บทที่ 2

วิธีดำเนินการวิจัย

2.1 การเก็บรวบรวมข้อมูล

ทำการเก็บข้อมูลจากการบันทึกผลการรักษาใน OPD Card, เวชระเบียนผู้ป่วยในและผลการตรวจสุขภาพ ซึ่งประกอบด้วย

2.1.1 ข้อมูลทางด้านประชากร

2.1.2 ผลการตรวจร่างกายและผลการรักษา

2.1.3 ผลการตรวจทางห้องปฏิบัติการ

โดยผลการตรวจทั้ง 2.1.2 และ 2.1.3 จะเป็นผลการตรวจที่ได้ระบุชัดว่าผู้ป่วยนี้มีอาการของโรคแทรกซ้อนเกิดขึ้นแล้ว

ได้คัดเลือกผู้ป่วยที่มีลักษณะดังนี้ออกจากการศึกษา

1. ผู้ที่มีโรคแทรกซ้อนหลายชนิดในขณะเดียวกัน
2. ผู้ที่มีข้อมูลไม่ครบตามต้องการ

ทั้งนี้ ยกเว้นในเรื่องส่วนสูงของผู้ป่วย ซึ่งจากการคัดลอกข้อมูลของผู้ป่วยโรคเบาหวานจากเวชระเบียนผู้ป่วยนั้น โดยส่วนใหญ่แล้วไม่ได้บันทึกส่วนสูงไว้ ดังนั้น จึงแก้ไขโดยการประมาณส่วนสูงจากผู้ป่วยโรคเบาหวานที่คลีนิคผู้ป่วยโรคเบาหวาน โรงพยาบาลจุฬาลงกรณ์เป็นชาย 80 คน หญิง 210 คน ซึ่งมีการบันทึกส่วนสูง น้ำหนักและอายุอย่างสมบูรณ์ วิธีการประมาณนั้นใช้วิธีการของ Stepwise Multiple Regression ประมาณส่วนสูงจากอายุและน้ำหนัก โดยแยกการประมาณของชายและหญิงออกจากกัน ผลการประมาณได้สมการประมาณส่วนสูงดังนี้คือ

ชาย : ส่วนสูง(ซม.) = 155.9726+0.2170(น้ำหนัก)-0.1454(อายุ)

หญิง : ส่วนสูง(ซม.) = 141.6374+0.1643(น้ำหนัก)+0.0131(อายุ)

2.2 วิธีวิเคราะห์ข้อมูล

ข้อมูลที่ได้จากการคัดลอกและจากการสัมภาษณ์นั้น จะถูกนำมาตรวจสอบความสมบูรณ์แล้ว

นำไปลงรหัสตามคู่มือลงรหัสซึ่งได้สร้างไว้ก่อนแล้ว จากนั้นจะนำไปวิเคราะห์โดยวิธีการดังนี้

2.2.1 สถิติเชิงพรรณา (Descriptive Statistics) เพื่อศึกษาลักษณะการกระจายตัวแปรต่าง ๆ ทุกตัว ทั้งในผู้ป่วยโรคหลอดโลหิตเลี้ยงหัวใจ ผู้ป่วยที่มีการเปลี่ยนแปลงที่จอรับภาพ และผู้ป่วยที่มีจอรับภาพปกติ โดยนำเสนอในรูปตาราง แสดงจำนวนและค่าร้อยละ

2.2.2 ทดสอบความสัมพันธ์ระหว่างปัจจัยทางด้านประชากรว่ามีความสัมพันธ์กับปัจจัยทางด้านกายภาพและผลตรวจทางห้องปฏิบัติการหรือไม่ โดยศึกษาจากทั้ง 2 โรคแทรกซ้อนดังกล่าว การวิเคราะห์และทดสอบความสัมพันธ์นี้ใช้วิธีการของ Canonical Analysis

2.2.3 เพื่อทำนายถึงปัจจัยที่ทำให้เกิดโรคแทรกซ้อนในระบบอวัยวะต่าง ๆ โดยใช้ Discriminant Analysis เพื่อหา

2.2.3.1 สมการทำนายถึงปัจจัยที่ทำให้ผู้ป่วยมีการเปลี่ยนแปลงที่จอรับภาพโดยใช้ผู้ป่วยที่มีจอรับภาพปกติเป็นกลุ่มเปรียบเทียบ

2.2.3.2 สมการทำนายถึงปัจจัยที่แบ่งแยกการเกิดโรคหลอดโลหิตเลี้ยงหัวใจและการเปลี่ยนแปลงที่จอรับภาพ

2.3 การวิเคราะห์สหสัมพันธ์คาโนนิคอล (Canonical Correlation Analysis)

ตามปกติ เมื่อเรามีตัวแปรสุ่มอยู่ 2 ตัว คือ X และ Y นั้น เราจะสามารถหาความสัมพันธ์ของตัวแปรทั้งสองได้ คือค่า ρ = สัมประสิทธิ์สัมพันธ์ (Correlation Coefficient) ได้จาก

$$\rho = \frac{\mathrm{Cov}\ (X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

ซึ่งค่า ρ นี้ สามารถอธิบายว่า X และ Y มีความสัมพันธ์กันมากน้อยเพียงไร

แต่ถ้าหากเรามีค่า Y ที่ขึ้นอยู่กับค่า X หลาย ๆ ตัว วิธีการที่เราจะสามารถหาความสัมพันธ์ระหว่าง Y และ X เหล่านั้น ก็จะใช้วิธีวิเคราะห์สหสัมพันธ์พหุคูณ (Multiple Correlation)

โดยที่ให้ X เป็น เมตริกซ์ ที่ประกอบด้วย X_i หลาย ๆ ตัว

$$X = X_1 \ X_2 \ \ldots\ldots\ X_p$$

และ $\hat{Y}$ เป็นค่าประมาณของผลรวมเชิงเส้น

$$\hat{Y}_j = \hat{\beta}_1 X_{1j} + \hat{\beta}_2 X_{2j} + \ldots\ldots\ldots + \hat{\beta}_p X_{pj} \quad ; j = 1, 2, \ldots\ldots, n$$

