บทที่ 2
เอกสารและผลงานวิจัยที่เกี่ยวข้อง

ในบทนี้ผู้วิจัยได้นำเสนอเอกสารและผลงานวิจัยที่เกี่ยวข้อง โดยแบ่งเนื้อหาออกเป็น 4 ตอน ดังนี้

ตอนที่ 1 ความรู้ทั่วไปเกี่ยวกับค่าแผกเหล่า
ตอนที่ 2 ความรู้ทั่วไปเกี่ยวกับการตรวจหาค่าแผกเหล่า
ตอนที่ 3 การตรวจหาค่าแผกเหล่าโดยใช้ชุดสี่ (tetrads)
ตอนที่ 4 งานวิจัยที่เกี่ยวข้อง

ตอนที่ 1 ความรู้ทั่วไปเกี่ยวกับค่าแผกเหล่า

1. ธรรมชาติ และจุดกำเนิดของค่าแผกเหล่า (The nature and origin of outlier)

Grubbs (1969) ได้ให้ความหมายของค่าแผกเหล่า ว่า เป็นค่าสังเกตที่เบี่ยงเบนจากชุดของข้อมูลทั้งหมด โดยค่าแผกเหล่าอาจเกิดจากการผิดพลาดของการอ่าน การบันทึก หรือ การคำนวณค่าในชุดข้อมูล ซึ่งความผิดพลาดเหล่านี้ถ้าผู้วิจัยสังเกตเห็น จะสามารถแก้ไขได้ เช่น การนำค่านั้นออกจากชุดข้อมูล หรือแทนค่าที่ถูกต้องลงไป แต่ในบางสถานการณ์ผู้วิจัยไม่สามารถอธิบายลักษณะที่แท้จริงของค่าแผกเหล่าได้ ฉะนั้นจึงไม่สามารถหาวิธีแก้ไขที่เหมาะสมมาใช้แต่แนวทางหนึ่งซึ่งผู้วิจัยสามารถทำได้ คือ การพิจารณาว่าค่าแผกเหล่านั้นเกิดขึ้นอย่างสุ่มหรือไม่ โดยการประเมินรูปแบบของความเบี่ยงเบนของชุดข้อมูลสุ่ม ที่สร้างโดยโมเดลความน่าจะเป็นของสมมุติฐานเบื้องต้น

จากการศึกษาของ Anscombe (1960), Grubb (1969), Barnett (1978), Bookman and Cook (1983), Barnett (1983) และ Hawkins (1980) สรุปได้ว่าแหล่งที่ทำให้เกิดความเบี่ยงเบนในการเก็บรวบรวมข้อมูล สามารถจำแนกได้เป็น 3 แหล่ง คือ

1. การเบี่ยงเบนภายใน (Inherent Variability) หมายถึง ค่าสังเกตที่มีความเบี่ยงเบน หรือความหลากหลาย เนื่องมาจากลักษณะทางธรรมชาติของประชากรที่ไม่สามารถควบคุมได้ เช่น การวัดความสูงของผู้ชาย ผลการเก็บรวบรวมข้อมูลจะแสดงให้เห็นความหลากหลายของความสูงตามลักษณะพื้นเมืองของประชากร (และมีเหตุผลที่จะทำให้ข้อมูลนั้นมีการแจกแจงแบบปกติ)

2. ความผิดพลาดของการวัด (Mesurement Error) ความเบี่ยงเบนประเภทนี้เกิดขึ้นเนื่องจากความไม่เหมาะสมของเครื่องมือ , ขาดความรอบคอบในการสังเกตค่า , ความผิดพลาดในการบันทึก รวมถึงลักษณะการวัดที่มีความคลาดเคลื่อน ซึ่งความเบี่ยงเบนประเภทนี้ผู้วิจัยสามารถควบคุมได้

3. ความผิดพลาดจากการดำเนินการ (Execution Error) เป็นความเบี่ยงเบนที่เกิดขึ้นเนื่องจากความไม่สมบูรณ์ในขั้นตอนของการเก็บรวบรวมข้อมูล การที่ผู้วิจัยขาดความเอาใจใส่ในการเก็บรวบรวมข้อมูล อาจทำให้ได้กลุ่มตัวอย่างที่ลำเอียง หรือกลุ่มตัวอย่างที่ไม่เป็นตัวแทนที่ดีของประชากรที่ศึกษา ซึ่งความระมัดระวังจะช่วยลดความเบี่ยงเบนนี้ได้ แต่ผู้วิจัยมักไม่ทราบว่าเกิดความผิดพลาดประเภทนี้ขึ้นในชุดข้อมูลที่ศึกษา จนในบางครั้งทำให้ผู้วิจัยเห็นว่ามีความเหมาะสมที่จะทำการเปลี่ยนโมเดลพื้นฐานของประชากร เพื่อให้ครอบคลุมลักษณะของสมาชิกที่ผิดปกตินั้น

ค่าแผกเหล่าที่เกิดขึ้นในชุดข้อมูล อาจมีเหตุผลสมบูรณ์ที่จะอธิบายได้ว่า ค่าเหล่านี้เป็นผลมาจากลักษณะการเบี่ยงเบนภายในโดยธรรมชาติที่ไม่สามารถควบคุมได้ ถ้าสามารถแสดงโดยใช้วิธีทางสถิติว่า ไม่มีเหตุผลพอที่จะอธิบายว่าผลกระทบนี้ทำให้เกิดความไม่เหมาะสมกับโมเดลเบื้องต้น

2. วิธีการทางสถิติสำหรับดำเนินการกับค่าแผกเหล่า (Relevant statistic procedure for handling outlier)

2.1 การปรับค่าแผกเหล่า (Accomodation of Outlier) เป็นวิธีการทางสถิติที่พยายามศึกษาประชากรโดยการอ้างอิงกลุ่มตัวอย่างที่ศึกษา โดยไม่สนใจกับลักษณะที่ผิดรูปไปเนื่องมาจากการมีค่าแผกเหล่า เช่น ผู้วิจัยต้องการประมาณค่า

พารามิเตอร์ในโมเดลเบื้องต้นแต่คาดว่ามีค่าแผกเหล่ารวมอยู่ในชุดข้อมูล ฉะนั้นการเลือกใช้ตัวประมาณ ควรเลือกใช้ตัวประมาณที่ไม่ค่อยตอบสนองต่อค่าแผกเหล่า เช่น ใช้ค่ามัธยฐาน (median) เป็นตัวประมาณ (Hample, 1974)

