บทที่ 2

ระเบียบวิธีที่ใช้ในการวิจัย

ในงานวิจัยนี้ใช้กฎการจำแนกกลุ่ม (Classification rule) หรือเกณฑ์การจัดเข้ากลุ่มที่เหมาะสม (Optimal classification rule) ของ แอนด์เดอร์สัน (Anderson's classification function) และอัตราความผิดพลาดที่เกิดขึ้นจากการใช้กฎที่ได้มา ดังนั้น ในบทนี้จึงเสนอรายละเอียดดังนี้

2.1 การวิเคราะห์การจัดกลุ่ม และ การจำแนกกลุ่ม (Discriminant and classification analysis)

2.2 อัตราความผิดพลาดที่มีเงื่อนไข (Conditional Error rate)

2.3 วิธีการประมาณค่าอัตราความผิดพลาด 4 วิธี ดังนี้

2.3.1 วิธี R หรือ Resubstitution Estimator

2.3.2 วิธี U หรือ Leave-one-out Estimator

2.3.3 วิธี B หรือ Bootstrap Estimator

2.3.4 วิธี DS หรือ Shrunken-D Estimator

2.1 การวิเคราะห์การจัดกลุ่ม และ กฎการจำแนกกลุ่ม

เป็นวิธีการวิเคราะห์ที่มีวัตถุประสงค์จะคัดเลือกตัวแปรกลุ่มหนึ่งหรือชุดหนึ่งที่นักวิจัยคิดว่าตัวแปรเหล่านี้มีความสัมพันธ์กับสิ่งที่ต้องการศึกษาจนถึงขั้นที่สามารถแยกประชากรออกเป็นกลุ่มต่างๆ ได้แล้วนำเอาตัวแปรชุดนี้มาใช้ในการประมาณการเป็นสมาชิกของกลุ่ม สมการที่ได้คือสมการการจัดกลุ่ม (Discriminant Function) และการระบุสมาชิกของกลุ่มคือกฎการจำแนกกลุ่ม (Classification Rule) อาจกล่าวได้ว่า การวิเคราะห์การจัดกลุ่มคือการจำแนกวัตถุออกเป็นกลุ่มๆ ตามธรรมชาติของวัตถุที่อยู่เป็นหมู่พวกเดียวกันโดยอาจใช้เครื่องมือในการบ่งชี้ เช่น ตัวเลข, ภาพ, หรือสมการพีชคณิต เพื่อให้เห็นความแตกต่างระหว่างกลุ่มอย่างเด่นชัดที่สุดเท่าที่จะทำได้ ส่วนกฎการจำแนกกลุ่มหมายถึงการนำเอาวัตถุใดๆ จัดเข้ากลุ่มใดกลุ่มหนึ่งที่สมการการจัดกลุ่มได้แบ่งไว้ให้แล้วจะทำได้ถูกต้องดีเพียงใดนั้นขึ้นอยู่กับกฎเกณฑ์หรือวิธีที่จะนำเอามาใช้ในการจัดวัตถุเข้ากลุ่มว่าวิธีไหนจะดีหรือเหมาะสมที่สุด

2.1.1 การวิเคราะห์การจัดกลุ่มและกฎการจำแนกกลุ่ม 2 กลุ่มประชากร

การแบ่งวัตถุออกเป็น 2 กลุ่ม เช่น กลุ่มผู้มีปัญหาทางกระเพาะอาหารกับกลุ่มที่ไม่มีปัญหา การซื้อและไม่ซื้อสินค้าของผู้บริโภค ผู้มีเครดิตดีกับไม่ดี เป็นต้น การแบ่งวัตถุ เช่น คน สัตว์ สิ่งของ (object or individual) ออกเป็น 2 กลุ่ม เรียกว่า ประชากรกลุ่มที่ 1 (π_1) และ ประชากรกลุ่มที่ 2 (π_2) โดยในแต่ละกลุ่มจะมีการบันทึกคุณลักษณะประจำตัวของวัตถุแต่ละหน่วยที่จะอธิบายลักษณะของกลุ่มได้เรียกคุณลักษณะเหล่านี้ว่าตัวแปรอิสระ (Independent variable) เช่น การแบ่งกลุ่มของผู้ที่ซื้อและไม่ซื้อสินค้าของผู้บริโภค เวกเตอร์ของตัวแปรที่ใช้แสดงคุณลักษณะของกลุ่ม คือ

X = [พื้นฐานการศึกษา, ระดับรายได้, ขนาดครอบครัว, ความถี่ของการเปลี่ยนผลิตภัณฑ์ยี่ห้ออื่นในอดีต]

= [X_1 , X_2 , X_3 , X_4],

เวกเตอร์ X อาจมาจากประชากรกลุ่มที่ 1 ซึ่งมีฟังก์ชันความหนาแน่น (Probability density function , pdf) เป็น $f_1(x)$ หรือมาจากประชากรกลุ่มที่ 2 ซึ่งมี pdf เป็น $f_2(x)$ ถ้ามีวัตถุใหม่เพิ่มเข้ามาจะมีวิธีใดที่สะดวกและมีประสิทธิภาพในการจัดวัตถุนั้นให้เข้ากลุ่มใดกลุ่มหนึ่ง

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ \cdot \\ x_k \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3n} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ x_{k1} & x_{k2} & x_{k3} & \cdots & x_{kn} \end{bmatrix}$$

เราสามารถมองเมตริกซ์ X ได้ 2 ลักษณะ คือ

1. Set of row vector

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ \cdot \\ x_k \end{bmatrix}$$

โดยที่

$$\begin{aligned} x_1 &= [\, x_{11} \quad x_{12} \quad x_{13} \quad \cdots \quad x_{1n} \,] \\ x_2 &= [\, x_{21} \quad x_{22} \quad x_{23} \quad \cdots \quad x_{2n} \,] \\ x_3 &= [\, x_{31} \quad x_{32} \quad x_{33} \quad \cdots \quad x_{3n} \,] \\ &\cdot \\ &\cdot \\ &\cdot \\ x_k &= [\, x_{k1} \quad x_{k2} \quad x_{k3} \quad \cdots \quad x_{kn} \,] \end{aligned}$$

กรณีนี้ x_i คือตัวแปรตัวที่ i

2. Set of column vector หรือ set of observation

$$X = [\, x_1 , x_2 , x_3 , \ldots , x_n \,]$$

โดยที่

$$x_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ . \\ . \\ . \\ x_{k1} \end{bmatrix} \quad x_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \\ . \\ . \\ . \\ x_{k2} \end{bmatrix} \quad x_3 = \begin{bmatrix} x_{13} \\ x_{23} \\ x_{33} \\ . \\ . \\ . \\ x_{k3} \end{bmatrix} \quad \ldots x_n = \begin{bmatrix} x_{1n} \\ x_{2n} \\ x_{3n} \\ . \\ . \\ . \\ x_{kn} \end{bmatrix}$$

ในกรณีนี้ x_j คือค่าสังเกตของวัตถุหน่วยที่ j มีคุณลักษณะ k รายการ แต่ละรายการคือตัวแปร x_i