เมื่อ $\hat{\beta}_s$ $s = 1, 2, \ldots, p$ เป็นค่าความชันของสมการ Multiple regression ซึ่งหาค่าได้จากการประมาณโดยวิธี least-squares อันจะทำให้เราสามารถหาค่าสัมประสิทธิ์สหสัมพันธ์โดยการหาค่าความสัมพันธ์อย่างง่าย(simple correlation) ระหว่าง Y และ $\hat{Y}$ นั่นเอง ยังมีอีกกรณีหนึ่งซึ่งนอกจากจะมีตัวแปร X หลาย ๆ ตัวแล้วยังมี Y หลาย ๆ ตัวคือ

$$Y = Y_1 \quad Y_2 \ldots\ldots\ldots\ldots\ldots Y_q$$

แล้ววิธีการที่จะหาความสัมพันธ์ระหว่าง X หลาย ๆ ตัวและ Y หลาย ๆ ตัวเหล่านั้นก็คือวิธีวิเคราะห์สหสัมพันธ์คาโนนิคอล จะได้ค่าความสัมพันธ์คาโนนิคอล(Canonical Correlation) ซึ่งหาได้จากความสัมพันธ์ระหว่างผลรวมเชิงเส้นของกลุ่มตัวแปร X และผลรวมเชิงเส้นของกลุ่มตัวแปร Y

เมื่อเรากำหนดค่า

$$\begin{aligned} X^* &= \alpha_1 X_1 + \alpha_2 X_2 + \ldots\ldots\ldots\ldots + \alpha_p X_p \\ Y^* &= \beta_1 Y_1 + \beta_2 Y_2 + \ldots\ldots\ldots\ldots + \beta_q Y_q \end{aligned} \quad \ldots\ldots\ldots\ldots(1)$$

สมการนี้เรายังไม่ทราบค่าของ α_i และ β_i ที่จะทำให้ความสัมพันธ์ระหว่าง Y^* และ X^* มีค่ามากที่สุดที่จะเป็นไปได้

ถ้าเราให้ ρ_c = Canonical Correlation Coefficient ระหว่าง X และ Y

$$= \frac{\text{Cov}(X^*, Y^*)}{\sqrt{\text{Var}(X^*)\ \text{Var}(Y^*)}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

และ $\rho_c^{(1)}$ เป็น Canonical Correlation Coefficient ที่มีค่ามากที่สุดเรียกชื่อว่า First Canonical Correreilation ซึ่งเกิดจาก X_1^* และ Y_1^*

$\rho_c^{(2)}$ เป็น Canonical Correlation Coefficient ที่มีค่ามากที่สุดเป็นอันดับสองเรียกชื่อว่า Second Canonical Correlation ซึ่งเกิดจาก X_2^* และ Y_2^*

$\rho_c^{(3)}$ เป็น Canonical Correlation Coefficient ที่มีค่ามากที่สุดเป็นอันดับสามเรียกชื่อว่า Third Canonical Correlation ซึ่งเกิดจาก X_3^* และ Y_3^*

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

ทำดังนี้เรื่อย ๆ ไป

X_1^* และ Y_1^* จะเป็น first canonical variables

X_2^* Y_2^* จะเป็น second canonical variables

X_3^* Y_3^* จะเป็น third canonical variables

. .

. .

. .

คู่ของ canonical variable ต่าง ๆ นั้น จะมีหลักการเหมือนกัน

กรณีที่ $q < p$ ก็จะมี Canonical Correlation ทั้งหมด q ค่า

และ Canonical Variables ทั้งหมด q คู่

ผลลัพธ์ต่าง ๆ เหล่านี้ จะได้มาจากหลักการต่อไปนี้คือ

เมื่อเรา assume ว่าตัวแปร X และ Y มีการแจกแจงดังนี้คือ

X มี mean เป็น μ_1 และ variance เท่ากับ Σ_{11}

Y มี mean เป็น μ_2 และ variance เท่ากับ Σ_{22}

ให้ $\Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right]$

โดยที่ Σ_{11} เป็น p X p variance matrix ของ X

Σ_{22} เป็น q X q variance matrix ของ Y

Σ_{12} เป็น p X q covariance matrix ของ X และ Y

Σ_{21} เป็น transpose matrix ของ Σ_{12}

เราสามารถเขียนรายละเอียดได้เป็น

$$\Sigma = \left[\begin{array}{ccc|ccc} \sigma_{x_1x_1} & \sigma_{x_1x_2} \cdots & \sigma_{x_1x_p} & \sigma_{x_1y_1} & \sigma_{x_1y_2} \cdots & \sigma_{x_1y_q} \\ \sigma_{x_2x_1} & \sigma_{x_2x_2} \cdots & \sigma_{x_2x_p} & \sigma_{x_2y_1} & \sigma_{x_2y_2} \cdots & \sigma_{x_2y_q} \\ \sigma_{x_px_1} & \sigma_{x_px_2} \cdots & \sigma_{x_px_p} & \sigma_{x_py_1} & \sigma_{x_py_2} \cdots & \sigma_{x_py_q} \\ \hline \sigma_{x_1y_1} & \sigma_{x_2y_1} \cdots & \sigma_{x_py_1} & \sigma_{y_1y_1} & \sigma_{y_1y_2} \cdots & \sigma_{y_1y_q} \\ \sigma_{x_1y_q} & \sigma_{x_2y_q} \cdots & \sigma_{x_py_q} & \sigma_{y_qy_1} & \sigma_{y_qy_2} \cdots & \sigma_{y_qy_q} \end{array}\right]$$

ถ้าเราแสดง Canonical variables ในรูปของ matrix จะได้ว่า

$$\underset{\sim}{X}^* = X\underset{\sim}{\alpha} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (1a)$$

$$\underset{\sim}{Y}^* = Y\underset{\sim}{\beta}$$

เมื่อ $\underset{\sim}{X}^*$ และ $\underset{\sim}{Y}^*$ เป็น vector ของ Canonical variable

X และ Y เป็น matrix ของตัวแปรสุ่มแรกเริ่ม (original random variables)

$\underset{\sim}{\alpha}$ เป็น Coefficient vectors ขนาด p X 1

$\underset{\sim}{\beta}$ เป็น Coefficient vectors ขนาด q X 1

ดังนั้น จากสมการ (2) เราสามารถเขียน Canonical Correlation Coefficient ของประชากรได้เป็น

$$\rho_c = \frac{\text{Cov }(X\underset{\sim}{\alpha}, Y\underset{\sim}{\beta})}{\sqrt{\text{Var }(X\underset{\sim}{\alpha})\text{Var}(Y\underset{\sim}{\beta})}}$$

