บทที่ 2

ระเบียบวิธีที่ใช้ในการวิจัย

ในการศึกษาเปรียบเทียบวิธีการจำแนกกลุ่ม 2 กลุ่ม ในการวิเคราะห์ตัวแปรพหุ มีวิธีการดังต่อไปนี้

2.1 การคัดเลือกตัวแปรที่จะนำไปใช้ในการวิเคราะห์

ในกรณีที่ใช้ข้อมูลจริงในการวิเคราะห์ ขั้นแรกจะคัดเลือกตัวแปรพื้นฐานที่คาดว่าจะมีอิทธิพลต่อการจำแนกกลุ่มโดยจะพิจารณาเป็น 2 กรณี คือ

2.1.1 กรณีตัวแปรอิสระเป็นข้อมูลเชิงปริมาณ จะพิจารณาความสัมพันธ์ของตัวแปรอิสระด้วยกัน (Independent Variable ; Y_i , Y_j) จากเมตริกซ์สหสัมพันธ์ (Correlation Matrix) ถ้าตัวแปรอิสระ Y_i กับ Y_j คู่ใดมีค่าสัมบูรณ์ของสหสัมพันธ์ $|r_{YiYj}|$ เข้าใกล้ 1 ถือว่าตัวแปรอิสระคู่นั้นมีสหสัมพันธ์กันค่อนข้างสูง จะใช้ตัวแปรอิสระตัวใดตัวหนึ่งแทนได้โดยตัดอีกตัวหนึ่งออกไป

การคำนวณค่าสหสัมพันธ์

กำหนดให้ Y_1 และ Y_2 เป็นตัวแปรเชิงปริมาณสหสัมพันธ์ r ระหว่างตัวแปร Y_1 และ Y_2 สามารถหาได้จากสูตร

$$r = \frac{\Sigma Y_1Y_2 - \Sigma Y_1\ \Sigma Y_2/N}{\sqrt{\{\left[\Sigma Y_1^2 - (\Sigma Y_1)^2/N\right]\left[\Sigma Y_2^2 - (\Sigma Y_2)^2/N\right]\}}}$$

หรือ

$$r = \frac{N\ \Sigma Y_1Y_2 - \Sigma Y_1\Sigma Y_2}{\sqrt{\left[N\ \Sigma Y_1^2 - (\Sigma Y_1)^2\right]\left[N\ \Sigma Y_2^2 - (\Sigma Y_2)^2\right]}}$$

ค่าสหสัมพันธ์มีคุณสมบัติดังนี้ คือ

1. ค่าสหสัมพันธ์ไม่มีหน่วยหรือมาตรา ทำให้สามารถเปรียบเทียบความสัมพันธ์ระหว่างคู่ของตัวแปรต่าง ๆ ได้โดยตรง

2. สัมประสิทธิ์สหสัมพันธ์มีค่าตั้งแต่ -1 ถึง +1 ตัวแปรทั้งคู่จะมีความสัมพันธ์กันโดยสมบูรณ์ ถ้า r เท่ากับ +1 หรือ -1 และตัวแปรทั้งคู่จะไม่มีความสัมพันธ์กันเลย ถ้า r เท่ากับ 0 ถ้า r มีค่าเป็นลบ แสดงว่าเมื่อ Y_1 เพิ่มขึ้น Y_2 จะลดลงหรือในทางตรงกันข้าม เมื่อ Y_1 ลดลง Y_2 จะเพิ่มขึ้น

การทดสอบนัยสำคัญทางสถิติของค่า r

หลังจากที่คำนวณค่าสัมประสิทธิ์สหสัมพันธ์แล้วต้องทดสอบความมีนัยสำคัญของสัมประสิทธิ์สหสัมพันธ์ด้วย เพื่ออ้างอิงกลับไปหากลุ่มประชากรว่าตัวแปรทั้ง 2 มีความสัมพันธ์กันจริง โดยการตั้งสมมติฐานดังนี้

H_o : $\rho = 0$

H_A : $\rho \neq 0$

หรือ H_o : ตัวแปรทั้งสองไม่มีความสัมพันธ์กัน

H_A : ตัวแปรทั้งสองมีความสัมพันธ์กัน

สถิติที่ใช้ในการทดสอบคือ

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

เมื่อ r = ค่าสัมประสิทธิ์สหสัมพันธ์

n = จำนวนตัวอย่างทั้งหมด

df = n-2

จะยอมรับสมมติฐาน (H_o) ถ้าค่า r จากการคำนวณมีค่าน้อยกว่าค่า r จากตารางที่ระดับนัยสำคัญที่กำหนด แต่ถ้าค่า r จากการคำนวณเท่ากับหรือมากกว่าค่า r จากตาราง จะปฏิเสธสมมติฐาน (H_o) หมายความว่ามีความสัมพันธ์ระหว่างตัวแปรคู่นั้น

2.1.2 กรณีตัวแปรอิสระเป็นข้อมูลเชิงคุณภาพแบบไบนารีจะใช้การทดสอบไคสแควร์ เพื่อทดสอบความเป็นอิสระระหว่างตัวแปรทั้งสอง (Test of Independence)

การแจกแจงอันเนื่องจากคุณลักษณะสองลักษณะของคนหรือสิ่งใดสิ่งหนึ่งสามารถแสดงด้วยตารางการณ์จร (Contingency Table) ซึ่งแบ่งเป็น r แถว และ c สดมภ์ การทดสอบไคสแควร์มีขั้นตอนในการทดสอบดังนี้

H_o : ตัวแปรทั้งสองเป็นอิสระต่อกัน

H_A : ตัวแปรทั้งสองไม่เป็นอิสระต่อกัน

สถิติที่ใช้ในการทดสอบคือ

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

เมื่อ O_{ij} = ความถี่ของข้อมูลที่ได้จากการสังเกตในแถวที่ i ของตัวแปรที่ 1 และสดมภ์ที่ j ของตัวแปรที่ 2

E_{ij} = ความถี่คาดหวังในแถวที่ i ของตัวแปรที่ 1 และสดมภ์ที่ j ของตัวแปรที่ 2 ภายใต้สมมติฐาน H_o ดังนั้นจะได้ว่า

$$E_{ij} = \frac{|X_{i.}|\,|X_{.j}|}{X_{..}}$$

$X_{i.}$ = ผลรวมของความถี่ในแถวที่ i ของตัวแปรที่ 1

$X_{.j}$ = ผลรวมของความถี่ในสดมภ์ที่ j ของตัวแปรที่ 2

$X_{..}$ = ผลรวมของความถี่ทั้งหมดหรือจำนวนข้อมูลทั้งหมด

r = จำนวนแถวของตัวแปรที่ 1

c = จำนวนสดมภ์ของตัวแปรที่ 2

df = $(r-1)(c-1)$

จะยอมรับสมมติฐาน (H_o) ถ้าค่า χ^2 จากการคำนวณมีค่าน้อยกว่าค่า χ^2 จากตารางที่ระดับนัยสำคัญที่กำหนด แต่ถ้า χ^2 จากการคำนวณเท่ากับหรือมากกว่าค่า χ^2 จากตารางจะปฏิเสธสมมติฐาน (H_o) หมายความว่าตัวแปรทั้งสองไม่เป็นอิสระต่อกัน

ในกรณีที่แต่ละตัวแปรที่นำมาทดสอบ เป็นตัวแปรชนิดที่แบ่งเป็น 2 ลักษณะเท่านั้น (Dichotomous Variable) นั่นคือ ข้อมูลอยู่ในตารางการณ์จร ขนาด 2x2 ซึ่งมีองค์ความเป็นอิสระ เท่ากับ 1 และมีค่าขนาดความถี่คาดหวังบางช่องของตารางการณ์จรมีค่าไม่ถึง 20 จะใช้วิธีการคำนวณค่าไคสแควร์ด้วยสูตรการปรับแก้ของเยทส์ (Yate's Correction) ดังนี้

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(|O_{ij}-E_{ij}| - 0.5)^2}{E_{ij}}$$

ในกรณีของตารางการณ์จร แบบ 2x2 ค่าไคสแควร์อาจคำนวณได้โดยใช้สูตร ดังนี้

$$\chi^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

ในกรณีที่ความถี่คาดหวังและจำนวนตัวอย่างทั้งหมดมีขนาดเล็ก ๆ จะต้องหาทางแก้ไขโดยใช้สูตรการปรับแก้ของเยทส์ เมื่อข้อมูลอยู่ในตารางการณ์จร ขนาด 2x2 ซึ่งมีองค์ความเป็นอิสระเท่ากับ 1 จะได้สูตรการปรับแก้ของเยทส์ เป็นดังนี้

$$\chi^2_{correction} = \frac{n(ad-bc-o.5n)^2}{(a+c)(b+d)(a+b)(c+d)}$$

เมื่อ n = ขนาดตัวอย่าง

a = ความถี่ของค่าสังเกตในแถวที่ 1 และสดมภ์ที่ 1

b = ความถี่ของค่าสังเกตในแถวที่ 1 และสดมภ์ที่ 2

c = ความถี่ของค่าสังเกตในแถวที่ 2 และสดมภ์ที่ 1

d = ความถี่ของค่าสังเกตในแถวที่ 2 และสดมภ์ที่ 2

ตารางที่ 2.1 แสดงลักษณะของข้อมูลในตารางการณ์จร ขนาด 2x2

ตัวแปรที่ 2 / ตัวแปรที่ 1	สดมภ์ที่ 1	สดมภ์ที่ 2	รวม
แถวที่ 1	a	b	a+b
แถวที่ 2	c	d	c+d
รวม	a+c	b+d	n

2.2 วิธีวิเคราะห์การจำแนกกลุ่ม

ในการศึกษาเปรียบเทียบวิธีการจำแนกกลุ่ม 2 กลุ่มในการวิเคราะห์ตัวแปรพหุจะใช้วิธีการวิเคราะห์ 2 วิธีด้วยกัน คือ

2.2.1 วิธี Linear Discriminant Function (LDF)

สามารถแบ่งเป็น 2 กรณี คือ

1. ใช้เฉพาะข้อมูลเชิงปริมาณในการวิเคราะห์
2. ใช้ทั้งข้อมูลเชิงปริมาณและข้อมูลไบนารีในการวิเคราะห์

2.2.2 วิธี Optimum Allocation Rule

วิธี Linear Discriminant Function (LDF)

ตัวแปรอิสระที่ต้องการศึกษามี P ตัว

$$\underset{\sim}{X}' = [X_1 \ X_2 \ \dots \ X_p]$$

นำค่าจากตัวแปรอิสระและตัวแปรตามไปวิเคราะห์เพื่อสร้างสมการจำแนกกลุ่ม

$$Y = A_1 X_1 + A_2 X_2 + \dots + A_p X_p \quad \dots\dots\dots\dots\dots(1)$$

เมื่อ

Y = คะแนนการจำแนกกลุ่ม

A_i = ค่าสัมประสิทธิ์ของสมการจำแนกกลุ่มระหว่างตัวแปรแต่ละตัวต่อตัวแปรตามที่ต้องการศึกษา (Discriminant Function Coefficient)

X_i = ค่าของตัวแปรอิสระแต่ละตัวในสมการ

$i = 1, 2, \ldots p$ คือ ลำดับของตัวแปรอิสระในสมการ

กำหนดให้

π_1 = ประชากรกลุ่มที่ 1

π_2 = ประชากรกลุ่มที่ 2

μ_{1Y} = ค่าเฉลี่ยของ Y ซึ่งได้มาจากค่า X ของ π_1

μ_{2Y} = ค่าเฉลี่ยของ Y ซึ่งได้มาจากค่า X ของ π_2

μ_1 = ค่าคาดหวังของตัวแปรอิสระ X ที่มาจาก $\pi_1 = E(X/\pi_1)$

μ_2 = ค่าคาดหวังของตัวแรปอิสระ X ที่มาจาก $\pi_2 = E(X/\pi_2)$

จากสมการ (1) พิจารณาผลรวมเชิงเส้น

$$\underset{(1x1)}{Y} = \underset{(1xp)}{A'} \underset{(px1)}{X} \qquad \text{..........(2)}$$

จะได้

$$\mu_{1Y} = E[Y/\pi_1] = E[A'X/\pi_1] = A'\mu_1 \qquad \text{..................(3)}$$

$$\mu_{2Y} = E[Y/\pi_2] = E[A'X/\pi_2] = A'\mu_2 \qquad \text{..................(4)}$$

ความแปรปรวนของ Y จากทั้ง 2 ประชากร คือ

$$\sigma_Y^2 = Var(A'X) = A' Cov(X) A$$

$$= A' \Sigma A \qquad \text{..............(5)}$$

การหาสมการที่เหมาะสมในการจำแนกกลุ่มประชากรทั้ง 2 กลุ่ม ออกจากกันได้ผลมากที่สุดก็คือ การหาค่า A ที่ทำให้ λ มีค่ามากที่สุด โดยที่

$$\lambda = \frac{(\text{ระยะทางระหว่างค่าเฉลี่ยของ } Y)^2}{\text{ความแปรปรวนของ } Y}$$

$$= \frac{(\mu_{1Y}-\mu_{2Y})^2}{A'\Sigma A} \qquad \ldots\ldots\ldots\ldots\ldots(6)$$

ในทางปฏิบัติค่า μ_1, μ_2 และ Σ เป็นค่าที่เรามักจะไม่ทราบ ดังนั้นถ้าเราสุ่มตัวอย่าง จาก π_1 และ π_2 มาจำนวน n_1 และ n_2 ตามลำดับแล้ววัดค่าสังเกต

$$X' = \left[\begin{array}{cccc:cccc} x_{111} & x_{112} & \cdots & x_{11n_1} & x_{211} & x_{212} & \cdots & x_{21n_2} \\ x_{121} & x_{122} & \cdots & x_{12n_1} & x_{221} & x_{222} & \cdots & x_{22n_2} \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ x_{1p1} & x_{1p2} & \cdots & x_{1pn_1} & x_{2p1} & x_{2p2} & \cdots & x_{2pn_2} \end{array}\right]$$

แบ่งเมตริกซ์ X ออกเป็น 2 ส่วน ดังนี้

$$X' = \left[X_1' \,\vdots\, X_2'\right]$$

โดยที่ X_1' และ X_2' เป็นเมตริกซ์ของค่าสังเกต ซึ่งแต่ละเมตริกซ์มีเวคเตอร์ค่าเฉลี่ยดังนี้

$$\bar{x}_1' = [\,\bar{x}_{11}\ \bar{x}_{12}\ \ldots\ \bar{x}_{1p}]$$

$$\bar{x}_2' = [\,\bar{x}_{21}\ \bar{x}_{22}\ \ldots\ \bar{x}_{2p}]$$

และประมาณ covariance matrix Σ ด้วย pooled covariance matrix S_*

$$\underset{(p\times p)}{S_*} = \left[\frac{(n_1-1)}{(n_1-1)+(n_2-1)}\right] S_1 + \left[\frac{(n_2-1)}{(n_1-1)+(n_2-1)}\right] S_2$$

เมื่อ

$$\underset{(p\times p)}{S_1} = \frac{1}{n_1-1} \sum_{j=1}^{n_1} (X_{1j}-\bar{X}_1)(X_{1j}-\bar{X}_1)'$$

$$\underset{(p\times p)}{S_2} = \frac{1}{n_2-1} \sum_{j=1}^{n_2} (X_{2j}-\bar{X}_2)(X_{2j}-\bar{X}_2)'$$

ดังนั้นจากสมการ (6) จะได้

$$\hat{\lambda} = \frac{A'(\bar{X}_1-\bar{X}_2)(\bar{X}_1-\bar{X}_2)'A}{A' S_* A} \qquad \text{..............(7)}$$

$\hat{\lambda}$ จะมีค่ามากที่สุดเมื่อ $\frac{\partial \hat{\lambda}}{\partial A} = 0$

จะได้ $c(\bar{X}_1-\bar{X}_2) - S_* A = 0$

$$\hat{A} = cS_*^{-1}(\bar{X}_1-\bar{X}_2)$$

กำหนดให้ $c = 1$ ดังนั้น $\hat{A} = S_*^{-1}(\bar{X}_1-\bar{X}_2)$

แทนค่า $\hat{A}$ ในสมการ (7) จะได้ $\hat{\lambda} = (\bar{X}_1-\bar{X}_2)' S_*^{-1} (\bar{X}_1-\bar{X}_2)$

$$= D^2$$

ซึ่งเรียกว่า Mahalanobis D^2 หรือ Generalized Distance

จากสมการ (2) จะได้ผลรวมเชิงเส้น

$$\hat{Y} = \hat{A}'X = (\bar{X}_1-\bar{X}_2)' S_*^{-1} X \qquad \text{...............(8)}$$

สมการ (8) นี้สามารถใช้เป็นเกณฑ์ในการแยกค่าสังเกตที่ได้มานั้นว่าจะอยู่ในประชากรกลุ่ม π_1 หรือ π_2 ได้โดยให้

$Y_o = (\bar{X}_1-\bar{X}_2)' S_*^{-1} X_o$ เป็นค่าของ Discriminant Function ของค่าสังเกต X_o และกำหนดให้ m เป็นจุดกึ่งกลางระหว่าง $\bar{Y}_1$ และ $\bar{Y}_2$

ค่า $m = \frac{1}{2}(\bar{Y}_1+\bar{Y}_2)$ โดย $\bar{Y}_1 = \hat{A}'\bar{X}_1$, $\bar{Y}_2 = \hat{A}'\bar{X}_2$

$$m = \frac{1}{2}(\bar{X}_1-\bar{X}_2)' S_*^{-1} (\bar{X}_1+\bar{X}_2)$$

ดังนั้น จะได้กฎการจำแนกกลุ่ม คือ

จัดค่าสังเกต (X_o) อยู่ใน π_1 ถ้า $Y_o \geq m$

จัดค่าสังเกต (X_o) อยู่ใน π_2 ถ้า $Y_o < m$

วิธี Optimum Allocation Rule

เป็นวิธีวิเคราะห์ทางสถิติเพื่อประยุกต์ใช้ในการวิเคราะห์การจำแนกกลุ่มในกรณีที่ตัวแปรอิสระมีรูปแบบผสมระหว่างตัวแปรแบบไบนารี และตัวแปรแบบต่อเนื่อง

กำหนดให้ π_1 แทนประชากรกลุ่มที่ 1

π_2 แทนประชากรกลุ่มที่ 2

และให้ $W' = [X' \; Y']$ เป็นเวคเตอร์ของตัวแปรที่ใช้ศึกษา

$X' = [X_1 \; X_2 \; . \; . \; . \; X_q]$ เป็นเวคเตอร์ของตัวแปรแบบไบนารีขนาด q

$Y' = [Y_1 \; Y_2 \; . \; . \; . \; Y_p]$ เป็นเวคเตอร์ของตัวแปรแบบต่อเนื่องขนาด p

เวคเตอร์ขนาด q ของตัวแปรแบบไบนารีสามารถเขียนในรูปของ multinomial ดังนี้

$$Z' = [Z_1 \; Z_2 \; ... \; Z_k] \quad ; \quad k = 2^q$$

ดังนั้นตัวแปรไบนารี X แต่ละตัว จะกำหนด Multinomial cell ได้ด้วย $X' = [X_1\ X_2 \ldots X_q]$ ซึ่งจะตกอยู่ใน cell ที่ c โดยที่

$$c = 1 + \sum_{i=1}^{q} X_i \cdot 2^{(i-1)}$$

ให้ Y ของประชากร π_i มีการแจกแจงแบบพหุปกติด้วยค่าเฉลี่ย $\mu_i^{(m)}$ ใน cell ที่ m $(m = 1,2,\ldots k\ ;\ i = 1,2)$ และมีความแปรปรวน Σ ร่วมกัน ใน cell ทุก ๆ cell ของทั้ง 2 ประชากร นั่นคือ

$$(Y/Z_m = 1\ ,\ Z_j = 0\ ,\ j \neq m = 1,2, \ldots k) \sim N(\mu_i^{(m)}, \Sigma)$$

ในประชากร π_i ความน่าจะเป็นที่จะเกิดค่าสังเกตใน cell ที่ m คือ P_{im}

โดยวิธี Optimum likelihood ratio จะจัดค่าค่าสังเกต $(x' Y')$ อยู่ใน π_1 ถ้า

$$(\mu_1^{(m)} - \mu_2^{(m)})' \Sigma^{-1} \{Y - \frac{1}{2}(\mu_1^{(m)} + \mu_2^{(m)})\} \geq \log \frac{P_{2m}}{P_{1m}}$$

เมื่อ $$m = 1 + \sum_{i=1}^{q} X_i \cdot 2^{(i-1)}$$

นอกนั้นจะจัดให้ค่าสังเกต $(X' Y')$ อยู่ใน π_2

ในทางปฏิบัติจะไม่ทราบค่าพารามิเตอร์ $P_{im}\ \mu_i^{(m)}$ และ Σ ดังนั้นจะสุ่มตัวอย่างขนาด n_1 และ n_2 จาก π_1 และ π_2 ตามลำดับแล้ววัดค่าสังเกต

ให้ n_{1m} แทนจำนวนของค่าสังเกตจาก π_1 ที่ตกอยู่ใน cell ที่ m

n_{2m} แทนจำนวนของค่าสังเกตจาก π_2 ที่ตกอยู่ใน cell ที่ m

$Y_{ji}^{(m)}$ แทนเวคเตอร์ของตัวแปรแบบต่อเนื่องของค่าสังเกต ที่ j ใน cell ที่ m ของตัวอย่างจาก π_i

P_{im} แทนค่าความน่าจะเป็นที่ค่าสังเกต Y จาก π_i ตกอยู่ใน cell ที่ m

จะได้ค่าประมาณของ P_{im} $\mu_i^{(m)}$ และ Σ ตามลำดับ โดยที่

$$\hat{P}_{im} = \frac{n_{im}}{n_i}$$

$$\hat{\mu}_i^{(m)} = \bar{Y}_i^{(m)} = \frac{1}{n_{im}} \sum_{i=1}^{n_{im}} Y_{ji}^{(m)}$$

$$\hat{\Sigma} = v = \frac{1}{n_1+n_2-2k} \sum_{i=1}^{2} \sum_{m=1}^{k} \sum_{j=1}^{n_{im}} (Y_{ji}^{(m)} - \bar{Y}_i^{(m)})(Y_{ji}^{(m)} - \bar{Y}^{(m)})'$$

$$(i = 1,2 \; ; \; m = 1,2, \ldots k)$$

สมการจำแนกประเภทที่ได้มานั้นจะมีความสามารถในการแบ่งแยกกลุ่มได้ดีหรือไม่นั้นสามารถดูได้จากค่า Misclassification Error ซึ่งทราบได้จากการที่เปรียบเทียบว่าสิ่งที่เราศึกษาอยู่นั้น เป็นสมาชิกของประชากรกลุ่มหนึ่งและเมื่อใช้สมการจำแนกประเภทมาจัดการจำแนกแล้วกลับไปเป็นสมาชิกของอีกกลุ่มหนึ่งซึ่งทำให้ได้ผลที่ผิดจากความเป็นจริงไป

ตารางที่ 2.2 แสดงความผิดพลาดที่ได้จากการจำแนกกลุ่ม

ประชากรเดิม	จัดเป็นประชากร		รวม
	π_1	π_2	
π_1	n_{11}	n_{12}	n_1
π_2	n_{21}	n_{22}	n_2
รวม	n_1	n_2	n

เมื่อ n_{11} = จำนวนค่าสังเกตที่เดิมอยู่ใน π_1 และเราจัดเข้าพวก π_1 (จัดถูก)

n_{12} = จำนวนค่าสังเกตที่เดิมอยู่ใน π_1 และเราจัดเข้าพวก π_2 (จัดผิด)

n_{21} = จำนวนค่าสังเกตที่เดิมอยู่ใน π_2 และเราจัดเข้าพวก π_1 (จัดผิด)

n_{22} = จำนวนค่าสังเกตที่เดิมอยู่ใน π_2 และเราจัดเข้าพวก π_2 (จัดถูก)

$n = n_{11}+n_{12}+n_{21}+n_{22}$

ดังนั้น ความผิดพลาดจากการจำแนก (%) $= \dfrac{(n_{12}+n_{21})}{n} \times 100$