บทที่ 2

ทฤษฎีที่เกี่ยวข้อง

รายละเอียดที่จะกล่าวถึงในบทที่ 2 สามารถแบ่งได้เป็น 3 ส่วน ดังนี้

ส่วนที่หนึ่ง จะเป็นการศึกษาทฤษฎีเกี่ยวกับการวิเคราะห์ข้อมูลตัวอย่างและนำผลจากการวิเคราะห์ไปสรุปลักษณะของประชากร ซึ่งถือว่าเป็นส่วนเริ่มต้นของการดำเนินงานวิจัยฉบับนี้ การวิเคราะห์ข้อมูลขั้นต้น หรือที่เรียกว่าสถิติพรรณนา (descriptive statistics) เป็นการสรุปลักษณะเบื้องต้นของข้อมูลที่เก็บรวบรวมมาได้ ได้แก่ การวิเคราะห์การกระจายของข้อมูล , การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณมากกว่า 2 ตัวขึ้นไป , การแบ่งกลุ่มตัวแปรด้วยเทคนิคการวิเคราะห์ปัจจัย (factor analysis) เพื่อใช้ในการศึกษาความเป็นไปได้ในการลดตัวแปรให้เหลือเฉพาะตัวแปรที่จำเป็นซึ่งมีความสัมพันธ์กันและเกี่ยวข้องกับการออกแบบระบบการจัดขนาดเสื้อ จากนั้นจึงนำผลที่ได้จากการวิเคราะห์ขั้นต้นนี้ไปทำการวิเคราะห์ขั้นสูงต่อไป เพื่อนำผลจากการวิเคราะห์ขั้นสูงไปประกอบการตัดสินใจได้ถูกต้องยิ่งขึ้น

หลังจากทำการวิเคราะห์ข้อมูลขั้นต้นแล้ว ผู้วิจัยจะต้องทำการเลือกจุดวัดที่เหมาะสมและมีความสัมพันธ์กันมาใช้ในการออกแบบระบบแต่เนื่องจากเทคนิคที่สามารถนำมาใช้ในการลดจำนวนตัวแปรนั้นมีอยู่หลายวิธี แต่ละวิธีล้วนมีเงื่อนไขหรือหลักเกณฑ์ที่แตกต่างกันไปตามความเหมาะสมของประเภทข้อมูล ผู้วิจัยจึงเลือกทำการศึกษาเฉพาะทฤษฎีที่น่าจะนำมาประยุกต์ใช้กับงานวิจัยนี้ได้ แล้วจึงเลือกเทคนิคที่เหมาะสมที่สุดมาใช้ลดจำนวนตัวแปร ดังนั้นส่วนที่สองของบทจึงกล่าวถึงทฤษฎีที่เกี่ยวข้องกับการแบ่งกลุ่มข้อมูลออกเป็นกลุ่มย่อยด้วยเทคนิคการวิเคราะห์การจัดกลุ่ม (cluster analysis) ซึ่งผู้วิจัยเห็นว่ามีความเหมาะสมกับลักษณะข้อมูลที่มีอยู่มากที่สุด ทฤษฎีของเทคนิคดังกล่าวบางส่วนจะถูกนำไปใช้ในการกำหนดรูปแบบของปัญหาเกี่ยวกับการออกแบบระบบจัดขนาดที่เหมาะสมที่สุดในเชิงคณิตศาสตร์และนำเสนอออกมาในรูปของปัญหาการหาความเหมาะสมที่สุด (optimization problem)

ในส่วนสุดท้าย จะกล่าวถึงรายละเอียดของอัลกอริทึมที่จะนำมาใช้แก้ปัญหาการหาความเหมาะสมที่สุด ดังนั้นในส่วนนี้จะสรุปทฤษฎีของ Nelder Mead simplex method อย่างคร่าวๆซึ่งใช้เป็นอัลกอริทึมหลักในการแก้ปัญหาภายในงานวิจัยฉบับนี้ เนื่องจากเป็นอัลกอริทึมที่สามารถแก้ปัญหาที่ไม่มีเงื่อนไขกำหนดได้เป็นอย่างดี

2.1 การนำเสนอข้อมูล

ข้อมูลที่นำมาวิเคราะห์โดยทั่วไปแล้วจะประกอบไปด้วยข้อมูลหลายประเภท การพิจารณาแบ่งประเภทของข้อมูลนั้น จะพิจารณาตามลักษณะต่างๆ ได้ดังนี้

2.1.1 แบ่งตามลักษณะของข้อมูล สามารถแบ่งได้เป็น 2 ประเภท คือ

- ข้อมูลเชิงปริมาณ (quantitative data) เป็นข้อมูลที่วัดค่าได้ว่ามีค่ามากหรือน้อย จึงแสดงเป็นตัวเลข เช่น รายได้ อายุ ความสูง น้ำหนัก ซึ่งแบ่งได้เป็น 2 แบบ คือ
  - ก. ข้อมูลแบบไม่ต่อเนื่อง (discrete data) หมายถึง ข้อมูลที่มีค่าเป็นจำนวนเต็มหรือจำนวนนับ เช่น จำนวนคน จำนวนสินค้า ฯลฯ ดังนั้นค่าของข้อมูลแบบนี้อาจเป็น 0, 1, 2....
  - ข. ข้อมูลแบบต่อเนื่อง (continuous data) หมายถึง ข้อมูลที่มีค่าได้ทุกค่าในช่วงที่กำหนดที่มีความหมาย เช่น รายได้ น้ำหนักคน ส่วนสูงคน ฯลฯ
- ข้อมูลเชิงคุณภาพ (quantitative data) หรือข้อมูลจำแนกประเภท (categorical data) หรือบางครั้งเรียกว่าข้อมูลเชิงกลุ่ม เป็นข้อมูลที่ไม่สามารถระบุค่าได้ว่ามากหรือน้อย มักเป็นข้อความ เช่น ความพอดีของเสื้อผ้า ที่บอกเป็นหลวมเกินไป , แคบเกินไป หรือพอดี เป็นต้น

2.1.2 แบ่งตามแหล่งที่มาของข้อมูล สามารถแบ่งได้เป็น 2 ชนิด คือ

- ข้อมูลปฐมภูมิ (primary data) เป็นข้อมูลที่ผู้ใช้หรือหน่วยงานที่ใช้เป็นผู้ทำการเก็บรวบรวมข้อมูลเอง ซึ่งอาจจะได้โดยการสัมภาษณ์ ทดลอง หรือสังเกตการณ์ ข้อมูลนี้จะเป็นข้อมูลที่มีรายละเอียดตรงตามที่ผู้ใช้ต้องการ แต่จะเสียเวลาและค่าใช้จ่ายมาก และเป็นข้อมูลที่ยังไม่ได้ทำการวิเคราะห์
- ข้อมูลทุติยภูมิ (secondary data) เป็นข้อมูลที่ผู้ใช้ไม่ได้ทำการเก็บเอง แต่มีผู้อื่นหรือหน่วยงานอื่นๆ ทำการเก็บข้อมูล ผู้ใช้เป็นเพียงผู้นำข้อมูลมาใช้เท่านั้น

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

2.2 การวิเคราะห์ข้อมูล

หลังจากที่ได้ทำการเก็บรวบรวมข้อมูลมาแล้ว จะต้องวิเคราะห์ข้อมูลเพื่อนำผลสรุปไปใช้ การวิเคราะห์ข้อมูลอาจทำได้ 2 ขั้นตอน ดังนี้

- การวิเคราะห์ข้อมูลขั้นต้น (descriptive statistics) เป็นการสรุปถึงลักษณะของข้อมูลที่มีอยู่ ดังนั้นผู้วิเคราะห์อาจเป็นผู้ที่ไม่มีความรู้ทางสถิติมาก่อนก็ได้ การวิเคราะห์ข้อมูลขั้นต้นอาจพิจารณาในรูปของการแจกแจงความถี่ การหาสัดส่วนหรือร้อยละ การวัดแนวโน้มสู่

ส่วนกลาง เช่น ค่าเฉลี่ย ค่ามัธยฐาน และค่าฐานนิยม การวัดการกระจายข้อมูล เช่น ค่าพิสัย ค่าแปรปรวน ค่าเบี่ยงเบนมาตรฐาน ฯลฯ

- การวิเคราะห์ข้อมูลขั้นสูง หรือเรียกว่า สถิติเชิงอนุมาน (inference statistics) เป็นการสรุปถึงลักษณะของประชากรโดยใช้ข้อมูลตัวอย่าง การวิเคราะห์ในขั้นนี้ ได้แก่ การประมาณค่า การวิเคราะห์ความแปรปรวน การวิเคราะห์ความถดถอยและสหสัมพันธ์ เป็นต้น

2.3 การสรุปลักษณะของข้อมูลด้วยค่าสถิติ

กรณีที่ข้อมูลเป็นข้อมูลเชิงปริมาณจะทำให้สามารถสรุปลักษณะที่สำคัญของข้อมูลชุดนั้นได้ โดยการคำนวณค่าสถิติซึ่งประกอบด้วย

2.3.1 การวัดแนวโน้มสู่ส่วนกลาง (central tendency) เป็นสถิติที่แสดงค่ากลางของข้อมูลเชิงปริมาณ ซึ่งประกอบด้วย

- Mean : ค่าเฉลี่ย
- Median : ค่ามัธยฐาน
- Mode : ฐานนิยม

2.3.2 การวัดการกระจายของข้อมูล (measure of variation) เป็นสถิติที่วัดการกระจายของข้อมูลเชิงปริมาณ ซึ่งสามารถเลือกได้หลายทางเลือก ดังนี้

- Std. Deviation : ค่าเบี่ยงเบนมาตรฐานของข้อมูล
- Variance : ค่าแปรปรวน
- Range : พิสัย
- Minimum : ค่าต่ำสุดของข้อมูล
- Maximum : ค่าสูงสุดของข้อมูล
- S.E. mean : ค่าคลาดเคลื่อนมาตรฐานของค่าเฉลี่ยตัวอย่าง

2.3.3 การแจกแจงของข้อมูล (distribution) ค่าสถิติที่แสดงการแจกแจงของข้อมูลเชิงปริมาณ

- Skewness : ค่าความเบ้ของการแจกแจงของข้อมูล
- Kurtosis : ค่าความโด่งของการแจกแจงของข้อมูล

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

2.4 การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณ 2 ตัว

โดยทั่ว ๆ ไปตัวแปรต่างๆ มักมีความสัมพันธ์กัน เช่น ยอดขายอาจจะมีความสัมพันธ์กับค่าโฆษณา รายจ่ายสัมพันธ์กับรายได้ ดังนั้นเมื่อตัวแปรตัวใดตัวหนึ่งเปลี่ยนไปย่อมมีผลกระทบต่อตัวแปรที่มีความสัมพันธ์กัน เช่น ถ้ารายได้ลดลง อาจทำให้รายจ่ายลดลงด้วย เป็นต้น

การศึกษาความสัมพันธ์ระหว่างตัวแปร แบ่งเป็น 4 ประเภท คือ

- การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเชิงกลุ่ม
- การวิเคราะห์ความสัมพันธ์ของตัวแปรเชิงปริมาณกับตัวแปรเชิงกลุ่ม
- การวิเคราะห์ความสัมพันธ์ของตัวแปรเชิงปริมาณ 2 ตัว
- การวิเคราะห์ความสัมพันธ์ของตัวแปรเชิงปริมาณมากกว่า 2 ตัวขึ้นไป

ในที่นี้จะทำการศึกษาความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณ 2 ตัว ซึ่งเทคนิคที่ใช้ในการวิเคราะห์ความสัมพันธ์จะมี 2 วิธี ดังนี้

1. การวิเคราะห์ความถดถอยอย่างง่าย (regression analysis)
2. การวิเคราะห์สหสัมพันธ์ (correlation analysis)

ทั้งการวิเคราะห์ความถดถอยและสหสัมพันธ์ เป็นการวัดความสัมพันธ์ของตัวแปรเชิงปริมาณในรูปเชิงเส้น ดังนั้นก่อนที่จะทำการวิเคราะห์จึงต้องมีการตรวจสอบรูปแบบความสัมพันธ์ก่อน

วัตถุประสงค์ของการวิเคราะห์ความถดถอยและสหสัมพันธ์

การวิเคราะห์ความสัมพันธ์ของตัวแปรต่างๆ มีวัตถุประสงค์ ดังนี้

1. เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรว่ามีความสัมพันธ์กันมากน้อยเพียงใด ถ้า X และ Y มีความสัมพันธ์กันมาก แสดงว่าถ้า X มีค่าเปลี่ยนแปลงไปจะมีผลกระทบต่อค่าของ Y เป็นอย่างมาก
2. ใช้ความสัมพันธ์ที่วิเคราะห์ได้มาประมาณค่าหรือพยากรณ์ค่า Y ในอนาคตเมื่อกำหนดค่า X

2.5 การตรวจสอบรูปแบบความสัมพันธ์ของตัวแปรเชิงปริมาณ 2 ตัว

ความสัมพันธ์ระหว่างตัวแปร 2 ตัวนั้น อาจจะมีหลายรูปแบบ เช่น สัมพันธ์กันในรูปเชิงเส้น เส้นโค้ง พาราโบลา เป็นต้น ดังนั้นก่อนที่จะคำนวณค่าสถิติเพื่อแสดงถึงระดับความสัมพันธ์นั้น จึงจำเป็นต้องทำการตรวจสอบลักษณะของความสัมพันธ์ก่อน (โดยทั่วไปจะนิยมตรวจสอบ

โดยการพล๊อตกราฟแสดงความสัมพันธ์ระหว่างตัวแปรทั้งสอง เนื่องจากเป็นวิธีที่ง่ายและสะดวกที่สุด) หลังจากตรวจสอบลักษณะของความสัมพันธ์เรียบร้อยแล้ว จึงจะสามารถเลือกเทคนิคของการวิเคราะห์ได้อย่างถูกต้อง

2.6 การวิเคราะห์ความถดถอยอย่างง่าย (simple regression analysis)

เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณ 2 ตัว หรือ ลักษณะที่สนใจศึกษา 2 ลักษณะ โดยที่ต้องทราบค่าของตัวแปรหนึ่งหรือต้องกำหนดค่าของตัวแปรตัวหนึ่งไว้ล่วงหน้า ซึ่งตัวแปรที่เราต้องกำหนดไว้ล่วงหน้านี้ เรียกว่า ตัวแปรอิสระ (independent variable) หรือ ตัวแปรต้นเหตุ (predictor variable) และมักจะใช้สัญลักษณ์ "X" ส่วนตัวแปรที่สามารถหาได้ตามความสัมพันธ์ระหว่างตัวแปรทั้งสองนี้ จะเรียกว่า ตัวแปรตาม (dependent variable) และใช้สัญลักษณ์ " Y "

2.6.1 การวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย (simple linear regression analysis)

เป็นการศึกษาถึงความสัมพันธ์ระหว่างตัวแปร 2 ตัว ที่มีความสัมพันธ์อยู่ในรูปเชิงเส้น ซึ่งสามารถแสดงความสัมพันธ์ในรูปของสมการเชิงเส้น ดังนี้

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad ; \quad i = 1,2,\ldots,N \qquad \ldots\ldots\ldots\ldots (2.1)$$

