บทที่ 2
แนวคิดและทฤษฎี

โดยทั่วไปในการเก็บรวบรวมข้อมูลเพื่อที่จะศึกษาถึงลักษณะประชากร มักจะมีข้อมูลบางค่าที่มีค่าแตกต่างไปจากข้อมูลส่วนใหญ่ โดยแตกต่างไปในทางมากกว่าหรือน้อยกว่าและข้อมูลที่มีค่าแตกต่างนี้อาจจะเป็นข้อมูลผิดปกติ(Outlier)หรือไม่ก็ได้ซึ่งจะทราบได้จากการตรวจสอบถ้าข้อมูลที่แตกต่างไปจากข้อมูลส่วนใหญ่นั้นเป็นข้อมูลผิดปกติ จะทำให้ผลการวิเคราะห์ในทางสถิติการหาค่าต่างๆ เช่น ค่าเฉลี่ย ค่าความแปรปรวน และอิทธิพลของข้อมูลชุดหนึ่งๆหรือหลายชุดไม่ตรงกับความเป็นจริง ทำให้เกิดความผิดพลาดในการนำผลการวิเคราะห์ไปใช้ ค่าเหล่านี้อาจมีสาเหตุมาจากความผิดพลาดในการเก็บรวบรวมข้อมูลหรือการเบี่ยงเบนจากข้อกำหนดของการวิเคราะห์ข้อมูล ซึ่งถ้าเราสามารถแก้ปัญหาข้อมูลผิดปกติเหล่านี้ได้ก็จะไม่เกิดผลกระทบต่อผลการวิเคราะห์ทางสถิติทำให้ได้ผลการวิเคราะห์ที่ถูกต้องไปใช้ สำหรับในการศึกษาครั้งนี้จะแก้ปัญหาดังกล่าวนี้โดยการแปลงข้อมูลด้วยเลขยกกำลังค่าต่างๆ โดยเปรียบเทียบจากสัดส่วนของจำนวนข้อมูลผิดปกติที่ลดลงหลังจากทำการแปลงข้อมูลต่อจำนวนชุดข้อมูลทั้งหมดเป็นหลักและพิจารณาจากค่า p-value ของสถิติทดสอบเอฟโดยคำนวณหาค่าสัดส่วนของการปฏิเสธสมมติฐานว่างซึ่งคำนวณจากการนับจำนวนชุดข้อมูลที่ปฏิเสธสมมติฐานว่างต่อจำนวนชุดข้อมูลทั้งหมดและหาค่าอำนาจการทดสอบ ในขั้นต้นจะกล่าวถึงตัวแบบสำหรับแผนแบบการทดลองสุ่มตลอด การวิเคราะห์ความแปรปรวนหรือการทดสอบเอฟและการแปลงข้อมูลโดยใช้เลขยกกำลัง

2.1 แผนแบบการทดลองสุ่มตลอด (Completely Randomized Design)

แผนการทดลองแบบสุ่มตลอดนี้เป็นแผนการทดลองที่ง่ายที่สุด เหมาะกับการทดลองที่ไม่สามารถแยกได้ว่าสิ่งทดลองที่นำมาใช้นั้นมีลักษณะแตกต่างกันอย่างไรก่อนการทดลอง เทคนิคการวิเคราะห์ความแปรปรวนสำหรับแผนการทดลองนี้จะแยกสาเหตุของความแปรผันของข้อมูลทั้งหมด เนื่องมาจากอิทธิพลของวิธีทดลองเพียงอย่างเดียว ไม่มีสาเหตุจากปัจจัยอื่นอีก จึงเรียกข้อมูลนี้ว่าข้อมูลแบบแจกแจงทางเดียว (One-way classification) ตามแผนการทดลองนี้แสดงว่าเมื่อหน่วยทดลองได้รับวิธีทดลองที่ต้องการทดสอบแล้ว ความแตกต่างของข้อมูลที่เก็บได้จากแต่ละหน่วยทดลองจะต้องเกิดจากอิทธิพลของวิธีทดลองที่ต่างกันเท่านั้น ดังนั้นเพื่อให้แผนการทดลองนี้มีประสิทธิภาพสูงสุด หน่วยทดลองที่นำมาใช้จึงควรมีลักษณะสม่ำเสมอกันหือคล้ายคลึงกันมากที่สุด (Homogeneous) หรือให้มีความแปรผันระหว่างหน่วยทดลองน้อยที่สุด หลักการสำคัญของแผนการทดลองนี้คือ การจัดวิธีทดลองให้กับหน่วยทดลองหรือจัดหน่วยทดลองให้กับให้กับวิธีทดลอง จะต้องเป็นไปโดยสุ่ม ไม่มีข้อ จำกัดเกี่ยวกับการสุ่ม และแผนการทดลองแบบสุ่ม

