8.55591	185.3262	1	185.3262	Between Groups
			21.66061	15	324.9091	Within Groups
				16	510.2353	Total

המסקנה הנובעת מפלט השונות כעת היא כי ברמת מובהקות 0.05, לא ניתן לקבוע כי תוחלת הציונים בכיתות כלכלה זהה לתחולת הציונים בכתת דן-דוד. כלומר כעת איששנו את טענת הסטודנטים.

<u>ניתוח שגיאות – בדיקת הנחות המודל</u>

מודל ניתוח השונות מניח כי:

- השגיאות לקוחות מפילוג נורמלי $\varepsilon_{ij} \sim N(0, \sigma^2)$
- השגיאות בלתי תלויות
- השונות של השגיאות אחידה על פני הרמות של הפקטור σ^2

<u>שיטות לבדיקת הנחות המודל:</u>

1. האם התצפיות / שאריות לקוחות מפילוג נורמלי? **שיטה: ניתוח שאריות - Q-Q Plot**

שיטת Q-Q plot נועדה לבדוק האם הרעש מפולג נורמלי. את המבחן נערוך לכל רמה בנפרד. במידה ומספר התצפיות קטן מ-5-10 נערוך את המבחן לכל התצפיות ביחד.

1. נחשב את השארית של כל תצפית לפי: $\varepsilon_{ij} = y_{ij} - \bar{y}_i$
2. נמיין את השאריות בסדר עולה (w = מיקום השארית ברשימה, כלומר עבור השארית הנמוכה ביותר $w = 1$, השארית הבאה אחריה $w = 2$ וכך הלאה)
3. נחלק את הטווח לאחוזים עפ"י הנוסחה: $\alpha = \frac{2w-1}{2n_i}$ (כלומר לכל שארית נתאים את האחוזון שלה)
4. נמצא את הערך הקריטי המתאים לכל שארית $\Phi^{-1}(\alpha)$. (הערך התיאורטי שמתאים לאחוזון)