
การเปรียบเทียบค่าเฉลี่ยภายหลัง

การวิเคราะห์ความแปรปรวน

รศ.นิภา ศรีไพโรจน์ *

จุดประสงค์ที่สำคัญของการเปรียบเทียบค่าเฉลี่ยของผลการทดลองภายหลังการวิเคราะห์ความแปรปรวน (Analysis of Variance) โดยใช้ F-test ที่เรียกว่า A Posteriori multiple comparisons หรือ Post Hoc Comparisons ก็คือเมื่อผลการทดสอบรวมมีนัยสำคัญทางสถิติในระดับใดระดับหนึ่งที่ผู้วิจัยกำหนดไว้ (α) การทดสอบในขั้นนี้ผู้วิจัยยังไม่ทราบแน่ชัดว่ามีค่าเฉลี่ยคู่ใดบ้างที่มีความแตกต่างกันอย่างมีนัยสำคัญจึงต้องมีการตรวจสอบในรายละเอียดต่อไป ซึ่งมีหลายวิธี แต่ละวิธีมีวัตถุประสงค์และคุณสมบัติเฉพาะในการเปรียบเทียบแตกต่างกันออกไป ผู้วิจัยจึงต้องเลือกใช้ให้เหมาะสมกับปัญหาที่ศึกษา เพื่อให้ผลการทดสอบคลาดเคลื่อนน้อยที่สุดและตอบปัญหาได้ชัดเจนที่สุด

วิธีการเปรียบเทียบความแตกต่างของค่าเฉลี่ยภายหลังการทดสอบรวมที่มีนัยสำคัญทางสถิติได้แก่

1. **LSD** (Least Significance Difference Test)

2. **FLSD** (Fisher Least Significance Difference Test)

3. **DMRT หรือ DC** (Duncan's New Multiple Range Test)

4. **S-N-K หรือ Newman-Keuls' MRT** (Student - Newman - Keuls Test

or Newman-Keuls Multiple Range Test)

5. **TUKEY A หรือ Tukey's หรือ HSD** (Tukey's Honestly Significance Difference Test : T_A)

6. **TUKEY B** (Tukey's Alternate Test : T_B)

7. **SCHEFFE** (Scheffé Test : S)

8. **DUNNETT** (Dunnett's Test : DN)

9. **SMC** (Stepwise Multiple Comparison Test)

10. **BSD หรือ Bayes LSD หรือ k-ratio LSD** (Bayes Significance Difference Test)

11. **BET** (Bayes Exact Test)

12. **Bonferroni t Statistic**

13. **Williams' Method**

การเลือกวิธีที่เหมาะสมสำหรับเปรียบเทียบค่าเฉลี่ย ผู้วิจัยจะต้องพิจารณาปัจจัยที่สำคัญดังนี้

1. ลักษณะความคลาดเคลื่อนทางสถิติ
2. ลักษณะการเปรียบเทียบของแต่ละวิธี
3. ขนาดของกลุ่มตัวอย่างในแต่ละกลุ่มที่

* รองศาสตราจารย์ ภาควิชาพื้นฐานของการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ มหาสารคาม

เปรียบเทียบ

1. ลักษณะความคลาดเคลื่อนทางสถิติ

ความคลาดเคลื่อนทางสถิติในทางวิเคราะห์ข้อมูล มี 3 ประเภทคือ

1.1 ความคลาดเคลื่อนประเภทที่ 1 (Type I error หรือ α error) หมายถึงความคลาดเคลื่อนที่เกิดจากการปฏิเสธ null hypothesis โดยที่ความจริงแล้ว null hypothesis นั้นถูกต้อง (Probability of rejecting a true null hypothesis)

ซึ่งเขียนในลักษณะของอัตราความคลาดเคลื่อนได้ดังนี้

1.1.1 อัตราความคลาดเคลื่อนต่อการเปรียบเทียบ (α_c)
(per-comparison or Comparison wise error rate)

$$= \frac{\text{จำนวนการเปรียบเทียบอย่างมีนัยสำคัญที่ผิด}}{\text{จำนวนทั้งหมดของการเปรียบเทียบ}}$$

1.1.2 อัตราความคลาดเคลื่อนต่อการทดลอง จำแนกเป็น

ก. อัตราความคลาดเคลื่อนต่อการทดลอง (α_e)
(per-experiment error rate)

$$= \frac{\text{จำนวนการเปรียบเทียบอย่างมีนัยสำคัญที่ผิด}}{\text{จำนวนทั้งหมดของการทดลอง}}$$

