การวิเคราะห์ความแปรปรวน

: ทำไมจึงเรียกการวิเคราะห์ความแปรปรวน (What is Analysis of Variance)

E48203

ดร.พหล ศักดิ์คะทัศน์
ผู้ช่วยศาสตราจารย์. ภาควิชาคณิตศาสตร์
และสถิติ คณะวิทยาศาสตร์.
มหาวิทยาลัยแม่โจ้.

บทนำ

ในการวิเคราะห์ความแปรปรวนที่มีการแจกแจงทางเดียว (One – way Classification Analysis of Variance) หรือ แผนการทดสอบแบบสุ่มสมบูรณ์ (Completely Randomized Design, CRD) มีรูปแบบสมการเชิงเส้นตรง (Linear Equation Model) คือ

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

เมื่อ Y_{ij} คือ ค่าสังเกตที่วัดได้

μ คือ ค่าเฉลี่ยรวมทั้งหมด

τ_i คือ อิทธิพลของ Treatment ที่ i โดยที่ i = 1, 2, 3, ..., t (t คือ จำนวนซ้ำ)

ε_{ij} คือ ความคลาดเคลื่อนของหน่วยทดลองที่ได้รับ Treatment ที่ i ซ้ำที่ j โดยที่ j = 1, 2, 3, ..., r (r คือ จำนวนซ้ำ)

สมมติฐานที่ต้องการทดสอบคือ

H_0 : $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_t$ (ค่าเฉลี่ยทุก treatment เท่ากัน)

H_1 : มีค่าเฉลี่ยอย่างน้อย 1 ค่าที่ต่างจากค่าอื่นๆ หรือ

H_2 : $\tau_i = 0$ (ไม่มีอิทธิพลของ treatments)

H_1 : มี τ_i อย่างน้อย 1 ค่าที่ไม่เท่ากับ 0

จากสมมติฐานหลัก H_0 เราบอกว่าค่าเฉลี่ยทุก treatments เท่ากัน หรือ อิทธิพลของ treatment ไม่มี จึงอาจมีบางคนที่สงสัยว่า เมื่อดูชื่อเรื่องแล้วเป็นการวิเคราะห์ความแปรปรวน จะเห็นว่าไม่เกี่ยวกับสมมติฐานที่จะทดสอบ เพราะจากสมมติฐานเป็นการทดสอบค่าเฉลี่ย ของข้อมูลหลายๆ กลุ่ม ดังนั้นก่อนอื่นจึงขอให้พิจารณาเกี่ยวกับลักษณะของข้อมูลในแผนการทดลองแบบสุ่มสมบูรณ์ก่อน

2. แหล่งความแปรปรวน ของข้อมูล

ในการทดลองแบบสุ่มสมบูรณ์นี้ หน่วยทดลองต้องสม่ำเสมอกันมากที่สุด แผนการทดลองนี้จึงจะมีประสิทธิภาพ ขอให้สังเกตข้อมูลตัวอย่างข้อมูลการทดลองปลูกข้าว 12 แปลง โดยข้าวที่ปลูกเป็นข้าวพันธุ์เดียวกัน อายุเท่ากัน สภาพแวดล้อมต่างๆ เหมือนกัน (สม่ำเสมอกัน) ทำการปลูกโดยไม่ใส่ปุ๋ยทุกแปลง ถ้าหน่วยทดลองสม่ำเสมอกันหรือได้รับการปฏิบัติเหมือนกันจริง ผลผลิตหรือค่าสังเกตที่ได้ต้องเท่ากัน ดังตารางที่ 1

	(i)			
	20	20	20	20
(j)	20	20	20	20
	20	20	20	20

ตารางที่ 1 หน่วยทดลองได้รับการปฏิบัติเหมือนกัน

จากตารางที่ 1 จะพบว่าผลผลิตข้าวแต่ละแปลงหรือ Y_{ij} = 20 ยังไม่มีอิทธิพล treatment และความแปรปรวนภายในกลุ่ม แต่เมื่อมีอิทธิพลของ treatments (ปุ๋ย) ดังตารางที่ 2 โดยที่ treatment ที่ 1 ไม่ใช้ปุ๋ยหรือปฏิบัติต่อหน่วยทดลองเหมือนเดิม จึงไม่ทำให้ผลผลิตเปลี่ยนแปลง ส่วน treatment ที่ 2, 3 และ 4 สมมติให้มีอิทธิพลของ treatment (ปุ๋ย) ที่ทำให้ผลผลิตเพิ่ม 2, 4 และ 5 ก.ก. ต่อแปลง ตามลำดับ ผลที่ได้ดังตารางที่ 2 ดังนี้

	Treatments (i)				
	1	2	3	4	←
	(0)	(+2)	(+4)	(+5)	อิทธิพลของ treatments
	20	22	24	25	
ซ้ำ (j)	20	22	24	25	
	20	22	24	25	

