บทที่ 3

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้ต้องการศึกษาและเปรียบเทียบวิธีการประมาณค่าสังเกตที่สูญหายในการวางแผนการทดลองแบบสุ่มในบล็อคสมบูรณ์ โดยเปรียบเทียบ 3 วิธี คือ วิธีกำลังสองน้อยสุด วิธีEM algorithm (Expectation Maximization) วิธี Imputation Method โดยมีขั้นตอนในการวิจัย 3 ขั้นตอน ซึ่งขั้นตอนแรกคือ การสร้างรูปแบบประชากรให้มีการแจกแจงแบบปกติ สองคือ การสุ่มตัดข้อมูลให้เหมือนกับการสูญหายจริง สามคือ การคำนวณค่าโดยใช้วิธีการต่าง ๆ

3.1 การสร้างรูปแบบการแจกแจงของประชากรแบบปกติ

ในการวิจัยครั้งนี้ได้ทำการสร้างการแจกแจงของประชากรแบบปกติ ด้วยเทคนิคมอนติคาร์โล ซึ่งเป็นวิธีหนึ่งในการจำลองตัวแบบทางคณิตศาสตร์ที่นิยมใช้ในปัจจุบันโดยหลักของมอนติคาร์โลนั้นต้องจำลองตัวเลขสุ่ม (Random Number) เพื่อช่วยในการหาคำตอบของปัญหาที่ต้องการศึกษา ซึ่งขั้นตอนของวิธีมอนติคาร์โลที่ใช้กันอยู่ในปัจจุบันแบ่งได้เป็น 2 ขั้นตอนดังนี้

1) การสร้างตัวเลขสุ่ม การใช้ตัวเลขสุ่มเป็นสิ่งสำคัญมากในวิธีมอนติคาร์โลทั้งนี้ก็เพราะว่าหลักของมอนติคาร์โลนั้นจะใช้ตัวเลขสุ่มมาช่วยในการหาคำตอบของปัญหา

2) การประยุกต์ปัญหาที่ต้องการศึกษามาใช้กับตัวเลขสุ่มโดยตรงแต่อาจมีขั้นตอนอื่น ๆ อีกหลายขั้นตอน ซึ่งขั้นตอนเหล่านี้บางขั้นตอนต้องใช้ตัวเลขสุ่ม การเขียนโปรแกรมในงานวิจัยครั้งนี้ใช้ภาษา S-plus 2000 และประมวลผลด้วยเครื่อง PC (Personal Computer) สร้างการแจกแจงแบบปกติใช้ตัวเลขสุ่มในฟังก์ชัน *morm* โดยมีรายละเอียดในการแจกแจงปกติดังนี้
ในการวิจัยครั้งนี้กำหนดให้

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

และกำหนดให้ τ_i , β_j , ε_{ij} เป็นตัวแปรสุ่มที่มีการแจกแจงแบบปกติที่เป็นอิสระกันด้วย $E(\tau_i) = E(\beta_j) = E(\varepsilon_{ij}) = 0$ และ $Var(\tau_i) = \sigma_\tau^2$, $Var(\beta_j) = \sigma_\beta^2$, $Var(\varepsilon_{ij}) = \sigma_\varepsilon^2$ โดย $\sigma_\tau^2 = \sigma_\beta^2 = h\sigma_\varepsilon^2$ ซึ่ง h เป็นจำนวนเต็มคงที่ดังนั้นเราจะได้ค่า y_{ij} ซึ่งเป็นค่าสังเกตในการทดลองนั้น ๆ