ข. อัตราความคลาดเคลื่อนต่อการทดลอง (α_{ew})
(experimentwise error rate)

$$= \frac{\text{จำนวนของการทดลองที่มีอย่างน้อย 1 คู่แตกต่างกันอย่างมีนัยสำคัญที่ผิด}}{\text{จำนวนทั้งหมดของการทดลอง}}$$

ความสัมพันธ์ระหว่างอัตราความคลาดเคลื่อนประเภทที่ 1.1.1 กับ 1.1.2 ปรากฏดังสมการ

$$\alpha_{ew} = 1 - (1 - \alpha_c)^c$$

เมื่อ c แทน จำนวนครั้งของการเปรียบเทียบค่าเฉลี่ยแต่ละคู่

($c = \frac{K(K-1)}{2}$ เมื่อ K แทน จำนวนค่าเฉลี่ยทั้งหมดของกลุ่มทดลองซึ่งมากกว่า 2 กลุ่มขึ้นไป)

และเมื่อผู้วิจัยกำหนดระดับนัยสำคัญที่ต่ำ ๆ ความน่าจะเป็นจะเท่ากับค่าประมาณ $c\alpha_c$ จะพบว่าอัตราความคลาดเคลื่อนที่เกิดขึ้น จะมีค่าสูงกว่าอัตราความคลาดเคลื่อนที่กำหนดไว้ นั่นคือ จำนวนครั้งของการเปรียบเทียบมีผลทำให้อัตราความคลาดเคลื่อนต่อการทดลองเพิ่มสูงขึ้นได้มากกว่าระดับนัยสำคัญที่กำหนดไว้

1.2 ความคลาดเคลื่อนประเภทที่ 2 (Type II error or β error) ความคลาดเคลื่อนที่เกิด

จากการยอมรับ null hypothesis โดยความจริงแล้ว null hypothesis ไม่ถูกต้อง (probability of failing to reject a false null hypothesis)

1.3 ความคลาดเคลื่อนประเภทที่ 3 (Type III error or γ error) หมายถึงความคลาดเคลื่อนที่เกิดจากการปฏิเสธ null hypothesis และการยอมรับ alternative hypothesis นั้นขัดแย้งกันเองในลักษณะตรงข้ามกับความเป็นจริง (probability of declaring one treatment superior to another when the reverse is actually true)

ในทางปฏิบัติความคลาดเคลื่อนประเภทที่ 3 เกิดขึ้นน้อยมาก และการวิเคราะห์ข้อมูลมักจะเกี่ยวข้องกับความคลาดเคลื่อนประเภทที่ 1 และ 2 นักสถิติจะกำหนดระดับนัยสำคัญเพื่อป้องกันความคลาดเคลื่อนประเภทที่ 1 (α) มากกว่าป้องกันความคลาดเคลื่อนประเภทที่ 2 (β)

สถิติที่ใช้ในการเปรียบเทียบตั้งแต่วิธีที่ 1 ถึง 9 เป็นสถิติที่ป้องกันความคลาดเคลื่อนประเภทที่ 1 เท่านั้น ส่วนวิธีที่ 10 และ 11 เป็นวิธีที่เกี่ยวข้องกับความคลาดเคลื่อนทั้งประเภทที่ 1 และ 2 ในที่นี้ผู้เขียนจะกล่าวถึงสถิติเปรียบเทียบวิธีที่ 1 ถึง 8 ซึ่งเป็นสถิติที่ป้องกันความคลาดเคลื่อนประเภทที่ 1

นักสถิติได้ให้ข้อเสนอแนะว่าถ้าผู้วิจัยต้องการเปรียบเทียบค่าเฉลี่ยเป็นบางคู่ ให้กำหนดความคลาดเคลื่อนแบบ Comparisonwise error rate (α_c) แล้วพิจารณาเลือกใช้สถิติดังนี้คือ วิธี LSD, FLSD, DMRT (Duncan), S-N-K และ Tukey B

ถ้าผู้วิจัยต้องการเปรียบเทียบค่าเฉลี่ยทุกคู่ของกลุ่มตัวอย่างทั้งหมด ให้กำหนดความคลาดเคลื่อนแบบ experimentwise error rate แล้ว

พิจารณาเลือกใช้สถิติดังนี้คือ วิธี FLSD, S-N-K, Tukey A Scheffé และ Dunnett

2. ลักษณะการเปรียบเทียบของแต่ละวิธี

ผู้วิจัยจำเป็นต้องเข้าใจคุณสมบัติเฉพาะของแต่ละวิธี เพื่อจะได้เลือกวิธีให้ถูกต้องเหมาะสมกับเรื่องที่ศึกษา แบ่งออกได้ 2 กลุ่มใหญ่ คือ

2.1 The Fixed Range Test ใช้ค่าพิสัยค่าเดียวสำหรับเปรียบเทียบทั้งหมด สถิติในกลุ่มนี้ได้แก่ LSD, FLSD TUKEY A (HSD), SCHEFFÉ และ DUNNETT โดยที่วิธี LSD และ FLSD กำหนดความคลาดเคลื่อนแบบ Comparisonwise error rate และใช้ t - distribution ส่วนวิธีของ Tukey Scheffé และ Dunnett กำหนดความคลาดเคลื่อนแบบ experimentwise error rate โดยใช้ค่า Studentized range, F และ Dunnett's distribution ตามลำดับ

วิธี LSD กำหนดระดับนัยสำคัญในการเปรียบเทียบแต่ละคู่ (α_c) เท่ากับระดับนัยสำคัญของการวิเคราะห์ ANOVA โดยใช้ F-test ในตอนแรกซึ่งเท่ากับ α เช่น α ใน F-test กำหนดไว้ = .05 เมื่อผลทดสอบมีนัยสำคัญก็จะใช้ระดับ $\alpha = .05$ เป็นค่า α_c สำหรับเปรียบเทียบแต่ละคู่ทุกคู่

วิธี FLSD กำหนดระดับนัยสำคัญในการเปรียบเทียบ แต่ละคู่เท่ากับระดับนัยสำคัญของการวิเคราะห์ ANOVA โดยใช้ F-test ทหารด้วยจำนวนคู่หรือจำนวนครั้งที่ใช้เปรียบเทียบทั้งหมด เช่น α ใน F-test กำหนดไว้ = .05 ต้องการเปรียบเทียบ 5 คู่ (หรือ 5 ครั้ง) เมื่อผลทดสอบมีนัยสำคัญก็จะใช้ระดับ $\alpha = .05 / 5 = .01$ เป็นค่า α_c สำหรับเปรียบเทียบแต่ละคู่

วิธีของ Tukey หรือ HSD หรือ TUKEY A เป็นวิธีที่คล้ายกับวิธี LSD แต่ถ้าค่า α_{cw} เป็นค่าเดียวกับ $\alpha_c = .05$ และ $K > 2$ สถิติ Tukey

A จะมีค่าสูงกว่า LSD ซึ่งจะทำให้จำนวนคู่ในการเปรียบเทียบที่มีนัยสำคัญของ TUKEY น้อยกว่าของ LSD

วิธีของ Scheffé เป็นวิธีที่ใช้เปรียบเทียบค่าเฉลี่ย โดยที่ขนาดของตัวอย่างในแต่ละกลุ่มไม่เท่ากัน เมื่อเปรียบเทียบแบบ Pairwise multiple Comparisons ที่ระดับนัยสำคัญของ $F(\alpha)$ และจำนวนค่าเฉลี่ย (K) มากกว่า 2 ค่า S จะสูงกว่าค่า T_A (Tukey A) และ LSD ซึ่งทำให้จำนวนคู่ในการเปรียบเทียบที่มีนัยสำคัญของ Scheffé น้อยกว่าของ Tukey และ LSD

วิธีของ Dunnett เป็นวิธีที่ใช้เปรียบเทียบค่าเฉลี่ยของกลุ่มควบคุมและกลุ่มทดลองทุกกลุ่ม ซึ่งมีขนาดตัวอย่างเท่ากัน โดยมีตารางสำหรับหาค่าวิกฤติ โดยคัดแปลงจาก Studentized range Statistic และทดสอบได้ทั้งแบบ One-tailed หรือ two-tailed test

2.2 The Multiple Range Tests ใช้ค่าพิสัยหลายค่าตามจำนวนครั้งของการเปรียบเทียบ หรือจำนวนชั้นของระยะห่างระหว่างค่าเฉลี่ยของคู่ที่เปรียบเทียบ สถิติในกลุ่มนี้ได้แก่ S-N-K, DMRT, และ TUKEY B

วิธี S-N-K เป็นวิธีที่ปรับปรุงมาจากวิธีของ Student (1927) Newman (1939) และ Keuls (1952) โดยใช้ค่าพิสัยที่เปลี่ยนแปลงไปตามจำนวนของค่าเฉลี่ยในการเปรียบเทียบ (มีจำนวนมากที่สุด $K-1$ ค่า) วิธี S-N-K กำหนดระดับนัยสำคัญของการเปรียบเทียบค่าเฉลี่ยเป็นคู่ ๆ โดยเรียงลำดับค่าเฉลี่ยจากน้อยไปมาก มีการกำหนดความคลาดเคลื่อนต่อการทดลอง (α_{ew}) และใช้ระดับนัยสำคัญนี้ระดับเดียวสำหรับทดสอบความแตกต่างในแต่ละขั้นของการเปรียบเทียบ เมื่อ $p = 2$ ค่า S-N- K_2 จะเท่ากับค่า LSD และเมื่อ $p = K$ ค่า S-N- K_K จะเท่ากับ

T_A และเมื่อ p มีค่าอยู่ระหว่าง 2 และ K ค่า S-N- K_p จะมีค่าอยู่ระหว่างค่าของ LSD และ T_A

วิธี DMRT ของ Duncan คล้ายกับวิธี S-N-K โดยถ้า K มากจะทำให้วิธีของ Duncan มีอำนาจการทดสอบสูง (ค่าวิกฤติของ Duncan จะน้อยกว่าค่าของ S-N-K และจะแตกต่างเพิ่มมากขึ้น เมื่อจำนวน p มากขึ้น) วิธีนี้จะให้อำนาจการทดสอบที่สูงกว่าวิธีของ S-N-K เมื่อมี p มาก ในกรณี ที่ $p=2$ ค่าวิกฤติของ Duncan จะเท่ากับค่าวิกฤติของ S-N-K และ LSD เมื่อ $p=K$ ค่าของ DMRT จะอยู่ระหว่างค่าของ LSD และ S-N- K_K Duncan ได้คัดแปลงการใช้ Studentized range statistic ให้เหมาะสมและสร้างตารางสำหรับค่าวิกฤติแบบ One-tailed Test

วิธีของ Tukey B เป็นวิธีที่ Tukey ได้ปรับปรุงแก้ไขจากวิธี Tukey A โดยที่ค่าพิสัยในวิธีนี้ได้จากการเฉลี่ยค่าพิสัยที่ใช้ในวิธี S-N-K และวิธี Tukey A เมื่อ $p=2$ ค่าวิกฤติของ Tukey B จะอยู่ในระหว่างค่าของ S-N-K และ Tukey A และเมื่อ $p=K$ ค่าวิกฤติของ Tukey B (T_B), S-N-K และ Tukey A (T_A) จะเท่ากันหมด

3. ขนาดของกลุ่มตัวอย่างในแต่ละกลุ่มที่เปรียบเทียบ

วิธีต่าง ๆ ที่กล่าวมา ยกเว้นวิธีของ Scheffé ใช้สำหรับการเปรียบเทียบค่าเฉลี่ยของกลุ่มทดลองที่มีขนาดตัวอย่างเท่ากัน แต่ในทางปฏิบัติผู้วิจัยจำนวนมากใช้ขนาดของกลุ่มทดลองไม่เท่ากัน ทั้งนี้เนื่องจากเหตุผลในเรื่องวิธีการสุ่มตัวอย่าง หรือในระหว่างการทดลองกลุ่มตัวอย่างมีจำนวนลดลงหรือเหตุผลอื่นใดก็ตาม การหาค่า Critical Values ของสถิติแต่ละตัวก็อนุโลมให้ใช้หลักการเดียวกันกับกรณีกลุ่ม

ตัวอย่างที่เปรียบเทียบมีขนาดเท่ากันแต่สูตรคำนวณ
ค่า $S_{\bar{x}}$ เปลี่ยนไปดังนี้

วิธี LSD, FLSD และ Dunnett

$$S_{\bar{x}} = \sqrt{MS_{\text{error}} \left(\frac{1}{n_1} + \frac{1}{n_j} \right)}$$

วิธี Tukey A , Tukey B, S-N-K และ Duncan

$$S_{\bar{x}} = \sqrt{\frac{S^2}{2} \left(\frac{1}{n_1} + \frac{1}{n_j} \right)}$$

หรือ

$$S_{\bar{x}} = \sqrt{\frac{MS_{\text{error}}}{\tilde{n}}}$$

\tilde{n} แทนค่า harmonic mean

$$\text{เมื่อ } \tilde{n} = \frac{K}{(1/n_1) + (1/n_2) + \dots + (1/n_k)}$$

การใช้สถิติเปรียบเทียบสำหรับขนาดของกลุ่มตัวอย่างที่ไม่เท่ากันยังมีการหาค่าคอบตอบไปว่า วิธีใดจะให้ผลถูกต้องมากที่สุดและขนาดของกลุ่มตัวอย่างในแต่ละกลุ่มที่เปรียบเทียบจะแตกต่างกันมากขนาดใด จึงจะทำให้ผลการเปรียบเทียบยังคงเชื่อถือได้

นักวิจัยได้เสนอแนะวิธีเลือกวิธีเปรียบเทียบค่าเฉลี่ยโดยพิจารณาจากจำนวนกลุ่มทดลอง โดยใช้เกณฑ์ดังนี้

| | | |
|-----------|---------|---|
| "น้อย" | หมายถึง | จำนวนกลุ่มทดลองที่อยู่ในระหว่าง 2-5 กลุ่ม |
| "ปานกลาง" | หมายถึง | จำนวนกลุ่มทดลองที่อยู่ในระหว่าง 6-10 กลุ่ม |
| "มาก" | หมายถึง | จำนวนกลุ่มทดลองที่อยู่ในระหว่าง 11-20 กลุ่ม |

| จำนวนกลุ่มทดลอง | เสนอแนะวิธีการเปรียบเทียบที่กำหนดความคลาดเคลื่อนแบบ | |
|-----------------|---|--|
| | Comparisonwise error rate | Experimentwise error rate |
| น้อย | LSD | FLSD |
| ปานกลาง | DMRT (Duncan) S-N-K Tukey B หรือ FLSD (อำนาจการทดสอบวิธี DMRT (Duncan) ใกล้เคียงกับวิธี S-N-K แต่มีอำนาจการทดสอบสูงกว่าวิธี Tukey B และ FLSD) | S-N-K หรือ Tukey A (วิธี S-N-K มีอำนาจการทดสอบสูงกว่าวิธี Tukey A แต่วิธี Tukey A สะดวกในการคำนวณมากกว่าวิธี S-N-K เพราะคำนวณค่าวิกฤติเพียงค่าเดียว) |
| มาก | DMRT (Duncan) S-N-K (DMRT มีอำนาจการทดสอบสูงกว่า S-N-K) | S-N-K Tukey A (เหมือนกรณีปานกลาง) |

นอกจากนี้นักสถิติยังได้เสนอแนะว่า วิธีของ Scheffé ใช้สำหรับเปรียบเทียบค่าเฉลี่ยของกลุ่มทดลองหลาย ๆ ค่าในลักษณะ Linear Contrast หรือใช้เปรียบเทียบค่าเฉลี่ยที่ปรับความแปรปรวนร่วม (means adjusted by Covariance) และไม่เสนอให้ใช้วิธีของ Scheffé เปรียบเทียบในลักษณะ Pairwise comparisons เพราะมีอำนาจการทดสอบต่ำกว่าวิธีอื่น ๆ ส่วนวิธีของ Dunnett ใช้สำหรับเปรียบเทียบค่าเฉลี่ยของกลุ่มควบคุมกับกลุ่มทดลองเท่านั้น

สำหรับรายละเอียดเกี่ยวกับวิธีคำนวณผู้เขียนขอตั้งข้อสมมติว่าผู้อ่านมีความรู้พื้นฐานในเรื่องดังกล่าวแล้ว หากต้องการศึกษาเพิ่มเติมอาจศึกษาได้จากตำราสถิติวิจัย

เอกสารอ้างอิง

- Bancroft, T.A **Topics in Intermediate Statistical Methods** Vol 1 : The Iowa State University Press., 1988.
- Carmer, S.G. and M.R. Swanson " An Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte Carlo Methods," **Journal of the American Statistical Association**, 1973.
- Geoffrey Keppel, **Design and Analysis : A Researcher's Handbook** Prentice-Hall Inc, 1982.
- Milton. J.S. and J.O. Tsokos **Statistical Methods in the Biological and Health Sciences** McGraw-Hill International Book Company, 1983.
- Steel, R.G. and J.H. Torrie. **Principles and Procedures of Statistics : A Biometrical Approach** McGraw-Hill International Book Company, 1981.
- Winer, B.J, **Statistical Principles in Experimental Design**, McGraw Hill International Book Company, 1971.