หรือ

$$\rho_c = \frac{\underset{\sim}{\alpha}'\Sigma_{12}\underset{\sim}{\beta}}{\sqrt{(\underset{\sim}{\alpha}'\Sigma_{11}\underset{\sim}{\alpha})(\underset{\sim}{\beta}'\Sigma_{22}\underset{\sim}{\beta})}} \quad \cdots\cdots\cdots\cdots (2a)$$

จะหาค่า ρ_c จากสมการ (2a) โดยที่เราต้องการให้ได้ ρ_c ที่มีค่าสูงสุด ดังนั้นจะ Maximize ρ_c เทียบกับ $\underset{\sim}{\alpha}$ และ $\underset{\sim}{\beta}$ ซึ่งวิธีดังกล่าวนี้จะทำได้ง่ายขึ้น เมื่อเราเพิ่มเงื่อนไขที่ว่า Canonical variable แต่ละตัวจะต้องมีคุณสมบัติว่า มีความแปรปรวนเป็น Unit variance คือ Variance = 1 นั่นคือ

$$\text{Var}(X^*) = \underset{\sim}{\alpha}'\Sigma_{11}\underset{\sim}{\alpha} = 1 \quad \cdots\cdots\cdots\cdots\cdots (3)$$

$$\text{Var}(Y^*) = \underset{\sim}{\beta}'\Sigma_{22}\underset{\sim}{\beta} = 1 \quad \cdots\cdots\cdots\cdots\cdots (4)$$

ซึ่งเงื่อนไขนี้จะทำให้ได้ $\rho_c = \underset{\sim}{\alpha}'\Sigma_{12}\underset{\sim}{\beta}$ ทันที ดังนั้น เราจึง Maximize ρ_c นี้โดยวิธี Least Square Method ซึ่งจะได้ necessary conditions ดังนี้

$$\Sigma_{12}\underset{\sim}{\beta}-\lambda\Sigma_{11}\underset{\sim}{\alpha} = 0 \quad \cdots\cdots\cdots\cdots\cdots (5)$$

$$\Sigma_{21}\underset{\sim}{\alpha}-\lambda\Sigma_{22}\underset{\sim}{\beta} = 0 \quad \cdots\cdots\cdots\cdots\cdots (6)$$

เมื่อ λ เป็น Lagrange multiplier

คูณสมการ (5) ด้วย λ ได้

$$\Sigma_{12}\lambda\underset{\sim}{\beta} = \lambda^2\Sigma_{11}\underset{\sim}{\alpha} \quad \cdots\cdots\cdots\cdots (7)$$

คูณสมการ (6) ด้วย Σ_{22}^{-1} ได้

$$\Sigma_{22}^{-1}\Sigma_{21}\underset{\sim}{\alpha} = \lambda\underset{\sim}{\beta} \quad \cdots\cdots\cdots\cdots\cdots (8)$$

(7) แทนใน (8) ได้

$$\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\underset{\sim}{\alpha} = \lambda^2\Sigma_{11}\underset{\sim}{\alpha}$$

คูณทั้งสองข้างด้วย Σ_{11}^{-1} และจัดเทอมใหม่

$$(\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \lambda^2 I)\underset{\sim}{\alpha} = 0 \quad \cdots (9)$$

เช่นเดียวกันจะได้

$$(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \lambda^2 I)\underset{\sim}{\beta} = 0 \quad \cdots (10)$$

ทั้งสมการ (9) และ (10) เป็น homogeneous equations ซึ่งเราสามารถหาคำตอบได้ แต่จากการที่เราได้กำหนดไว้ก่อนแล้วว่า $q \leq p$ เสมอ ดังนั้น จึงมี degree of the characteristic equation ทั้งหมด q ค่า ซึ่งถ้าหากเราทราบค่า λ^2 และ β แล้ว เราก็จะสามารถหา α ได้จากสมการ (5)

เนื่องจาก

$$\underset{\sim}{\alpha} = \frac{\Sigma_{11}^{-1}\Sigma_{12}\underset{\sim}{\beta}}{\lambda} \quad \cdots\cdots\cdots\cdots (11)$$

จากสมการ (5) คูณด้วย $\underset{\sim}{\alpha}'$ ได้

$$\underset{\sim}{\alpha}'\Sigma_{12}\underset{\sim}{\beta} = \lambda\underset{\sim}{\alpha}'\Sigma_{11}\underset{\sim}{\alpha}$$

จากสมการ (6) คูณด้วย β' ได้

$$\underset{\sim}{\beta}'\Sigma_{21}\underset{\sim}{\alpha} = \lambda\underset{\sim}{\beta}'\Sigma_{22}\underset{\sim}{\beta}$$

เนื่องจาก $\underset{\sim}{\alpha}'\Sigma_{11}\underset{\sim}{\alpha} = \underset{\sim}{\beta}'\Sigma_{22}\underset{\sim}{\beta} = 1$

ดังนั้น จะได้ $\underset{\sim}{\alpha}'\Sigma_{12}\underset{\sim}{\beta} = \underset{\sim}{\beta}'\Sigma_{21}\underset{\sim}{\alpha} = \lambda$ ......(12)

ดังนั้น จากสมการ (2a) จะได้ว่า $\underset{\sim}{\alpha}'\Sigma_{12}\underset{\sim}{\beta} = \lambda$ เป็น Canonical Correlation ภายใต้เงื่อนไขของ (3) และ (4) นั่นคือ

$$\rho_c = \lambda = \underset{\sim}{\alpha}'\Sigma_{12}\underset{\sim}{\beta}$$

ถ้าหากถอดรากที่สองของ characteristic root ที่มีค่ามากสุดของสมการ (9) หรือ (10) ก็จะได้ first canonical correlation

$$\lambda^2 = \text{characteristic root}$$

จากสมการ (10) สามารถหาค่า λ^2 และ $\underset{\sim}{\beta}$ ได้ และจาก (11) สามารถหาค่า $\underset{\sim}{\alpha}$ ได้

ให้ $\lambda_1^2 > \lambda_2^2 > \ldots\ldots > \lambda_q^2$ เป็น characteristic roots ของสมการ (10)

$\underset{\sim}{\beta}_1, \underset{\sim}{\beta}_2, \ldots\ldots\ldots, \underset{\sim}{\beta}_q$ เป็น characteristic vectors ที่สัมพันธ์กับ characteristic root λ_i^2 ดังกล่าว