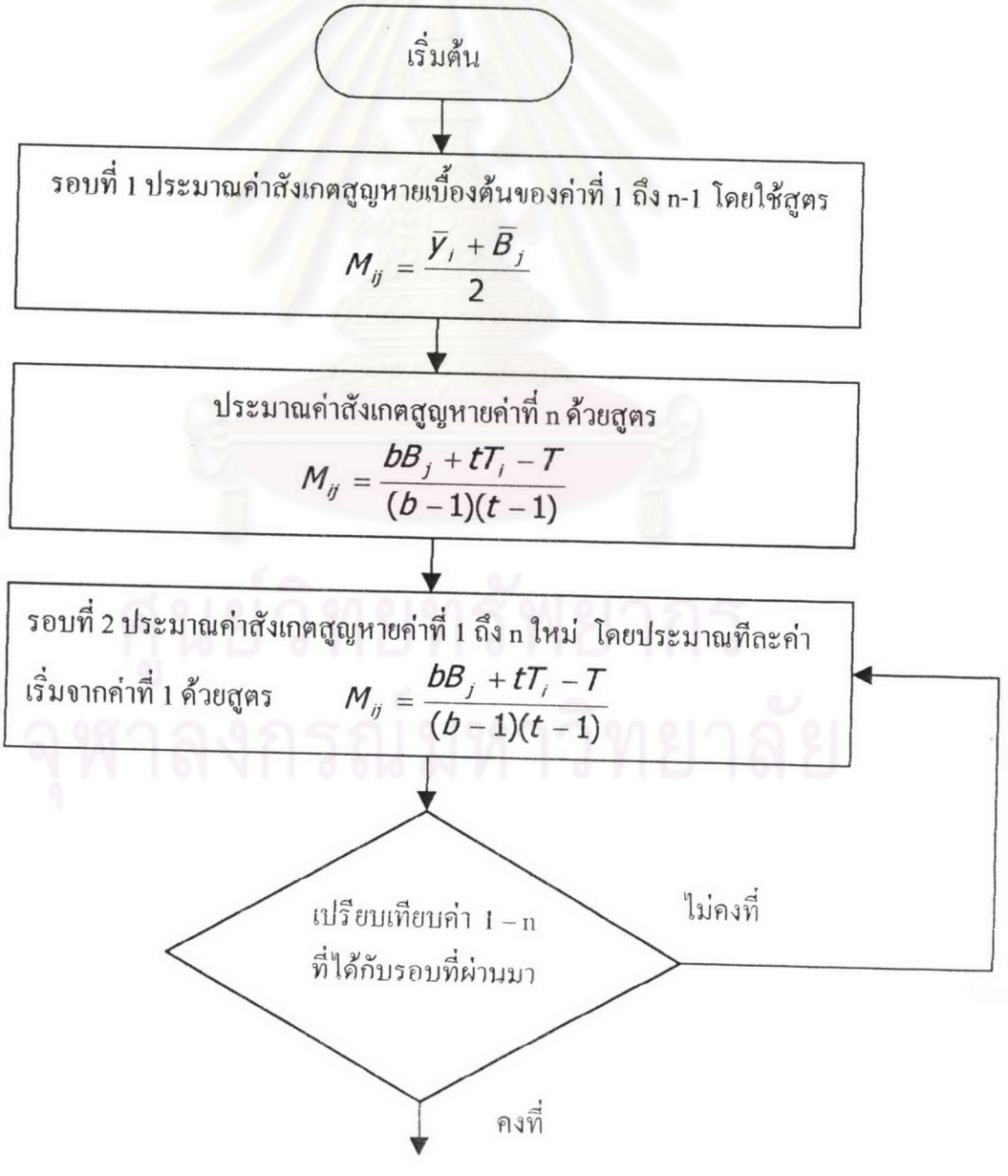
3.2 สร้างข้อมูลให้เกิดการสูญหาย

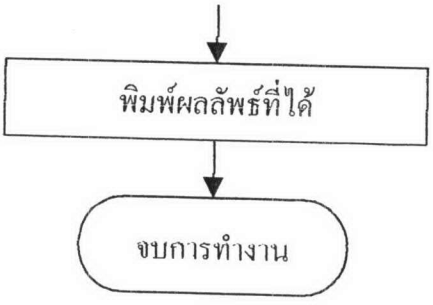
เมื่อสร้างข้อมูลเสร็จแล้วในขั้นตอนนี้จะสุ่มตัดข้อมูลออกด้วยฟังก์ชันการแจกแจงแบบยูนิฟอร์ม นำข้อมูลที่ตัดออกไปเก็บไว้เพื่อเปรียบเทียบกับค่าใหม่ที่จะประมาณขึ้น

3.3 การคำนวณค่าที่สูญหาย

เมื่อสร้างข้อมูล y_{ij} ที่เป็นไปตามข้อกำหนดข้างต้นและสุ่มตัดข้อมูลให้เกิดการสูญหายเรียบร้อยแล้วนำ ข้อมูลที่เหลือไปคำนวณหาค่าประมาณด้วยวิธีการต่าง ๆ

3.3.1 วิธีประมาณค่าสูญหายโดยวิธีกำลังสองน้อยสุด (least square method)