กำหนดสัญลักษณ์ต่างๆ ดังนี้

π_1 = ประชากรกลุ่มที่ 1
π_2 = ประชากรกลุ่มที่ 2
$f_1(x)$ = pdf. ของเวกเตอร์ x จากประชากรกลุ่มที่ 1
$f_2(x)$ = pdf. ของเวกเตอร์ x จากประชากรกลุ่มที่ 2
$c(2|1)$ = ความสูญเสียที่เกิดจากการจัดวัตถุเข้ากลุ่มที่ 2 เมื่อวัตถุนั้นมาจากกลุ่มที่ 1
$c(1|2)$ = ความสูญเสียที่เกิดจากการจัดวัตถุเข้ากลุ่มที่ 1 เมื่อวัตถุนั้นมาจากกลุ่มที่ 2
q_1 = ความน่าจะเป็นที่วัตถุนั้นจะมาจากกลุ่มที่ 1 (prior probability of π_1)
q_2 = ความน่าจะเป็นที่วัตถุนั้นจะมาจากกลุ่มที่ 2 (prior probability of π_2)
R_1 = เขตการตัดสินใจ จัดวัตถุเข้าพวกกลุ่มที่ 1
R_2 = เขตการตัดสินใจ จัดวัตถุเข้าพวกกลุ่มที่ 2
k = ขนาดของตัวแปรอิสระ
Ω = Sample Space = $R_1 \cup R_2$

นิยาม Pr(จัดเข้าพวกถูก) และ Pr(จัดเข้าพวกผิด) ได้ดังนี้

$$\Pr(\text{จัดเข้าพวกกลุ่มที่ 1 ถูกต้อง}) = \Pr(\text{วัตถุมาจากกลุ่ม1 จัดเข้าพวกกลุ่ม 1})$$
$$= \Pr(x \in R_1 \mid \pi_1).q_1$$
$$= \Pr(1|1).q_1$$

$$\Pr(\text{จัดเข้าพวกกลุ่มที่ 1 ผิด}) = \Pr(\text{วัตถุมาจากกลุ่ม2 จัดเข้าพวกกลุ่ม 1})$$
$$= \Pr(x \in R_1 \mid \pi_2).q_2$$
$$= \Pr(1|2).q_2$$

$$\Pr(\text{จัดเข้าพวกกลุ่มที่ 2 ถูกต้อง}) = \Pr(\text{วัตถุมาจากกลุ่ม2 จัดเข้าพวกกลุ่ม 2})$$
$$= \Pr(x \in R_2 \mid \pi_2).q_2$$
$$= \Pr(2|2).q_2$$

$$\Pr(\text{จัดเข้าพวกกลุ่มที่ 2 ผิด}) = \Pr(\text{วัตถุมาจากกลุ่ม1 จัดเข้าพวกกลุ่ม 2})$$
$$= \Pr(x \in R_2 \mid \pi_1).q_1$$
$$= \Pr(2|1).q_1$$

จากตารางของฟังก์ชันความสูญเสีย (Cost of missclassificstion) แสดงได้ดังนี้

true population	classification as 1	2
1	0	c(2|1)
2	c(1|2)	0

ดังนั้นจะได้ค่าคาดหวังของฟังก์ชันความสูญเสีย (Expected Cost of Misclassification , ECM) ดังนี้

$$ECM = c(2|1)P(2|1).q_1 + c(1|2)P(1|2).q_2$$

เนื่องจากกฎการจำแนกกลุ่มที่ดี ควรจะให้ค่าของ ECM มีค่าต่ำที่สุดเท่าที่จะทำได้ (Minimize ECM) ดังนั้นจึงสรุปเป็นทฤษฎี 2.1 ได้ดังนี้

ทฤษฎี 2.1 เขตการตัดสินใจ R_1 และ R_2 ที่มีผลให้ ECM มีค่าต่ำสุดคือเขตต่อไปนี้

1. จัดวัตถุเข้าพวกกลุ่มที่ 1 ถ้า

$$\frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)\,q_2}{c(2|1)\,q_1}$$

2. จัดวัตถุเข้าพวกกลุ่มที่ 2 ถ้า

$$\frac{f_1(x)}{f_2(x)} < \frac{c(1|2)\,q_2}{c(2|1)\,q_1}$$

<u>พิสูจน์</u>

จาก ECM $= c(1|2)P(1|2).q_2 + c(2|1)P(2|1)\cdot q_1$

$= c(1|2)q_2 \int_{R_1} f_2(x)dx + c(2|1)q_1 \int_{R_2} f_1(x)dx$

แต่

$\Omega = R_1 \cup R_2$

และ

$$1 = \int f_1(x)dx = \int_{R_1} f_1(x)dx + \int_{R_2} f_1(x)dx$$

$$\int_{R_2} f_1(x)dx = 1 - \int_{R_1} f_1(x)dx$$

ดังนั้นจะได้

$$ECM = c(1|2)q_2 \int_{R_1} f_2(x)dx + c(2|1)q_1\left[1 - \int_{R_1} f_1(x)dx\right]$$

$$= \int_{R_1}[c(1|2)q_2 f_2(x) - c(2|1)q_1 f_1(x)]dx + c(2|1)q_1$$

$$= \int_{R_1} kdx + c(2|1)q_1$$

จะพบว่า ECM เป็นบวกเสมอเพราะว่า q_1 , q_2 , $c(2|1)$ และ $c(1|2)$ เป็นบวกเสมอขณะที่ $f_1(x)$ และ $f_2(x)$ ไม่เป็นลบ (nonnegative) โดยที่ $f_1(x)$ และ $f_2(x)$ จะผันแปรไปตามเวกเตอร์ X

กำหนดให้ค่า $K = c(1|2)q_2 f_2(x) - c(2|1)q_1 f_1(x)$
ในที่นี้ ค่า K จะมีผลกระทบโดยตรงต่อค่า ECM ดังนั้นค่า ECM จะมีค่าต่ำสุดก็ต่อเมื่อ

$$c(1|2)q_2 f_2(x) \leqslant c(2|1)q_1 f_1(x)$$

ดังนั้นจะจัดวัตถุเข้าพวกกลุ่มที่ 1 ถ้า

$$\frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2).q_2}{c(2|1).q_1}$$

สำหรับการจัดเข้าพวกกลุ่มที่ 2 ก็พิสูจน์ได้ทำนองเดียวกัน

ในกรณีที่กล่าวถึงมาแล้วสามารถจำแนก ทฤษฎี 2.1 ได้ดังนี้

1. กรณีที่ค่าความน่าจะเป็นของวัตถุที่มาจากกลุ่ม 1 เท่ากับค่าความน่าจะเป็นของวัตถุที่มาจากกลุ่ม 2 (Equal Prior Probabilities) นั่นคือ $q_2/q_1 = 1$

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)}$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)}{c(2|1)}$$

2. กรณีที่ค่าความสูญเสียที่เกิดจากกลุ่มที่ 1 เท่ากับค่าความสูญเสียที่เกิดจากกลุ่มที่ 2 (Equal Cost of Missclassification)
นั่นคือ $c(1|2)/c(2|1) = 1$

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{q_2}{q_1}$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \frac{q_2}{q_1}$$

3. กรณีที่ค่าความน่าจะเป็นของวัตถุที่มาจากกลุ่ม 1 เท่ากับค่าความน่าจะเป็นของวัตถุที่มาจากกลุ่ม 2 และ ค่าความสูญเสียที่เกิดจากกลุ่มที่ 1 เท่ากับค่าความสูญเสียที่เกิดจากกลุ่มที่ 2

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq 1$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < 1$$