โดยที่ Y คือ ตัวแปรตาม (dependent variable) เนื่องจากค่าของ Y ขึ้นอยู่กับค่าของ X

X คือ ตัวแปรอิสระ (independent variable)

e_i คือ ความคลาดเคลื่อนอย่างสุ่ม (random error)

β_0 คือ ส่วนตัดแกน Y หรือ คือค่าของ Y เมื่อ X เป็นศูนย์

β_1 คือ ความชัน (slope) ของเส้นตรง ซึ่งเป็นค่าที่แสดงถึงอัตราการเปลี่ยนแปลงของ Y เมื่อ X เปลี่ยนไป 1 หน่วยและจะเรียก β_1 ว่า " สัมประสิทธิ์ความถดถอย " (regression coefficient)

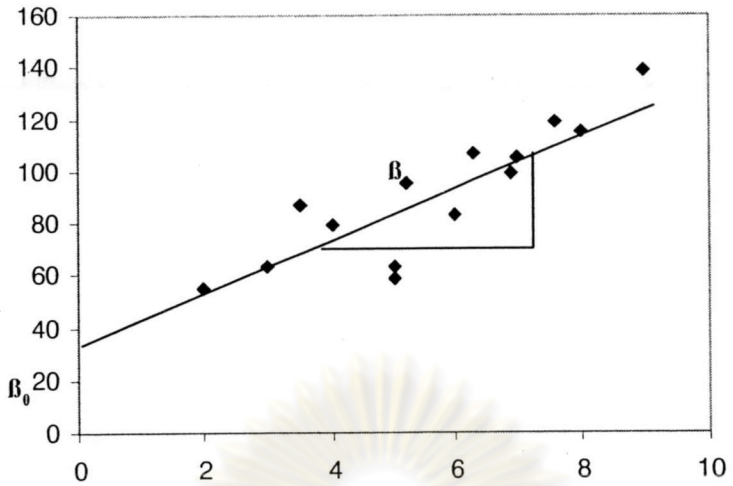
ค่าของ β_1 อาจจะเป็น $\beta_1 > 0$ แสดงว่า X และ Y มีความสัมพันธ์ในทางเดียวกัน คือ ถ้า X เพิ่ม Y จะเพิ่มด้วย แต่ถ้า X ลดลง Y ก็จะลดลง

ก. $\beta_1 < 0$ แสดงว่า X และ Y มีความสัมพันธ์ในทางตรงข้ามกัน คือ ถ้า X เพิ่มขึ้น Y จะลดลง แต่ถ้า X ลดลง Y ก็จะเพิ่มขึ้น

ข. $\beta_1 = 0$ แสดงว่า X และ Y ไม่มีความสัมพันธ์กันเลย

ค. β_1 มีค่าเข้าใกล้ศูนย์ แสดงว่า X และ Y มีความสัมพันธ์กันน้อย

สำหรับหน่วยของ β_0 และ β_1 จะมีหน่วยเหมือนกับหน่วยของ Y เสมอ



รูปที่ 2.1 : แสดงค่า เมื่อ X และ Y สัมพันธ์กันในรูปเชิงเส้น

2.6.2 การประมาณค่าพารามิเตอร์ของสมการความถดถอยเชิงเส้นอย่างง่าย

เมื่อพิจารณาจากแผนภาพการกระจาย ซึ่งแสดงความสัมพันธ์ระหว่าง X และ Y หาก พบว่า X และ Y สัมพันธ์กันในรูปเส้นตรง จะต้องคำนวณหาค่า $ß_0$ และ $ß_1$ ซึ่งจะทำให้ทราบถึงความสัมพันธ์ระหว่าง X และ Y ว่ามีความสัมพันธ์ตามกันหรือตรงข้ามกันและความสัมพันธ์นั้นมากหรือน้อยเพียงใด ถ้า $ß_1$ มีค่ามาก แสดงว่า Y มีความสัมพันธ์กับ X มากด้วย

การที่จะหาค่า $ß_0$ และ $ß_1$ ได้ จำเป็นต้องทราบค่า X และ Y ทุกค่าที่ได้เกิดขึ้นแล้วในอดีต ซึ่งเป็นไปได้ยากในทางปฏิบัติ ดังนั้นจึงใช้ข้อมูลตัวอย่างขนาด n ในการประมาณค่า $ß_0$ และ $ß_1$ ดังนั้นค่าประมาณของ Y คือ

$$Y_i = ß_0 + ß_1 x_i + e_i \quad \text{..... (2.2)}$$

หรือ $\hat{Y}_i = a + bX_i$

โดยที่ $\hat{\beta}_0 = a$ และ $\hat{\beta}_1 = b$

2.6.3 การทดสอบความเหมาะสมของสมการความถดถอยเชิงเส้นอย่างง่าย

การทดสอบว่า สมการที่ 2.2 เป็นสมการที่แสดงความสัมพันธ์ระหว่าง X และ Y จริงหรือไม่ จะต้องมีการทดสอบสมมติฐาน

$H_0 : ß_1 = 0$ หรือ H_0 : Y ไม่มีความสัมพันธ์กับ X ในรูปเชิงเส้น

$H_1 : ß_1 \neq 0$ หรือ H_0 : Y มีความสัมพันธ์กับ X ในรูปเชิงเส้น

สถิติที่ใช้ในการทดสอบ คือ

1. F-test จากการวิเคราะห์ความแปรปรวนแบบทางเดียว (1-way ANOVA)
2. t-test : $t = \frac{b_1}{S_{b1}}$

- **การใช้สถิติทดสอบ F**

ในการศึกษาการวิเคราะห์ความแปรปรวนแบบทางเดียว จะแยกความผันแปรของ Y เป็น 2 ส่วน คือ

1. ความผันแปรของ Y ที่เกิดจากอิทธิพลของตัวแปร X (regression sum square)
2. ความผันแปรของ Y ที่เกิดจากอิทธิพลของปัจจัย (ตัวแปร) อื่นๆ ที่สัมพันธ์กับตัวแปร Y (residual sum square or error sum square)

   หรือ Total Sum Square = Regression Sum Square + Residual Sum Square

ตารางที่ 2.1 : 1-way ANOVA สำหรับการวิเคราะห์ความถดถอย

	Sum of Square	*Df*	*Mean Square*	*F*	*Significance*
Regression	SSR	1	MSR = SSR	$\frac{MSR}{MSE}$	$P(F > F_{cal})$
Residual (Error)	SSE	n-2	$MSE = \frac{SSE}{n-2}$		
Total	SST	n-1			

จะปฏิเสธ H_0 ถ้า $F > F_{1-\alpha}$ ที่ degree of freedom 1 , n-2 หรือ significance น้อยกว่าระดับนัยสำคัญที่กำหนด

- **การใช้สถิติทดสอบ t**

สำหรับสมมติฐานของการทดสอบเมื่อใช้สถิติทดสอบ t จะแบ่งเป็น 3 แบบ ดังนี้

ก. $H_0 : \beta_1 = 0$ หรือ H_0 : Y ไม่มีความสัมพันธ์กับ X ในรูปเชิงเส้น

   $H_1 : \beta_1 \neq 0$ หรือ H_0 : Y มีความสัมพันธ์กับ X ในรูปเชิงเส้น

ข. $H_0 : \beta_1 \leq 0$ หรือ H_0 : Y ไม่มีความสัมพันธ์กับ X ในรูปเชิงเส้น

   $H_1 : \beta_1 > 0$ หรือ H_0 : Y มีความสัมพันธ์กับ X ในรูปเชิงเส้นด้านบวก

ค. $H_0 : \beta_1 \geq 0$ หรือ H_0 : Y ไม่มีความสัมพันธ์กับ X ในรูปเชิงเส้น

   $H_1 : \beta_1 < 0$ หรือ H_0 : Y มีความสัมพันธ์กับ X ในรูปเชิงเส้นด้านลบ

สถิติทดสอบ : $t = \frac{b_1}{S_{b1}}$

ตารางที่ 2.2 : ตารางแสดงเขตปฏิเสธสมมติฐาน H_0

สมมติฐาน	*เขตปฏิเสธสมมติฐาน H_0*
ก. $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$	เมื่อ $t > t_{1.}\alpha/2$ หรือ $t < -t_{1.}\alpha/2$ ที่ df. = n-1 หรือ Significance ของ t น้อยกว่าระดับนับสำคัญที่กำหนด - ถ้าปฏิเสธ H_0 จะยอมรับ $H_1 : \beta_1 \neq 0$ จึงสรุปได้ว่าตัวแปร X และ Y มีความสัมพันธ์ในรูปเชิงเส้นที่ระดับนัยสำคัญที่กำหนด - ถ้ายอมรับ $H_0 : \beta_1 = 0$ จะสรุปได้ว่าตัวแปร X และ Y ไม่มีความสัมพันธ์ในรูปเชิงเส้นที่ระดับนัยสำคัญที่กำหนด
ข. $H_0 : \beta_1 \leq 0$ $H_1 : \beta_1 > 0$	เมื่อ $t > t_{1-\alpha}$ ที่ df. = n-1 หรือเมื่อ $\frac{significance}{2}$ น้อยกว่าระดับนัยสำคัญที่กำหนด และ t มีค่าเป็นบวก - ถ้าปฏิเสธ H_0 จะยอมรับ $H_1 : \beta_1 > 0$ หรือตัวแปร X และ Y มีความสัมพันธ์ในรูปเชิงเส้นและความสัมพันธ์เป็นบวก - ถ้ายอมรับ $H_0 : \beta_1 \leq 0$ หมายถึง X และ Y ไม่มีความสัมพันธ์ในรูปเชิงเส้น
ค. $H_0 : \beta_1 \geq 0$ $H_1 : \beta_1 < 0$	เมื่อ $t < -t_{1-\alpha}$ ที่ df. = n-1 หรือ $\frac{significance}{2}$ น้อยกว่าระดับนัยสำคัญที่กำหนด และ t มีค่าเป็นลบ - ถ้าปฏิเสธ H_0 จะยอมรับ $H_1 : \beta_1 < 0$ นั่นคือ X และ Y มีความสัมพันธ์ในรูปเชิงเส้นและความสัมพันธ์เป็นลบ - ถ้ายอมรับ $H_0 : \beta_1 \geq 0$ หมายถึง ตัวแปร X และ Y ไม่มีความสัมพันธ์ในรูปเชิงเส้น

2.6.4 สัมประสิทธิ์การตัดสินใจ (coefficient of determinant : R Square)

สัมประสิทธิ์การตัดสินใจ หมายถึง สัดส่วนที่ตัวแปร X สามารถอธิบายการเปลี่ยนแปลงของตัวแปร Y ได้ ดังนั้นถ้า R^2 มีค่ามากแสดงว่า Y และ X มีความสัมพันธ์กันมากหรือ X สามารถอธิบายการเปลี่ยนแปลงของค่า Y ได้มากโดยที่

$$R^2 = \frac{\text{ความแปรปรวนของ Y ที่เกิดจาก X}}{\text{ความแปรปรวนของ Y ทั้งหมด}}$$

$$= \frac{SSR}{SST}$$

ดังนั้น $0 \leq R^2 \leq 1$ เนื่องจาก SST > SSR

หาก R^2 มีค่าเข้าใกล้ 1 แสดงว่าเปอร์เซ้นต์ที่ X สามารถอธิบายการเปลี่ยนแปลงของ Y มีค่ามาก หรือ X และ Y มีความสัมพันธ์กันมาก แต่ถ้า R^2 มีค่าเข้าใกล้ 0 แสดงว่าเปอร์เซ็นต์ที่ X สามารถอธิบายการเปลี่ยนแปลงของ Y มีค่าน้อย

2.6.5 การตรวจสอบเงื่อนไขของการวิเคราะห์ความถดถอยเชิงเส้น

สมมติฐานหรือเงื่อนไขของการวิเคราะห์ความถดถอย มี 4 ข้อ ซึ่งเป็นเงื่อนไขเกี่ยวกับความคลาดเคลื่อน (error of residual) การนำสมการ ไปประยุกต์ใช้ทั้งในแง่แสดงระดับและทิศทางของความสัมพันธ์ระหว่างตัวแปร X และ Y หรือพยากรณ์ค่า Y เมื่อกำหนดค่า X ได้นั้น จะต้องมั่นใจในความถูกต้องของสมการ $\hat{Y} = a + bX$ ก่อน โดยจะต้องตรวจสอบเงื่อนไขของการวิเคราะห์ความถดถอยเกี่ยวกับค่าคลาดเคลื่อน ($e_i = Y_i - \hat{Y}$) ดังนี้

1. ค่าเฉลี่ยของค่าคลาดเคลื่อนจะต้องเท่ากับ 0 (E(e) = 0)
2. ค่าแปรปรวนของ e คือ σ^2 ซึ่งจะต้องคงที่ทุกค่าของ X
3. ค่าความคลาดเคลื่อนต้องเป็นอิสระกัน
4. ค่าคลาดเคลื่อนต้องมีการแจกแจงแบบปกติ

เงื่อนไขดังกล่าวทั้งสี่ข้อจะต้องเป็นจริงจึงจะสามารถใช้ทดสอบ F และ t ในการทดสอบความสัมพันธ์ระหว่างตัวแปร X และ Y ได้ และก่อนที่จะทำการคำนวณหาค่าสัมประสิทธิ์ความถดถอย (a และ b) จะต้องตรวจสอบว่าความสัมพันธ์ของ X และ Y อยู่ในรูปเชิงเส้นจริงหรือไม่ ในกรณีที่ตรวจสอบค่าคลาดเคลื่อนแล้วพบว่า ไม่เป็นไปตามเงื่อนไขจะต้องพิจารณารูปแบบความสัมพันธ์ของตัวแปรซึ่งอาจจะอยู่ในรูปแบบอื่นที่ไม่ใช่เชิงเส้น

2.7 สัมประสิทธิ์สหสัมพันธ์ (correlation coefficient)

จากหัวข้อ 2.6 เป็นการวิเคราะห์เพื่อหาความสัมพันธ์ของตัวแปร X และ Y และเป็นการทดสอบว่า X และ Y มีความสัมพันธ์กันในรูปเชิงเส้นหรือไม่ โดยการทดสอบสัมประสิทธิ์ความถดถอย (β_1) โดยใช้ค่าประมาณ (b) ซึ่งไม่สามารถระบุได้ว่า X และ Y มีความสัมพันธ์กันมากหรือน้อย เนื่องจาก b ที่ได้มีหน่วยตามค่าของ Y เช่น Y มีหน่วยเป็นบาท ค่า b จะมีหน่วยเป็นบาทด้วย แต่ถ้าค่า Y มีหน่วยเป็นล้านบาท ค่า b ก็จะมีหน่วยเป็นล้านบาทด้วย เมื่อค่า b ที่มีหน่วยเป็นบาทมีค่ามากกว่าค่า b ที่มีหน่วยเป็นล้านบาท ไม่ได้หมายความว่าความสัมพันธ์ของ X และ Y ที่ Y มีหน่วยเป็นบาทจะมีความสัมพันธ์มากกว่ากรณีที่ Y มีหน่วยเป็นล้านบาท นั่นคือการที่กำหนดให้หน่วยของ Y แตกต่างกัน จะทำให้ค่า b แตกต่างกันด้วย (X และ Y ไม่จำเป็นต้องมีหน่วยเหมือนกัน) ดังนั้นหากผู้วิจัยใช้หน่วยแตกต่างกัน จะได้ค่า b ที่แตกต่างกันด้วยและค่า b ไม่มีขีดจำกัด (คือ ไม่ทราบค่าสูงสุดและค่าต่ำสุดของ b) จึงมีการหาค่าสถิติที่สามารถแสดงความสัมพันธ์ระหว่าง X และ Y ที่สามารถแสดงว่าความสัมพันธ์นั้นมากหรือน้อยเพียงใด