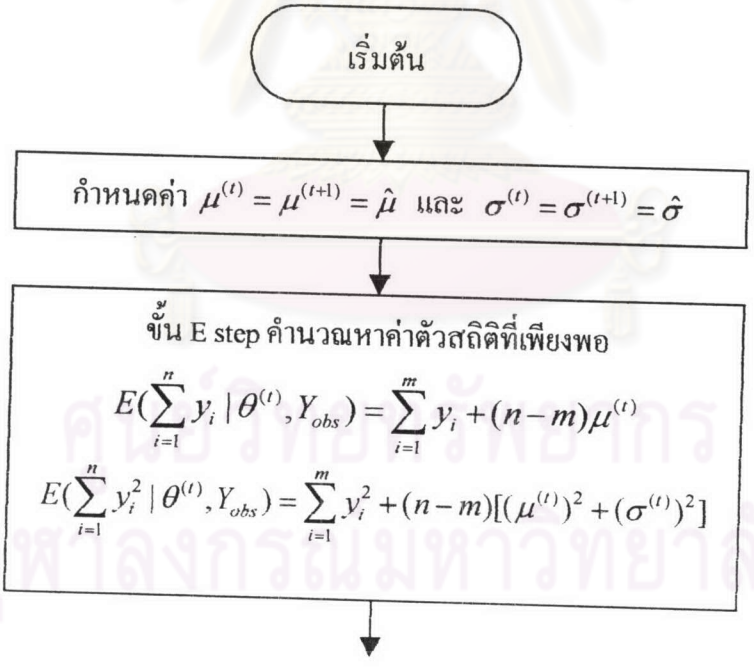


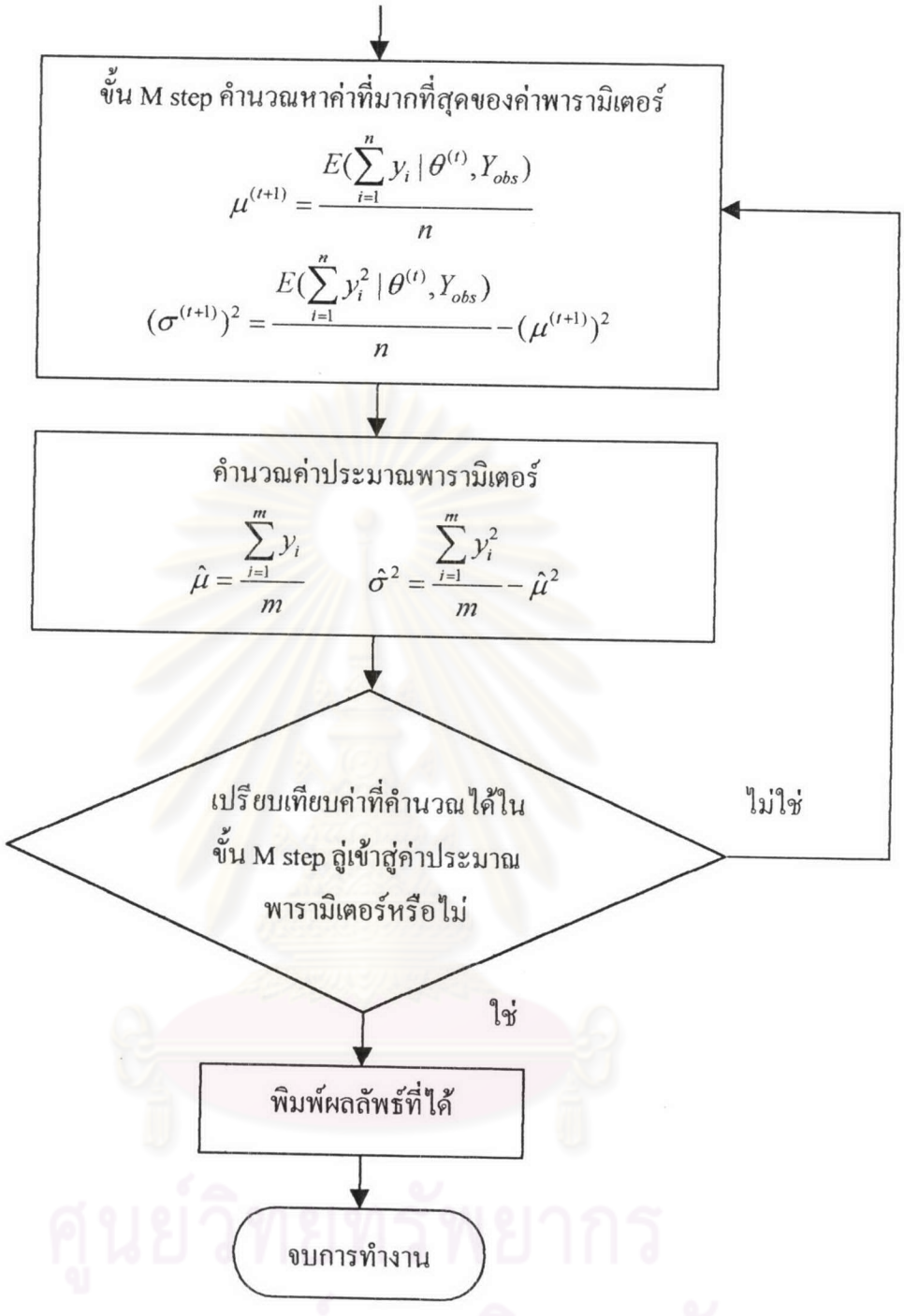


3.3.2 วิธีประมาณค่าสูญหายโดยวิธี EM algorithm (Expectation Maximization)

สมมติว่าข้อมูล Y มีองค์ประกอบ 2 ส่วน คือ $Y = (Y_{mis}, Y_{obs})$ และสมมติว่ามีข้อมูลทั้งหมด n ค่า m ค่าเป็นค่าที่เก็บมาได้ ส่วนที่เหลืออีก $n - m$ ค่าเป็นข้อมูลที่หายไป EM algorithm นี้แบ่งได้เป็น 2 ขั้นตอน คือขั้นหาค่าคาดหวัง E step เป็นขั้นหาค่าตัวสถิติพอเพียง และขั้นหาค่ามากที่สุด M step เพื่อหาค่าที่มากที่สุดของค่าพารามิเตอร์ที่ไม่เพิ่มขึ้นอีกเมื่อเทียบกับรอบที่ t-1

พิจารณาการแจกแจงของข้อมูลที่เหลืออยู่ว่ามีการแจกแจงแบบใดหากมีการแจกแจงแบบปกติ จะมีขั้นตอนดังนี้





3.3.3 วิธีประมาณค่าสูญหายโดยวิธี Imputation (Imputation Method)

วิธี Imputation Method เป็นวิธีที่ใช้กับกรณีข้อมูลที่หายเป็นไปอย่างสุ่ม ถ้ามีข้อมูลสูญหาย m ค่า ให้สร้างชุดข้อมูลขึ้นมาใหม่โดยสุ่มจากข้อมูลที่เหลืออยู่ สร้างให้จำนวนแถวเท่ากับจำนวนที่

ข้อมูลหาย จำนวนคอลัมน์มีค่าอยู่ระหว่าง 2 ถึง 10 คอลัมน์ จากนั้นคำนวณหาค่าเฉลี่ยในแต่ละชุดของข้อมูลโดยนำค่าเฉลี่ยของข้อมูลชุดที่มีความแปรปรวนต่ำสุดมาเป็นตัวแทนของข้อมูลที่สูญหาย