ในการวิเคราะห์การจำแนกกลุ่มโดยส่วนมากจะมีข้อตกลงเบื้องต้นของการแจกแจงคือการแจกแจงต้องเป็นการแจกแจงแบบปกติ ดังนั้นการแจกแจงของเวกเตอร์ X คือ การแจกแจงแบบพหุปกติ (Multivariate normal distribution)

2.1.2 กฎการจำแนกกลุ่ม 2 กลุ่ม กรณีประชากรมีการแจกแจงปกติ (Classification with two multivariate normal populations)

ในงานวิจัยนี้ มีข้อตกลงในการกำหนดความแปรปรวนร่วมของประชากรกลุ่ม 1 เท่ากับกลุ่ม 2 โดยกำหนดให้

$$\Sigma_1 = \Sigma_2 = \Sigma$$

$$X' = [x_1 , x_2 , x_3 , . . , x_k]$$

มีฟังก์ชันความหนาแน่นร่วม (joint pdf) เป็น การแจกแจงปกติดังนี้

$$f_1(x) = \frac{1}{(2\pi)^{k/2}\left|\Sigma\right|^{1/2}} \exp[-1/2(x-\mu_1)'\Sigma^{-1}(x-\mu_1)]$$

ดังนั้น

$$\frac{f_1(x)}{f_2(x)} = \exp[-1/2(x-\mu_1)'\Sigma^{-1}(x-\mu_1)+1/2(x-\mu_2)'\Sigma^{-1}(x-\mu_2)]$$

$$= \exp[(\mu_1-\mu_2)'\Sigma^{-1}x-1/2(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)]$$

จากทฤษฎี 2.1 จะได้กฎการจำแนกกลุ่ม 2.1.2 ดังนี้

$$R_1 : \exp[(\mu_1-\mu_2)'\Sigma^{-1}x_o 1/2(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)] \geqslant \frac{c(1|2)P_2}{c(2|1)P_1}$$

$$R_2 : \exp[(\mu_1-\mu_2)'\Sigma^{-1}x_o 1/2(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)] < \frac{c(1|2)P_2}{c(2|1)P_1}$$

หมายเหตุ x_o ในที่นี้ คือ ค่าสังเกตของวัตถุหน่วยใหม่

โดยทั่วไปไม่สามารถนำเอากฎนี้ไปใช้ได้เพราะไม่ทราบค่าพารามิเตอร์ μ_1, μ_2 และ Σ ดังนั้นแอนเดอร์สัน (Anderson) จึงแนะนำให้แทนค่าพารามิเตอร์ของประชากรด้วยค่าที่ได้จากตัวอย่าง คือ แทน μ_1 ด้วย $\bar{x}_1$ แทน μ_2 ด้วย $\bar{x}_2$ และ แทน Σ ด้วย S
โดยที่

$$\bar{x}_1 = \sum_{i=1}^{n_1} (x_1)/n_1$$

$$\bar{x}_2 = \sum_{i=n_1+1}^{n_1+n_2} (x_1)/n_2$$

$$S^{-1} = \frac{1}{n_1+n_2-2} \left[\sum_{i=1}^{n_1} (x_1-\bar{x}_1)(x_1-\bar{x}_1)' + \sum_{i=n_1+1}^{n_1+n_2} (x_1-\bar{x}_2)(x_1-\bar{x}_2)' \right]$$

ค่าประมาณของกฎการจำแนกกลุ่ม 2.1.2

จัด x_o ให้เข้ากลุ่มที่ 1 ถ้า

$$(\bar{x}_1-\bar{x}_2)'S^{-1}x_o - 1/2(\bar{x}_1-\bar{x}_2)'S^{-1}(\bar{x}_1-\bar{x}_2) \geq \ln\left|\frac{c(1|2)q_2}{c(2|1)q_1}\right|$$

จัด x_o ให้เข้ากลุ่มที่ 2 ถ้า เป็นอย่างอื่น

กำหนดให้

$$W = (\bar{x}_1-\bar{x}_2)'S^{-1}x_o - 1/2(\bar{x}_1-\bar{x}_2)'S^{-1}(\bar{x}_1-\bar{x}_2)$$

เรียกตัวสถิติ W นี้ว่า Anderson's classification function

ในงานวิจัยนี้ใช้ตัวสถิติ W ในกรณีที่ อัตราส่วนของค่าความสูญเสีย และ prior probability มีค่าเท่ากับ 1 ดังนั้น

$$\ln\left|\frac{c(1|2)q_2}{c(2|1)q_1}\right| = 0$$

ดังนั้นกฎการจำแนกกลุ่มที่ใช้ในงานวิจัยนี้คือ

จัด x_0 ให้เข้ากลุ่มที่ 1 ถ้า $W(x) \geq 0$
จัด x_0 ให้เข้ากลุ่มที่ 2 ถ้า $W(x) < 0$

2.2 อัตราความผิดพลาดที่มีเงื่อนไข (Conditional Error Rate)

เมื่อมีการจำแนกวัตถุให้กับกลุ่มต่างๆ ตามกฎการจำแนกกลุ่มที่สร้างขึ้นจากตัวอย่าง นักวิจัยก็ต้องการที่จะทราบค่าของอัตราความผิดพลาดที่เกิดขึ้นจากการใช้กฎนั้น ซึ่งขึ้นอยู่กับค่าที่ได้จากตัวอย่างคือ $\bar{x}_1$, $\bar{x}_2$ และ S ดังนั้นจึงเรียกอัตราความผิดพลาดชนิดนี้ว่า อัตราความผิดพลาดที่มีเงื่อนไข (Conditional Error Rate) การหาค่าที่แน่นอนของอัตราความผิดพลาดนี้หาค่าได้ลำบากเนื่องจากเราไม่ทราบลักษณะฟังก์ชันการแจกแจกของ W ดังนั้นจึงมีการใช้ Asymptotic expansion เข้ามาช่วยซึ่ง แอนด์เดอร์สัน ได้แสดงการพิสูจน์ดังนี้

พิสูจน์ กำหนดให้

$$U = x'\Sigma^{-1}(\mu_1-\mu_2) - 1/2(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$$

โดยมีข้อตกลงเบื้องต้นคือ x มีการแจกแจงแบบ $N(\mu_1,\Sigma)$ หรือ $N(\mu_2,\Sigma)$ ดังนั้น U จะมีการแจกแจงปกติด้วยค่าเฉลี่ย และ ความแปรปรวน ดังนี้

$$E_1(U) = \mu_1'\Sigma^{-1}(\mu_1-\mu_2) - 1/2\,(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$$

$$= 1/2\,(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$$

$$Var_1(U) = E_1(\mu_1-\mu_2)'\Sigma^{-1}(x-\mu_1)(x-\mu_1)'\Sigma^{-1}(\mu_1-\mu_2)$$

$$= (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$$

หมายเหตุ ค่าความแปรปรวนนี้ก็คือค่า Mahalanobis Square Distence, Δ^2 ระหว่างประชากรกลุ่มที่ 1 และ กลุ่มที่ 2