$\underset{\sim}{\alpha}_1, \underset{\sim}{\alpha}_2, \ldots\ldots\ldots, \underset{\sim}{\alpha}_q$ เป็น characteristic vectors ซึ่งคำนวณได้จากการแทนค่า λ_i และ $\underset{\sim}{\beta}_i$ ลงใน (11)

ดังนั้น Canonical Correlation Coefficient ที่ i คือ λ_i

และ set ของ canonical variables ก็คือ $\underset{\sim}{X}_i^* = X\underset{\sim}{\alpha}_i$

$$\underset{\sim}{Y}_i^* = Y\underset{\sim}{\beta}_i$$

Canonical Correlation Coefficient นี้มีคุณสมบัติเช่นเดียวกับ simple correlation coefficient กล่าวคือ ค่า absolute value ของ canonical correlation coefficient จะมีค่าอยู่ระหว่าง 0 และ 1 และจะไม่เปลี่ยนค่า (invariant) แม้ว่าจะเปลี่ยนหน่วยการวัดก็ตาม

<u>การคำนวณ</u> เนื่องจากคุณสมบัติดังกล่าวข้างต้นของ Canonical Correlation Coefficients ดังนั้น เมื่อไม่ทราบค่า Σ เราจะประมาณ Σ ด้วย S แทน

$$S = \left[\begin{array}{c|c} S_{11} & S_{12} \\ \hline S_{21} & S_{22} \end{array}\right]$$

เมื่อ S เป็น covariance matrix ของตัวอย่าง หรือ

$$R = \left[\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right]$$

เมื่อ R เป็น correlation matrix ของตัวอย่าง

R_{11} เป็น matrix ขนาด p X p ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปร X p ตัว

R_{22} เป็น matrix ขนาด q X q ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปร Y q ตัว

R_{12} เป็น matrix ขนาด p X q ซึ่งแสดงความสัมพันธ์ระหว่าง X และ Y

$R_{21} = R'_{12}$

ค่าของ Canonical Correlation Coefficients ที่คำนวณจาก S หรือ R นั้นจะได้ผลเช่นเดียวกัน แต่ค่า α_i และ β_i นั้นจะได้ไม่เหมือนกัน กล่าวคือ

ถ้าหากคำนวณ α_i และ β_i จาก S matrix แล้ว ค่าที่ได้นั้นจะเป็นค่าตัวแปรแรกเริ่ม (original variables)

ถ้าหากคำนวณ จาก R matrix ค่าที่ได้จะเปลี่ยนอยู่ในรูปตัวแปรมาตรฐาน (Standardized original variables)

อย่างไรก็ตามเราก็สามารถที่จะเปลี่ยนค่าสัมประสิทธิ์ที่คำนวณได้จากการใช้ R นั้นไปเป็นค่าสัมประสิทธิ์ที่ได้จาก S ได้ ดังจะเห็นได้จาก

$$\begin{bmatrix} \frac{1}{\sqrt{S_{11}}} & 0 \\ 0 & \frac{1}{\sqrt{S_{22}}} \end{bmatrix} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{S_{11}}} & 0 \\ 0 & \frac{1}{\sqrt{S_{22}}} \end{bmatrix} = \begin{bmatrix} 1 & R_{12} \\ R_{21} & 1 \end{bmatrix}$$

ดังนั้น เมื่อเราเอา diagonal matrix ของส่วนกลับของค่า standard deviation มาคูณทั้งหน้าและหลังของ Covariance Matrix แล้ว เราก็จะสามารถหาค่าของ correlation matrix α , β ได้ คือ

ถ้า $\alpha^*_{i1}, \alpha^*_{i2} \ldots\ldots\ldots, \alpha^*_{ip}$ และ $\beta^*_{i1}, \beta^*_{i2} \ldots\ldots\ldots\ldots, \beta^*_{iq}$;

$i = 1, 2, \ldots\ldots\ldots, q$

เป็นค่าที่คำนวณได้จากค่า R แล้ว เราจะสามารถหาค่า α, β ซึ่งคำนวณได้จาก S ได้โดยการหารแต่ละค่าสัมประสิทธิ์เหล่านั้นด้วยค่า standard deviation ของตัวแปรแรกเริ่ม (original variable) นั่นคือ เราเพียงคำนวณ

$$\alpha^*_{i1}/S_{x_{11}}, \alpha^*_{i2}/S_{x_{22}}, \alpha^*_{i3}/S_{x_{33}}, \ldots\ldots\ldots\ldots, \alpha^*_{ip}/S_{x_{pp}}$$

และ $$\beta^*_{i1}/S_{y_{11}}, \beta^*_{i2}/S_{y_{22}}, \ldots\ldots\ldots\ldots, \beta^*_{ip}/S_{y_{qq}}$$

2.3.1 <u>การทดสอบนัยสำคัญของการวิเคราะห์สหสัมพันธ์คาโนนิคอล</u>

เมื่อทำการวิเคราะห์สหสัมพันธ์คาโนนิคอลเสร็จเรียบร้อยแล้ว สิ่งที่เราจะได้คือ

(1) ρ_c

(2) $\underset{\sim}{\alpha}$ และ $\underset{\sim}{\beta}$

ค่าของ $\underset{\sim}{\alpha}$ และ $\underset{\sim}{\beta}$ ก็จะสอดคล้องกับ λ^2 ของเมตริกซ์ $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ โดยที่ λ^2 ก็คือความแปรปรวนที่ร่วมกันของตัวแปรทั้งสองชุดนั่นเอง ด้วยเหตุนี้ก็จะมี $\underset{\sim}{\alpha}$ และ $\underset{\sim}{\beta}$ อยู่หลายชุดด้วยกัน ดังนั้นสิ่งที่เราจะต้องคำนึงถึงคือ Canonical variables ทั้ง q คู่นั้นมีความสัมพันธ์กันหรือไม่ คือต้องมีการทดสอบสมมติฐานที่ว่า ตัวแปรทั้งสองชุดนั้นมีความสัมพันธ์กันหรือไม่ นั่นคือทดสอบว่า Σ_{12} หรือ $\Sigma_{21} = 0$ หรือไม่

$$H_0 : \Sigma_{12} = 0 \qquad\qquad H_0 : \Sigma_{21} = 0$$

หรือ

$$H_a : \Sigma_{12} \neq 0 \qquad\qquad H_a : \Sigma_{21} \neq 0$$

จะใช้วิธีการทดสอบของ Bartlette ดังนี้

2.3.1.1 <u>วิธีทดสอบของ Bartlett</u>

โดยกำหนด Λ_o (lambda) จาก

$$\Lambda_o = \frac{1}{\prod_{i=1}^{q}\left[1+\frac{\lambda_i^2}{(1-\lambda_i)^2}\right]}$$

$$= \frac{1}{\prod_{i=1}^{q}\left[\frac{1}{(1-\lambda_i^2)}\right]}$$

$$= \prod_{i=1}^{q}(1-\lambda_i^2)$$

ค่า Λ_o ที่คำนวณได้ จะนำมาหาค่า Barlett V_o จากสูตร

$$V_o = -\left[n-1-\frac{1}{2}(p+q+1)\right] \ln \Lambda_o$$

ซึ่ง V_0 จะมีการกระจายเป็น ไค-สแควร์ โดยมีองศาอิสระ (d.f) = pq ที่ระดับนัยสำคัญ α จะนำมาเปรียบเทียบกับ $\chi^2_{\alpha,pq}$ จากตาราง

ถ้าค่า V_0 ที่คำนวณได้มากกว่า χ^2_{table} แล้ว เราจะสรุปได้ว่า $\rho_c^{(1)}$ ซึ่งเป็น first canonical correlation coefficient นั้น ไม่เท่ากับ 0 และเราจะทดสอบความสัมพันธ์ของ ρ_c ของคู่ที่เหลือต่อไป คือ

$$\Lambda_1 = \prod_{i=2}^{q} (1-\lambda_i^2)$$

อันจะนำไปสู่ Barlett V_1 ดังนี้

$$V_1 = -\left[n-2-\frac{1}{2}(p+q+1)\right] \ln\Lambda_1$$

ซึ่งจะมีการกระจายเป็น χ^2 ด้วยองศาอิสระ = (p-1)(q-1) การสรุปผลจะทำเหมือน V_0

ดังนั้นถ้าหากมี k canonical correlation coefficients ที่ทดสอบแล้วปรากฏว่ามีนัยสำคัญส่วนที่เหลือก็จะถูกทดสอบว่า (q-k) canonical correlation coefficient เหล่านั้นจะเท่ากับ 0 หรือไม่ กล่าวคือ

$$\Lambda_k = \prod_{i=(k+1)}^{q} (1-\lambda_i^2)$$

ซึ่งจะได้ $$V_k = -\left[n-1-k-\frac{1}{2}(p+q+1)\right] \ln\Lambda_k$$

มีการกระจายเป็น χ^2 ด้วยองศาอิสระ = (p-k)(q-k)

ในการประมวลผลด้วยโปรแกรมสำเร็จรูป SPSS นั้น ผลลัพธ์ที่ได้จะมีค่า significance หรือ p-value ซึ่งเป็นค่าของความน่าจะเป็นที่จะไม่ยอมรับ H_0 เมื่อ H_0 จริง ในกรณีที่เราจะทดสอบสมมติฐาน ณ ระดับนัยสำคัญ α = 0.05 นั้น เราอาจใช้ค่า p-value มาเปรียบเทียบกับ α = 0.05 เพื่อใช้เป็นเกณฑ์ในการตัดสินว่าจะยอมรับสมมติฐาน H_0 หรือไม่

โดยที่ ถ้าค่า p value < 0.05 จะไม่ยอมรับ H_0

p value > 0.05 จะยอมรับ H_0

2.4 <u>การวิเคราะห์จำแนกประเภท</u> (Discriminant Analysis)

เป็นวิธีการวิเคราะห์ทางสถิติโดยมีจุดมุ่งหมายที่จะคัดเลือกตัวแปรชุดหนึ่ง ซึ่งนักวิจัยคิดว่าตัวแปรชุดนี้มีความสัมพันธ์กับสิ่งที่ต้องการศึกษา จนถึงขั้นที่ใช้ตัวแปรชุดนี้เป็นตัวแบ่งแยกประชากรออกเป็นกลุ่มต่าง ๆ ได้อย่างชัดเจน โดยประชากรแต่ละกลุ่มนั้นจะมีตัวแปรที่สัมพันธ์กับชุดตัวแปรที่ใช้แบ่งแยกกลุ่มนั้น ซึ่งเราเรียกชุดตัวแปรที่มีความสัมพันธ์ในรูปของสมการว่า สมการจำแนกประเภท

จำนวนสมการที่ได้จะน้อยกว่าจำนวนกลุ่มของประชากรที่ต้องการจำแนก 1 กลุ่มเสมอหรือถ้ามีจำนวนกลุ่มมากกว่าจำนวนตัวแปร จำนวนสมการจะเท่ากับจำนวนตัวแปรที่จะใช้จำแนกประเภท (discriminant variables)

2.4.1 <u>ขั้นตอนในการวิเคราะห์จำแนกประเภท</u> มี 2 ส่วนคือ

1. การคัดเลือกตัวแปรชุดหนึ่ง เพื่อใช้ในการแบ่งแยกประชากรออกเป็นกลุ่มต่าง ๆ กันได้อย่างชัดเจน สมการนี้คือ สมการจำแนกประเภท (Discriminant function)

2. การจำแนกตัวอย่างที่ได้ศึกษามานั้นเข้าเป็นสมาชิกของประชากรกลุ่มหนึ่ง ๆ ได้โดยอาศัยสมการจำแนกประเภทเป็นเครื่องช่วยระบุจำแนกประเภท

<u>ในกรณีที่มีประชากร 2 กลุ่ม</u> จะใช้วิธีการของ Fisher มาดำเนินการขั้นตอนทั้ง 2 ดังกล่าวคือ กำหนดให้

π_1 เป็นประชากรกลุ่มที่ 1

π_2 เป็นประชากรกลุ่มที่ 2

เราศึกษาตัวแปร X ทั้งหมด p ตัวคือ

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

วิธีการของ Fisher นั้นจะแปลงตัวแปร X เหล่านี้ไปเป็นค่าตัวแปรเพียงตัวเดียวคือ Y โดยที่ Y_1 และ Y_2 นี้เป็นค่าสังเกตที่ได้จากประชากร π_1 และ π_2 ตามลำดับ

กำหนดให้ μ_{1y} = ค่าเฉลี่ยของค่า Y ซึ่งได้มาจากค่า X ของประชากร π_1

μ_{2y} = ค่าเฉลี่ยของค่า Y ซึ่งได้มาจากค่า X ของประชากร π_2

$$\mu_1 = E(X/\P_1) = \text{ค่าคาดหวังของตัวแปร X ที่มาจากประชากร } \P_1$$
$$\mu_2 = E(X/\P_2) = \text{ค่าคาดหวังของตัวแปร X ที่มาจากประชากร } \P_2 \quad \cdots(1)$$

และโควาเรียนซ์ เมตริกซ์ คือ

$$\Sigma = E(X-\mu_i)(X-\mu_i)' \quad ; \quad i = 1, 2 \quad \cdots\cdots(2)$$

เราพิจารผลรวมเชิงเส้น

$$\underset{(1X1)}{Y} = \underset{(1Xp)}{\beta'}\underset{(pX1)}{X} \quad \cdots\cdots(3)$$

ดังนั้น จะได้
$$\mu_{1y} = E(Y/\P_1) = E(\beta'X/\P_1) = \beta'\mu_1$$
$$\mu_{2y} = E(Y/\P_2) = E(\beta'X/\P_2) = \beta'\mu_2 \quad \cdots\cdots(4)$$

และค่าความแปรปรวนของ Y จากทั้ง 2 ประชากรนั้นคือ

$$\sigma_y^2 = Var(\beta'X) = \beta'Cov(X)\beta$$
$$= \beta'\Sigma\beta \quad \cdots\cdots(5)$$

วิธีการของ Fisher คือจะพยายามหาผลรวมเชิงเส้นของค่า X ซึ่งจะทำให้ระยะทางระหว่าง μ_{1y} และ μ_{2y} มีค่ามากที่สุดที่จะเป็นไปได้

ค่าผลรวมเชิงเส้นซึ่งดีที่สุดจะต้องมีคุณสมบัติในการแบ่งแยกประชากรทั้ง 2 กลุ่มออกจากกันได้มากที่สุด ซึ่งวิธีการที่จะได้ผลรวมเชิงเส้นที่ดีนั้นก็โดยการหาค่า β ที่ทำให้อัตราส่วน $\frac{(\text{ระยะทางระหว่างค่าเฉลี่ยของ Y})^2}{\text{ความแปรปรวนของ Y}}$ มีค่ามากที่สุด

ให้
$$\ell = \frac{(\text{ระยะทางระหว่างค่าเฉลี่ยของ Y})^2}{\text{ความแปรปรวนของ Y}}$$
$$= \frac{(\mu_{1y}-\mu_{2y})^2}{\sigma_y^2}$$

$$= \frac{(\beta'\mu_1-\beta'\mu_2)^2}{\beta'\Sigma\beta}$$

$$= \frac{\beta'(\mu_1-\mu_2)(\mu_1-\mu_2)'\beta}{\beta'\Sigma\beta} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

ค่า β ที่ได้จะมีค่ามากที่สุดโดย

$$\frac{\partial \ell}{\partial \beta} = 0$$

จะได้ $c(\mu_1-\mu_2)-\Sigma\beta = 0$

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

กำหนดให้ $c = 1$

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

ดังนั้น จะได้ผลรวมเชิงเส้นเป็น

$$Y = \beta'X = (\mu_1-\mu_2)'\Sigma^{-1}X \quad \ldots\ldots\ldots\ldots\ldots(7)$$

โดยที่เราสมมตุว่า Σ เป็น positive definite เมตริกซ์

ค่า Y ที่ได้นี้เรียกว่า Fisher's linear discriminant function

สมการ $Y = (\mu_1-\mu_2)'\Sigma^{-1}X$ นี้เราสามารถใช้เป็นตัวจำแนกค่าสังเกตที่ได้มานั้นว่าจะอยู่ในประชากรกลุ่ม Π_1 หรือ Π_2 ได้ โดยให้

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

เป็นค่าของ discriminant function ของค่าสังเกต X_o และให้

$$m = \frac{1}{2}(\mu_{1y}+\mu_{2y})$$

$$= \frac{1}{2}(\beta'\mu_1+\beta'\mu_2)$$

$$= \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) \qquad \cdots\cdots\cdots\cdots\cdots(8)$$

เป็นจุดกึ่งกลางระหว่างค่าเฉลี่ยของค่า Y จากทั้ง 2 ประชากร

เราสามารถเขียนกฎการจำแนกประเภทหรือกล่าวได้ดังนี้ คือ

จะจัด X_o ให้อยู่ใน π_1 ถ้าหากว่า $Y_o = (\mu_1-\mu_2)'\Sigma^{-1}X_o \geq m$

$\cdots(9)$

จะจัด X_o ให้อยู่ใน π_2 ถ้าหากว่า $Y_o = (\mu_1-\mu_2)'\Sigma^{-1}X_o < m$

ในทางปฏิบัตินั้นค่า μ_1, μ_2 และ Σ เป็นค่าที่เรามักจะไม่ทราบ ดังนั้นถ้าเราสุ่มตัวอย่างจาก π_1 และ π_2 มาเป็นจำนวน n_1 และ n_2 แล้ว โดยที่เราวัดค่าสังเกต

$$X' = \left[x_1, x_2, \ldots\ldots\ldots, x_p\right]$$ เราจะได้ว่า

$$\underset{(pXn_1)}{X_1} = \left[x_{11}, x_{12}, \ldots\ldots\ldots, x_{1n_1}\right]$$

$$\cdots\cdots\cdots\cdots\cdots(10)$$

$$\underset{(pXn_2)}{X_2} = \left[x_{21}, x_{22}, \ldots\ldots\ldots, x_{2n_2}\right]$$

ดังนั้น จากตัวอย่างดังกล่าวจะได้เวคเตอร์ของค่าเฉลี่ยและเมตริกซ์ของโควา-เรียนซ์เป็น

$$\underset{(pX1)}{\overline{X}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1} X_{1j} \quad ; \quad \underset{(pX1)}{\overline{X}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j}$$

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

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

เนื่องจากเราได้สมมติไว้ก่อนแล้วว่า เมตริกซ์ของโควาเรียนซ์ของประชากรทั้ง 2 เป็นค่าเดียวกันคือ Σ ดังนั้น จากตัวอย่างจะได้เมตริกซ์โควาเรียนซ์เป็น S_1 และ S_2 ซึ่งเมื่อ

เรานำมารวมกัน (pooled) ก็จะได้ S pooled เป็นค่าตัวประมาณที่ไม่เอนเอียงของ Σ ดังนี้คือ

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

$$= \frac{(n_1-1)S_1+(n_2-1)S_2}{(n_1+n_2-2)} \quad \ldots\ldots\ldots\ldots\ldots\ldots(12)$$

ดังนั้น ค่าของ $\overline{X}_1$, $\overline{X}_2$ และ S_{pooled} จะแทนลงไปใน (7) ทำให้ได้ Fisher's Sample Linear Discriminant Function ดังนี้คือ

$$Y' = \hat{\beta}'X$$

$$= (\overline{X}_1-\overline{X}_2)'S^{-1}_{pooled}\ X \quad \ldots\ldots\ldots\ldots\ldots\ldots(13)$$

และค่าจุดกึ่งกลางระหว่างตัวแปร $\overline{Y}_1 = \hat{\beta}'\overline{X}_1$ และ $\overline{Y}_2 = \hat{\beta}'\overline{X}_2$ จะเขียนได้เป็น

$$\hat{m} = \frac{1}{2}(\overline{Y}_1+\overline{Y}_2)$$

$$= \frac{1}{2}(\overline{X}_1-\overline{X}_2)'S^{-1}_{pooled}(\overline{X}_1+\overline{X}_2) \quad \ldots\ldots\ldots(14)$$

กฎการจำแนกประเภทจะเขียนได้ว่า ถ้า

$$Y_o = (\overline{X}_1-\overline{X}_2)'S^{-1}_{pooled}\ X_o$$

Y_o จะถูกจัดให้อยู่ใน Π_1 ถ้าหากว่า $Y_o \geqslant \hat{m}$

Y_o จะถูกจัดให้อยู่ใน Π_2 ถ้าหากว่า $Y_o < \hat{m}$

จากสมการ เราสามารถเขียนได้เป็น

$$Y = \beta_1X_1+\beta_2X_2+\beta_3X_3+\ \ldots\ldots\ldots\ +\beta_pX_p$$

ค่า β_1, β_2, ..., β_p จะเป็นค่าแสดงถึงความสำคัญของตัวแปร X_1, X_2, ..., X_p ซึ่งโดยปกติ X_i แต่ละค่านั้นจะมีหน่วยไม่เหมือนกัน การเปรียบเทียบค่า β_i เหล่านี้จะทำได้ก็โดยการปรับค่า β_i เหล่านี้ให้เป็นมาตรฐาน (standardized) เสียก่อน

การแปลงค่า β_i เหล่านี้สามารถทำได้โดยเอาสมาชิกในแนวเส้นทแยงมุม (diagonal) ของเมตริกซ์ W มาถอดรากที่สอง แล้วนำไปคูณกับค่า β_i ทุก ๆ ค่าตามสูตรต่อไปนี้

$$\beta_i^* = (\sqrt{W_{ii}})(\beta_i)$$

เมื่อ W คือ $(\underset{\sim}{X}-\underset{\sim}{\bar{X}}_{g.})(\underset{\sim}{X}-\underset{\sim}{\bar{X}}_{g.})'$

$$\underset{\sim}{X} = \left[\begin{array}{ccc|ccc} X_{111} & X_{121} \ldots\ldots & X_{1n1} & X_{112} & X_{122} \ldots\ldots & X_{1n2} \\ X_{211} & X_{221} \ldots\ldots & X_{2n1} & X_{212} & X_{222} \ldots\ldots & X_{2n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{p11} & X_{p21} \ldots\ldots & X_{pn1} & X_{p12} & X_{p22} \ldots\ldots & X_{pn2} \end{array}\right]$$

กลุ่มที่ 1 กลุ่มที่ 2

$$\underset{\sim}{\bar{X}}_{g.} = \left[\begin{array}{cc|cc} \bar{X}_{11} \ldots\ldots\ldots & \bar{X}_{11} & \bar{X}_{12} \ldots\ldots\ldots\ldots & \bar{X}_{12} \\ \bar{X}_{21} \ldots\ldots\ldots & \bar{X}_{21} & \bar{X}_{22} \ldots\ldots\ldots\ldots & \bar{X}_{22} \\ \vdots & \vdots & \vdots & \vdots \\ \bar{X}_{p1} \ldots\ldots\ldots & \bar{X}_{p1} & \bar{X}_{p1} \ldots\ldots\ldots\ldots & \bar{X}_{p2} \end{array}\right]$$

β_i^* คือ Standardized discriminant weight ของ discriminant function

β_i คือ Raw discriminant weight ของ discriminant function

W_{ii} คือ diagonal element ของ W เมตริกซ์ โดยที่ $i = 1, 2, \ldots, p$

สมการจำแนกประเภทที่ได้มานั้น จะมีความสามารถในการแบ่งแยกกลุ่มได้ดีหรือ

ไม่นั้น เราสามารถดูได้จากค่า miss classification error ซึ่งเราทราบได้จากการที่เปรียบเทียบว่าสิ่งที่เราศึกษาอยู่นั้นเป็นสมาชิกของประชากรกลุ่มหนึ่ง และเมื่อใช้สมการจำแนกประเภทมาจัดการจำแนกแล้วกลับไปเป็นสมาชิกของอีกกลุ่มหนึ่ง ซึ่งทำให้ได้ผลที่ผิดจากความเป็นจริงไป

2.4.2 <u>วิธีการคัดเลือกตัวแปรเข้าไปในสมการจำแนกประเภท</u> (Selection of Variables)

จากที่กล่าวมาแล้วข้างต้นนั้นเป็นหลักการโดยทั่ว ๆ ไปของวิธีการสร้างสมการจำแนกประเภท ในกรณีที่หากว่าในการวิจัยนั้น มีตัวแปรเป็นจำนวนมาก การคัดเลือกตัวแปรให้เหลือจำนวนน้อยที่สุด แต่มีความสามารถในการใช้เป็นตัวจำแนกมากที่สุดนั้น เราสามารถทำได้โดยใช้การคัดเลือกตัวแปรทีละตัว โดยที่จะหาตัวแปรที่ดีที่สุดตัวแรก และตัวแปรที่ดีที่สุดตัวที่สองที่จะทำให้สมการจำแนกประเภทมีความสามารถจำแนกประเภทได้ดีที่สุด จากนั้นก็จะเลือกตัวที่สามและตัวต่อๆไป ที่จะช่วยการจำแนกให้ดีขึ้นตามลำดับ ในแต่ละขั้นตอนตัวแปรที่ได้รับการคัดเลือกมาก่อนแล้วนั้นอาจถูกตัดทิ้งไปหากพบว่าเมื่อนำมารวมกับตัวแปรตัวอื่น ๆ แล้วไม่ช่วยให้สมการจำแนกประเภทได้ดีขึ้น วิธีการนี้เรียกว่าวิธีการสร้างสมการจำแนกประเภทแบบขั้นตอน (Stepwise discriminant analysis)

วิธีการคัดเลือกตัวแปรแบบขั้นตอนนี้เป็นวิธีการคัดเลือกตัวแปรอิสระทีละตัวตามอำนาจการจำแนกของตัวแปรนั้น โดยที่ตัวแปรตัวแรกที่ถูกเลือกเข้าไปนั้นจะมีอำนาจในการจำแนกสูงสุด เมื่อเทียบกับตัวแปรอื่น ๆ ตัวแปรที่เหลือก็จะถูกคัดเลือกเช่นเดียวกัน ทำเช่นนี้จนกระทั่งถึงเกณฑ์หนึ่ง ๆ ซึ่งเราได้กำหนดไว้ว่าตัวแปรที่เหลือนั้นจะไม่สามารถเข้าไปในสมการจำแนกได้อีกแล้ว วิธีการนี้จะคล้ายคลึงกับวิธีการวิเคราะห์ถดถอยพหุแบบขั้นตอน (stepwise multiple regression analysis) ซึ่งโดยปกติเกณฑ์ที่ใช้คือ

กำหนดให้ $\hat{\ell}_p = (\bar{X}_1-\bar{X}_2)'S^{-1}_{pooled}(\bar{X}_1-\bar{X}_2)$ เป็นค่าที่เกิดจากตัวแปร p ตัว

$\hat{\ell}_{p+q} = (\bar{X}_1-\bar{X}_2)'S^{-}_{pooled}(\bar{X}_1-\bar{X}_2)$ เป็นค่าที่เกิดจากตัวแปร p+q ตัว

เราจะกล่าวได้ว่าตัวแปร q ตัวที่เพิ่มขึ้นมานั้น จะไม่มีอำนาจในการจำแนก นั่นคือ ไม่ได้ช่วยให้สมการจำแนกเดิมที่ใช้ตัวแปรอยู่ p ตัวแล้วนั้นมีประสิทธิภาพดีขึ้น ถ้าหากว่า

$$F = \frac{n_1+n_2-p-q-1}{q} \times \frac{m(\hat{\ell}_{p+q}-\hat{\ell}_p)}{1+m\hat{\ell}_p} < F_{\alpha,q,n_1+n_2-p-q-1}$$

โดยที่ n_1 = จำนวนตัวอย่างจากประชากร Π_1

n_2 = จำนวนตัวอย่างจากประชากร Π_2

$$m = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-2)}$$

q = ค่าตัวแปรที่จะเข้าไปในสมการจำแนก ในที่นี้โดยวิธีแบบขั้นตอน q จึงมีค่าเป็น 1 เสมอ

p = จำนวนตัวแปรที่มีอยู่เดิมแล้วในสมการจำแนก

α = ระดับนัยสำคัญที่กำหนดไว้

จากวิธีการดังกล่าวข้างต้น จะทำให้ได้สมการจำแนกประเภทและค่า eigenvalue ซึ่งเป็นค่าที่ได้จากขั้นตอนการหาสมการวิเคราะห์จำแนกประเภทโดยเป็นค่าความแปรปรวนของค่า Y ซึ่งได้จากการแปลงรูป (Transform) จากค่า X ต่าง ๆ $(X_1, X_2, \ldots\ldots, X_p)$ เป็นค่าที่ใช้วัดความสำคัญเชิงเปรียบเทียบของสมการที่ได้ โดยที่สมการวิเคราะห์จำแนกประเภทที่ได้มานั้นได้ตามลำดับความสำคัญของค่า eigenvalue ต่าง ๆ จากมากที่สุดและรองลงไปตามลำดับ ค่าผลรวมของ eigenvalue ทั้งหมดนั้นเป็นค่าของความแปรปรวนรวม (Total variance) ทั้งหมดของตัวแปรจำแนกประเภท (ตัวแปร X) ค่า eigenvalue แต่ละค่าจึงคิดเป็นอัตราส่วนร้อยของค่ารวมของ eigenvalue ทั้งหมด ทำให้สามารถใช้ค่านี้ไปอ้างอิงถึงความสำคัญเชิงเปรียบเทียบของสมการจำแนกประเภทได้

2.4.3 <u>การทดสอบนัยสำคัญของสมการจำแนกประเภท</u>

เมื่อเราได้สมการจำแนกประเภทมาแล้ว เราอยากทราบว่าสมการนี้สามารถมีอำนาจในการจำแนกกลุ่มได้อย่างมีนัยสำคัญทางสถิติหรือไม่ เราสามารถทดสอบได้จาก

$$V_m = \left[N-1-(p+k)/2\right] \ell n(1+\lambda_m)$$

โดยที่ V_m จะมีการกระจายเป็นแบบ Chi-Square ที่มี d.f. = p+k-2m

V คือ Bartlett V Statistic

N คือ จำนวนตัวอย่างทั้งหมดจากประชากรทั้ง 2 กลุ่ม

p คือ จำนวนตัวแปรทั้งหมด

k คือ จำนวนกลุ่ม

λ_m คือ Eigenvalues ต่าง ๆ เช่น ค่าตัวที่ 1, 2,......., m

ค่า λ_m นี้ คำนวณได้จาก

$$|A-\lambda_m I| = 0$$

เมื่อ A คือ โควาเรียนซ์ เมตริกซ์ของ X

ค่า λ_m นี้จะเลือกเอาเฉพาะค่าที่ไม่เป็นศูนย์

ถ้าค่า V_m ที่คำนวณได้มากกว่าค่า Chi-Square ที่ระดับนัยสำคัญที่กำหนดไว้แล้ว แสดงว่าสมการนี้สามารถใช้จำแนกกลุ่มได้อย่างมีนัยสำคัญทางสถิติ