ค่าประมาณแบบจุดของ q^* คือ

$$q^* = \frac{\sum_{j=1}^{m} q_i}{m}$$

q^* = ค่าเฉลี่ยของข้อมูลแต่ละชุด

ตัวอย่าง กรณีข้อมูลสูญหาย 2 ค่า

จำนวนแถวที่สร้างขึ้นใหม่จะมีค่าเท่ากับจำนวนข้อมูลที่สูญหายไป

จำนวนคอลัมน์ที่สร้างขึ้นมีค่าอยู่ระหว่าง 2 ถึง 10 ค่า จะให้ผลการประมาณค่าสูญหายที่ดี

Y1
Y5

ค่าเฉลี่ยชุดที่ 1

Y6
Y9

ค่าเฉลี่ยชุดที่ 2

Y11
Y14

ค่าเฉลี่ยชุดที่ 3

พิจารณาเลือกค่าเฉลี่ยชุดที่มีความแปรปรวนต่ำสุดเป็นตัวแทนของข้อมูลที่สูญหาย

3.3.4 การคำนวณหาค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน

หาค่าประมาณจากวิธีการต่าง ๆ และนำมาเปรียบเทียบกับค่าจริง

$$MSE = \frac{\sum_{i=1}^{n} (Y - \hat{Y})^2}{n}$$

Y = ค่าจริงที่ได้จากการจำลอง

$\hat{Y}$ = ค่าประมาณจากการใช้วิธีการประมาณค่า

n = จำนวนรอบจากการทดลอง

ดังนั้นวิธีการใดให้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน MSE ต่ำกว่าเป็นวิธีที่ดีกว่า นั่นแสดงว่าค่าประมาณที่ได้มีค่าใกล้เคียงค่าจริงที่สูญหายไปมากกว่า

โดยมีลำดับขั้นตอนการทำงานของโปรแกรมดังนี้

ขั้นตอนในการทำงานของโปรแกรม