ดังนั้นจึงกล่าวได้ว่า U มีการแจกแจงแบบ $N(1/2\Delta^2, \Delta^2)$ ถ้า X แจกแจงแบบ $N(\mu_1, \Sigma)$ แต่ถ้า X แจกแจงแบบ $N(\mu_2, \Sigma)$ แล้ว

$$E_2(U) = \mu_2' \Sigma^{-1}(\mu_1-\mu_2) - 1/2\ (\mu_1+\mu_2)' \Sigma^{-1}(\mu_1-\mu_2)$$

$$= 1/2(\mu_2-\mu_1)' \Sigma^{-1}(\mu_1-\mu_2)$$

$$= -(1/2)\Delta^2$$

หมายเหตุ ค่าความแปรปรวน $Var_2(U)$ จะเท่ากับ $Var_1(U)$

ดังนั้น จึงกล่าวได้ว่า U มีการแจกแจงแบบ $N(-1/2\Delta^2, \Delta^2)$ ถ้า X มีการแจกแจงแบบ $N(\mu_2, \Sigma)$

ทฤษฎีที่กล่าวถึงค่าความน่าจะเป็นของการจำแนกกลุ่มผิด แอนด์เดอร์สันได้สรุปไว้ดังนี้

ทฤษฎี 2.2.1 ขณะที่ $n_1 \to \infty$, $n_2 \to \infty$ และ $n_1/n_2 \to$ a positive limit $(N = n_1+n_2)$

(1) $$Pr\{(w-1/2\Delta^2)/\Delta < U \mid \pi_1\}$$

$$= (U)-\phi(U)\{1/(2n_1\Delta^2)[U^3+(p-3)U-p\Delta]$$
$$+ 1/(2n_2\Delta^2)[U^3+2\Delta U^2+(p-3+\Delta^2)U+(p-2)\Delta]$$
$$+ 1/(4N)[4U^3+4\Delta U^2+(6p-6+\Delta^2)U+2(p-1)\Delta]\}+O(n^{-2})$$

หมายเหตุ

1. $n = n_1 + n_2 - 2$
2. $Pr\{-(w+1/2\Delta^2)/\Delta < U \mid \pi_2\}$ คือสมการที่ 1 ที่แทนค่า n_2 ด้วย n_1
3. ค่าของ $O(n^{-2})$ คือ เทอมของลำดับ (order) ที่ 2
4. กฎของ W คือ

จัดค่า x ให้กับกลุ่มที่ 1 ถ้า $W(x) \geq K$
จัดค่า x ให้กับกลุ่มที่ 2 ถ้า $W(x) < K$

<u>บทแทรก 2.2.1</u> กำหนดให้ $U = (c-(1/2)\Delta^2)/\Delta$ และ $-(c+(1/2)\Delta^2)/\Delta$ สำหรับ $K = 0$, $U = -(1/2)\Delta$ และ $n_1 = n_2$

$$(2) \qquad \Pr\{w \leqslant 0 \mid \pi_1 , \lim_{N \to \infty} n_1/n_2 = 1\}$$

$$= \Phi(-1/2\Delta)+1/2\phi((1/2)\Delta)[(p-1)/\Delta+p\Delta/4]+O(n^{-1})$$

$$= \Pr\{w \geqslant 0 \mid \pi_2 , \lim_{N \to \infty} n_1/n_2 = 1\}$$

ปกติเราจะไม่ทราบค่าของ Δ^2 ดังนั้นจึงใช้ค่า Sample Mahalanobis Squared Distance มีค่าดังนี้

$$(3) \qquad D^2 = (\bar{x}_1-\bar{x}_2)'S^{-1}(\bar{x}_1-\bar{x}_2)$$

มาประมาณค่าของ Population Mahalanobis Squared Distance, Δ^2 ค่าคาดหวังของ D^2 คือ

$$(4) \qquad E(D^2) = n/(n-K-1)[\Delta^2+p(1/n_1+1/n_2)]$$

<u>ทฤษฎี 2.2.2</u> ถ้า n_1/n_2 -> a positive limit ขณะที่ $N \to \infty$ จากทฤษฎีที่ 2.2.1 จะได้

$$(5) \qquad \Pr\{(w - 1/2\ D^2)/D \leqslant u \mid \pi_1\}$$

$$= \Phi(u)-\phi(u)\{1/n_1(u/2 - (p-1)/\Delta) +1/N[u^3/4 + (p-3/4)u]\}+O(n^{-2})$$

<u>หมายเหตุ</u> $\Pr\{-(w + 1/2D^2)/D \leqslant u \mid \pi_2\}$ คือ สมการที่ (5) ซึ่งแทน n_1 ด้วย n_2

พิจารณาค่าความน่าจะเป็นของการจำแนกกลุ่มผิดจากการใช้กฎการจำแนกกลุ่มที่ได้ ดังนั้นภายใต้เงื่อนไขของ $\bar{x}_1$, $\bar{x}_2$ และ S พร้อมกับ W มีการแจกแจงปกติ ภายใต้ ค่าเฉลี่ยที่มีเงื่อนไข (Conditional mean) ดังนี้

$$(6)\qquad E(W \mid \pi_1, \bar{x}_1, \bar{x}_2, S^{-1}) = [\mu_1 - 1/2(\bar{x}_1+\bar{x}_2)]' S^{-1}(\bar{x}_1-\bar{x}_2)$$

$$= \mu_1(\bar{x}_1, \bar{x}_2, S^{-1})$$

เมื่อ x มาจากประชากรกลุ่ม 1 หรือกลุ่ม 2 (π_i, i = 1,2)

ค่าความแปรปรวนภายใต้เงื่อนไข (Conditional variance) ของ $\bar{x}_1$ $\bar{x}_2$ และ S^{-1} คือ

$$(7)\qquad Var(W \mid \pi_1, \bar{x}_1, \bar{x}_2, S) = (\bar{x}_1-\bar{x}_2)' S^{-1}\Sigma S^{-1}(\bar{x}_1-\bar{x}_2)$$

$$= \sigma^2(\bar{x}_1, \bar{x}_2, S^{-1})$$

ซึ่งทั้งค่าเฉลี่ย และความแปรปรวนต่างก็เป็นฟังก์ชันของ $\bar{x}_1$, $\bar{x}_2$ และ S^{-1} กล่าวคือ

$$(8)\qquad \underset{n_1, n_2 \to \infty}{Plim}\ \mu_1(\bar{x}_1, \bar{x}_2, S) = (-1)^{1-1}(1/2)\Delta^2,$$

$$\underset{n_1, n_2 \to \infty}{Plim}\ \sigma^2(\bar{x}_1, \bar{x}_2, S) = \Delta^2$$

ดังนั้นเมื่อ K คือจุดแบ่งหรือจุดตัด (cut of point) ค่าความน่าจะเป็นของการจำแนกกลุ่มผิดบนเงื่อนไขของ $\bar{x}_1, \bar{x}_2$ และ S^{-1} (Probability of misclassification conditional on x_1, x_2 and S^{-1}) คือ

$$(9)\qquad P(2|1, K, \bar{x}_1, \bar{x}_2, S^{-1}) = [(K-\mu_1(\bar{x}_1, \bar{x}_2, S^{-1}))/\sigma(\bar{x}_1, \bar{x}_2, S^{-1})]$$

$$(10)\qquad P(1|2, K, \bar{x}_1, \bar{x}_2, S^{-1}) = 1- [(K-\mu_1(\bar{x}_1, \bar{x}_2, S^{-1}))/\sigma(\bar{x}_1, \bar{x}_2, S^{-1})]$$