ตารางที่ 2 หน่วยทดลองได้รับ treatments ที่แตกต่างกันหรือมีความแปรปรวนระหว่าง treatment แต่ยังไม่มีความแปรปรวนภายในกลุ่มหรือภายใน treatment (error)

ดังนั้น $Y_{ij} = \mu + \tau_i$

เช่น ผลผลิตข้าวใน treatment ที่ 3 จะมีผลผลิตเท่ากับ 20 + 4 = 24 เป็นต้น

แต่ในสภาพความเป็นจริง แม้ว่าเราจะเลือกข้าวที่มาปลูกสม่ำเสมอกันแล้วก็ตาม อาจมีความแตกต่างกันบ้างที่ผู้ทดลองไม่สามารถควบคุมได้ทั่วถึง เช่น การได้รับแสง ปริมาณน้ำที่ได้รับไม่สม่ำเสมอ มีแมลงมากัดกินบางต้น เป็นต้น ดังนั้นผลผลิตข้าวจากการทดลองจริงๆ แล้วอาจจะเป็นได้ดังตารางที่ 3 ดังนี้

	Treatments (i)				
	1	2	3	4	←
	(0)	(+2)	(+4)	(+5)	อิทธิพลของ treatments
	19.2	22.5	24.8	26.3	
ซ้ำ (j)	21.3	23.1	23.4	25.2	ความแปรปรวนภายใน treatment
	18.6	21.2	24.1	24.5	(error)

ตารางที่ 3 หน่วยทดลองที่ได้รับ treatments ที่ต่างกันและมีความแปรปรวนภายในกลุ่ม (สภาพการทดลองโดยทั่วไป)

ดังนั้น $Y_{ij} = 20 + \tau_i + \varepsilon_{ij}$

จะพบว่าภายใน treatment เดียวกัน มีความแตกต่างหรือความแปรปรวนระหว่างหน่วยทดลองเกิดขึ้น นั่นก็คือ มีความคลาดเคลื่อนภายในกลุ่ม (error) เกิดขึ้น

ดังนั้นรูปแบบสมการเชิงเส้นตรง ของแผนการทดลองแบบสุ่มสมบูรณ์ ค่าสังเกต $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

จะพบว่าการวิเคราะห์ข้อมูลของแผนการทดลองแบบสุ่มสมบูรณ์นั้น ความแปรปรวนทั้งหมดแบ่งออกเป็น 2 ส่วน คือ ความแปรปรวนระหว่าง treatments หรือ ความแปรปรวนระหว่างกลุ่ม กับความแปรปรวนระหว่างหน่วยทดลองภายใน treatments เดียวกัน หรือความแปรปรวนภายในกลุ่ม (error)

3. ตารางการวิเคราะห์ความแปรปรวน (ตาราง ANOVA) และการทดสอบ

แหล่งของความแปรปรวน ที่ทำให้ค่าสังเกตที่ได้ไม่เท่ากันทุกค่า ก็คือ อิทธิพลของ treatments และ error

SOV	df.	SS	MS	F
Treatments	t – 1	SStrt	MStrt	MStrt/MSE
Error	n – t	SSE	MSE	
Total	n – 1			

เมื่อ n คือจำนวนข้อมูลหรือหน่วยทดลองทั้งหมด

ถ้าพิจารณาค่า Expected Mean Square (EMS) ของ treatments และ error จะพบว่าแต่ละส่วนประกอบด้วยความคลาดเคลื่อนดังนี้

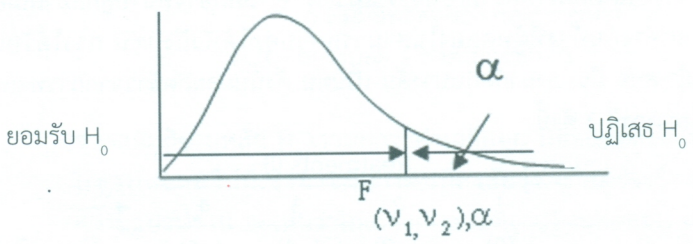
SOV	EMS
Treatments	$\sigma^2 + r\sigma_t^2$
Error	σ^2
Total	

ซึ่ง SS(total) = SS(Treatments) + SS(error)

หรือ

$$\sum(Y_{ij} - \bar{Y})^2 = r\sum(\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum(Y_{ij} - \bar{Y}_{i.})^2$$

(SStotal) (SStrt) (SSE)



เมื่อ $_{V1}$ คือ degree of freedom ของ treatments เท่ากับ t – 1

$_{V2}$ คือ degree of freedom ของตัวหาร error เท่ากับ n – t

$\bar{Y}$ เป็นค่าเฉลี่ยรวมทั้งหมดของตัวอย่าง และใช้ประมาณค่า μ หรือค่าเฉลี่ยประชากร

$\bar{Y}_{i.}$ เป็นค่าเฉลี่ยของ treatment ที่ i

SStotal หรือ ความคลาดเคลื่อนทั้งหมด คือ ผลรวมของกำลังสองของผลต่างระหว่างค่าสังเกตแต่ละค่า กับค่าเฉลี่ยทั้งหมด = $\sum(Y_{ij}-\bar{Y})^2$

SSE คือ ผลรวมของกำลังสองของผลต่างระหว่าง ค่าสังเกตแต่ละค่า กับค่าเฉลี่ยของ treatment นั้นๆ ทั้งหมดทุก treatment = $\sum(Y_{ij}-\bar{Y})^2$ = EMS (error) = σ^2

SStrt หรือ ความคลาดเคลื่อนระหว่าง treatment คือ ผลรวมของกำลังสองของผลต่างระหว่าง treatment กับค่าเฉลี่ยทั้งหมด =r $\sum(Y_{ij}-\bar{Y})^2$ = EMS (error) = $\sigma^2+r\sigma_t^2$

ในการวิเคราะห์ข้อมูลแบบสุ่มสมบูรณ์นั้น ถ้าความแปรปรวนระหว่าง treatments ซึ่งค่า EMS ประกอบด้วย $\sigma^2+r\sigma^2$ เปรียบเทียบกับความแปรปรวนของ Error ซึ่งประกอบด้วย σ^2. หรือเท่ากับอัตราส่วนของ $(\sigma^2+r\sigma_t^2)/\sigma^2$ ถ้าค่าเฉลี่ยระหว่าง treatments ใกล้เคียงกันมากๆ หรือความแปรปรวนเนื่องจาก treatments น้อย ค่า $r\sigma^2$จะมาก อัตราส่วน$(\sigma^2+r\sigma_t^2)/\sigma^2$จะมากกว่า 1 มากๆ ซึ่งค่าอัตราส่วน $(\sigma^2+r\sigma_t^2)/\sigma^2$ -เรียกว่า F – ratio ซึ่งในการคำนวณจะใช้ค่า MStrt/MSE แทน

ถ้าค่า F – ratio มีค่ามากเกินจุดวิกฤติ หรือ F > $F_\alpha(V_1,V_2)$ เราจะสรุปว่าความคลาดเคลื่อนระหว่าง Treatments มีค่ามากกว่าความคลาดเคลื่อนระหว่างหน่วยทดลอง (error) เกินจุดที่จะยอมรับได้ (จุดวิกฤติ) เราจะปฏิเสธสมมติฐานหลักว่า ค่าเฉลี่ยแต่ละ treatments เท่ากัน ซึ่งก็คือ ความคลาดเคลื่อนระหว่าง treatments มีค่ามากนั่นเองหรือสรุปได้ว่าค่าเฉลี่ยแต่ละ treatments แตกต่างกัน

4. สรุป

จากสมมติฐานที่ทดสอบว่า ค่าเฉลี่ยเท่ากันหรือไม่มีความแตกต่างระหว่างค่าเฉลี่ยสมมติฐาน

H_0 : $\mu_1=\mu_2=\mu_3=\ldots=\mu_t$ (ค่าเฉลี่ยทุก treatment เท่ากัน)

H_1 : τ_i มีค่าเฉลี่ยอย่างน้อย 1 ค่าที่ต่างจากค่าอื่นๆ

หรือ

H_0 : τ_i = 0 (ไม่มีอิทธิพลของ treatments)

H_1 : มี τ_i อย่างน้อย 1 ค่าที่ไม่เท่ากับ 0

เราอาจจะเขียนสมมติฐานในรูปแบบของการทดสอบความแปรปรวนเหมือนกับชื่อเรื่องการวิเคราะห์ความแปรปรวนก็ได้ คือ

H_0 : σ_t^2= 0 หรือความแปรปรวนของ treatments เท่ากับ 0 หรือไม่มีความแตกต่างระหว่าง treatment หรือค่าเฉลี่ยเท่ากัน และ

H_1 : $\sigma_t^2 \neq 0$

จากบทความที่กล่าวมา หลายท่านคงเกิดความเข้าใจบ้างว่าการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของข้อมูลตั้งแต่ 2 กลุ่มขึ้นไป ซึ่งเราเรียกว่าการวิเคราะห์ความแปรปรวนนั้นแท้ที่จริงใช้เทคนิคการวิเคราะห์ความแปรปรวนมาสรุปผลต่างค่าเฉลี่ย

บรรณานุกรม

พหล ศักดิ์คะทัศน์. (2548). **หลักสถิติ.** เอกสารประกอบการสอนวิชาหลักสถิติ ภาควิชาคณิตศาสตร์ และสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยแม่โจ้. เชียงใหม่: มหาวิทยาลัยแม่โจ้.

Cochran, W. G. and G. M. Cox. (1957). **Experimental Design.** New York: John Wiley and Sons.

Cooper, B. E. (1975). **Statistics for Experimentalist.** Oxford: Pergamon.