2.2 การทดสอบความไม่สอดคล้อง (Discordancy Test) เป็นวิธีการทางสถิติอีกวิธีหนึ่ง สำหรับตรวจหาค่าแผกเหล่า ฉะนั้นในรายงานการวิจัยส่วนใหญ่ การทดสอบนี้จึงถูกอ้างถึงว่าเป็นการทดสอบเพื่อปฏิเสธค่าแผกเหล่า แต่ที่จริงแล้วการปฏิเสธไม่ใช่แนวทางเดียวที่สามารถทำได้ โดยอาจจำแนกการทดสอบความไม่สอดคล้องได้ดังนี้

2.2.1 การทดสอบความไม่สอดคล้องของค่าแผกเหล่าในกลุ่มตัวอย่างที่ศึกษาตัวแปรเดียว (Univariate Sample)

ก. เมื่อกลุ่มตัวอย่างมีการแจกแจงแบบแกมมา (gamma) และเอ็กซ์โปเนนเชียล (exponential) เช่น

- การตรวจหาค่าแผกเหล่า 1 ค่าด้านมาก เมื่อมีการแจกแจงแบบแกมมา

$$T_{\gamma 1} = \frac{Xn}{Xj}$$

เมื่อ X_n = ค่าแผกเหล่า

X_j = ผลรวมของค่าสังเกต

- การตรวจหาค่าแผกเหล่า 1 ค่าด้านมาก เมื่อมีการแจกแจงแบบเอ็กซ์โปเนนเชียล (exponential)

$$T_{\gamma 2} = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)}}$$

เมื่อ $X_{(n)}$ = ค่าแผกเหล่า

ข. เมื่อกลุ่มตัวอย่างมีการแจกแจงแบบปกติ เช่น

- การตรวจหาค่าแผกเหล่า 1 ค่าด้านมาก

$$T_{N1} = \frac{X_{(n)} - \bar{x}}{S}$$

เมื่อ $X_{(n)}$ = ค่าแผกเหล่า

$\bar{x}$ = ค่าเฉลี่ย

S = ส่วนเบี่ยงเบนมาตราฐาน

- การตรวจหาค่าแผกเหล่าหลายค่า (k > 2) ด้านมาก

$$T_{N2} = \frac{X(n-k+1)+.....+X(n)-kx}{S}$$

เมื่อ $x_{(n)}$ = ค่าแผกเหล่า

S = ส่วนเบี่ยงเบนมาตรฐาน

ค. เมื่อกลุ่มตัวอย่างมีการแจกแจงแบบอื่น ๆ

- การตรวจสอบค่าแผกเหล่า 1 ค่าด้านน้อย เมื่อมีการแจกแจงแบบ truncated exponential

$$t_{n2} = \frac{x_{(2)}-x_{(n)}}{x_{(n)}-x_{(1)}}$$

เมื่อ $x_{(n)}$ = ค่าแผกเหล่า

2.2.2 การทดสอบความไม่สอดคล้องของค่าแผกเหล่าในกลุ่มตัวอย่างที่ศึกษาตัวแปรพหุ (Multivariate Sample)

ก. เมื่อกลุ่มตัวอย่างมีการแจกแจงแบบปกติ

เราสามารถอธิบายได้ว่า ค่าแผกเหล่า $(X_{(n)})$ คือ ค่าสังเกตใด ๆ (X_i) ที่ทำให้

$R_i(\bar{x}, V) = (X_i - \bar{x})' V^{-1} (X_i - \bar{x})$ มีค่ามากที่สุด

หรือกล่าวได้ว่า $X_{(n)}$ เป็นค่าแผกเหล่า

ถ้า $R_{(n)}(\bar{x}, V) = (X_{(n)} - \bar{x})' V^{-1} (X_{(n)} - \bar{x}) = \max_{j=1,\ldots,n} R_j(\bar{x}, V)$ มีนัยสำคัญทางสถิติ

ข. เมื่อกลุ่มตัวอย่างมีการแจกแจงแบบเอ็กซ์โปเนนเชียล

$R(X) = X_1 + X_2 - \theta X_1 X_2$ อธิบายได้ว่า

ค่าสังเกตใด ๆ (X_{1i}, X_{2i}) ที่ทำให้ค่า $R(X)$ มีค่ามากที่สุดจะเป็นค่าแผกเหล่าด้านมาก

3. จุดประสงค์ในการตรวจหาค่าแผกเหล่า

ถ้าผู้วิจัยสนใจศึกษาเฉพาะลักษณะของโมเดลเบื้องต้น : F โดยไม่สนใจการเกิดและธรรมชาติของการปนเปื้อนในชุดข้อมูล ค่าแผกเหล่าจะเป็นเพียงค่าที่สร้างความรำคาญ โดยผู้วิจัยคาดว่าจะใช้วิธีการทางสถิติที่มีความแกร่งในการวิเคราะห์ข้อมูลชุดนั้น เพื่อให้ค่าแผกเหล่ามีผลกระทบต่อชุดข้อมูลน้อยที่สุด ซึ่งจุดประสงค์นี้คือการปรับข้อมูล (accomodation) และต่อมาได้มีวิธีทางสถิติมากมายที่มุ่งประยุกต์ใช้ชุดข้อมูล เพื่อประเมินลักษณะความไม่สอดคล้องของค่าแผกเหล่านั้น ซึ่งถ้าสามารถตรวจพบว่า มีข้อมูล 1 ค่า หรือมากกว่านั้นเป็นค่าแผกเหล่าที่ไม่สอดคล้องกับชุดข้อมูล แนวทางที่ผู้วิจัยสามารถปฏิบัติได้ นอกจากการใช้วิธีการปรับข้อมูล คือ การปฏิเสธ (หรือแทนที่) ค่าแผกเหล่า การปรับโมเดลเพื่อให้ครอบคลุมค่าแผกเหล่า ทำให้ค่าแผกเหล่าเป็นค่าที่สอดคล้องกับโมเดล ตลอดจนใช้ค่าแผกเหล่าเป็นตัวบ่งชี้ปัจจัยที่ไม่คาดหวัง (non - discordant)