ซึ่งในงานวิจัยนี้ พิจารณาค่า K = 0 และ X มาจากประชากรกลุ่มที่ 1 ดังนี้

$$(11)\qquad P(2|1) = \Phi\left[\frac{-\{\mu_1 - (\bar{x}_1 + \bar{x}_2)/2\}' S^{-1}(\bar{x}_1 - \bar{x}_2)}{\{(\bar{x}_1 - \bar{x}_2)' S^{-1}\Sigma S^{-1}(\bar{x}_1 - \bar{x}_2)\}^{1/2}}\right]$$

หมายเหตุ

$\Phi(t)$ = Cumulative distribution function ที่tของ N(0,1)

$\phi(t)$ = Probability density function ที่ t ของ N(0,1)

$$(12) \qquad p(1|2) = \Phi\left[\frac{\{\mu_2 - (\bar{x}_1+\bar{x}_2)/2\}' S^{-1}(\bar{x}_1 - \bar{x}_2)}{\{(\bar{x}_1-\bar{x}_2)' S^{-1}\Sigma S^{-1}(\bar{x}_1-\bar{x}_2)\}^{1/2}}\right]$$

2.3 การประมาณค่าอัตราความผิดพลาด (Estimation of Error Rates)

เนื่องจากอัตราความผิดพลาดอยู่ในรูปของฟังก์ชันที่ไม่ทราบค่าพารามิเตอร์ดังนั้นจึงมีผู้เสนอตัวประมาณค่าอัตราความผิดพลาดขึ้นมามากมาย แต่ในงานวิจัยนี้ขอนำเอาวิธีที่นิยมใช้กันโดยทั่วไปเป็นที่แพร่หลายมาศึกษาเปรียบเทียบ ดังนี้
วิธี R ไม่มีข้อจำกัดใดๆเกี่ยวกับการแจกแจงของประชากร และวิธี DS จะต้องอยู่ภายใต้การแจกแจงปกติ นอกจากนั้นยังมีตัวประมาณที่น่าสนใจถูกเสนอขึ้นมาใหม่อีก 2 ตัว คือ วิธี U และ วิธี B รายละเอียดของตัวประมาณใน วิธี R , วิธี U วิธี B และ วิธี DS มีดังนี้

2.3.1 ตัวประมาณในวิธี R หรือ (Resubstitution Estimator)

วิธี R นี้ถูกเสนอขึ้นมาโดย สมิท (Smith,1946) ซึ่งวิธีนี้เป็นสัดส่วนของค่าสังเกตจากตัวอย่างที่ถูกจัดกลุ่มให้ผิดกับขนาดตัวอย่างของกลุ่ม
กำหนดให้

$\hat{\alpha}^R$ คือ ตัวประมาณอัตราความผิดพลาดที่มีเงื่อนไข (Conditional Error Rate)

x_j ($j = 1,2,3, \ldots, n_1$) คือ เวกเตอร์ของตัวแปรสุ่ม X ที่มาจากประชากรกลุ่มที่ 1

$$h(x) = \begin{cases} 1 & \text{ถ้า } W(x) \leqslant 0 \\ 0 & \text{ถ้า } W(x) > 0 \end{cases}$$

ดังนั้น

$$(13) \qquad \hat{\alpha}^{R} = \sum_{j=1}^{n_1} h(x_j)/n_1$$

วิธีนี้จะให้ค่าที่ต่ำกว่าความเป็นจริง (underestimate) เพราะวิธีนี้นำเอาข้อมูลเดิม (original) มาใช้สร้างสมการการจำแนกกลุ่ม และนำเอาข้อมูลเดิมนี้ไปประมาณค่าอัตราความผิดพลาดจึงทำให้เกิดความเอนเอียง (biasness)

จากบทความของ ฮิล (Hill,1966) ได้สรุปรูปแบบของอสมการต่างๆ ที่เกี่ยวข้องกับวิธี R ดังนี้

ถ้า Λ คือ กฎไลลิฮูด (likelihood rule) ในการจำแนกกลุ่มดังนี้

$l(x) > K$ จัด x ให้เข้ากลุ่มที่ 1
$l(x) < K$ จัด x ให้เข้ากลุ่มที่ 2
$l(x) = K$ จัด x ให้เข้ากลุ่มที่ 1 ด้วยความน่าจะเป็น γ และให้เข้ากลุ่มที่ 2 ด้วย ความน่าจะเป็น $1 - \gamma$

แล้ว จะได้

$$E\{\hat{M}(\Lambda^*)\} < M(\Lambda) < E\{M(\Lambda^*)\}$$

<u>ความหมายของสัญลักษณ์</u>

q_1, q_2 = Prior Probability
$\Lambda(x)$ = ความน่าจะเป็นที่ค่าลังเกต x จากกลุ่ม 1 ถูกจัดให้อยู่กลุ่ม 2 โดยใช้กฎ Λ
$\alpha_1(\Lambda)$ = $E_1\{\Lambda(x)\}$ (optimum error of allocation for π_1)
α_2 = $E_2\{1-\Lambda(x)\}$
$M(\Lambda)$ = $q_1\alpha_1(\Lambda) + q_2\alpha_2(\Lambda)$ = ความน่าจะเป็นในการจำแนกกลุ่มผิดเฉลี่ยที่เหมาะสม (optimum average risk)
$\hat{M}(\Lambda)$ = $q_1\alpha_1(\Lambda) + q_2\alpha_2(\Lambda)$ = ตัวประมาณที่ไม่เอนเอียงของ $M(\Lambda)$
$M(\Lambda^*)$ = ความน่าจะเป็นในการจำแนกกลุ่มผิดเฉลี่ยที่แท้จริง (actual average risk)
$\hat{M}(\Lambda^*)$ = ตัวประมาณของ $M(\Lambda^*)$ ได้จากการแทนค่าพารามิเตอร์ที่ประมาณได้จากตัวอย่างลงใน $\hat{M}(\Lambda^*)$ (apparent averagr risk)

อธิบายได้ว่า ค่าเฉลี่ยของความน่าจะเป็นที่แท้จริง (actualprobability) ของการจำแนกกลุ่มผิดจากประชากรกลุ่มที่ 1 จะมีค่ามากกว่าความน่าจะเป็นที่เหมาะสม (optimum probability) และเมื่อ $\hat{M}(\lambda)$ เป็นตัวประมาณที่ไม่เอนเอียงของ $M(\lambda)$ ค่าเฉลี่ยที่ได้รับโดยการแทนตัวประมาณพารามิเตอร์ใน $M(\lambda^*)$ จะมีค่าน้อยกว่าความน่าจะเป็นที่เหมาะสม นั่นคือ ตัวประมาณที่ได้ ประมาณค่าได้ต่ำกว่าที่เป็นจริง (under estimate)
ถ้า m_1/n_1 คืออัตราส่วนของจำนวนตัวอย่างที่มาจากกลุ่มที่ 1ถูกจำแนกกลุ่มผิดจากการใช้กฎ λ^* แล้วจะได้

$$E(m_1/n_1) = Pr\{ \hat{1}(x_1) < K \} + 1/2\ Pr\{ 1(x_1) = K \}$$

$$\alpha_1(\lambda) = Pr\{ 1(x_1) < K \} + 1/2Pr\{ 1(x_1) = K \}$$

$$= Pr\{ 1(x) < K \} + 1/2Pr\{ 1(x) = K \}$$

$$E\{\alpha_1(\lambda^*) = Pr\{ \hat{1}(x) < K \} + 1/2Pr\{ \hat{1}(x) = K \}$$

กล่าวได้ว่า

$$E(m_1/n_1) < E\{ \alpha_1(\lambda^*) \}$$

$$E(m_1/n_1) < \alpha_1(\lambda) < E\{ \alpha_1(\lambda^*) \}$$

ถ้ากำหนดให้ $X \sim N(\mu_1, 1)$ ใน π_1 และ $N(\mu_2, 1)$ ใน π_2
ค่าคาดหวังของ $\alpha_1(\lambda^*)$, $\alpha_1(\lambda^*)$ และ m_1/n_1 แสดงได้ในเทอมของ $\Phi(h,k;\rho)$ ดังนี้
กำหนดให้