สำหรับสถิติที่ใช้วัดความสัมพันธ์ระหว่าง X และ Y ว่ามีความสัมพันธ์กันมากหรือน้อยนั้น จะเรียกว่า สัมประสิทธิ์สหสัมพันธ์ (ρ) ซึ่งในกรณีที่ค่าของ Y ขึ้นกับ X เพียงตัวเดียวจะเรียกว่า **"สัมประสิทธิ์สหสัมพันธ์อย่างง่าย" (simple correlation coefficient)** โดยที่ ρ จะไม่มีหน่วย จึงสามารถใช้วัดความสัมพันธ์ระหว่าง Y และ X ได้ว่ามีความสัมพันธ์มากหรือน้อยเพียงใด เนื่องจากค่า ρ จะมีค่าสูงสุดเป็น 1 และต่ำสุดเป็น -1

นอกจากนั้นความถดถอยและสหสัมพันธ์ยังมีวิธีการในการเก็บข้อมูลที่แตกต่างกัน คือ สำหรับเรื่องความถดถอยนั้น ตัวแปรอิสระ X จะต้องถูกกำหนดไว้ล่วงหน้า ในขณะที่ตัวแปรตามหรือ Y จะเป็นตัวแปรสุ่ม แต่ในการวิเคราะห์สหสัมพันธ์นั้น ทั้ง X และ Y จะต้องเป็นตัวแปรสุ่มทั้งคู่

ในการวิเคราะห์ข้อมูลตัวอย่าง จะทำการประมาณค่า ρ ด้วยค่า r โดยที่ค่า r คือสัมประสิทธิ์สหสัมพันธ์ตัวอย่าง

ความหมายของค่า r

1. ค่า r เป็นลบ แสดงว่า X และ Y มีความสัมพันธ์กันในทิศทางตรงข้าม คือ ถ้า X เพิ่ม Y จะลด แต่ถ้า X ลด Y ก็จะเพิ่ม
2. ค่า r เป็นบวก แสดงว่า X และ Y มีความสัมพันธ์กันในทิศทางเดียวกัน คือ ถ้า X เพิ่ม Y จะเพิ่ม แต่ถ้า X ลด Y ก็จะลดด้วย
3. ถ้า r มีค่าเข้าใกล้ 1 หมายถึง X และ Y สัมพันธ์ในทิศทางเดียวกันและมีความสัมพันธ์กันมาก

4. ถ้า r มีค่าเข้าใกล้ -1 หมายถึง X และ Y สัมพันธ์ในทิศทางตรงกันข้ามและมีความสัมพันธ์กันมาก
5. ถ้า r = 0 แสดงว่า X และ Y ไม่มีความสัมพันธ์กันเลย
6. ถ้า r เข้าใกล้ 0 แสดงว่า X และ Y มีความสัมพันธ์กันน้อย

2.8 การตรวจสอบการแจกแจงของข้อมูล

2.8.1 เหตุผลที่ต้องทำการตรวจสอบการแจกแจงของข้อมูล

ในการทำวิจัย ข้อมูลที่ได้มาส่วนใหญ่เป็นข้อมูลตัวอย่าง ดังนั้นจึงต้องทำการคำนวณค่าสถิติ หรือใช้เทคนิคการวิเคราะห์ทางสถิติ เพื่อสรุปลักษณะของประชากร จึงจำเป็นที่จะต้องมีการตรวจสอบการแจกแจงหรือลักษณะของข้อมูลตัวอย่าง เพื่อที่จะได้อ้างอิงถึงลักษณะของประชากรต่อไป การตรวจสอบการแจกแจงของข้อมูล แบ่งเป็น

1. ข้อมูลเชิงคุณภาพหรือเชิงกลุ่ม การตรวจสอบการแจกแจง คือ การสร้างตารางการแจกแจงเพื่อพิจารณาถึงความถี่ และร้อยละ และการพิจารณาจาก bar chart

2. ในการวิเคราะห์ข้อมูลเชิงปริมาณด้วยเทคนิคทางสถิติต่างๆ โดยเฉพาะอย่างยิ่งการประมาณค่าแบบช่วง การทดสอบสมมติฐาน และ การวิเคราะห์ความถดถอย (regression analysis) มีเงื่อนไขว่า ตัวแปรหรือประชากรที่จะนำมาวิเคราะห์จะต้องมีการแจกแจงแบบปกติ

ส่วนเทคนิคการวิเคราะห์ความแปรปรวน จะต้องมี 2 เงื่อนไข คือ

1. ประชากรหรือตัวแปรที่ศึกษา ต้องมีการแจกแจงแบบปกติทุกตัวแปรหรือทุกประชากร
2. ความแปรปรวนของทุกประชากรต้องเท่ากัน

การศึกษาสถิติเป็นการศึกษาถึงลักษณะที่สำคัญของประชากร แต่ส่วนใหญ่ข้อมูลที่มีอยู่หรือเก็บรวบรวมมาได้มักจะเป็นข้อมูลตัวอย่าง (บางส่วนของประชากร) จึงต้องทำการวิเคราะห์ข้อมูลตัวอย่างนั้นเพื่ออ้างอิงถึงลักษณะของประชากร หรือบางครั้งอาจจะต้องการพยากรณ์ค่าของพารามิเตอร์ในอนาคตโดยการใช้ข้อมูลตัวอย่าง โดยเทคนิคการวิเคราะห์ข้อมูลจะมีเงื่อนไขเกี่ยวกับการแจกแจงของประชากรกำหนดอยู่ หากผลการตรวจสอบสามารถสรุปได้ว่าประชากรมีการแจกแจงแบบปกติ จึงจะสามารถใช้การทดสอบสมมติฐานที่ใช้พารามิเตอร์ทางสถิติได้

2.8.2 วิธีการตรวจสอบการแจกแจงของข้อมูลเชิงปริมาณ

สำหรับวิธีการตรวจสอบการแจกแจงของข้อมูลว่ามีการแจกแจงแบบปกติหรือใกล้เคียงแบบปกติหรือไม่ สามารถแบ่งได้เป็น 2 วิธีใหญ่ คือ

1. การตรวจสอบโดยการใช้กราฟ ซึ่งกราฟที่ใช้ในการตรวจสอบการกระจายของข้อมูล ได้แก่

   - ฮิสโตแกรม
   - แผนภาพลำต้นและใบ (Stem and Leaf)
   - Boxplot
   - Normal probability plot
   - Detrended normal plot

2. การทดสอบโดยการใช้สถิติทดสอบ ประกอบด้วย

   - Lilliefor's test ใช้ตรวจสอบว่าการแจกแจงของประชากรเป็นแบบปกติหรือใกล้เคียงแบบปกติหรือไม่
   - Levene's test ใช้ตรวจสอบว่าค่าแปรปรวนของแต่ละประชากร (กลุ่ม) เท่ากันหรือไม่

ในที่นี้จะขอกล่าวถึงรายละเอียดของการตรวจสอบการแจกแจงของข้อมูลเชิงปริมาณโดยใช้สถิติทดสอบเท่านั้น

2.8.3 การตรวจสอบการแจกแจงของข้อมูลโดยสถิติทดสอบ

จากการตรวจสอบการแจกแจงข้อมูลเชิงปริมาณด้วยกราฟ ผู้วิเคราะห์จะต้องพิจารณาจากกราฟและสรุปเองว่าข้อมูลมีความสมมาตรหรือมีการแจกแจงแบบปกติหรือไม่ จึงเป็นไปได้ที่ผู้วิเคราะห์จะสรุปผลต่างกัน ดังนั้นการใช้สถิติทดสอบเพื่อทดสอบว่า ข้อมูลมีการแจกแจงแบบปกติหรือไม่จึงเป็นสิ่งที่ควรพิจารณาร่วมกับการวิเคราะห์จากกราฟเสมอ

สำหรับสถิติที่ใช้ในการทดสอบนั้น มี 2 ประเภท คือ Kolmogorov-Smirnov test และ Shapiro-Wilk test

- <u>Kolmogorov-Smirnov test</u>

เป็นสถิติทดสอบที่ใช้ทดสอบการแจกแจงของประชากรว่าเป็นแบบปกติหรือไม่ หลักการของการทดสอบนี้ คือ การเปรียบเทียบค่าฟังก์ชั่นการแจกแจงสะสมของข้อมูลตัวอย่างกับค่าฟังก์ชั่นการแจกแจงสะสมของข้อมูลภายใต้สมมติฐานที่ว่า "ประชากร/ข้อมูลมีการแจกแจงแบบปกติหรือไม่" ถ้าค่าความแตกต่างต่ำ แสดงว่า การแจกแจงเป็นแบบปกติ

สมมติฐานของการทดสอบ คือ

H_0 : สุ่มตัวอย่างจากประชากรที่มีการแจกแจงแบบปกติ

H_1 : สุ่มตัวอย่างจากประชากรที่ไม่มีการแจกแจงแบบปกติ

สถิติทดสอบ คือ Kolmogorov

เขตปฏิเสธ H_0 : จะปฏิเสธ H_0 ถ้าค่า significance น้อยกว่าระดับนัยสำคัญที่กำหนด

- Shapiro-Wilk Test

เป็นสถิติที่ใช้ทดสอบการแจกแจงของตัวแปรเชิงปริมาณว่ามีการแจกแจงแบบปกติหรือไม่ สมมติฐานของการทดสอบ คือ

H_0 : สุ่มตัวอย่างจากประชากรที่มีการแจกแจงแบบปกติ

H_1 : สุ่มตัวอย่างจากประชากรที่ไม่มีการแจกแจงแบบปกติ

สถิติทดสอบ คือ Shapior-Wilk

เขตปฏิเสธ H_0 : จะปฏิเสธ H_0 ถ้าค่า Significance น้อยกว่าระดับนัยสำคัญที่กำหนด

หมายเหตุ ความแตกต่างระหว่าง Shapiro-Wilks test และ Kolmogorov-Smirnov test คือ Kolmogorov-Smirnov test จะใช้ในกรณีที่ไม่ทราบค่าเฉลี่ยและค่าแปรปรวนของประชากร จึงต้องใช้ค่าเฉลี่ยและค่าแปรปรวนของตัวอย่างแทน ส่วน Shapiro-Wilks test สามารถใช้ได้กับปัญหาที่อาจจะทราบหรือไม่ทราบค่าเฉลี่ยหรือค่าแปรปรวนของประชากรก็ได้ แต่ขนาดตัวอย่างจะต้องไม่เกิน 50 หน่วย

ดังนั้นงานวิจัยฉบับนี้ จึงใช้ Kolmogorov-Smirnov test เป็นสถิติทดสอบการแจกแจงของข้อมูล เนื่องจากมีขนาดตัวอย่าง 2000 ตัวอย่าง (มากกว่า 50)

2.9 การวิเคราะห์ปัจจัย (factor analysis)

การวิเคราะห์ปัจจัย (หรือ factor analysis) เป็นเทคนิคการวิเคราะห์ตัวแปรหลายตัว (multivariate statistical technique) อย่างหนึ่งที่สามารถนำไปใช้ได้แทบทุกวงการวิชาการ

การวิเคราะห์ปัจจัย เดิมมีผู้เรียกกันว่า การวิเคราะห์ตัวประกอบ ซึ่งการเรียกเช่นนี้อาจจะถูกต้อง เพราะเทคนิคหลักของการวิเคราะห์ปัจจัย คือ การหาองค์ประกอบหลัก (principal components) ของตัวแปรต่างๆ แต่ในปัจจุบันนี้การวิเคราะห์ปัจจัยนี้มีเทคนิคต่างๆ มากมายในการแยกปัจจัยต่างๆ จากข้อมูลที่นอกเหนือไปจากวิธีการตัวประกอบหรือองค์ประกอบหลัก ดังนั้นจึงเปลี่ยนมาใช้ชื่อภาษาไทยว่า การวิเคราะห์ปัจจัย แทนชื่อว่า การวิเคราะห์ตัวประกอบ

สำหรับเทคนิคการวิเคราะห์ปัจจัย คือ วิธีการทางสถิติที่ใช้ลดจำนวนข้อมูลที่มีอยู่มากๆ ให้น้อยลงเพื่อให้อยู่ในสภาพที่สามารถเข้าใจและจัดการกับข้อมูลที่น้อยลงได้ง่ายขึ้น เทคนิคดังกล่าวไม่มีการแบ่งตัวแปรออกเป็นตัวแปรอิสระและตัวแปรตาม เนื่องจากไม่ใช่เทคนิคการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ

เมื่อผู้วิจัยมีจำนวนตัวแปรมากๆ ไม่ทราบว่าตัวแปรต่างๆ เหล่านี้มีการจัดกลุ่มหรือมีแบบแผนอย่างไร เทคนิคการวิเคราะห์ปัจจัยจะทำการวิเคราะห์จัดกลุ่มหรือหาแบบแผนการรวมกันของตัวแปรเหล่านี้ให้เป็นกลุ่มๆ โดยจำนวนกลุ่มจะมีน้อยกว่าจำนวนตัวแปรมากๆ เช่นเดิมอาจมีตัวแปร 24 ตัว อาจจะได้กลุ่มของปัจจัยเพียง 3-4 ปัจจัย

2.9.1 เป้าหมายของการวิเคราะห์ปัจจัย

การวิเคราะห์ปัจจัยมีเป้าหมายที่สำคัญ 2 ประการ คือ

1. การลดจำนวนตัวแปร (ข้อมูล) ให้น้อยลง (data reduction)

เป้าหมายหลักของการลดจำนวนตัวแปรให้น้อยลงทำได้โดยอาศัยความสัมพันธ์ระหว่างตัวแปร หรือความร่วมกันระหว่างตัวแปรเป็นฐานในการเปลี่ยนสภาพตัวแปรให้เป็นปัจจัยเพียงไม่กี่ปัจจัย โดยแต่ละปัจจัยจะประกอบด้วยตัวแปรหลายตัว ในการเปลี่ยนสภาพตัวแปรให้เป็นปัจจัยจะต้องพยายามให้ปัจจัยที่ได้สามารถอธิบายการผันแปรของตัวแปรให้ได้มากที่สุด หรือสามารถอธิบายการผันแปรของตัวแปรทั้งหมดได้มากที่สุดไม่ต่ำกว่าร้อยละ 75