4. แหล่งของค่าแผกเหล่าที่พบโดยทั่วไป อาจจำแนกได้ 5 แหล่ง คือ

4.1 ค่าแผกเหล่า ที่เกิดจากการรวมค่าสังเกตอื่นที่ไม่ใช่ประชากรที่ศึกษา มาเป็นกลุ่มตัวอย่าง

4.2 ค่าแผกเหล่า อาจเป็นค่าที่ถูกพิจารณาว่าเหมาะสม เมื่อพิจารณาความสัมพันธ์กับเรื่องที่ศึกษา แต่แท้จริงเป็นค่าสังเกตของโมเดลทางเลือก (Alternative Model)

4.3 ค่าแผกเหล่า อาจเป็นค่าที่คลาดเคลื่อนจากการสังเกตหรือการบันทึกข้อมูล

4.4 ค่าแผกเหล่า อาจเป็นผลมาจากค่าสุดโต่งในองค์ประกอบที่คลาดเคลื่อนของโมเดล การวัดแบบดั้งเดิม เช่น ผลการเดาถูกทุกข้อในข้อสอบแบบเลือกตอบ

4.5 ค่าแผกเหล่า อาจมาจากผลของความผิดพลาดในการเตรียมข้อมูลสำหรับวิเคราะห์ เช่น จาก การวัดแบบดั้งเดิม

คะแนนสังเกต = คะแนนจริง + ความคลาดเคลื่อน

(O = T + E)

แสดงให้เห็นว่า มีการยอมรับคะแนนที่สูงหรือมากกว่าคะแนนจริง เป็นข้อมูลสำหรับวิเคราะห์ และหาผลสรุป

5. ปัญหาของการตรวจหาค่าแผกเหล่า สามารถจำแนกได้เป็น 2 ประเภท คือ

1. การตรวจหาค่าแผกเหล่า ที่ตัดสินว่ามีจำนวนค่าแผกเหล่าน้อยกว่า จำนวนค่าแผกเหล่าที่มีอยู่จริง (masking)

2. การตรวจหาค่าแผกเหล่า ที่ตัดสินว่ามีจำนวนค่าแผกเหล่ามากกว่า จำนวนค่าแผกเหล่าที่มีอยู่จริง (swamping)

ตอนที่ 2 ความรู้ทั่วไปเกี่ยวกับการตรวจหาค่าแผกเหล่า

ในระยะเริ่มแรกได้มีการศึกษาปัญหาค่าแผกเหล่าในชุดข้อมูลปกติ คือ X_1 , X_2 , ...,X_n และสามารถแสดงในรูปสับเซตได้ 2 ลักษณะ คือ

$J_1 = \{X_{i1}, \ldots\ldots, X_{im}\}$ เมื่อ $X_{iq} \sim N(\mu, \sigma^2)$, $q = 1, \ldots\ldots, m$

เรียก J_1 ว่า "ค่าปกติ" (inlier) หรือ สับเซต m

$J_2 = \{X_{i1}, \ldots\ldots, X_{ik}\}$ เมื่อ $X_{il} \sim N(\mu_{il}, \sigma^2_{il})$, $l = 1, \ldots\ldots, k$

เรียก J_2 ว่า "ค่าแผกเหล่า" (outlier) หรือสับเซต k

โดยที่ $0 \leq k \leq n$ และ $m + k = n$

1. กระบวนการทางสถิติ ที่มีอำนาจสูงในการตรวจหาค่าแผกเหล่า 1 ค่า ในกลุ่มตัวอย่างที่มีการแจกแจงปกติ ขนาด n โดยกระบวนการเหล่านี้จะมีรูปแบบทั่วไป คือ

(1) หา X_α โดย $Pr(S(X) > X_\alpha /$ ไม่มีค่าแผกเหล่าปรากฏ$) = \alpha$ สำหรับค่าสถิติ $S(X)$

(2) กำหนดว่ามีค่าแผกเหล่าปรากฏ ถ้า $S(x) > X_\alpha$ หรือไม่ปรากฏค่าแผกเหล่า ถ้า $S(X) \leq X_\alpha$

ตัวอย่างของสถิติ $S(x)$ คือ

(a) Extreme Studentized Deviate (ESD)

$$Sx = \max_{i = 1, \ldots\ldots, m} |x_i - \bar{x}| / S$$

(b) Studentized Range (STR) $= (X_{(n)} - X_{(1)}) / S$

(c) Kurtosis (KUR) $= n \sum_{i=1}^{n} (x_i - \bar{x})^4 / \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^2$

(d) R - Statistic (RTS) $= \max_{i = 1, \ldots, n} |X_i - a| / b$

เมื่อ $a = \sum_{i=k+1}^{n-k} x_{(i)} / (n - 2k)$ = trimmed mean

$b = \sum_{i=k+1}^{n-k} (x_{(i)} - a)^2 / (n - 2k - 1)$

Paulson (1975) อธิบายว่า ESD เป็นวิธีที่เหมาะสมที่สุดในการตรวจหาค่าแผกเหล่า 1 ค่า โดยเฉพาะการศึกษาปัญหาค่าแผกเหล่าในกรณีกลุ่มตัวอย่าง 2 กลุ่มที่มีปัญหาการเลื่อน (2 - sample slippage problem)

นอกจากนี้ได้มีนักสถิติบางท่าน ศึกษาวิธีการตรวจหาค่าแผกเหล่า มากกว่า 1 ค่า เช่น

Grubbs (1969) เสนอสถิติ

$S_{n-1,n}2 / S^2 = \sum_{i=1}^{n-2} (x_{(i)} - \bar{x}_2)^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2$ สำหรับตรวจสอบหาค่าแผกเหล่า 2 ค่าด้านมาก (1.1)

เมื่อ $\bar{x}_2 = \sum_{i=1}^{n-2} x_{(i)} / (n-2)$

$S_{1,2}^2 / S^2$ สำหรับตรวจหาค่าแผกเหล่า 2 ค่า ด้านน้อย (1.2)