$U_1 = Z+\mu_1-(1/2)(\bar{x}_1+\bar{x}_2)$ เมื่อ $Z \sim N(0,1)$
$U_2 = X_j-(1/2)(\bar{x}_1+\bar{x}_2)$ เมื่อ X_j เป็นตัวอย่างที่งจากกลุ่ม 1
$U_3 = Z+x_1-(1/2)(\bar{x}_1+\bar{x}_2)$
$V = (1/2)(\bar{x}_1-\bar{x}_2)$

จะได้

$$E\{\alpha_1(\lambda^*)\} = Pr\{U_1 V < 0\}$$
$$E(m_1/n_1) = Pr\{U_2 V < 0\}$$
$$E\{\hat{\alpha}_1(\lambda^*)\} = Pr\{U_3 V < 0\}$$

<u>หมายเหตุ</u> ดูพิสูจน์ได้ในบทความของ ฮิล (Hill,1966)

2.3.2 ตัวประมาณในวิธี U หรือ (Leave-one-out Estimator)

วิธีนี้เสนอโดยลาเคนบรุช (Lachenbruch,1967) ซึ่งแนะนำให้ตัดค่าสังเกตหนึ่งค่าออกจากกลุ่มตัวอย่างกลุ่มใดกลุ่มหนึ่ง ในงานวิจัยนี้พิจารณาการตัดค่าสังเกตออกจากกลุ่มตัวอย่างที่ 1 แล้วสร้างกฎการจำแนกกลุ่มโดยอาศัยข้อมูลจากขนาดตัวอย่างที่เหลือจากนั้นจึงนำเอาค่าสังเกตที่ตัดออกหรือกักเอาไว้ (holdout) จัดเข้าพวก ถ้าเข้าพวกผิดให้บันทึกไว้ จากนั้นจึงสลับไปจัดหรือกักค่าสังเกตหน่วยอื่นแล้วดำเนินการสร้างกฎการจำแนกกลุ่มจากข้อมูลตัวอย่างที่เหลือ จากนั้นจึงนำเอาค่าสังเกตที่กักไว้มาจัดเข้าพวก ถ้าจัดเข้าพวกผิดให้บันทึกไว้ ดำเนินการเช่นนี้จนครบทุกหน่วยหรือจนครบทุกๆ ตัวอย่างแล้วนับจำนวนค่าสังเกตหรือตัวอย่างที่จัดเข้าพวกผิดจะได้ค่าประมาณที่ไม่มีความเอนเอียงของอัตราความผิดพลาดที่มีเงื่อนไข

ขั้นตอนการหาอัตราส่วนการจัดเข้าพวกผิด

1. เริ่มจากกลุ่มตัวอย่างที่ 1 ซึ่งมี n_1 หน่วยให้กักค่าสังเกตไว้ 1 หน่วยแล้วสร้างกฎการจำแนกกลุ่ม จากค่าสังเกต $n_1 - 1$, n_2
2. จัดค่าสังเกตที่กักเอาไว้เข้าพวกโดยใช้กฎการจำแนกกลุ่มที่สร้างขึ้นในข้อ 1 ถ้าจัดเข้าพวกผิดให้นับไว้ โดยกำหนดให้ฟังก์ชันนับคือ $h^{j}(x_j) = 1$ เมื่อ $j = 1,2,3, \ldots , n_1$
3. ดำเนินการข้อ 1 และข้อ 2 (โดยเปลี่ยนไปกักค่าสังเกตอื่น) เรื่อยๆ ไปจนค่าสังเกตทุกหน่วยได้รับการจัดเข้าพวกและบันทึก $h^{j}(x_j)$ ไว้ทุกๆ หน่วย
4. จะได้ตัวประมาณอัตราความผิดพลาดที่มีเงื่อนไข ด้วยวิธี U คือ อัตราส่วนของผลรวมของ $h^{j}(x_j)$ กับขนาดตัวอย่างกลุ่มที่ 1 ดังนี้

$$(14) \qquad \hat{\alpha}^{u} = \sum_{j=1}^{n_1} h^{j}(x_j)/n_1$$

เนื่องจากการคำนวนหาดิสคริมิแนนท์ฟังก์ชันแต่ละครั้งจำเป็นต้องคำนวนหาอินเวอรส์ (inverse) เมตริกซ์ของ S^{-1} ตัวใหม่เสมอ ดังนั้นจึงใช้วิธีการหาอินเวอรส์ของ บาธเลนต์ (Bartelett) ที่มีการคำนวนง่ายขึ้น ดังนี้

กำหนดให้

$$S = \text{sample covariance matrix}$$

$$(n_1+n_2-2)S = C_1 + C_2 \qquad (15)$$

โดยที่

$$C_g = \sum_{i=1}^{n_1} (x_1-x_g)(x_1-x_g)' \quad , g = 1,2$$

g^{th} group

สมมติว่าค่าสังเกตที่ j ถูกกักไว้จากกลุ่ม g

$$u_j = x_j-\bar{x}_g$$

ดังนั้น

$$\bar{x}_{g(j)} = \bar{x}_g-u_j/(n_g-1) \qquad (16)$$

และ

$$x_1 - \bar{x}_g = x_1-\bar{x}_{g(j)}+u_j/(n_g-1) \qquad (17)$$

เมื่อ

$$C_g = C_{g(j)}+n_g u_j u'_j/(n_g-1)$$

ดังนั้น

$$(n_1+n_2-3)S_{(j)} = C_{g(j)}+C_{3-g} \qquad (18)$$

$$= (n_1+n_2-2)S-n_g u_j u'_j/(n_g-1)$$

<u>หมายเหตุ</u>

$S_{(j)}$ = Pooled covariance matrix เมื่อกักค่าสังเกตหน่วยที่ j

จาก Barttett's identity

$$B = A+UV' \qquad B^{-1} = A^{-1}-A^{-1}UV'A^{-1}/(1+V'A^{-1}U) \qquad (20)$$

โดยที่

A หมายถึง nonsingular matrix

U,V หมายถึง column vectors

ในที่นี้กำหนดให้

$$B = (n_1+n_2-3)S_{(j)} \qquad A = (n_1+n_2-2)S$$

$$U = \sqrt{n_g/(n_g-1)}u_j \qquad V = -U$$

กำหนด

$$C_g = n_g/((n_g-1)(n_1+n_2-2))$$

ดังนั้น

$$V'A^{-1}U = C_k u_j' S^{-1} u_j$$

และ

$$A^{-1}UV'A^{-1} = \frac{C_k S^{-1}u_j u_j' S^{-1}}{(n_1+n_2-2)}$$ แทนค่าลงในสมการที่ (20)

จะได้

$$(21)\ S^{-1}_{(j)} = \frac{(n_1+n_2-3)}{(n_1+n_2-2)}\left[S^{-1} + \frac{C_k S^{-1}u_j u_j' S^{-1}}{1- C_k u_j' S^{-1} u_j}\right]$$

ดังนั้น Discriminant function ที่ปรับปรุงแล้ว คือ

$$(X_j - 1/2(\bar{X}_1+\bar{X}_2)_{(j)})' S^{-1}_{(j)}(\bar{X}_1-\bar{X}_2)_{(j)} \qquad (22)$$

จากสมการที่ (16) จะได้

$$(\bar{x}_1+\bar{x}_2)_{(j)} = \bar{x}_1+\bar{x}_2-u_j/(n_2-1) \qquad (23)$$

$$(\bar{x}_1-\bar{x}_2)_{(j)} = \bar{x}_1-\bar{x}_2+(-1)^{\bar{}}u_j/(n_2-1) \qquad (24)$$

จากบทความของลาเคนบรุชได้กล่าวถึงค่าเฉลี่ยและความแปรปรวนของตัวประมาณวิธี U ดังนี้ ถ้า x_j มาจากกลุ่มที่ 1 ได้กฎการจำแนกกลุ่มจากสมการการจำแนกกลุ่มดังนี้

$$D_2(x_j) < 0$$

$$E\{D_2(x_j) < 0 \mid G_1\} = P_1(n_1-1,n_2)$$

นั่นคือค่าคาดหวังของความน่าจะเป็นในการจำแนกกลุ่มผิดในกลุ่มที่ 1 คือความน่าจะเป็นสำหรับสมการการจำแนกกลุ่มที่คำนวนจากขนาดตัวอย่าง n_1-1 กับ n_2 สำหรับ j ทุกๆ ตัว

ดังนั้น ถ้ากำหนดตัวแปรสุ่ม Z_j โดยที่

$$Z_j = \begin{cases} 1 & \text{ถ้า } x_j \text{ (มาจากกลุ่มที่ 1) ถูกจำแนกกลุ่มผิด} \\ 0 & \text{ถ้าเป็นอย่างอื่น} \end{cases}$$

จะได้

$$E(Z_j) = P_1(n_1-1,n_2)$$

และ

$$P_1^* = \Sigma_{G_1} Z_j/n_1$$

$$P_2^* = \Sigma_{G_2} Z_j/n_2$$

เป็นตัวประมาณที่ไม่เอนเอียง ของ $P_1(n_1-1,n_2)$ และ $P_2(n_1,n_2-1)$ ตามลำดับ

ค่าความแปรปรวนของตัวประมาณ เมื่อ x_j มาจากกลุ่มที่ 1 คือ

$$VAR(P_1^*) = VAR(\Sigma Z_j/n_1^2) \quad (j=1 \text{ ถึง } n_1)$$

$$= [\Sigma VAR(Z_j)+\Sigma\Sigma COV(Z_i,Z_j)]/n_1^2 \quad (j=1 \text{ ถึง } n_1,\ i\neq j)$$

เนื่องจาก i และ j ไม่ขึ้นแก่กันหรือเป็นอิสระกัน

$$VAR(Z_1) = P_1Q_1 \quad ; \quad COV(Z_1,Z_j) = P_1Q_1\rho$$

เมื่อ ρ คือ ความสัมพันธ์ ระหว่าง Z_1 กับ Z_j ดังนั้น

$$VAR(P_1^*) = \frac{(P_1Q_1)}{n_1}[1+(n_1-1)\rho]$$

ถ้า ρ มีค่าเป็น 0 ค่าของ $VAR(P_1^*) = (P_1Q_1)/n_1 > 0$
ดังนั้นจะได้ค่าของ ρ จาก

$$1+(n_1-1)\rho > 0$$

$$\rho > -1/(n_1-1)$$

ดังนั้นเราสามารถประมาณ ρ ได้จากตัวอย่างเป็นจำนวนซ้ำๆ กัน M ครั้ง โดยใช้ค่าเดิมของ n_1 n_2 และ P_1 จะได้

$$P_1 = \Sigma P_{1i}/M \quad , i = 1,2,3, \ldots , M$$

ดังนั้นตัวประมาณของ $VAR(P_1^*)$ คือ

$$\Sigma(P_{1i}-P_1)^2/M-1 \quad (i=1 \text{ ถึง } M)$$

จากผลการวิจัยของลาเคนบรชได้แสดงให้เห็นว่า ค่าของ correlation มีค่าน้อยมากสามารถตัดทิ้งได้ โดยที่ค่าของ correlation จะลดลงขณะที่ n_1 มีขนาดใหญ่ โดยปกติ $(n_1-1)\rho < 0.5$

2.3.3 วิธี B หรือ (Boostrap estimator)

วิธีนี้เสนอโดย เอฟรอน (Efron,1979) โดยมีหลักเกณฑ์ดังนี้ทำการสุ่มตัวอย่างจากข้อมูลที่เก็บรวบรวมมาแบบใส่คืน (with replacement) ขนาดเท่ากับจำนวนตัวอย่างหรือข้อมูลที่มีอยู่เพื่อสร้างข้อมูลชุดใหม่แล้วนำมาใช้ในการประมาณค่าพารามิเตอร์ที่สนใจ สำหรับการหาตัวประมาณด้วยวิธีนี้สามารถทำได้ 3 แบบ คือ

1. โดยการคำนวณหาจากสูตร
2. โดยใช้เทคนิคมอนติคาร์โล
3. โดยใช้การกระจายของอนุกรมเทย์เลอร์ (Taylor series)

ส่วนมากจะใช้เทคนิคมอนติคาร์โล โดยการนำเอาคอมพิวเตอร์มาเป็นเครื่องมือช่วย สำหรับจำนวนครั้งที่ทำการสุ่มตัวอย่าง (bootstrap sampling) ควรจะอยู่ในช่วง 50 - 200 ครั้ง ก็เพียงพอที่จะทำให้ได้ตัวประมาณที่ดี ในงานวิจัยนี้ได้นำเอาวิธี B มาใช้ในการประมาณค่าอัตราความผิดพลาดที่มีเงื่อนไขซึ่งมีรายละเอียดในการประมาณค่าดังนี้
จากการใช้วิธี R ประมาณค่าอัตราความผิดพลาดได้ตัวประมาณดังนี้
$A_1 = m_1/n_1$ ซึ่งเป็นอัตราส่วนของจำนวนค่าสังเกตที่ถูกจำแนกกลุ่มให้ผิดกับขนาดตัวอย่างในกลุ่มที่ i แต่ถ้าขนาดตัวอย่างไม่ใหญ่พอ A_1 จะประมาณค่าได้ต่ำกว่าความเป็นจริง (under estimate) ดังนั้น Efron จึงได้เสนอวิธีการประมาณค่าความเอนเอียงที่ถูกต้อง (bias correction) ซึ่งตัวประมาณที่ได้คือ $A_1 + b_1$ เมื่อพิจารณาจากตัวอย่างที่มาจากกลุ่มที่ 1 ซึ่ง b_1 คือค่า bias correction ตัวประมาณ b_1 คือ $\hat{b}_1$ ซึ่ง $\hat{b}_1 = E(P_1 - A_1)$ เมื่อ E คือค่าคาดหวังบนกลุ่มตัวอย่างที่ 1

ขั้นตอนการประมาณค่า b_1 มีดังนี้

1. สุ่มตัวอย่างใหม่จากกลุ่มตัวอย่างเดิม 2 กลุ่ม ให้สัญลักษณ์เป็น H^*_1 และ H^*_2 โดยแต่ละค่าสังเกตใหม่มีโอกาสถูกเลือกเท่าๆ กันคือ $1/n_1$ ซึ่งเป็นการสุ่มแบบใส่คืน (with replacement)

2. นำเอาข้อมูลที่ได้ใหม่มาคำนวนหา ดิสคริมิแนนท์ฟังก์ชั่น และ กฎการจำแนกกลุ่มได้ $W^*(x)$ จาก H^*_1 และ H^*_2 ซึ่งขั้นตอนการหา $W^*(x)$ เหมือนกับการหา $W(x)$

3. หาผลต่างของ $d = A^{**}_1 - A^{*}_1$ เมื่อ
A^{**}_1 คือ อัตราส่วนระหว่างจำนวนสมาชิกในกลุ่มตัวอย่างเดิมถูกจำแนกกลุ่มผิดโดยใช้ $W^{*}(x)$ กับขนาดตัวอย่างของกลุ่มที่ 1 (เนื่องจากพิจารณากลุ่มที่ 1)
A^{*}_1 คือ อัตราส่วนระหว่างจำนวนสมาชิกในกลุ่มตัวอย่างใหม่ถูกจำแนกกลุ่มผิดโดยใช้ $W^{*}(x)$ กับขนาดตัวอย่างของกลุ่มที่ 1

4. ค่าคาดหวังของ d คือค่าประมาณของ b_1 นั่นคือ $\hat{b}_1 = \bar{d}$ การหา $\bar{d}$ ก็คือการเฉลี่ยค่า d ประมาณ 50 - 200 ครั้ง นั่นคือต้องทำขั้นตอนที่ 1 - 3 เป็นจำนวน 50 - 200 ครั้ง ดังนั้นตัวประมาณด้วยวิธี B ที่ได้คือ

$$\alpha^{B} = \alpha^{R} + \hat{b}_1 \qquad (25)$$

จากบทความของเอฟรอนได้กล่าวถึงค่าเฉลี่ย และ ความแปรปรวนของวิธี B ดังนี้
กำหนดให้

$X_i = x_i$, $X_i \sim ind\ F$, $i = 1,2,3, \ldots , m$
$Y_j = y_j$, $Y_j \sim ind\ G$, $j = 1,2,2, \ldots , n$

ค่าสังเกต x มาจาก F ถ้า $x \in A$ หรือ ค่าสังเกต x มาจาก G ถ้า $x \in B$
ตัวประมาณวิธี R คือ

$$\widehat{error}_F = \#\{ x_i \in B \}/m$$

เนื่องจากตัวประมาณวิธี R จะประมาณค่า $error_F = Prob\{ x \in B \}$ (actual error rate) ได้ต่ำกว่าที่เป็นจริง ดังนั้นจึงสนใจการแจกแจงของความแตกต่างระหว่างค่าจริงกับค่าประมาณ ดังนี้

$$R((x,y) , (F,G)) = error_F - \widehat{error}_F$$

เนื่องจากในกลุ่มประชากร F และ G จะมีการพิจารณาที่คล้ายคลึงกัน ดังนั้นในที่นี้จะพิจารณาในกลุ่มประชากร F
ลุ่มตัวอย่างแบบใส่คืน จาก X และ Y โดยวิธี บูตสแทรบ (bootstrap samples) จะได้ค่าของ X_i^{*} และ Y_j^{*} ดังนี้

$X_i^* = x_i^* \quad , \quad X_i^* \sim \text{ind } F \quad , \quad i = 1,2,3, \ldots , m$

$Y_j^* = y_j^* \quad , \quad Y_j^* \sim \text{ind } G \quad , \quad j = 1,2,3, \ldots , n$

ซึ่งการแจกแจงของ $\hat{F}$ และ $\hat{G}$ มีการแจกแจงที่สอดคล้องกับ F และ G จะได้

$$R^* = R((x^*,y^*),(\hat{F},\hat{G})) = \#\{ x_i \in B^* \}/m - \#\{x_i^* \in B^* \}/m$$

สมการนี้คือความแตกต่างระหว่างอัตราความผิดพลาดที่เป็นจริง (actual error rate) ในกลุ่ม $\hat{F}$ และอัตราความผิดพลาดที่เป็นจริงในกลุ่ม F
ต่อไปให้ทำการสุ่มตัวอย่างหา R^* เป็นจำนวน I ครั้งจนได้
$R_1^*, R_2^*, R_3^*, \ldots, R_I^*$ ทำให้ได้ $E(R^*)$ เป็นตัวประมาณของ $E_{F,G}(R)$ ดังนี้

$$E(R^*) = \sum_{j=1}^{I} R_j^*/I$$

และ $VAR(R^*)$ เป็นตัวประมาณของ $VAR_{F,G}(R)$ ดังนี้

$$VAR(R^*) = \{ \sum_{j=1}^{I} [R_j^* - E(R^*)]^2/(I-1) \}$$

2.3.4 วิธี DS หรือ (Shrunken-D estimator)

วิธีนี้ มอร์ริสัน (Morrison,1976) ได้พัฒนามาจากวิธี D หรือ Plug-in estimator เนื่องจากวิธี D มีข้อเสียคือการประมาณค่าตัวประมาณจะประมาณค่าได้ต่ำกว่าความเป็นจริง (under estimate) ดังนั้น มอร์ริสัน จึงได้พัฒนาการใช้ค่าจากตัวอย่างที่เหมาะสมหรือไม่เอนเอียงของ μ_1 , μ_2 และ Σ ซึ่งจะปรากฏในรูปของ การใช้ Σ^{-1} ค่าประมาณของ μ_1 และ μ_2 คือ $\bar{X}_1$ และ $\bar{X}_2$ ตามลำดับ แต่ค่าประมาณของ Σ^{-1} ใช้คุณสมบัติของ Wishart distribution สามารถแสดงได้ดังนี้

$$E(S) = (N-2)\Sigma/(N-k-3) \qquad (26)$$

ดังนั้น

$$\Sigma = (N-2)S/(N-k-3) \qquad (27)$$

เป็นตัวประมาณที่ไม่เอนเอียงของ Σ โดยที่ $N = n_1 + n_2$ แทนค่า x_1 และ Σ ลงในสมการ (11) จะได้

$$\alpha^{DS} = \Phi\left[\frac{1/2(\bar{x}_1+\bar{x}_2)'S^{-1}(\bar{x}_1-\bar{x}_2)-x_1'S^{-1}(\bar{x}_1-\bar{x}_2)}{((\bar{x}_1-\bar{x}_2)'S^{-1}((N-k)S)/(N-k-3)S^{-1}(\bar{x}_1-\bar{x}_2))^{1/2}}\right]$$

$$= \Phi[-\{(N-k-3)(\bar{x}_1-\bar{x}_2)'S^{-1}(\bar{x}_1-\bar{x}_2)/(N-2)\}/2]$$

$$= \Phi[-(DS)^{1/2}/2] \qquad (28)$$

เมื่อ

$DS = (N-k-3)D^2/(N-2)$

$D^2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)$ •

k = จำนวนตัวแปรอิสระ