สมมติว่ามีตัวแปรอยู่ 20 ตัว เป็นตัวแปรที่ถูกแปลงค่าเป็นค่ามาตรฐาน คือ ค่าที่มีค่าเฉลี่ยเท่ากับ 0 และค่าการผันแปรเท่ากับ 1 (ซึ่งสามารถทำได้โดยการหาค่าเฉลี่ยของตัวแปรแต่ละตัวแล้วนำเอาค่าเฉลี่ยนั้นมาลบออกจากค่าจริง หารด้วยค่าเบี่ยงเบนมาตรฐาน) การผันแปรรวมทั้งหมดของตัวแปร 20 ตัวนี้ เรียกว่า ค่าไอเค็น (Eigenvalue) ซึ่งเท่ากับค่าการผันแปรของแต่ละตัวแปรรวมกัน หากภายหลังการวิเคราะห์ปัจจัยแล้วพบว่าสามารถแยกปัจจัยออกมาได้ 4 ปัจจัย ปัจจัยดังกล่าวสามารถอธิบายตัวแปรทั้งหมดได้ครบถ้วน นั่นก็หมายความว่าปัจจัยเหล่านี้ อธิบายค่าการผันแปรของตัวแปรได้ทั้งหมด 20 ค่าไอเค็น

ในทางปฏิบัติ ปัจจัยที่สกัดได้จากการวิเคราะห์อาจจะต้องมีหลายปัจจัยจึงจะสามารถอธิบายการผันแปรได้ครบ ดังนั้นในบางครั้งเพื่อความสะดวกในการจัดการวิเคราะห์ จึงไม่จำเป็นที่จะต้องเลือกปัจจัยทุกตัวเพียงเพื่อให้สามารถอธิบายการผันแปรของตัวแปรได้ทั้งหมด แต่อาจจะเลือกมาเพียงบางปัจจัยที่สามารถอธิบายการผันแปรของตัวแปรได้มากเพียงพอ (เช่น 80% หรือมากกว่านั้น)

การที่เทคนิคการวิเคราะห์ปัจจัยสามารถหาปัจจัยเพียงไม่กี่ปัจจัยมาแทน หรือมาอธิบายการผันแปรของตัวแปรจำนวนมากได้ ทำให้ผู้วิจัยสามารถนำปัจจัยดังกล่าวเหล่านั้นไปใช้ได้สะดวกกว่าการใช้ตัวแปรจำนวนมาก เพราะจำนวนข้อมูล (จำนวนปัจจัย) ที่ได้จากการวิเคราะห์นั้นมีจำนวนน้อยกว่าการใช้ตัวแปรเดิมการก่อนวิเคราะห์ปัจจัย นอกจากนั้น เทคนิคการวิเคราะห์บางวิธี เช่น การวิเคราะห์การถดถอย (regression analysis) เมื่อนำตัวแปรจำนวนมากมาใช้ในการวิเคราะห์มักจะก่อให้เกิดปัญหาความสัมพันธ์ระหว่างตัวแปรอิสระ (multicollinearity problems)

แต่เมื่อสกัดตัวแปรเหล่านี้ให้เหลือเพียงไม่กี่ปัจจัยที่ไม่มีความสัมพันธ์กัน (เช่นใช้วิธีสกัดแบบองค์ประกอบและหมุนแกนปัจจัยแบบมุมฉาก) ก็จะทำให้ปัจจัยเหล่านี้แทนตัวแปรทั้งหมดได้

การที่ตัวแปรที่สัมพันธ์กันหรือมีส่วนที่มีความร่วมกัน (communality) สูงมารวมกันเป็นปัจจัยเดียวกัน ทำให้ปัจจัยแต่ละปัจจัยมีความหมายมากขึ้น ความหมายของปัจจัยเหล่านี้ขึ้นอยู่กับคุณลักษณะหรือคุณสมบัติของตัวแปรเหล่านี้ที่มารวมกันเป็นปัจจัยเดียวกัน

2. การยืนยันทดสอบความถูกต้องของมาตรวัด (confirmatory factor analysis)

ในการวิจัย โดยมากจะมีผู้ที่พยายามสร้างหรือได้สร้างมาตรวัดขึ้นมาใหม่ซึ่งประกอบด้วยตัวแปรหลายตัว ทั้งนี้นอกจากจะรุบุว่ามาตรวัดใหม่นั้นประกอบด้วยตัวแปรของอะไรแล้ว ยังอาจจะระบุน้ำหนักมากน้อยของแต่ละตัวแปรที่นำมารวมกันอีกด้วย ปัญหาคือ มาตรวัดที่สร้างขึ้นมานั้นควรรวมตัวแปรต่างๆ เหล่านี้หรือไม่ และการให้น้ำหนักของตัวแปรเหล่านี้มีความถูกต้องหรือไม่ เทคนิคการวิเคราะห์ปัจจัยจะเป็นเทคนิคหนึ่งที่จะทดสอบความถูกต้องของการสร้างมาตรวัดดังกล่าว

2.9.2 ประโยชน์ของเทคนิคการวิเคราะห์ปัจจับ

1. ลดจำนวนตัวแปรโดยการรวมตัวแปรหลายๆ ตัวให้อยู่ในปัจจัยเดียวกัน ปัจจัยที่ได้ถือเป็นตัวแปรใหม่ ที่สามารถหาค่าข้อมูลของปัจจัยที่สร้างขึ้นได้ เรียกว่า factor score จึงสามารถนำปัจจัยดังกล่าวไปเป็นตัวแปรสำหรับการวิเคราะห์ทางสถิติต่อไป เช่น

- การวิเคราะห์ความถดถอย และสหสัมพันธ์ (regression and correlation analysis)
- การวิเคราะห์ความแปรปรวน (ANOVA)
- การทดสอบสมมติฐาน t-test , Z-test
- การวิเคราะห์จำแนกกลุ่ม (discriminant analysis)

เป็นต้น

2. ใช้ในการแก้ปัญหาการที่ตัวแปรอิสระของเทคนิคการวิเคราะห์ความถดถอยมีความสัมพันธ์กัน (multicollinearity)

วิธีการหนึ่งในการแก้ปัญหา multicollinearity คือ การรวมตัวแปรอิสระที่มีความสัมพันธ์เข้าไว้ด้วยกัน โดยการสร้างเป็นตัวแปรใหม่ หรือเรียกว่าปัจจัย โดยใช้เทคนิคการวิเคราะห์ปัจจัยแล้วนำปัจจัยดังกล่าวไปเป็นตัวแปรอิสระในการวิเคราะห์ความถดถอยต่อไป

3. ทำให้เห็นโครงสร้างความสัมพันธ์ของตัวแปรที่ศึกษา เนื่องจากเทคนิคการวิเคราะห์ปัจจัยจะหาค่าสัมประสิทธิ์สหสัมพันธ์ (correlation) ของตัวแปรทีละคู่แล้วรวมตัวแปรที่สัมพันธ์กันมากไว้ในปัจจัยเดียวกัน จึงสามารถวิเคราะห์ถึงโครงสร้างที่แสดงความสัมพันธ์ของตัวแปรต่างๆ ที่อยู่ในปัจจัยเดียวกันได้

4. ทำให้สามารถอธิบายความหมายของแต่ละปัจจัยได้ ตามความหมายของตัวแปรต่างๆ ที่อยู่ในปัจจัยนั้น ทำให้สามารถนำไปใช้ในด้านการวางแผนได้

2.9.3 หลักเกณฑ์ของเทคนิคการวิเคราะห์ปัจจัย

เทคนิคการวิเคราะห์ปัจจัย คือ เทคนิคที่ใช้ในการเปลี่ยนตัวแปรเดิม เป็นตัวแปรหรือปัจจัยใหม่ที่ไม่มีความสัมพันธ์กัน โดยที่ปัจจัยที่ได้เป็น linear combination ของตัวแปรเดิม โดยจะพยายามนำรายละเอียดจากตัวแปรเดิมต่างๆ มาไว้ในปัจจัยให้มากที่สุด เช่น ปัจจัยที่ 1 ประกอบไปด้วยตัวแปร 3 ตัว คือ X_3 , X_8 , X_{13} หมายความว่าตัวแปรทั้งสาม คือ ตัวแปรที่มีความสัมพันธ์กันมาก และสามารถนำรายละเอียดจากตัวแปร X_3 , X_8 และ X_{13} มาไว้ใน factor 1 ได้มากที่สุด โดยเขียนสมการเชิงเส้นแสดงความสัมพันธ์ได้ดังนี้

$$F_1 = W_3X_3 + W_8X_8 + W_{13}X_{13} + e$$

สำหรับการวัดปริมาณรายละเอียดของแต่ละแฟคเตอร์ จะวัดจากค่าแปรปรวน (variance) ของปัจจัยนั้นๆ

ในกรณีที่ตัวแปร X's ต่างๆ มักจะมีหน่วยต่างกัน เช่น อายุ มีหน่วยเป็นปี , รายได้มีหน่วยเป็นบาท , ประสบการณ์มีหน่วยเป็นปี เป็นต้น ซึ่งจะทำให้ตัวแปรที่มีค่ามาก เช่น รายได้มีความสำคัญมากที่สุด (มีค่าสัมประสิทธิ์ หรือค่าน้ำหนักมากที่สุด) ดังนั้นก่อนที่จะทำการวิเคราะห์โดยเทคนิควิเคราะห์ปัจจัยควรทำการปรับตัวแปรทุกตัวให้มีค่าเฉลี่ยเป็นศูนย์ และมีความแปรปรวนเป็น 1 หรือทำการ standardized ข้อมูลก่อนนั่นเอง

สำหรับสมการที่ใช้ในการประมาณค่าปัจจัยที่ j คือ

$$F_j = W_{j1}X_1 + W_{j2}X_2 + ...+ W_{jp}X_{ps} + e$$

โดยที่ X_j = ตัวแปรที่ j

W_j = สัมประสิทธิ์ของตัวแปรที่ j

นอกจากนั้นยังสามารถแสดงความสัมพันธ์ของค่าตัวแปร X ซึ่งเป็น linear combination ของแฟคเตอร์ต่างๆ ดังนี้

$$\begin{aligned} Z_1 &= L_{11}F_1 + L_{12}F_2 + ... + L_{1m}F_m + e_1 \\ Z_2 &= L_{21}F_1 + L_{22}F_2 + ... + L_{2m}F_m + e_2 \\ &\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{..... (2.3)} \\ Z_p &= L_{p1}F_1 + L_{p2}F_2 + ... + L_{pm}F_m + e_p \end{aligned}$$

โดยที่ Z_i = ตัวแปร X_i ที่ทำการ standardized แล้ว ; j = 1 , 2 , ..., p

p = จำนวนตัวแปร

m = จำนวนแฟคเตอร์ ; $m < p$

$F_1, \ldots, F_m$ = common factors

e = unique factor = error

L_{ij} = ค่าสัมประสิทธิ์ หรือเรียกว่า factor loading

factor loading หรือ ค่าสัมประสิทธิ์ เป็นค่าสัมประสิทธิ์สหสัมพันธ์ (r) เนื่องจากตัวแปร X's ได้ถูก standardized เป็น Z_i แล้ว ซึ่งทำให้ค่า factor loading มีค่าระหว่าง -1 ถึง 1 ค่า factor loading จะใช้ในการจัดกลุ่มตัวแปรว่าควรอยู่ในปัจจัยใด

2.9.4 ขั้นตอนการวิเคราะห์เทคนิคการวิเคราะห์ปัจจัย

การวิเคราะห์ข้อมูลด้วยเทคนิคการวิเคราะห์ปัจจัย เพื่อการจัดกลุ่มหรือจำแนกกลุ่มตัวแปรนั้น สามารถแบ่งออกเป็น 4 ขั้นตอน ดังนี้

ขั้นที่ 1 : ทำการสร้างเมตริกซ์สัมประสิทธิ์สหสัมพันธ์ (correlation matrix) ของตัวแปรทุกคู่ โดยในขั้นตอนแรกนี้จะมีการพิจารณาความสัมพันธ์ของตัวแปรทุกคู่โดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์

- ถ้าค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรคู่ใดที่มีค่าใกล้ +1 หรือ -1 แสดงว่าตัวแปรคู่นั้นมีความสัมพันธ์กันมาก และควรจะจัดอยู่ในแฟคเตอร์เดียวกัน
- ถ้าค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรคู่ใดมีค่าใกล้ศูนย์ แสดงว่าตัวแปรคู่นั้นไม่มีความสัมพันธ์กันหรือสัมพันธ์กันน้อยมาก ควรจัดให้อยู่คนละแฟคเตอร์
- ถ้ามีตัวแปรที่ไม่มีความสัมพันธ์กับตัวแปรอื่นๆ หรือมีความสัมพันธ์กับตัวแปรอื่นๆ ที่เหลือน้อยมาก ควรตัดตัวแปรนั้นออกจากการวิเคราะห์

ขั้นที่ 2 : การสกัดปัจจัย (factor extraction)

วัตถุประสงค์ของการสกัดปัจจัย คือ การหาจำนวนแฟคเตอร์ที่สามารถใช้แทนตัวแปรทั้งหมดทุกตัวได้ วิธีการสกัดปัจจัยซึ่งเป็นวิธีที่นิยมกันมากที่สุด คือ วิธี Principal Component Analysis : PCA

วิธี PCA เป็นเทคนิคที่มีวัตถุประสงค์ที่จะนำรายละเอียดของตัวแปรที่มีจำนวนตัวแปรมากๆ มาไว้ในปัจจัยที่มีเพียงไม่กี่ปัจจัย โดยจะพิจารณาจากรายละเอียดทั้งหมดจากแต่ละตัวแปร

ในการวิเคราะห์ PCA จะสร้าง linear combination ของตัวแปร โดยที่

- ปัจจัยที่ 1 จะเป็น linear combination แรก และมีรายละเอียดจากตัวแปรที่สุด หรือกล่าวได้ว่ามีความแปรปรวนสูงสุด
- ปัจจัยที่ 2 ก็เป็น linear combination ของตัวแปรเช่นกัน และสามารถนำรายละเอียดที่เหลืออยู่มากที่สุดจากตัวแปร โดยที่ ปัจจัยที่ 2 จะต้องตั้งฉาก (orthogonal) กับ ปัจจัย

ที่ 1 หรือกล่าวว่า ปัจจัยที่ 2 ไม่มีความสัมพันธ์กับปัจจัยที่ 1 ซึ่งเป็นการแก้ปัญหาการเกิด multicollinearity

- ปัจจัยที่ 3 ก็เป็น linear combination ของตัวแปรเช่นกัน ไม่มีความสัมพันธ์กับปัจจัยที่ 1 และ 2 และสามารถนำ information ที่เหลือจากตัวแปรได้มากที่สุด
- ในทำนองเดียวกัน การสร้างปัจจัยที่ 4 , 5 ,.....ก็ใช้หลักเกณฑ์ดังที่กล่าวข้างต้น

ในขั้นที่ 2 นี้จะทำให้สามารถประมาณค่า factor loading ได้ แล้วใช้ factor loading เป็นตัวพิจารณาว่ามีตัวแปรใดบ้างที่ควรจะจัดอยู่ในปัจจัยเดียวกัน ในบางกรณีค่า factor loading มีค่ากลางๆ เช่น ถ้ามี 2 ปัจจัย แล้วพบว่า factor loading ของตัวแปรหนึ่ง ในปัจจัยที่ 1 เป็น 0.42 และในปัจจัยที่ 2 เป็น 0.51 ทำให้ไม่แน่ใจว่าควรจัดตัวแปรนี้อยู่ในปัจจัยใดดี เมื่อเป็นเช่นนี้ก็ควรจะทำการหมุนแกน ซึ่งจะอยู่ในขั้นตอนที่ 3