$S_{1,n}^2 / S^2$ สำหรับตรวจหาค่าแผกเหล่า 1 ค่า ด้านน้อย (1.3)
และ 1 ค่าด้านมาก

Tietjen and moore (1973) ศึกษาวิธีการใน (1.2) และพัฒนาวิธีการสร้างตรวจหาค่าแผกเหล่า k ค่าด้านมาก โดยกำหนดให้

$$L_k = \sum_{i=1}^{n-k} (x_{(i)} - \bar{x}_{(k)})^2 / \sum_{i=1}^{n} (x_{(i)} - \bar{x})^2 \quad (1.4)$$

เมื่อ $\bar{x}_k = \sum_{i=1}^{n-k} x_{(i)} / (n - k)$

และพัฒนา $S_{1,n}^2 / S^2$ เพื่อตรวจหาค่าแผกเหล่า k ค่า ทั้งทางด้านน้อยและด้านมากของค่าเฉลี่ย โดยกำหนดให้

$$E_k^* = \sum_{i=1}^{n-k} (z_{(i)} - \bar{z}_k)^2 \Big/ \sum_{i=1}^{n} (z_i - \bar{z})^2 \qquad (1.5)$$

เมื่อ $Z_i = |x_i - \bar{x}|$, $i = 1, \ldots\ldots\ldots, n$

$$\bar{z}_k = \sum_{i=1}^{n-k} z_{(i)} / (n-k)$$

และMurphy (1951) อธิบายว่า ผลรวมของสองค่าที่มากที่สุดของค่ามาตรฐานของส่วนที่เหลือ (Studentized residual) สามารถใช้สถิติสำหรับตรวจหาค่าแผกเหล่า 2 ค่าทางขวาได้

2. กฎทั่วไปสำหรับกระบวนการตรวจหาค่าแผกเหล่าหลายค่า (General Formulation of Many Outlier Procedure)

พิจารณาสับเซต $I_0, \ldots\ldots\ldots, I_n$

เมื่อ $I_0 = \{x_1, \ldots\ldots, x_n\}$

$$I_{t+1} = I_t - x^{(t)}$$

โดยที่ $|I_t|$ = จำนวนค่าสังเกตใน I_t และ $x^{(t)}$ กำหนดโดย

$$\max_{x_i \in I_t} |x_i - \bar{x}(I_t)| = |x^{(t)} - \bar{x}(I_t)|$$

$$\bar{x}(I_t) = \sum_{x_i \in I_t} x_i / |I_t| \; ; \; t = 0, \ldots\ldots\ldots, k-1 \qquad (2.1)$$

ดังนั้น I_{t+1} จัดรูปแบบโดย ลบค่าที่ไกลที่สุดจากค่าเฉลี่ยของชุด I_t

ให้ $R_t = S(I_{t-1})$

เมื่อ $S(I_{t-1})$ แทนประเภทสถิติที่ใช้ตรวจหาค่าแผกเหล่า 1 ค่าใน (1.1) และประยุกต์ให้ I_{t-1} เป็นค่าตัวอย่าง (Sample point) เมื่อ $t = 1, \ldots\ldots\ldots, k$ จากการพิจารณาลักษณะการแจกแจงร่วม (Joint Distribution) ของ $R_1, \ldots\ldots\ldots, R_k$ และพิจารณาลักษณะการแจกแจงหลัก (Marginal Distrbution) ของ $R_1, \ldots\ldots\ldots, R_k$ จะสามารถคำนวณหาค่า β, $\lambda_1(\beta), \ldots\ldots, \lambda_k(\beta)$ เช่น

$\Pr(R_i > \lambda_i(\beta)) = \beta$; i =1 , ... , k (2.2)

และ $\Pr\left\{\bigcup_{i=1}^{k}[R_i > \lambda_i(\beta)]\right\} = \alpha$

เมื่อ α = ระดับนัยสำคัญ

ถ้าพบว่า $\left\{\bigcap_{i=1}^{k}[R_i \le \lambda_i(\beta)]\right\}$ จริง แสดงว่า ไม่มีค่าแผกเหล่า แต่ถ้า มีอย่างน้อย 1 ค่าของ i ที่ทำให้ $\{[R_i > \lambda_i(\beta)]\}$ จริง , i = 1, ... k

เช่น ถ้า $\left\{\bigcup_{i=1}^{k}[R_i > \lambda_i(\beta)]\right\}$ เป็นจริง

และ $l = \max_{i=1,\ldots,k}\{i: R_i > \lambda_i(\beta)\}$

จะได้ว่า $X^{(0)}$, $X^{(1)}$, ... , $X^{(l-1)}$ เป็นค่าแผกเหล่า

<u>จุดเด่นของขบวนการตรวจหาค่าแผกเหล่าหลายค่า</u> คือ

1. สามารถหาค่าแผกเหล่าตั้งแต่ 1 - k ค่า
2. ไม่เสียอำนาจการทดสอบมาก เมื่อเปรียบเทียบกับวิธีการที่ใช้สำหรับตรวจหาค่าแผกเหล่าที่เฉพาะเจาะจง (เช่น หาค่าแผกเหล่า 1 - 2 ค่า)

ตอนที่ 3 การตรวจหาค่าแผกเหล่าโดยใช้ชุดสี

เนื่องจากวิธีการตรวจหาค่าแผกเหล่าโดยใช้ค่าชุดสี และค่าชุดสีร่วมกับแผนภาพกึ่งปกติ เป็นวิธีการตรวจหาค่าแผกเหล่าที่พัฒนามาจากการตรวจหาค่าแผกเหล่าโดยอาศัยค่าส่วนที่เหลือ (residual) ซึ่งยังเป็นวิธีที่สามารถตรวจหาค่าแผกเหล่าได้ถูกต้องเพียง 1 ค่า ฉะนั้นในส่วนนี้ผู้วิจัยจึงขอเสนอแนวคิดเกี่ยวกับการตรวจหาค่าแผกเหล่าโดยใช้ ค่าส่วนที่เหลือ และวิธีการตรวจหาโดยใช้ค่าชุดสี และค่าชุดสีร่วมกับแผนภาพกึ่งปกติ ซึ่งจะตรวจหาค่าแผกเหล่าได้หลายค่าและมีความถูกต้องกว่าการใช้ค่าส่วนที่เหลือ ตลอดจนเสนอวิธีการพัฒนาเกณฑ์ที่ใช้ตัดสินจำนวนค่าแผกเหล่าที่ตรวจหาได้จากการใช้ชุดสี เพื่อความสะดวกและรวดเร็วในการนำไปใช้ตัดสินจำนวนค่าแผกเหล่าได้อย่างถูกต้อง