ตลอดนี้สามารถใช้กับการทดลองที่มีวิธีทดลองจำนวนมากๆได้ และแต่ละวิธีทดลองไม่จำเป็นต้องใช้จำนวนหน่วยทดลองเท่ากันหรือทำซ้ำเท่ากัน

สำหรับแผนการทดลองแบบสุ่มตลอด มีรูปแบบเชิงเส้น (Linear model) ที่ใช้แทนค่าสังเกตแต่ละค่าในแผนการทดลองที่กำหนดขึ้น เป็นตัวแบบผลบวก (Additive model) ดังนี้

ตัวแบบสำหรับข้อมูลตอบสนองในกรณีที่กระทำวิธีทดลองกับหน่วยทดลอง คือ

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

หรือ $y_{ij} = \mu_i + \varepsilon_{ij}$ โดยที่ $\mu_i = \mu + \tau_i$

เมื่อ $i = 1, 2, \ldots, k$

$j = 1, 2, \ldots, n$

y_{ij} คือ ข้อมูลตอบสนองของหน่วยทดลองที่ j ที่ได้รับวิธีทดลองที่ i

μ คือ พารามิเตอร์แทนค่าเฉลี่ยรวม

τ_i คือ พารามิเตอร์แทนอิทธิพลจากปัจจัยทดลองที่ i

ε_{ij} คือ ค่าความคลาดเคลื่อนของหน่วยทดลองที่ j ที่ได้รับวิธีทดลองที่ i

k คือ จำนวนวิธีทดลอง

n คือ จำนวนซ้ำของหน่วยทดลองในแต่ละวิธีทดลอง

2.2 การวิเคราะห์ความแปรปรวนหรือการทดสอบเอฟสำหรับแผนการทดลองสุ่มตลอด (The Analysis of Variance for Completely Randomized Design)

การวิเคราะห์ความแปรปรวนหรือการทดสอบเอฟสำหรับแผนแบบการทดลองสุ่มตลอดเพื่อทดสอบเกี่ยวกับความแตกต่างระหว่างอิทธิพลของวิธีทดลอง แสดงดังตารางที่ 2.1 ดังนี้

ตารางที่ 2.1 ตารางวิเคราะห์ความแปรปรวนสำหรับแผนแบบทดลองสุ่มตลอดปัจจัยคงที่เมื่อไม่มีหน่วยตัวอย่างย่อยและจำนวนซ้ำของแต่ละวิธีทดลองเท่ากัน

แหล่งของความผันแปร	องศาความเป็นอิสระ	ผลรวมกำลังสอง	ผลรวมกำลังสองเฉลี่ย	ค่าเอฟ
ปัจจัยทดลอง	$k-1$	SSTr	$MSTr = \frac{SSTr}{k-1}$	$\frac{MSTr}{MSE}$
ความคลาดเคลื่อน	$k(n-1)$	SSE	$MSE = \frac{SSE}{k(n-1)}$	
รวม	$kn-1$	SST		

โดยที่

y_{ij} = ค่าของข้อมูลจากหน่วยทดลองที่ j ที่ได้รับวิธีทดลองที่ i

$y_{i.}$ = ผลรวมของข้อมูลจากหน่วยทดลองที่ได้รับวิธีทดลองที่ i $= \sum_{j=1}^{n} y_{ij}$

$\bar{y}_{i.}$ = ค่าเฉลี่ยของวิธีทดลองที่ i $= \frac{y_{i.}}{n}$

$y_{..}$ = ผลรวมของข้อมูลทั้งหมด $= \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \sum_{i=1}^{k} y_{i.}$

$\bar{y}_{..}$ = ค่าเฉลี่ยของข้อมูลทั้งหมด $= \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}}{nk} = \frac{\sum_{i=1}^{k} y_{i.}}{k}$

SST = ผลบวกกำลังสองทั้งหมด (Sum Square of Totals)

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{nk}$$

SSTr = ผลบวกกำลังสองของวิธีทดลอง (Sum Squares of Treatment)

$$= n\sum_{i=1}^{k}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\frac{y_{i.}^2}{n} - \frac{y_{..}^2}{nk}$$

SSE = ผลบวกกำลังสองของความคลาดเคลื่อนของการทดลอง

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2 = SST - SSTr$$

<u>สมมติฐานสำหรับการทดสอบ</u>

สำหรับปัจจัยทดลองเป็นปัจจัยคงที่ (Fixed factor)

$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

$H_1 : \mu_i \neq \mu_j$ มีอย่างน้อย 1 คู่ของ $i \neq j$

ในการวิเคราะห์ความแปรปรวนนี้อาจตั้งสมมติฐานที่ต้องการทดสอบในรูปแบบอิทธิพลของวิธีทดลอง (Treatment effect)

$H_0 : \tau_1 = \tau_2 = \tau_3 = \ldots = \tau_k = 0$

$H_1 : \tau_i$ มีบางค่าที่ไม่เท่ากับ 0

เกณฑ์ในการตัดสินใจของการทดสอบเอฟ

ในการวิเคราะห์ความแปรปรวนหรือการทดสอบเอฟ จะปฏิเสธสมมติฐานว่างเมื่อค่าเอฟจากการคำนวณมีค่ามากกว่าค่าเอฟที่ได้จากการเปิดตารางเอฟที่ระดับนัยสำคัญ α และองศาความเป็นอิสระ $\nu_1 = k-1$ และ $\nu_2 = k(n-1)$ ภายใต้สมมติฐานว่าง สามารถเขียนแทนด้วย $F_{\alpha[k-1,k(n-1)]}$ และสำหรับภายใต้สมมติฐานแย้ง การแจกแจงของเอฟจะเป็นการแจกแจงแบบเอฟห่างศูนย์กลาง (Non-central F distribution) ที่มีระดับนัยสำคัญ α และองศาความเป็นอิสระ $\nu_1 = k-1$ และ $\nu_2 = k(n-1)$ และมีพารามิเตอร์ห่างศูนย์กลาง

$$\lambda = \frac{n\sum_{i=1}^{k}\tau_i^2}{\sigma^2}$$

สามารถเขียนแทนด้วย $F_{\alpha[k-1,k(n-1);\lambda]}$ และเรียก λ นี้ว่า พารามิเตอร์ห่างศูนย์กลาง (Noncentral parameter) โดยที่ภายใต้สมมติฐานว่าง H_0 พารามิเตอร์ห่างศูนย์กลาง λ เท่ากับ 0 หรือพิจารณาจากค่า p-value ซึ่งค่า p-value จะใช้เปรียบเทียบกับระดับนัยสำคัญ α ที่กำหนดไว้ โดยถ้าพบว่าถ้าค่าp-valu น้อยกว่าระดับนัยสำคัญ α ที่กำหนดไว้จะปฏิเสธสมมติฐานว่าง และถ้าค่าp-value น้อยกว่าระดับนัยสำคัญ α ที่กำหนดไว้ จะไม่สามารถปฏิเสธสมมติฐานว่าง

2.3 การแปลงข้อมูลโดยใช้เลขยกกำลัง (Power transformation)

เป็นการแปลงข้อมูลด้วยพารามิเตอร์ยกกำลัง ซึ่งสามารถที่จะนำเลขจำนวนจริงใดๆมาเป็นเลขยกกำลัง ซึ่งการแปลงข้อมูลแบบนี้สามารถแบ่งได้เป็น 2 รูปแบบคือ

รูปแบบที่ 1

$$z = y^{(\lambda)} = \begin{cases} y^{\lambda} & ;\text{if } \lambda \neq 0 \\ \log y & ;\text{if } \lambda = 0 \end{cases}$$

รูปแบบที่ 2

$$z = y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda} & ;\text{if } \lambda \neq 0 \\ \log y & ;\text{if } \lambda = 0 \end{cases}$$

การแปลงข้อมูลในรูปแบบที่ 1 จะเป็นการยกกำลังแบบง่ายๆแบบทั่วไปซึ่งภายหลังได้มีการปรับปรุงรูปแบบการแปลงข้อมูลแบบที่ 1 เป็นการแปลงข้อมูลรูปแบบที่ 2 โดย Box และ Cox ในปี ค.ศ. 1964 สำหรับในรูปแบบที่ 2นี้นอกจากจะนำไปใช้ในการแก้ไขปัญหาเกี่ยวกับข้อตกลงเบื้องต้นในเรื่องการแจกแจงแล้ว สามารถนำไปใช้ในการแก้ปัญหาข้อตกลงเบื้องต้นอื่นๆ เช่น การแก้ไขปัญหาเกี่ยวกับข้อมูลผิดปกติที่เกิดขึ้นในแผนแบบการทดลอง

2.4 การตรวจสอบข้อมูลผิดปกติในแผนแบบการทดลอง

การตรวจสอบข้อมูลตอบสนองที่เป็นค่าผิดปกติในแผนแบบการทดลอง จะใช้เทคนิค Boxplot เป็นเกณฑ์ในการตรวจสอบ โดยที่เทคนิค Boxplot จะให้รายละเอียดของค่าสถิติเพื่อตรวจสอบสำหรับการแจกแจง นั่นคือจะ plot ค่ามัธยฐาน เปอร์เซนต์ไทล์ที่ 25 เปอร์เซนต์ไทล์ที่ 75 และให้ค่าข้อมูลที่เป็นค่าผิดปกติ(Outlier) นั่นคือ ค่าที่สูงมากหรือค่าที่ต่ำมากจากค่ากลาง

การสร้าง Boxplot จะใช้ค่าสถิติ 5 ค่าด้วยกันคือ

1. ค่าต่ำสุดของข้อมูลที่ยังไม่ต่ำผิดปกติ
2. ควอไทล์ที่ 1 (Q_1)
3. ค่ามัธยฐานหรือควอไทล์ที่ 2 (Q_2)
4. ควอไทล์ที่ 3 (Q_3)
5. ค่าสูงสุดของข้อมูลที่ยังไม่สูงผิดปกติ

โดยที่ ค่าต่ำสุด= เปอร์เซนต์ไทล์ที่ 25 ของข้อมูล หรือควอไทล์ที่ 1(Q_1)

ค่ากลาง= ค่ามัธยฐานหรือเปอร์เซนต์ไทล์ที่ 50 ของข้อมูล หรือควอไทล์ที่ 2(Q_2)

ค่าสูงสุด= เปอร์เซนต์ไทล์ที่ 75 ของข้อมูล หรือควอไทล์ที่ 3(Q_3)

ความกว้างของ Box = $Q_3 - Q_1$ = IQR (Interquartile Range)

ค่าสูงสุดของข้อมูลที่ยังไม่สูงผิดปกติ คือ ค่าสูงสุดของข้อมูลชุดนั้นๆที่มีค่าไม่เกิน $Q_3+1.5IQR$

ค่าต่ำสุดของข้อมูลที่ยังไม่ต่ำผิดปกติ คือ ค่าต่ำสุดของข้อมูลชุดนั้นๆที่มีค่าไม่ต่ำกว่า $Q_1-1.5IQR$

ค่าของข้อมูลตอบสนองค่าใดมีค่าสูงกว่า $Q_3 + 1.5IQR$ และ มีค่าต่ำกว่า $Q_1 - 1.5IQR$ จะถือว่าข้อมูลตอบสนองค่านั้นเป็นค่าผิดปกติ(Outlying Observation) โดยที่ IQR คือ ค่าพิสัยควอไทล์ (Inter Quartile Range) มีค่าเท่ากับ $Q_3 - Q_1$

2.5 เกณฑ์ที่ใช้ในการเปรียบเทียบวิธีการแปลงข้อมูล

สำหรับเกณฑ์ที่ใช้ในการพิจารณาวิธีการแปลงข้อมูลที่เหมาะสมในการแก้ปัญหาข้อมูลที่มีค่าผิดปกติมีดังนี้

1. ค่าสัดส่วนของจำนวนค่าผิดปกติที่ลดลงภายหลังการแปลงข้อมูล

2. ค่าสัดส่วนของข้อมูลภายหลังการแปลงข้อมูลมีการแจกแจงแบบปกติและความแปรปรวนเท่ากัน

3. ค่าสัดส่วนของการปฏิเสธสมมติฐานว่างของการทดสอบเอฟ

4. ค่าอำนาจการทดสอบของการทดสอบเอฟ

โดยวิธีที่เหมาะสมที่สุดในการแก้ปัญหาข้อมูลที่มีค่าผิดปกติแต่ละสถานการณ์ที่ศึกษาจะต้องมีค่าสัดส่วนของจำนวนค่าผิดปกติที่ลดลงภายหลังการแปลงข้อมูล ค่าสัดส่วนของข้อมูลภายหลังการแปลงมีการแจกแจงแบบปกติและความแปรปรวนเท่ากันที่สูง สามารถควบคุมความผิดพลาดประเภทที่ 1 ได้และมีค่าอำนาจการทดสอบของการทดสอบเอฟที่สูงเช่นกัน