ขั้นที่ 3 : การหมุนแกนปัจจัย (factor rotation)

ดังที่กล่าวไว้ในขั้นที่ 2 ว่ากรณีที่ค่า factor loading มีค่ากลางๆ ทำให้ไม่สามารถจัดตัวแปรว่าควรจะอยู่ในปัจจัยใดได้นั้น จะต้องทำการหมุนแกน ดังนั้นวัตถุประสงค์ของการหมุนแกนปัจจัยก็คือ เพื่อให้ค่า factor loading ของตัวแปรมีค่ามากขึ้นหรือลดลงจนกระทั่งทำให้ทราบว่าตัวแปรนั้นควรจะอยู่ในปัจจัยใดหรือไม่ควรจะอยู่ในปัจจัยใด

วิธีหมุนแกนปัจจัย มี 2 วิธี

- **Orthogonal Rotation**

เป็นการหมุนแกนโดยยังคงทำให้ปัจจัยตั้งฉากกัน หรือเป็นอิสระกันแต่ทำให้ค่า factor loading เพิ่มขึ้นหรือลดลง

- **Oblique Rotation**

เป็นการหมุนแกนปัจจัยไปในลักษณะที่ปัจจัยไม่ตั้งฉากกันหรือปัจจัยไม่เป็นอิสระกันนั่นเอง

ขั้นที่ 4 : การคำนวณค่า factor score

เมื่อสามารถจัดตัวแปรที่มีอยู่จำนวนมากให้เป็นกลุ่มตัวแปรไม่กี่กลุ่มได้แล้ว ก็จะสามารถคำนวณหาค่า factor score ของแต่ละกรณีได้ เช่น หากมี 2 ปัจจัยก็สามารถคำนวณหาค่า factor score ของทั้ง 2 ปัจจัยได้ และถือว่าทั้ง 2 ปัจจัยเป็นตัวแปรใหม่ที่จะนำไปวิเคราะห์ต่อไป

2.9.5 เงื่อนไขของเทคนิคการวิเคราะห์ปัจจัย

1. ปัจจัย (F) และ error (e) ในสมการที่ 2.3 จะต้องอิสระกัน

2. ตัวแปรควรเป็นตัวแปรเชิงปริมาณ และกรณีมีที่ตัวแปรเชิงกลุ่มผสมอยู่ด้วย จะต้องเปลี่ยนตัวแปรเชิงกลุ่มให้อยู่ในรูปตัวแปรเทียม (dummy variable)

3. ความสัมพันธ์ระหว่างปัจจัยและตัวแปรอยู่ในรูปเชิงเส้น (Linear) เท่านั้น

4. สำหรับเทคนิค Principal Component Analysis ตัวแปรแต่ละตัวหรือข้อมูลไม่จำเป็นต้องมีการแจงแจงแบบปกติ แต่หากตัวแปรบางตัวมีการแจกแจงเบ้ค่อนข้างมาก และมีค่าผิดปกติ (Outlier) ผลลัพธ์ที่ได้อาจจะไม่ถูกต้อง

5. จำนวนข้อมูล ควรจะมีมากกว่าจำนวนตัวแปร (ควรจะมากกว่าจำนวนตัวแปรอย่างน้อย 10 เท่า)

2.10 การจำแนกกลุ่มข้อมูลด้วยเทคนิควิเคราะห์การจัดกลุ่ม

จากที่ได้กล่าวในตอนแรกของบทว่า ในการแก้ปัญหาเกี่ยวกับการออกแบบระบบจัดขนาดของงานวิจัยนี้จำเป็นที่จะต้องทำการสร้างหรือกำหนดปัญหาดังกล่าวให้อยู่ในรูปแบบของปัญหาการหาความเหมาะสมที่สุดให้ได้ก่อน เป็นที่ทราบกันดีว่าองค์ประกอบสำคัญของปัญหาการหาความเหมาะสมที่สุดที่จะขาดไม่ได้ ก็คือ ฟังก์ชั่นวัตถุประสงค์ (objective function) และ เงื่อนไข (constraints) นั่นเอง ดังนั้นก่อนที่จะทำการกำหนดฟังก์ชั่นวัตถุประสงค์และเงื่อนไขได้นั้น จึงจำเป็นที่จะต้องทำการศึกษาเกี่ยวกับแนวคิด , ทฤษฎีที่เกี่ยวข้องกับการจำแนกกลุ่มข้อมูล เพื่อใช้เป็นแนวทางในการกำหนดรูปแบบปัญหาของระบบจัดขนาดให้เป็นปัญหาการหาความเหมาะสมที่สุด

" การวิเคราะห์การจัดกลุ่ม" เป็นเทคนิคการจำแนกหรือแบ่งกลุ่มของข้อมูลตัวอย่างออกเป็นกลุ่มย่อยๆ ตั้งแต่ 2 กลุ่มขึ้นไป

ข้อมูลที่จัดอยู่ในกลุ่มเดียวกันก็จะต้องมีลักษณะที่คล้ายคลึงกันภายในกลุ่ม ส่วนข้อมูลที่อยู่ต่างกลุ่มกันก็จะมีลักษณะที่แตกต่างกัน ดังนั้นการพิจารณาเลือกลักษณะหรือตัวแปรที่จะนำมาใช้ในการแบ่งกลุ่มข้อมูลจึงมีความสำคัญ นอกจากนั้น ข้อมูลใดข้อมูลหนึ่งจะต้องตกอยู่ในกลุ่มหนึ่งเพียงกลุ่มเดียว **(จากหลักการนี้ ผู้วิจัยพบว่าสามารถนำมาใช้ในการแก้ปัญหาเกี่ยวกับการออกแบบระบบจัดขนาดได้ โดยใช้เป็นเกณฑ์ในการตัดสินว่าคนใด ควรจะใส่เสื้อผ้าขนาดใดหรือควรจะถูกจัดเข้าไปอยู่ในกลุ่มขนาดใด ดังนั้นผู้วิจัยจึงทำการศึกษาเทคนิควิเคราะห์การจัดกลุ่มควบคู่กับเทคนิคการวิเคราะห์ปัจจัยด้วย)**

2.10.1 เทคนิคย่อยของการวิเคราะห์การจัดกลุ่ม

เทคนิคการวิเคราะห์การจัดกลุ่ม แบ่งออกเป็น 2 เทคนิคย่อย คือ

- **Hierarchical cluster analysis** เป็นเทคนิคที่นิยมใช้กันมากในการแบ่งกลุ่มข้อมูลหรือกลุ่มตัวแปร โดยมีคุณสมบัติดังนี้

  1. ในกรณีที่ใช้เทคนิคดังกล่าวในการแบ่งกลุ่มข้อมูลนั้น จำนวนข้อมูลจะต้องไม่มากนัก (จำนวนข้อมูลควรต่ำกว่า 200 หากมีข้อมูลตั้งแต่ 200 ขึ้นไปควรใช้วิธี K-means cluster) และจำนวนของตัวแปรต้องไม่มากเช่นกัน
  2. ไม่จำเป็นต้องทราบว่าตัวแปรใดอยู่ในกลุ่มใดก่อน
  3. ใช้ได้กับข้อมูลที่อยู่ในรูปความถี่ หรือ Binary

- **K-means cluster analysis** เป็นเทคนิคการจำแนกกลุ่มข้อมูลออกเป็นกลุ่มย่อย โดยมีคุณสมบัติดังนี้

  1. ใช้ในกรณีที่จำนวนข้อมูลมีมากและจำเป็นต้องกำหนดจำนวนกลุ่มหรือจำนวน cluster ที่ต้องการขึ้นมาก่อน
  2. เป็นเทคนิคที่จะมีการทำงานหลายรอบ (iteration) โดยในแต่ละรอบจะมีการรวมตัวแปรให้ไปอยู่ในกลุ่มใดกลุ่มหนึ่ง โดยเลือกกลุ่มที่ตัวแปรนั้นมีระยะห่างจากค่ากลางของกลุ่มน้อยที่สุด แล้วคำนวณค่ากลางของกลุ่มใหม่ จะทำเช่นนี้จนกระทั่งค่ากลางของกลุ่มไม่เปลี่ยนแปลงหรือครบจำนวนรอบที่กำหนดไว้
  3. ชนิดของตัวแปรที่ใช้ในเทคนิคนี้ จะต้องเป็นตัวแปรเชิงปริมาณ คือ เป็นสเกลอันตรภาค (interval scale) หรือ สเกลอัตราส่วน (ratio scale)

2.10.2 ความแตกต่างระหว่าง Hierarchical กับวิธี K-mean

- เทคนิค K-means ใช้เมื่อมีจำนวนข้อมูลมากๆ โดยทั่วไปนิยมใช้เมื่อมีจำนวนข้อมูลมากกว่า 200 ข้อมูล เนื่องจากเมื่อจำนวนข้อมูลมีมาก เทคนิค K-means จะง่ายกว่าและใช้ระยะเวลาในการคำนวณน้อยกว่า
- เทคนิค K-means ผู้ใช้จะต้องกำหนดจำนวนกลุ่มที่แน่นอนไว้ล่วงหน้าก่อน
- เทคนิค Hierarchical ผู้วิเคราะห์จะทำการ standardized ข้อมูลก่อนหรือไม่ก็ได้ แต่วิธี K-means จะต้องทำการ standardized ข้อมูลก่อน
- วิธี K-means จะหาระยะห่างโดยวิธี euclidean distance โดยอัตโนมัติ ในขณะที่ Hierarchical จะใช้วิธีการคำนวณระยะห่างหรือความคล้ายวิธีใดก็ได้

จากคุณสมบัติของเทคนิคทั้งสอง เมื่อพิจารณาแล้วจะพบว่า K-means cluster analysis จะมีความสอดคล้องกับงานวิจัยนี้มากที่สุด เนื่องจากจำนวนข้อมูลทางด้านสรีระจัดเป็นข้อมูลเชิงปริมาณและมีมากราว 3,000 ข้อมูล อีกทั้งในการแก้ปัญหาการหาความเหมาะสมที่สุดนั้น

จำเป็นต้องทำการกำหนดจำนวนกลุ่มของขนาดที่ต้องการขึ้นมาก่อนที่จะทำการแก้ปัญหาด้วย ดังนั้นผู้วิจัยคาดว่าจะสามารถนำเทคนิคดังกล่าวไปใช้ในการกำหนดรูปแบบของปัญหาการหาความเหมาะสมที่สุดได้ ในที่นี้จึงทำการศึกษาหลักการและทฤษฎีของ K-means cluster analysis เพียงอย่างเดียว

2.10.3 หลักการทำงานของการวิเคราะห์การจัดกลุ่ม

เนื่องจากวัตถุประสงค์หลักของการจัดกลุ่ม คือ เพื่อใช้ในการกำหนดโครงสร้างของกลุ่มข้อมูลโดยการกลุ่มให้กับข้อมูลที่มีลักษณะคล้ายกันมากที่สุดเข้าไว้ด้วยกัน แต่การที่จะทำให้งานนี้ประสบความสำเร็จได้นั้น จะต้องสามารถตอบคำถามทั้ง 3 ข้อนี้ให้ได้ก่อน นั่นคือ

- การวัดความคล้ายหรือความแตกต่างของข้อมูลนั้น (How do we measure similarity?) จะวัดได้อย่างไรและใช้อะไรเป็นตัววัด
- ใช้อะไรเป็นเกณฑ์ในการฟอร์มกลุ่มของข้อมูล (How do we form clusters?)

ไม่ว่าความเหมือนของข้อมูลจะถูกวัดด้วยวิธีใดก็ตาม แต่วิธีการที่ใช้ในการกำหนดกลุ่มของข้อมูลนั้นจะต้องสามารถทำให้ข้อมูลที่มีลักษณะคล้ายคลึงกันมากที่สุดถูกรวมเข้ามาอยู่ภายในกลุ่มเดียวกันได้

- จำนวนกลุ่มที่จะสร้างขึ้นควรมีจำนวนเท่าใดจึงจะเหมาะสม (How many groups do we form?)

เพราะหากจำนวนกลุ่มที่ต้องการสร้างขึ้นมีจำนวนน้อยๆ ระดับของความคล้ายกัน (level of similarity) ภายในกลุ่มก็จะแย่ลดลงด้วย เมื่อเป็นเช่นนี้ความสมดุลระหว่างทั้งสองสิ่งจึงเป็นเรื่องที่จะต้องถูกพิจารณาเป็นอย่างดี เพื่อให้ได้จำนวนกลุ่มที่น้อยที่สุดแต่ยังคงอยู่ภายใต้ระดับของความคล้ายที่ยอมรับได้

2.10.4 ขั้นตอนของการวิเคราะห์การจัดกลุ่ม

ในการวิเคราะห์การจัดกลุ่มนั้น สามารถแบ่งออกได้เป็น 6 ขั้นตอน ดังนี้

ขั้นที่ 1 : ผู้วิจัยจะต้องทราบถึงวัตถุประสงค์ของการวิเคราะห์การจัดกลุ่มก่อน ว่ามีวัตถุประสงค์เพื่ออะไร

วัตถุประสงค์หลักของ การวิเคราะห์การจัดกลุ่มจะมีอยู่ 3 ข้อใหญ่ๆ ได้แก่

- เพื่อทำการสำรวจและทำการกำหนดรูปแบบของการแบ่งประเภทของกลุ่มข้อมูล (Exploratory purposes and the formation of a taxonomy)
- เพื่อจัดข้อมูลให้อยู่ในสภาพที่ง่ายต่อการวิเคราะห์ (data simplification) คือ ทำการลดจำนวนข้อมูลที่มีลักษณะคล้ายกันเข้าไว้ด้วยกัน เช่น เดิมมีข้อมูลอยู่ 100 ข้อมูล 20 ตัวแปร รวมเป็นข้อมูลทั้งหมด 2000 ค่า หากจัดกลุ่มตัวแปร 20 ตัวให้เป็นกลุ่มเพียง

3 กลุ่ม ก็จะทำให้เหลือข้อมูลแค่ 300 ค่า ดังนั้นแทนที่ผู้วิจัยจะต้องทำการวิเคราะห์ข้อมูลทั้งหมดที่มีอยู่ ก็สามารถเลือกมาเฉพาะบางส่วนได้

- เพื่อทำการระบุความสัมพันธ์ของข้อมูล (relationship identification) ในการกำหนดโครงสร้างอย่างง่ายๆให้กับกลุ่มข้อมูล ผู้วิจัยจะเห็นภาพความสัมพันธ์ หรือความคล้ายและความแตกต่างของข้อมูลได้ด้วยเช่นกัน

การเลือกตัวแปรที่จะนำมาจัดกลุ่ม (selection of clustering variables)

การเลือกตัวแปรที่จะนำมาจัดกลุ่มนั้น จะต้องทำภายใต้แนวความคิดและทฤษฎีที่ถูกต้อง เนื่องจากตัวแปรที่ถูกเลือกนั้นจะมีผลต่อการจัดกลุ่มของข้อมูลด้วย โดยตัวแปรจะเป็นตัวกำหนดคุณลักษณะของข้อมูลที่จะต้องถูกจัดกลุ่ม อีกทั้งยังมีความสัมพันธ์ต่อวัตถุประสงค์ (objectives) ของการวิเคราะห์การจัดกลุ่มด้วย ดังนั้นเทคนิคดังกล่าวจะไม่มีความหมายใดเลยหากผู้วิจัยทำการแบ่งกลุ่มข้อมูลภายใต้ตัวแปรที่ไม่มีความเกี่ยวข้องกันอยู่