1. การตรวจหาค่าแผกเหล่าจากค่าส่วนที่เหลือ

โมเดล

$$\text{ให้ } Y = \{Y_{ij} / 1 \le i \le n, 1 \le j \le m\}$$

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \delta_{ij}$$

และกำหนดให้มีสับเซต T

โดยที่ $\delta_{ij} = 0$ สำหรับ $(i,j) \notin T$

$\delta_{ij} \neq 0$ สำหรับ $(i,j) \in T$

อธิบายได้ว่า ค่าสังเกตใดที่ไม่ได้อยู่ในเซต T เป็นค่าสังเกตปกติ (inliers)

ค่าสังเกตใดที่อยู่ในชุดเซต T เป็นค่าสังเกตแผกเหล่า (outliers)

การคำนวณ

$$Yij^* = Yij - \bar{Y}i. - \bar{Y}.j + \bar{Y}$$

เมื่อ Yij^* = ค่าประมาณ ส่วนที่เหลือของเซลล์ (i , j)

$$E(Yij^*) = \mu_{ij} = \delta_{ij} - \delta_{i.} - \delta_{.j} + \delta_{..}$$

เมื่อ $E(Yij^*)$ = ค่าคาดหวังของส่วนที่เหลือของเซลล์(i , j)

โดยที่ค่าคาดหวังของส่วนที่เหลือในเซลล์ที่เป็นข้อมูลแผกเหล่า อาจจะมีค่าเท่ากับหรือไม่เท่ากับศูนย์ก็ได้

เกณฑ์ที่พิจารณาค่าสังเกตแผกเหล่า

ในระยะเริ่มแรกเกณฑ์ที่ใช้พิจารณาข้อมูลแผกเหล่า คือ พิจารณาว่าเซลล์ใดที่มีค่าส่วนที่เหลือสูงผิดปกติ หรือ มีค่าแตกต่างจากเซลล์อื่นมาก ๆ เซลล์นั้นจะเป็นค่าแผกเหล่า

ตัวอย่างการคำนวณ

ตารางแสดง จำนวนผลผลิตของส้ม เมื่อใช้ปุ๋ยชนิด A และ B ในปริมาณที่แตกต่างกัน

	ปริมาณสาร B				รวม
ปริมาณสาร A	35	32	37	40	144
	29	29	34	36	128
	25	29	30	20	104
	19	25	25	35	104
	22	20	29	29	100
รวม	130	135	155	160	580

จากตาราง คำนวณหาค่าประมาณของส่วนที่เหลือ ได้ดังนี้

$$\text{จาก } Y_{ij*} = Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$$

$$Y_{11} = 35 - 36 - 26 + 29 = 2$$

$$\vdots \qquad\qquad \vdots$$

$$Y_{54} = 29 - 25 - 32 + 29 = 1$$

จะได้ ค่าประมาณส่วนที่เหลือ ทั้งหมด 20 ค่า คือ

2	-2	-1	1
0	-1	0	1
2	5	2	-9
-4	1	-3	6
0	-3	2	1

เมื่อพิจารณาจากค่าประมาณของส่วนที่เหลือ พบว่า - 9 เป็นค่าที่แตกต่างจากค่าประมาณของส่วนที่เหลืออื่นมากที่สุด จึงตัดสินได้ว่า $Y_{34} = 20$ เป็นค่าแผกเหล่า

แต่จะเห็นได้ว่า การค้นหาค่าแผกเหล่าโดยใช้ค่าส่วนที่เหลือนี้จะตรวจหาค่าแผกเหล่าได้อย่างชัดเจนเมื่อเกิดค่าแผกเหล่าเพียงค่าเดียวเท่านั้น ซึ่งต่อมาได้มีผู้พัฒนาเกณฑ์ที่ใช้ตัดสินค่าแผกเหล่า จากส่วนที่เหลือ ดังต่อไปนี้

Quenouille (1953)

$$\tau_1^2 = I\,(S_o - S_m) / S_o$$

เมื่อค่าวิกฤตของ τ_1^2 คือ $nP\,(F_{1,I} > \tau_1^2)$

Daniel (1960)

$$t^2 = \varepsilon^2_{max}\,[\,S_m / (I - 1)\,]$$

เมื่อค่าวิกฤตของ t^2 คือ $P\,(F_{1,I-1} > nt^2 / I^2)$

Goldsmith and Boddy (1973)

$$\tau_2^2 = (I - 1)\,(S_o - S_m) / S_m$$

เมื่อค่าวิกฤตของ τ_2^2 คือ $1.25\,(I - 1)\,P\,(F_{1,I-1} > \tau_2^2)$

โดยที่ Sc คือ ผลบวกกำลังสองของค่าส่วนที่เหลือของชุดข้อมูลทั้ง n ค่า

Sm คือ ผลบวกกำลังสองของค่าส่วนที่เหลือของชุดข้อมูลที่ตัดค่าสังเกตที่มีค่ากำลังสองของค่าส่วนที่เหลือมากที่สุดออก

I คือ ชั้นแห่งความเป็นอิสระ

จากการศึกษาของ John and Prescott (1975) พบว่าวิธีการทั้ง 3 วิธี คือ t^2, τ_1^2 และ τ_2^2 มีความสามารถในการตรวจสอบค่าสังเกตไม่แตกต่างกัน กล่าวคือ สามารถแสดงค่าแผกเหล่าได้เพียงค่าเดียว คือ ค่าสังเกตที่มีค่าของส่วนที่เหลือสูงสุดนั่นเอง