ขั้นที่ 2 : การออกแบบการวิจัยจัดกลุ่ม

ภายหลังจากการกำหนดวัตถุประสงค์ของการวิจัยและทำการเลือกตัวแปรที่เกี่ยวข้องได้แล้ว ขั้นต่อไปคือการกำหนดรูปแบบของการจัดกลุ่มคร่าวๆ โดยแบ่งออกเป็นข้อย่อยๆ ดังนี้

- *การค้นหา outlier (detecting outliers)*

เนื่องจากเทคนิคการวิเคราะห์การจัดกลุ่มนี้ค่อนข้างจะมีความไวต่อการนำเอาข้อมูลที่ไม่มีความเกี่ยวข้องกันเข้ามาพิจารณาและข้อมูลที่มีความแตกต่างจากข้อมูลทั่วๆไป หรือที่เรียกกันว่า outlier มาก เนื่องจาก outlier เป็นตัวแทนของตัวแปรที่มีความผิดปกติที่ไม่ได้สะท้อนให้เห็นถึงลักษณะที่แท้จริงของประชากรอื่นๆ ทำให้โครงสร้างของข้อมูลที่แท้จริงถูกทำลายไปหรือทำให้ได้โครงสร้างของกลุ่มประชากรที่ผิดเพี้ยนไป เมื่อเป็นเช่นนี้จึงอาจทำการกรองข้อมูลเบื้องต้นเพื่อหา outlier ก่อน โดยวิธีการที่ง่ายที่สุดก็คือ การพลอตกราฟ นั่นเอง ในกรณีที่จำนวนข้อมูลหรือกลุ่มตัวอย่างมีมาก การพลอตกราฟอาจไม่สะดวกนักและอาจมองเห็นภาพของ outlier ได้ไม่ชัดเจน เมื่อเป็นเช่นนี้ ผู้วิจัยก็สามารถทำการกำหนดได้ว่าข้อมูลตัวใดที่มีลักษณะต่างจากกลุ่มตัวอย่างอื่นๆ มากนัก ให้ถือว่าเป็น outlier

เมื่อข้อมูลใดถูกระบุว่าเป็น outlier แล้วจะต้องพิจารณาต่อไปว่าข้อมูลนั้นมีผลกระทบใดหรือไม่ หากตัวแปรดังกล่าวไม่ได้ทำให้การวิเคราะห์ผิดเพี้ยนไปจากความเป็นจริงก็ไม่จำเป็นต้องลบข้อมูลตัวนั้นทิ้งไป

ผู้วิจัยควรจะตระหนักด้วยว่าในการลบข้อมูลบางตัวทิ้งไปเพียงเพราะข้อมูลตัวนั้นเป็น outlier ที่ตรวจพบก็อาจส่งผลต่อโครงสร้างของข้อมูลที่แท้จริงได้เช่นกัน ดังนั้น outlier จึง เป็นเรื่องที่ต้องระมัดระวังเป็นพิเศษ

- *การวัดความคล้ายของตัวแปร (similarity measures)*

แนวคิดของความคล้ายกัน ถือว่าเป็นพื้นฐานสำคัญของการวิเคราะห์การจัดกลุ่ม

interobject similarity คือ การวัดความเหมือน , ความคล้ายกัน ระหว่างข้อมูลที่จะถูกนำมาจัดกลุ่ม โดยข้อมูลใดๆ จะถูกเปรียบเทียบกับข้อมูลตัวอื่นๆที่เหลือ ด้วยดัชนีที่ใช้วัดความคล้าย (similarity measure) จากนั้นข้อมูลที่คล้ายกันมากๆ ก็จะถูกรวมอยู่ในกลุ่มเดียวกัน

interobject similarity สามารถวัดได้ 3 แบบ คือ

1. correlational measure เป็นตัววัดความคล้ายคลึงกันของรูปแบบไม่เกี่ยวข้องกับขนาด ดังนั้นจึงไม่ค่อยนำมาใช้กับการวิเคราะห์จัดกลุ่ม ข้อมูลที่นำมาพิจารณาจะต้องเป็นข้อมูลเชิงปริมาณ (quantitative data) เท่านั้น

หากค่า correlational measure มีค่ามาก แสดงว่าข้อมูลนั้นมีรูปแบบที่คล้ายคลึงกันมาก

2. distance measure เป็นตัววัดความคล้ายที่พิจารณาจากขนาด (magnitude) ในลักษณะที่ว่าหากข้อมูลใดที่อยู่ใกล้กันก็ย่อมมีความคล้ายคลึงกัน ยิ่งค่าระยะห่างมาก ความเหมือนหรือความคล้ายคลึงกันก็ยิ่งน้อย ตัววัดนี้ใช้ได้กับข้อมูลเชิงปริมาณ เช่นเดียวกับ correlational measure

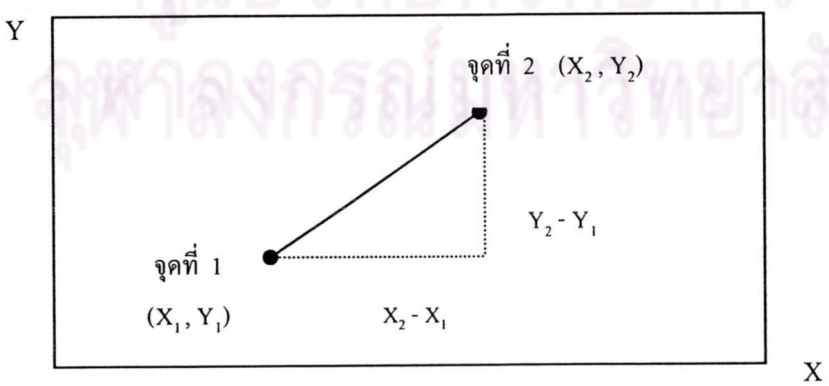
3. association measure เป็นตัววัดความคล้ายคลึงกันที่ใช้ได้กับข้อมูลเชิงคุณภาพหรือ qualitative data เท่านั้น

เมื่อเป็นเช่นนี้ ในที่นี้จึงเลือกศึกษาเฉพาะ distance measure เท่านั้น เนื่องจากสอดคล้องกับงานวิจัยฉบับนี้มากที่สุด

ประเภทของ distance measures (Type of distance measures)

สำหรับประเภทของ distance measure นั้นมีอยู่มากมายหลายประเภท แต่ที่นิยมมากที่สุด คือ Euclidean distance

การคำนวณหา Euclidean distance สามารถหาได้ ดังแสดงในภาพที่ 2.2 ซึ่งเป็นการหา Euclidean distance ของจุดสองจุดที่มีพิกัดเป็น (X_1 , Y_1) และ (X_2 , Y_2) ตามลำดับ



$$\text{Euclidean distance} = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

ภาพที่ 2.2 : แสดงตัวอย่างการหาค่า Euclidean measure ของข้อมูล 2 ข้อมูล

Euclidean measure สามารถแบ่งได้เป็น 2 ประเภท คือ

1. Simple Euclidean distance ซึ่งมีวิธีการคำนวณดังสมการ

$$\text{distance } (X, Y) = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

2. Squared Euclidean distance (หรือ absolute Euclidean distance) เป็นการหาค่า sum of squared ของผลต่างระหว่างจุด 2 จุด ดังนั้นสูตรที่ใช้ในการคำนวณจึงคล้ายกับการหา simple Euclidean distance เพียงแต่ไม่มีการถอดรากที่ 2 เท่านั้นเอง จึงมีข้อดีตรงที่ช่วยลดระยะเวลาในการประมวลผล และเป็น distance measure ที่นิยมนำไปใช้กับการจัดกลุ่มแบบ centroid และ Ward's methods

$$\text{distance } (X, Y) = (X_2 - X_1)^2 + (Y_2 - Y_1)^2$$

- Standardizing the data

หลังจากผู้วิจัยสามารถเลือกดัชนีที่ใช้วัดความคล้ายได้แล้ว จำเป็นอย่างยิ่งที่จะต้องพิจารณาต่อไปว่า ข้อมูลที่จะนำมาคำนวณหาความคล้ายนั้นจำเป็นที่จะต้องทำการ standardized ข้อมูลก่อนหรือไม่ โดยดูจาก distance measure ว่ามีความไวต่อสเกลหรือขนาดของตัวแปรต่างๆ มากน้อยเพียงใด โดยทั่วไปตัวแปรที่มีความเบี่ยงเบนมาตรฐานสูงมาก (larger standard deviation) จะมีผลกระทบต่อค่า similarity มากกว่าตัวแปรที่มีความเบี่ยงเบนมาตรฐานน้อยๆ เช่น สมมติว่าต้องการแบ่งกลุ่มของประชากรภายใต้ตัวแปร 3 ตัว คือ อายุ , รายได้ , และระดับการศึกษา โดยอายุมีหน่วยเป็นปี รายได้มีหน่วยเป็นบาท ส่วนระดับการศึกษาจะวัดออกมาเป็นระดับแบ่งเป็น 7 ระดับ เมื่อทำการพลอตกราฟ 3 มิติ ย่อมเป็นที่แน่นอนว่าระยะห่างระหว่างจุดหรือข้อมูลของแต่ละคนนั้น จะต้องขึ้นอยู่กับความแตกต่างของรายได้เป็นหลัก เนื่องจากความแตกต่างของระดับการศึกษาที่เป็นไปได้มีช่วงตั้งแต่ 1 ถึง 7 เท่านั้น ในขณะที่รายได้มีช่วงของความแตกต่างมากกว่าระดับการศึกษาเป็นพันๆ เท่า เมื่อเป็นเช่นนี้ ก็จำเป็นที่จะต้องมีการ standardized ข้อมูลก่อนที่จะทำการคำนวณหา distance measures แต่ถ้าหากว่าตัวแปรที่พิจารณาไม่มีความไวต่อความแตกต่างของหน่วยวัดหรือขนาดของข้อมูล ผู้วิจัยก็ไม่จำเป็นที่จะต้องทำการ standardized ข้อมูลก็ได้

การ standardized ข้อมูลวิธีหนึ่ง คือ การเปลี่ยนค่าของข้อมูลให้อยู่ในรูปของ Z scores หรือค่ามาตรฐาน โดยการนำค่าจริงลบด้วยค่าเฉลี่ยของข้อมูลทั้งหมดแล้วหารด้วยค่าเบี่ยงเบนมาตรฐาน วิธีการนี้จะเป็นการแปลงข้อมูลดิบให้อยู่ในรูปของค่ามาตรฐานโดยมีค่าเฉลียเป็น 0 และมีส่วนเบี่ยงเบนมาตรฐานเป็น 1

สรุปข้อดีของการ standardized ข้อมูล

1. ช่วยให้การเปรียบเทียบข้อมูลนั้นทำบนสเกลเดียวกัน

2. ช่วยลดผลกระทบที่เกิดจากความแตกต่างของหน่วยของกลุ่มตัวอย่าง ระหว่างตัวแปรและภายในตัวแปรเดียวกัน

ขั้นที่ 3 : สมมติฐานที่เกี่ยวข้องกับการวิเคราะห์การจัดกลุ่ม (assumptions in cluster analysis)

เทคนิคการวิเคราะห์การจัดกลุ่มไม่ใช่เทคนิคการอนุมานทางสถิติที่ใช้พารามิเตอร์จากกลุ่มข้อมูลตัวอย่างมากำหนดเป็นตัวแทนของประชากรทั้งหมด แต่เป็นวิธีการในการกำหนดโครงสร้างของกลุ่มตัวอย่าง ดังนั้นจึงจำเป็นต้องมีการกำหนดสมมติฐานก่อนที่จะทำการวิเคราะห์ ดังนี้

1. การสุ่มตัวอย่างขึ้นมาเพื่อหวังว่ากลุ่ม (clusters) ที่จะได้จากการแบ่งข้อมูลเหล่านี้จะสามารถใช้เป็นตัวแทนของโครงสร้างของประชากรได้นั้น จะต้องมั่นใจได้ว่าตัวอย่างที่สุ่มขึ้นมานั้น เป็นตัวแทนของประชากรที่มีอยู่จริง
2. ตัวแปรที่จะนำมาวิเคราะห์กลุ่มจะต้องไม่มี multicollinearity กับตัวแปรอื่น

ขั้นที่ 4 : การจัดกลุ่มให้กับข้อมูล (deriving clusters)

จากตัวแปรที่ถูกเลือกขึ้นมาและค่าของระยะห่างหรือดัชนีวัดความคล้ายที่คำนวณได้ ขั้นต่อไป คือการแบ่งกลุ่มให้กับข้อมูลต่างๆ ตามอัลกอริทึมที่เหมาะสม

ดังที่ได้กล่าวในตอนต้นว่าอัลกอริทึมในการจัดกลุ่มนั้นมี 2 ประเภท คือ hierarchical และ nonhierarchical ซึ่งก็จะมีหลักเกณฑ์ในการรวมกลุ่มมากมาย ดังต่อไปนี้

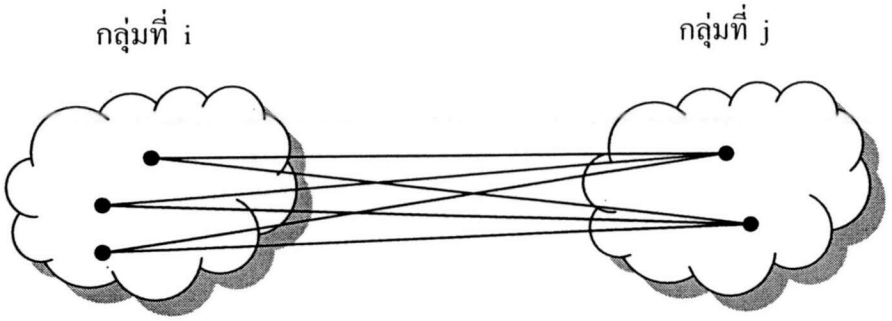
- หลักเกณฑ์ในการรวมกลุ่ม

หลักเกณฑ์ในการรวมกลุ่มในแต่ละขั้นตอนข้างต้นนั้นมีหลายวิธี ดังนี้

1. Between-groups linkage หรือที่เรียกว่าวิธี average linkage between groups หรือเรียกว่า UPGMA (Unweighted Pair – Group Method Using Arithmetic Average)

วิธีนี้ จะทำการคำนวณหาระยะห่างเฉลี่ยของทุกคู่ของข้อมูล โดยที่ข้อมูลหนึ่งอยู่ในกลุ่มที่ i ส่วนอีกข้อมูลหนึ่งจะอยู่ในกลุ่มที่ j , $i \neq j$

หากกลุ่มที่ i มีระยะห่างเฉลี่ยจากกลุ่มที่ j สั้นกว่าระยะห่างจาก กลุ่มอื่นๆ จะนำกลุ่ม i และ j รวมกันเป็นกลุ่มเดียวกัน



ภาพที่ 2.3 : แสดง average linkage

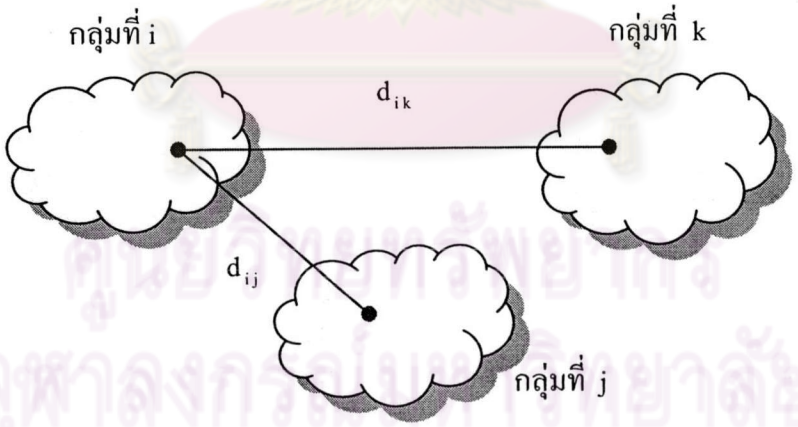
2. within-group linkage technique หรือ average linkage within groups method

วิธีนี้จะรวม กลุ่มเข้าด้วยกันถ้าระยะห่างเฉลี่ยระหว่างทุกข้อมูลในกลุ่มนั้นๆ มีค่าน้อยที่สุด

3. nearest neighbour หรือเรียกว่า single linkage

เมื่อให้ d_{ik} เป็นระยะห่างที่สั้นที่สุดของกลุ่ม i และ k ส่วน d_{ij} เป็นระยะห่างที่สั้นที่สุดระหว่างกลุ่ม i และ j

ดังรูป จะรวมกลุ่ม i และ j เข้าด้วยกัน เนื่องจาก $d_{ij} < d_{ik}$



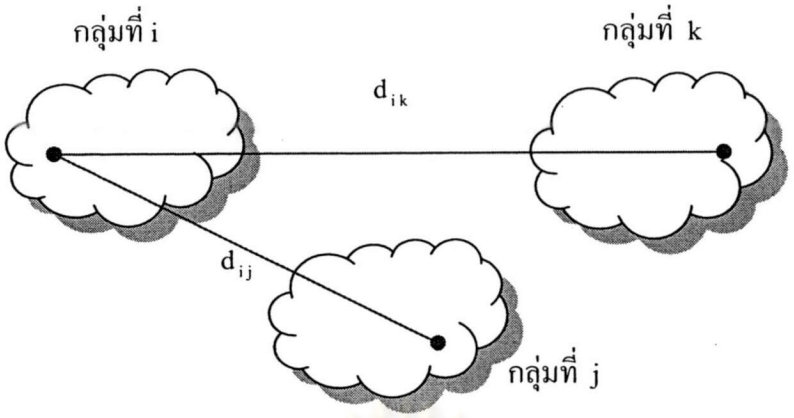
ภาพที่ 2.4 : แสดง single linkage

4. furthest neighbour technique หรือเรียกว่า complete linkage

d_{ik} = ระยะห่างที่ยาวที่สุดของกลุ่มที่ i และ k

d_{ij} = ระยะห่างที่ยาวที่สุดของกลุ่มที่ i และ j

ในที่นี้ $d_{ij} < d_{ik}$ จึงรวมกลุ่มที่ i และ j เข้าเป็นกลุ่มเดียวกัน

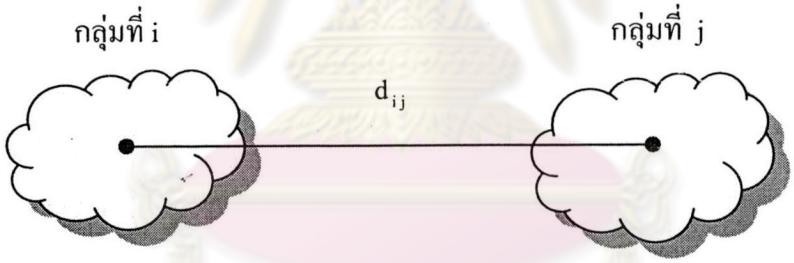


ภาพที่ 2.5 : แสดง complete linkage

5. centroid clustering

วิธีการนี้จะคำนวณหาระยะห่างระหว่าง centroid ของกลุ่มทีละคู่ ในที่นี้จะเรียกค่าเฉลี่ยหรือค่ากลางของแต่ละกลุ่มว่า centroid ของกลุ่ม เนื่องจากการแบ่งกลุ่ม case จะพิจารณาจากตัวแปรหลายๆ ตัวพร้อมๆ กัน จึงเรียกค่ากลางหรือค่าเฉลี่ยว่า centroid

ถ้าระยะห่างระหว่าง centroid ของกลุ่มคู่ใดต่ำจะรวมกลุ่มคู่นั้นเข้าเป็นกลุ่มเดียวกัน



ภาพที่ 2.6 : แสดง centroid clustering

6. median clustering

วิธีนี้จะรวมกลุ่ม 2 กลุ่มเข้าด้วยกันโดยให้แต่ละกลุ่มสำคัญเท่ากัน (ให้น้ำหนักเท่ากัน) ในขณะที่วิธีของ centroid clustering จะให้ความสำคัญแก่กลุ่มที่มีขนาดใหญ่มากกว่ากลุ่มที่มีขนาดเล็ก (ให้น้ำหนักไม่เท่ากัน)

median clustering จะใช้ค่า median เป็นค่ากลางของ centroid ถ้าระยะห่างระหว่างค่า median ของกลุ่มคู่ใดต่ำจะรวมกลุ่มคู่นั้นเข้าด้วยกัน

7. Ward's method

หลักการของวิธีนี้จะพิจารณาจากค่า sum of the squared within-cluster distance โดยจะรวมกลุ่มที่ทำให้ค่า sum of square within-cluster distance เพิ่มขึ้นน้อยที่สุด โดยค่า square within-cluster distance คือค่า square Euclidean distance ของแต่ละข้อมูลกับ cluster mean

- <u>หลักการจำแนกข้อมูลออกเป็นกลุ่มย่อย (K-means clustering)</u>

สำหรับเทคนิค K-means clustering ซึ่งมีการจำแนกข้อมูลออกเป็นกลุ่มย่อย โดยใช้เมื่อมีจำนวนข้อมูลมาก ผู้วิเคราะห์จะต้องกำหนดจำนวนกลุ่มหรือจำนวน cluster ที่ต้องการก่อน เช่น กำหนดให้มี k กลุ่ม

เทคนิค K-Means จะมีการทำงานหลายๆ รอบ (iteration) โดยในแต่ละรอบจะมีการรวมข้อมูลให้ไปอยู่ในกลุ่มใดกลุ่มหนึ่ง โดยเลือกกลุ่มที่ข้อมูลนั้นมีระยะห่างจากค่ากลางของกลุ่มน้อยที่สุด แล้วคำนวณค่ากลางของกลุ่มใหม่ ทำเช่นนี้จนกระทั่งค่ากลางของกลุ่มไม่มีการเปลี่ยนแปลงหรือครบจำนวนรอบที่กำหนดไว้

ชนิดตัวแปรที่ใช้ในเทคนิค K-Means clustering

ตัวแปรที่ใช้ในเทคนิคนี้จะต้องเป็นตัวแปรเชิงปริมาณ คือ เป็นสเกลอันตรภาค (interval scale) หรือ สเกลอัตราส่วน (ratio scale) โดยไม่สามารถใช้กับข้อมูลที่อยู่ในรูปความถี่หรือ binary เหมือนเทคนิค hierarchical

<u>ขั้นที่ 5</u> : การตีความหมายของกลุ่มที่ได้ (interpretation of the clusters)

ในขั้นตอนของการตีความนี้จะเป็นการตรวจสอบกลุ่มแต่ละกลุ่มในรูปของ cluster variate เพื่อทำการกำหนดชื่อหรือคุณลักษณะที่อธิบายธรรมชาติของกลุ่มได้อย่างถูกต้อง

เมื่อเริ่มทำการตีความหมายของกลุ่มที่ได้ ตัววัดที่ถูกนำมาใช้ในการพิจารณาบ่อยที่สุด คือ ค่า centroid ของกลุ่ม ในกรณีที่กลุ่มถูกดำเนินการมาจากข้อมูลจริง ผู้วิเคราะห์ก็สามารถตีความผลลัพธ์ที่ได้โดยตรง แต่ถ้าข้อมูลนั้นถูก standardized หรือเป็นข้อมูลที่ได้มาจากการวิเคราะห์ปัจจัยมาก่อน ผู้วิเคราะห์จะต้องกลับไปคิดเป็นคะแนนจริงสำหรับตัวแปรเดิมและคำนวณค่าเฉลี่ยจากข้อมูลเหล่านั้นก่อน

<u>ขั้นที่ 6</u> : การตรวจสอบความถูกต้องของกลุ่มที่ได้ (validation of the clusters)

หลังจากทำการแบ่งกลุ่มโดยการวิเคราะห์การจัดกลุ่มแล้ว ผู้วิเคราะห์ควรจะทำการตรวจสอบหรือพิสูจน์ให้เห็นว่า การแบ่งกลุ่มดังกล่าวสามารถนำไปใช้ในทางปฏิบัติได้จริง ถึงแม้ว่าในขณะนี้จะยังไม่มีวิธีการใดที่จะสามารถพิสูจน์ความถูกต้องของเทคนิคดังกล่าวได้จริง

แต่ก็มีกระบวนการหลายๆ กระบวนการที่ถูกเสนอเพื่อใช้เป็นหลักการในการประเมินความถูกต้องของ cluster ที่ได้

การตรวจสอบความถูกต้อง (validation) ในที่นี้ได้รวมถึงความพยายามในการรับประกันว่ากลุ่มที่แบ่งได้ด้วยการวิเคราะห์การจัดกลุ่มนั้นเป็นตัวแทนของประชากรโดยทั่วไปและสามารถนำไปใช้กับข้อมูลอื่นๆ โดยมีความเสถียร (stable) ในตัวเองเสมอ

กระบวนการยืนยันความถูกต้องที่นิยมใช้กันมาก คือ การทำการวิเคราะห์การแบ่งกลุ่มกับกลุ่มข้อมูลตัวอย่างอีกชุดหนึ่ง แล้วทำการเปรียบเทียบผลลัพธ์ที่ได้ว่าสอดคล้องกับผลลัพธ์ที่ได้ในตอนแรกหรือไม่ แต่บ่อยครั้งทีเดียวที่กระบวนการดังกล่าวไม่สามารถนำมาใช้ในทางปฏิบัติได้ เนื่องจากข้อจำกัดเรื่องของเวลาและค่าใช้จ่ายในการเก็บข้อมูลเพิ่มเติม ในกรณีที่การวิเคราะห์นั้นเกี่ยวข้องกับตัวแปรหลายตัว ในกรณีนี้ผู้วิเคราะห์อาจจะแบ่งกลุ่มข้อมูลที่มีออกเป็น 2 ส่วน ส่วนแรกสำหรับใช้ในการวิเคราะห์การจัดกลุ่มและส่วนที่ 2 สำหรับใช้ในการยืนยันความถูกต้อง (validation)

2.11 Nelder – Mead Simplex Algorithm : (NMSIMP)

ในปี 1965 Nelder และ Mead ได้นำเสนอ Downhill simplex method ซึ่งเป็นเทคนิคเกี่ยวกับการหาความเหมาะสมที่สุดโดยไม่ต้องใช้การหาค่าอนุพันธ์ (derivatives) และไม่ต้องมีสมมติฐานเกี่ยวกับ สมการเป้าหมาย (objective function) ว่าจะต้องมีอนุพันธ์อย่างต่อเนื่อง

NMSIMP จัดเป็นเทคนิคการแก้ปัญหาชั้นทั่วไปประเภท direct search ที่นิยมมากในการ minimize แบบ multidimensional unconstrained ถึงแม้ว่าจะมีการนำไปใช้กันอย่างกว้างขวางแต่วิธีการดังกล่าวก็ยังไม่มีผลลัพธ์ทางทฤษฎีที่สามารถพิสูจน์อัลกอริทึมนี้ได้อย่างชัดเจน

ประเภทของข้อกำหนด ที่สามารถนำมาใช้ได้กับอัลกอริทึมนี้ แบ่งเป็น 2 กลุ่ม ดังนี้

- ปัญหาประเภทที่ไม่มีเงื่อนไข (unconstrained problems) หรือปัญหาประเภทที่มีเงื่อนไขจำกัดขอบเขต (boundary constrained problems)

Nelder–Mead simplex algorithm แบบดั้งเดิม (original version) สามารถนำมาใช้กับปัญหาลักษณะนี้ได้ โดยเป็นอัลกอริทึมที่จะไม่นำเอาค่าที่เป็นไปไม่ได้ (infeasible points) มาคำนวณวัตถุประสงค์เป้าหมาย (objective function) อัลกอริทึมแบบเก่าจะสามารถนำมาใช้ได้ทันทีเมื่อปัญหาที่กำลังสนใจจัดอยู่ในกลุ่มนี้ เพราะอัลกอริทึม แบบเก่าจะสามารถแก้ไขปัญหาได้เร็วกว่า

- ปัญหาที่มีข้อกำหนด แบบเส้นตรง (linearly constrained problems)

อัลกอริทึมที่จะนำมาใช้กับปัญหากลุ่มที่ 2 ได้จะเป็น version Powell's (1992) ซึ่งได้ทำการดัดแปลงมาจากแบบเดิมเล็กน้อย เรียกว่า COBYLA (Constrained Optimization By Linear Approximations)

เนื่องจากปัญหาที่กำลังศึกษาในงานวิจัยฉบับนี้ เป็นปัญหาประเภทไม่มีเงื่อนไขและจะไม่นำเอาข้อมูลทางด้านสรีระของคนที่ไม่สามารถจัดเข้าไปอยู่ในระบบได้ (หรือจัดเป็น infeasible points) มาคำนวณค่าวัตถุประสงค์เป้าหมายด้วย ดังนั้นรูปแบบของอัลกอริทึมที่จะนำมาใช้ในการแก้ปัญหา คือ Nelder-Mead simplex algorithm : แบบดั้งเดิม

2.12 Nelder-Mead simplex algorithm : original version

เทคนิคการหาความเหมาะสมที่สุดนี้ เป็นวิธีการซึ่งอธิบายการ minimize function ของตัวแปรจำนวน n ตัวแปรที่อาศัยหลักการเปรียบเทียบค่าฟังก์ชันของจุดต่างๆ จำนวน n+1 จุดของซิมเพล็กซ์แล้วทำการแทนที่จุดที่มีค่ามากที่สุดด้วยจุดใหม่ซึ่งให้ค่าวัตถุประสงค์เป้าหมายต่ำกว่า

ซิมเพล็กซ์จะทำการปรับตัวเองเข้าสู่พื้นที่ที่เป็น local โดยการขยายพื้นที่ลงไปตามระนาบที่ลาดเอียง , การเปลี่ยนทิศทางไปตามหุบเขา (valley) ที่พบโดยบังเอิญ และการย่อพื้นที่บริเวณ neighborhood จนกระทั่งได้ค่าวัตถุประสงค์เป้าหมายที่ต่ำสุดค่าสุดท้าย (final minimum) ได้ วิธีการนี้ได้แสดงให้เห็นถึงประสิทธิภาพและการคำนวณที่กะทัดรัด จึงเป็นที่ต้องการสำหรับการแก้ปัญหาเกี่ยวกับการประมาณทางสถิติอย่างมาก