2. การตรวจหาค่าแยกเหล่าโดยใช้ค่าชุดสี่

ค่าชุดสี่ (Tetrads) เป็นวิธีการตรวจหาค่าแยกเหล่า ที่พัฒนามาจากข้อบกพร่องของการตรวจหาค่าแยกเหล่าโดยใช้ค่าส่วนที่เหลือ เนื่องจากวิธีดังกล่าวจะใช้เฉพาะ ค่าผลรวมในแต่ละแถว ค่าผลรวมในแต่ละหลัก และค่าผลรวมทั้งหมดเท่านั้น เพื่อ คำนวณหาค่าส่วนที่เหลือในแต่ละเซลล์ ($Y_{ij*} = Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} - \bar{Y}...$) แต่ในการคำนวณค่าชุดสี่จะใช้ทุกเซลล์ของข้อมูลมาคำนวณ เพื่อศึกษาลักษณะความเบี่ยงเบนของข้อมูลใดๆจากชุดข้อมูลทั้งหมด ซึ่งค่าชุดสี่สามารถอธิบายได้จากสูตร

$$T_{ij.eg} = Y_{ij} - Y_{ej} - Y_{ig} - Y_{eg}$$

เมื่อ $i, e = 1,2,.....n$

$j, g = 1,2,.....m$

สำหรับเซลล์ใด ๆ เมื่อ

$i \neq e$, $j \neq g$ $T_{ij.eg}$ เรียกว่า **ชุดสี่ที่เหมาะสม** (proper tetrads)

$i = e$ หรือ $j = g$ $T_{ij.eg}$ เรียกว่า **ชุดสี่ที่ไม่เหมาะสม** (improper tetrads) และมีค่าเท่ากับศูนย์เสมอ

ค่าคาดหวังของค่าชุดสี่ คือ

$$\tau_{ij.eg} = \mu_{ij} - \mu_{ej} - \mu_{ig} + \mu_{eg}$$

ถ้าเซลล์ทั้งสี่ตามความหมายของชุดสี่ มีลักษณะเชิงบวก จะได้ว่า $\tau_{ij.eg} = 0$ และจะเรียกว่า **ค่าชุดสี่ที่ถูกต้อง** (Clean Tetrads) ถ้าทั้ง 3 เซลล์ คือ (e,j), (i,g) และ (e,g) มีลักษณะเชิงบวก จะได้ว่า ค่าของชุดสี่ เป็นค่าประมาณที่ไม่เอียงเอนของ δ_{ij} และมีการแจกแจงแบบปกติ มีค่าเฉลี่ยเท่ากับ δ_{ij} และมีค่าความแปรปรวนเท่ากับ $4\sigma^2$

$$\tau_{ij.eg} \sim N(\delta_{ij}, 4\sigma^2)$$

และเรียกว่า **ค่าชุดสี่ที่ปนเปื้อน** (contaminated tetrad) ถ้าเซลล์ใดเซลล์หนึ่ง จากทั้ง

3 เซลล์ดังกล่าว มีลักษณะที่ไม่เป็นเชิงบวกและสามารถคำนวณค่าส่วนที่เหลือ จากค่าเฉลี่ยของค่าชุดสี่ ดังนี้

$$Y_{ij\bullet} = \sum_{e} \sum_{g} T_{ij.eg} / nm$$

วิธีการตรวจหาค่าค่าแผกเหล่าโดยใช้ค่าชุดสี่ สามารถสรุปได้ 4 ขั้นตอน คือ

ขั้นที่ 1 คำนวณค่าชุดสี่ ($T_{ij.eg}$) ของแต่ละเซลล์ (i , j)

ตัวอย่าง การตรวจหาค่าแผกเหล่าจากตาราง 2 ทางขนาด 3 x 3

i \ j	1	2	3
1	9	10	13
2	8	6	32
3	50	12	14

จาก $T_{ij.eg} = Y_{ij} - Y_{ej} - Y_{ig} + Y_{eg}$

เมื่อ i = 1 , j = 1 i = 1 , j = 2 i = 1 , j = 3

e = 2 , 3 , g = 2 , 3 e = 2 , 3 ; 1 = 1 , 3 e = 2 , 3 ; g = 1 , 2

$T_{11,22} = 9 - 8 - 10 + 6 = -3$ $T_{12,21} = 3$ $T_{13,21} = -20$

$T_{11,23} = 9 - 8 - 13 + 32 = 17$ $T_{12,23} = 23$ $T_{13,22} = -23$

$T_{11,32} = 9 - 50 - 10 + 12 = -39$ $T_{12,31} = 39$ $T_{13,31} = 40$

$T_{11,33} = 9 - 50 - 13 - 14 = 40$ $T_{12,33} = -1$ $T_{13,32} = 1$

เมื่อ i = 2 , j = 1 i = 2 , j = 2 i = 2 , j = 3

e = 1 , 3 ; g = 2 , 3 e = 1 , 3 ; g = 1 , 3 e = 1 , 3 ; g = 1 , 2

$T_{21,12} = 3$ $T_{22,11} = -3$ $T_{23,11} = 20$

$T_{21,13} = -20$ $T_{22,13} = -23$ $T_{23,12} = 14$

$T_{21,32} = -36$ $T_{22,31} = 36$ $T_{23,31} = 60$

$T_{21,33} = 60$ $T_{22,33} = -24$ $T_{23,32} = 24$

เมื่อ $i=3 ; j = 1$ $i = 3 ; j = 2$ $i = 3 ; j = 3$

$e = 1,2 ; g = 2,3$ $e = 1,2 ; g = 1,3$ $e = 1,2 ; g = 1,2$

$T_{31,12} = 39$ $T_{32,11} = -39$ $T_{33,11} = -40$

$T_{31,13} = 40$ $T_{32,13} = 1$ $T_{33,12} = -1$

$T_{31,22} = 36$ $T_{32,21} = -36$ $T_{33,21} = -60$

$T_{31,23} = 60$ $T_{32,23} = 24$ $T_{33,22} = -24$

ขั้นที่ 2 หาค่ามัธยฐานของค่าชุดสี่ $(Q_{2(i,j)})$

จากค่าชุดสี่ $(T_{ij,eg})$ ของแต่ละเซลล์ สามารถหาค่า $Q_{2(i,j)}$ ได้ดังนี้

$Q_{2(1,1)} = 7$ $Q_{2(1,2)} = 13$ $Q_{2(1,3)} = -9.5$

$Q_{2(2,1)} = -8.5$ $Q_{2(2,2)} = -13$ $Q_{2(2,3)} = 22$

$Q_{2(3,1)} = 39.5$ $Q_{2(3,2)} = -17.5$ $Q_{2(3,3)} = -32$

ขั้นที่ 3. นำค่าสัมบูรณ์ของค่ามัธยฐานของค่าชุดสี่ $(|Q_{2(i,j)}|)$ มาจัดลำดับ

จากค่า $Q_{2(i,j)}$ ของแต่ละเซลล์ สามารถหาค่า $(|Q_{2(i,j)}|)$ ได้ดังนี้

$|Q_{2(1,1)}| = 7$ $|Q_{2(1,2)}| = 13$ $|Q_{2(1,3)}| = 9.5$

$|Q_{2(2,1)}| = 8.5$ $|Q_{2(2,2)}| = 13$ $|Q_{2(2,3)}| = 22$

$|Q_{2(3,1)}| = 39.5$ $|Q_{2(3,2)}| = 17.5$ $|Q_{2(3,3)}| = 32$

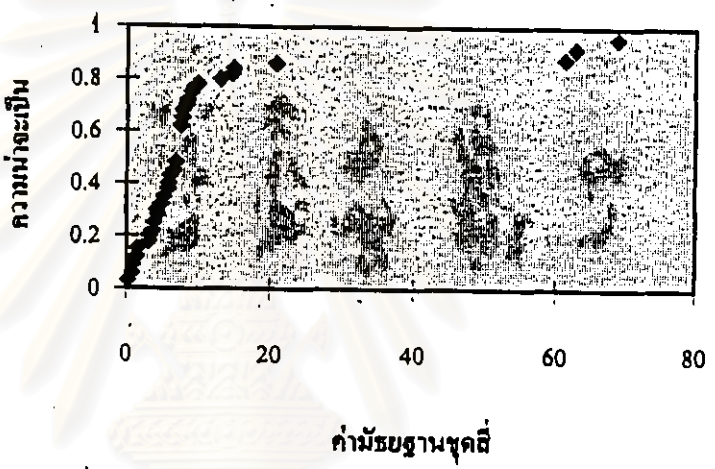
ฉะนั้น สามารถจัดลำดับ $|Q_{2(i,j)}|$ ได้ดังนี้

39.5 , 32 , 22 , 17.55 , 13 , 13 , 9.5 , 8.5 , 7

ขั้นที่ 4. พิจารณาเซลล์ที่เป็นค่าแผกเหล่า ซึ่งจากการศึกษาของ Bradu และ Hawkins เสนอว่า เซลล์ที่มีค่า ($|Q_{2(i,j)}|$) มาก ๆ จะเป็นค่าแผกเหล่า

3. การตรวจหาค่าแผกเหล่าโดยใช้ค่าชุดสี่ร่วมกับแผนภูมิเส้นกึ่งปกติ

การตรวจหาค่าแผกเหล่าโดยใช้วิธีนี้มีขั้นตอนในการคำนวณเช่นเดียวกับวิธีที่สอง แต่ภายหลังจากคำนวณค่าสัมบูรณ์ของค่ามัธยฐานได้ จะนำค่านี้ไปทำแผนภูมิเส้นกึ่งปกติ (Half Normal plots)



ค่าสังเกตใดที่เป็นค่าสังเกตแผกเหล่า จุดของค่าสังเกตนั้นจะกระจายออกจากเส้นของแผนภูมิเส้นกึ่งปกติดังกล่าว

การพัฒนาเกณฑ์สำหรับตัดสินจำนวนค่าแผกเหล่าที่คำนวณได้จากค่าชุดสี่

เกณฑ์ที่พัฒนาขึ้นสำหรับการวิจัยครั้งนี้ เป็นการพัฒนาเกณฑ์โดยใช้พื้นฐานของค่าชุดสี่ และค่าขอบเขตบนของช่วงความเชื่อมั่น 99.00 % , 99.50 % , 99.90 % , 99.99 % และ 99.999 % ของค่าส่วนเบี่ยงเบนมาตรฐานของกลุ่มตัวอย่าง เป็นหลักในการพัฒนา ซึ่งสามารถสรุปเป็นขั้นตอนได้ดังต่อไปนี้

การพัฒนาเกณฑ์การตัดสินค่าแผกเหล่า

ขั้นที่ 1 คำนวณค่าเฉลี่ยเลขคณิต ส่วนเบี่ยงเบนมาตรฐานและตำแหน่งมัธยฐานของชุดข้อมูลมีค่าแผกเหล่ารวมอยู่

i \ j	1	2	3
1	9	10	13
2	6	8	32
3	50	12	14

จากตาราง $\bar{x}$ = 17.11

S.D. = 14.49

ขั้นที่ 2 คำนวณค่า X_1 , X_2 , X_3, X_4 และ X_5 จากสูตร

$$X_i = \bar{x} + Z\ S.D.$$

ระดับความเชื่อมั่น 99.00 % : $X_1 = \bar{x} + 2.33$ S.D. จะได้ X_1 = 50.87

ระดับความเชื่อมั่น 99.50 % : $X_2 = \bar{x} + 2.58$ S.D. จะได้ X_2 = 54.49

ระดับความเชื่อมั่น 99.90 % : $X_3 = \bar{x} + 3.09$ S.D. จะได้ X_3 = 61.88

ระดับความเชื่อมั่น 99.99 % : $X_4 = \bar{x} + 3.72$ S.D. จะได้ X_4 = 71.01

ระดับความเชื่อมั่น 99.999 % : $X_5 = \bar{x} + 4.27$ S.D. จะได้ X_5 = 78.98

ขั้นตอนที่ 3 นำค่า X_i คำนวณได้ในขั้นที่ 2 แทนลงในตำแหน่งของค่ามัธยฐานของชุดข้อมูลทีละค่า

ขั้นตอนที่ 4 คำนวณค่าชุดสี่ จากสูตร $T_{ij.eg} = Y_{ij} - Y_{ej} - Y_{ig} + Y_{eg}$ ของเซลล์ที่ได้จากขั้นที่ 3

เมื่อ X_1 = 50.87

$T_{32.11}$ = -0.13

$T_{32.13}$ = 31.87

$T_{32.21}$ = -1.13

$T_{32.23}$ = 60.87

เมื่อ X_2 = 54.49

$T_{32.11}$ = 3.49

$T_{32.13}$ = 43.49

$T_{32.21}$ = 2.49

$T_{32.23}$ = 60.49

เมื่อ X_3 = 61.88

$T_{32.11}$ = 10.88

$T_{32.13}$ = 50.88

$T_{32.21}$ = 9.88

$T_{32.23}$ = 71.88

เมื่อ $X_4 = 71.01$ เมื่อ $X_5 = 78.98$

$T_{32.11} = 20.01$ $T_{32.11} = 27.98$

$T_{32.13} = 60.01$ $T_{32.13} = 67.98$

$T_{32.21} = 19.01$ $T_{32.21} = 26.98$

$T_{32.32} = 81.01$ $T_{32.32} = 88.98$

ขั้นที่ 5 หาค่ามัธยฐานของค่าชุดสี่ และค่าสัมบูรณ์ของค่ามัธยฐานของค่าชุดสี่

ที่ระดับความเชื่อมั่น 99.00% หรือเมื่อ $z_1 = 2.33$ จะได้ DC 1 $|Q_{2.1}| = 20.50$

ที่ระดับความเชื่อมั่น 99.50% หรือเมื่อ $z_2 = 2.58$ จะได้ DC 2 $|Q_{2.2}| = 23.49$

ที่ระดับความเชื่อมั่น 99.90% หรือเมื่อ $z_3 = 3.09$ จะได้ DC 3 $|Q_{2.3}| = 30.88$

ที่ระดับความเชื่อมั่น 99.99% หรือเมื่อ $z_4 = 3.72$ จะได้ DC 4 $|Q_{2.4}| = 40.01$

ที่ระดับความเชื่อมั่น 99.999% หรือเมื่อ $z_5 = 4.27$ จะได้ DC 5 $|Q_{2.5}| = 47.98$

จากค่า DC 1 , DC 2 , DC 3 , DC 4 และ DC 5 ที่ได้ จะเป็นเกณฑ์สำหรับตัดสินจำนวนค่าแผกเหล่า ที่ได้จากการคำนวณค่าชุดสี่ โดยค่าสัมบูรณ์ของค่ามัธยฐานของค่าชุดสี่ของค่าสังเกตใด ๆ ที่มีค่ามากกว่า ค่าสัมบูรณ์ของค่ามัธยฐานของค่าชุดสี่ที่สร้างขึ้น (DC i) จะตัดสินได้ว่าค่าสังเกตค่านั้นเป็นค่าแผกเหล่า

ตอนที่ 4 งานวิจัยที่เกี่ยวข้อง

Daniel (1959) ศึกษาการสร้างแผนภูมิเส้นกึ่งปกติ (half normal plot) ในแผนแบบการทดลองแบบแฟคทอเรียล (factorial design) ขนาด 2^p เมื่อ p คือจำนวนแฟคเตอร์ โดยพิจารณาลำดับของค่าสัมบูรณ์ของค่าผลกระทบรวม (effect totals) และไม่คำนึงว่ามีผลกระทบจริง (real effects) เกิดขึ้น ผลของการสร้างกราฟระหว่าง ค่าลำดับ กับ ค่าความน่าจะเป็นที่เหมาะสม พบว่า ค่าสังเกตปกติจะเรียงตัวเป็นเส้นตรง ผ่านจุดกำเนิด ค่าสังเกตที่เบี่ยงเบนออกจากเส้นตรง จะบ่งชี้ว่าเป็นค่าแผกเหล่า โดยใช้แผนภูมิเส้นกึ่งปกตินี้ จะแสดงผลที่ชัดเจน กรณีนี้เกิดค่าแผกเหล่าเพียงค่าเดียว กล่าวคือ เมื่อมีค่าแผกเหล่าเกิดขึ้น ค่าสัมบูรณ์ของค่าผลกระทบรวมจะมีขนาดใหญ่ และกราฟจะไม่ผ่านจุดกำเนิด ทำให้ผู้วิจัยทราบว่ามี

ค่าแผกเหล่า เกิดขึ้นในข้อมูล แต่ไม่สามารถบอกได้ว่าค่าสังเกตใดเป็นค่าแผกเหล่า และเมื่อค่าแผกเหล่ามากกว่า 2 ค่า วิธีการสร้างแผนภูมิเส้นกึ่งปกติจะยังไม่สามารถตรวจหา ค่าแผกเหล่าที่เกิดขึ้นได้อย่างชัดเจน

Bradu and Hawkins (1982) ได้ศึกษาวิธีการตรวจหาค่าแผกเหล่าโดยใช้ค่าชุดสี่ (tetrads) และพบว่า การตรวจหาค่าแผกเหล่า โดยใช้การสร้างแผนภูมิเส้นกึ่งปกติ (half normal plots) ของค่าส่วนที่เหลือที่ Danel (1959) , Birnbaum (1959) และ Wilk กับ Gnanadesikan (1968) ศึกษา ยังไม่สามารถตรวจหาค่าแผกเหล่าได้ถูกต้องมากกว่าการพิจารณาค่าส่วนที่เหลือเพียงอย่างเดียว และไม่สามารถตรวจหาค่าแผกเหล่าหลาย ๆ ค่าได้ แต่การสร้างแผนภูมิเส้นกึ่งปกติของค่ามัธยฐานของค่าชุดสี่ จะสามารถตรวจหาค่าแผกเหล่าหลาย ๆ ค่าได้ชัดเจนกว่า

John M.ORR (1996) ได้ทำการสำรวจวิธีการอันหลากหลายที่ใช้สำหรับตรวจหาและจัดการกับค่าแผกเหล่าและศึกษาลักษณะความแปรปรวนของข้อมูลขนาดใหญ่ เมื่อตัดค่าแผกเหล่าออกจากชุดข้อมูล โดยทำการศึกษาในชุดข้อมูลที่มีการแจกแจงแบบปกติ 2 ตัวแปร ผลการศึกษา พบว่า

1. ไม่มีข้อตกลงที่แน่นอนระหว่างนักวิจัย ที่จะตัดค่าแผกเหล่าออกจากชุดข้อมูลที่ศึกษา
2. นักวิจัยส่วนใหญ่รายงานว่า การตรวจหาค่าแผกเหล่าโดยใช้การสังเกตดีกว่าการตรวจหาโดยใช้เทคนิควิธีการทางคณิตศาสตร์
3. การกำจัดค่าแผกเหล่าออกจากชุดข้อมูล ส่งผลต่อขนาดอิทธิพลในแต่ละกรณีศึกษา แต่พบว่าค่าแผกเหล่าจะไม่มีผลต่าความแปรปรวนที่ศึกษากับข้อมูลขนาดใหญ่