2.12.1 วิธีการทำงานของ Nelder-Mead Simplex Method

เริ่มจากการพิจารณา minimize function ของตัวแปร n ตัวโดยไม่มีเงื่อนไขใดๆ ทั้งสิ้น

<u>กำหนดให้</u>

- $P_0, P_1, P_2 \ldots\ldots, P_n$ คือ จุด n+1 จุด ใน n มิติ ซึ่งถูกกำหนดให้เป็นซิมเพล็กซ์ปัจจุบัน [คำว่า ซิมเพล็กซ์ ในที่นี้ไม่ได้หมายถึง ซิมเพล็กซ์ ทั่วๆ ไป]
- Y_i คือ ค่าของวัตถุประสงค์เป้าหมาย (objective function) ณ จุด P_i ให้
  - h คือ suffix แทน high เช่น $Y_h = \max(Y_i)$ และ
  - l คือ suffix แทน low เช่น $Y_l = \min(Y_i)$
- $\overline{P}$ คือ จุดศูนย์กลาง (Centroid) ของเซตจุดที่ดีที่สุด n จุดแรก หรือเท่ากับ $\sum_{i=1}^{n} P_i$ / n ของจุดซึ่ง $i \neq h$

- $[P_i\ P_j]$ คือ ระยะห่างระหว่าง P_i ถึง P_j

ในแต่ละขั้นตอนของกระบวนการ P_h จะถูกแทนที่ด้วยจุดใหม่ตามการดำเนินการ 3 แบบ ได้แก่

1. การสะท้อน (reflection)
2. การย่อส่วน (contraction)
3. การขยายพื้นที่ (expansion)

การสะท้อนของจุด P_h คือ การแทนที่จุด P_h ด้วย P^* โดยพิกัดของจุด P^* สามารถหาได้จากความสัมพันธ์

$$P^* = (1+\alpha)\bar{P} - \alpha P_h$$

เมื่อ α คือ ค่าคงที่ ที่มีค่าเป็นบวก เรียกว่า สัมประสิทธิ์การสะท้อน (Reflection Coefficient)

ดังนั้น P^* คือ จุดที่อยู่บนเส้นซึ่งเชื่อมระหว่าง $\bar{P}$ และ P_h (บนด้านซึ่งอยู่ตรงข้ามกับ P_h) โดย $[P^*\bar{P}] = \alpha [P_h P^*]$

ถ้าค่า Y^* อยู่ระหว่าง Y_h และ Y_l แล้ว P_h จะถูกแทนที่ด้วย P^* แล้วจึงเริ่มซิมเพล็กซ์สำหรับรอบใหม่ต่อไป

ถ้า $Y^* < Y_l$ หรือหมายความว่า การสะท้อนในครั้งนั้นทำให้เกิดค่า minimum ค่าใหม่แล้ว ขั้นต่อไปก็คือการขยาย P^* ไปสู่ P^{**} จากความสัมพันธ์

$$P^{**} = \gamma P^* + (1-\gamma)\bar{P}$$

เมื่อ γ คือ ค่าคงที่ แทน สัมประสิทธิ์การขยาย (Expansion Coefficient)

γ คือ สัดส่วนของระยะระหว่าง $[P^{**}\bar{P}]$ และ $[P^*\bar{P}]$

ถ้าค่า $Y^{**} < Y_l$ จะแทนค่า P_h ด้วย P^{**} แล้วเริ่มการพิจารณารอบใหม่

ถ้าค่า $Y^{**} \geq Y_l$ แสดงว่าการขยายพื้นที่ในครั้งนี้ล้มเหลว ให้แทนค่า P_h ด้วย P^* ก่อนที่จะทำการเริ่มต้นใหม่ ในกรณีที่ทำการสะท้อน P ไปที่ P^* แล้วพบว่า $Y^* \geq Y_i$ สำหรับทุกๆ ค่า i ที่ $i \neq h$ เช่นในการแทน P ด้วย P^* แล้วทำให้ Y^* เป็นค่า maximum แล้ว ให้ทำการกำหนด P_h ใหม่ ซึ่งอาจเป็นได้ทั้งค่า P_h เก่าหรือ P^* ค่าใดก็ได้ที่ให้ค่าฟังก์ชัน Y ต่ำกว่า และฟอร์มจุด P^{**} ตามความสัมพันธ์

$$P^{**} = \beta P_h + (1 - \beta)\overline{P}$$

เมื่อ β คือ ค่าคงที่ แทน สัมประสิทธิ์การย่อ (contraction coefficient) ซึ่งมีค่าอยู่ระหว่าง 0 -1 แสดงอัตราส่วนระหว่างระยะห่างของ $[P^{**}\ \overline{P}\ ]$ กับ $[P\overline{P}]$

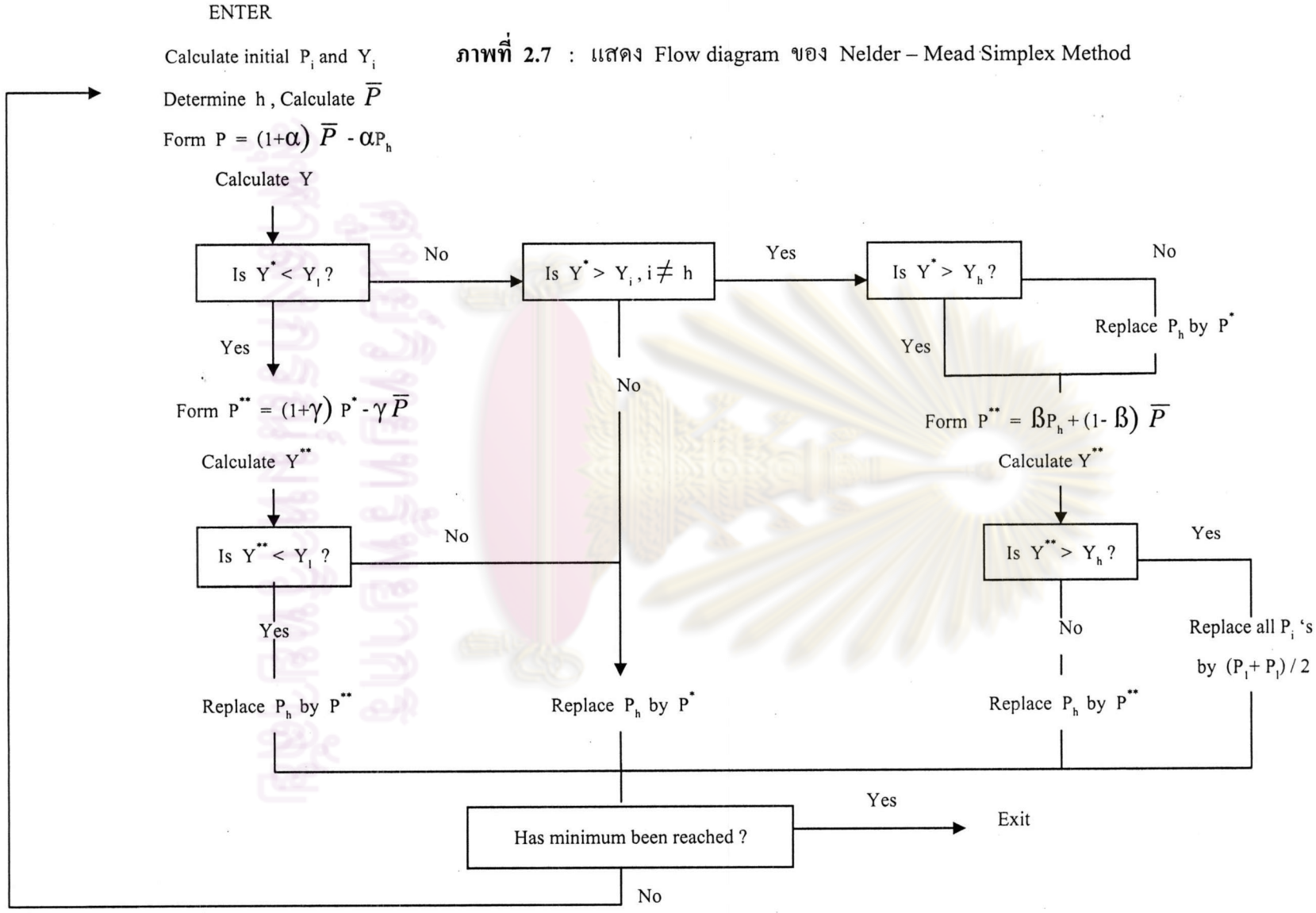
ถ้าค่า Y^{**} น้อยกว่าค่าต่ำสุดระหว่าง Y_h กับ Y^* ให้ทำการแทนค่า P_h ด้วย P^{**} แล้วเริ่มต้นพิจารณาใหม่

แต่ถ้า $Y^{**} > \min(Y_h, Y^*)$ แล้วแสดงว่าการย่อส่วนครั้งนี้ล้มเหลว ให้แทนค่าของ P_i ใหม่ด้วย $(P_i + P_l)/2$ แล้วเริ่มต้นพิจารณาใหม่

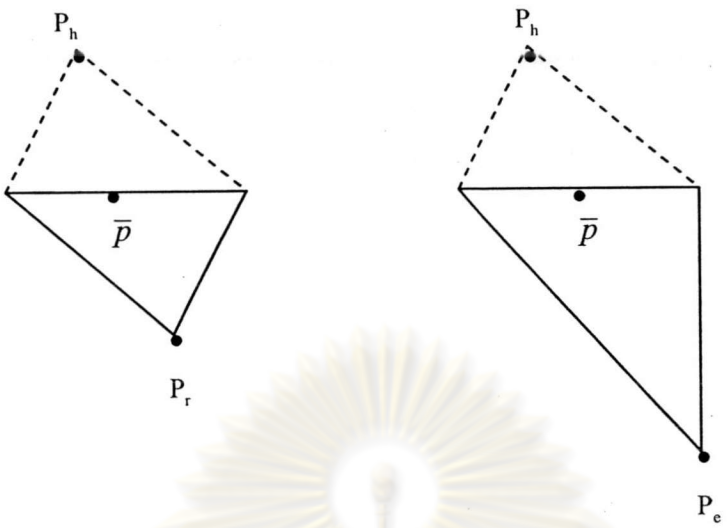
ค่าสัมประสิทธิ์ α, β, γ เป็นปัจจัยซึ่งแสดงปริมาณของซิมเพล็กซ์ที่จะถูกเปลี่ยนแปลงไปด้วยกระบวนการสะท้อน, การย่อ และการขยายตามลำดับ วิธีการทั้งหมดสามารถสรุปได้ตามแผนภูมิการไหล (Flow diagram) ดังภาพที่ 2.7

ค่าสัมประสิทธิ์ α, β, γ กำหนดให้มีค่าเท่ากับ 1 , 0.5 และ 2 ตามลำดับ [J.A.Nelder Mead and R.Mead (1965) , Jeffrey C. Lagarias, James A. Reeds, Margaret H.Wright and Paul E. Wright (1997)]

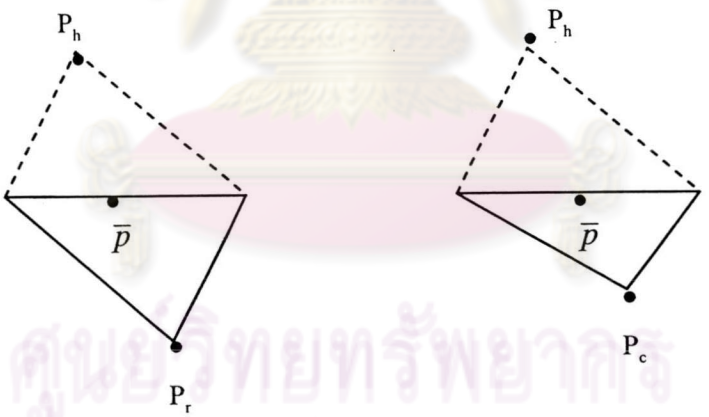
ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย



ภาพที่ 2.7 : แสดง Flow diagram ของ Nelder – Mead Simplex Method



ภาพที่ 2.8 : แสดง Nelder-Mead smplex หลังจากขั้นตอนการสะท้อนและการขยาย โดยซิมเพล็กซ์เดิมถูกแสดงด้วยเส้นประ



ภาพที่ 2.9 : แสดง Nelder-Mead simplex หลังจากขั้นตอนการสะท้อนและการย่อส่วน โดยซิมเพล็กซ์เดิมถูกแสดงด้วยเส้นประ

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

<u>หมายเหตุ</u> สามารถดูตัวอย่างการดำเนินการของอัลกอริทึมได้ในภาคผนวก ค

2.13 บทสรุปเกี่ยวกับทฤษฎีที่เกี่ยวข้อง

ภายหลังจากที่ได้ทำการศึกษาทฤษฎีที่เกี่ยวข้องกับการวิเคราะห์ตัวแปรหลายตัว (Multivariate Analysis) อันได้แก่

1. การวิเคราะห์ความสัมพันธ์ของตัวแปรเชิงปริมาณมากกว่า 2 ตัวขึ้นไป ซึ่งในที่นี้เลือกศึกษา 2 เทคนิค คือ
   - การวิเคราะห์ความถดถอยอย่างง่าย (simple regression analysis)
   - การวิเคราะห์สหสัมพันธ์ (correlation analysis)
2. การจำแนกกลุ่มข้อมูลหรือตัวแปรด้วยเทคนิคการวิเคราะห์ปัจจัย (factor analysis)
3. การจำแนกกลุ่มข้อมูลหรือตัวแปรด้วยเทคนิคการวิเคราะห์การจัดกลุ่ม (cluster analysis)

ผู้วิจัยจะนำเทคนิคการวิเคราะห์ปัจจัย (factor analysis) และการวิเคราะห์สหสัมพันธ์ (correlation analysis) มาใช้ในการวิเคราะห์ข้อมูลเพื่อการจัดกลุ่มตัวแปร เพราะวัตถุประสงค์หลักของการวิเคราะห์ความสัมพันธ์ระหว่างจุดวัดในงานวิจัยนี้เพื่อต้องการลดจำนวนตัวแปรหรือจำนวนจุดวัดต่างๆ ให้น้อยลงเพื่อนำไปใช้เป็นจุดวัดสำหรับการออกแบบระบบการจัดขนาดในขั้นต่อไป

สำหรับเทคนิคการวิเคราะห์การจัดกลุ่ม (cluster analysis) ผู้วิจัยจะนำหลักการดังกล่าวมาเป็นรูปแบบหนึ่งในการออกแบบระบบการจัดขนาดเพื่อหาค่าจุดวัดซึ่งเป็นตัวแทนของแต่ละขนาดภายในระบบ ผลลัพธ์ที่ได้จะถูกนำมาเปรียบเทียบกับการจัดขนาดในระบบปัจจุบันและระบบที่ได้จากการแก้ปัญหาด้วยวิธีการหาความเหมาะสมที่สุด

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย