บทที่ 3

การดำเนินการวิจัย

3.1 ขั้นตอนการดำเนินงานวิจัย

การวิจัยครั้งนี้ จะทำการศึกษาประสิทธิภาพของตัวประมาณค่ายอดรวมประชากร โดยอาศัยข้อมูลที่รวบรวมโดยหน่วยงานต่างๆ เช่น สำนักทะเบียนกลาง กรมการปกครอง, สำนักนโยบายและแผนกรุงเทพมหานคร, สำนักงานเศรษฐกิจการเกษตร เป็นต้น ข้อมูลมีใช้เป็นตัวอย่างเป็นข้อมูลจากการจดทะเบียน, รายงานการสำรวจตัวอย่าง เช่น การสำรวจแรงงานโดยสำนักงานสถิติแห่งชาติซึ่งกระทำ 3 รอบใน 1 ปี ซึ่งจะนำข้อมูลเหล่านี้ มาใช้เป็นข้อมูลในอดีตหรือค่าของตัวแปรช่วยที่มีความสัมพันธ์กับตัวแปรที่สนใจศึกษาและใช้แทนค่าวัดขนาด และเป็นค่าของตัวแปรที่สนใจศึกษาหรือค่าจริงของประชากร เพื่อนำมาเปรียบเทียบกับค่าประมาณค่ายอดรวมประชากรที่เป็นไปได้กับค่าจริง ในการสุ่มตัวอย่างแบบกลุ่มชั้นเดียว (one-stage cluster sampling) เมื่อแต่ละกลุ่มมีโอกาสถูกเลือกไม่เท่ากันแบบไม่ใส่คืน โดยข้อมูลส่วนใหญ่เป็นข้อมูลระดับเขต และระดับจังหวัด ดังนั้นจึงใช้เขตหรือจังหวัดเป็นหน่วยตัวอย่างในการเก็บรวบรวมข้อมูล ซึ่งมีลักษณะเป็นกลุ่มที่ประกอบด้วยหน่วยที่ให้ข้อมูลโดยตรงหลายๆหน่วยรวมกัน เมื่อได้ n กลุ่มหรือจังหวัดแล้วจะเก็บรวบรวมข้อมูลจากทุกหน่วยที่อยู่ในกลุ่มตัวอย่างเหล่านั้น

ขั้นตอนการดำเนินงานวิจัย มีดังนี้คือ

1. การเก็บรวบรวมข้อมูลที่ได้จัดทำเป็นรายงานหรือมีการบันทึกไว้แล้วจากหน่วยงานต่างๆ
2. กำหนดตัวแปรที่สนใจศึกษาและตัวแปรช่วยที่มีความสัมพันธ์กับตัวแปรที่สนใจศึกษา , หน่วยตัวอย่าง เลือกเฉพาะชุดข้อมูลที่มีค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร x และ y อยู่ระหว่าง 0.50-0.99
3. เรียงลำดับข้อมูลจากน้อยไปมากตามค่าของตัวแปร x แล้วให้ค่าอันดับกับหน่วยประชากร และให้ค่าอันดับตามวิธีการปรับใหม่โดยใช้ค่าสะสมของค่ารากที่สองของความถี่
4. กำหนดขนาดตัวอย่าง เพื่อใช้เป็นขอบเขตการวิจัย
5. คำนวณค่า MAPE ของตัวประมาณค่ายอดรวมประชากร ภายใต้แผนการเลือกตัวอย่างในงานวิจัยนี้
6. เลือกชุดข้อมูลต่อไปและดำเนินงานตามข้อ 2 ถึงข้อ 5 จนกระทั่งได้ข้อมูลที่เพียงพอต่อการวิเคราะห์ความแปรปรวน

7. เปรียบเทียบประสิทธิภาพของตัวประมาณค่ายอดรวมประชากรโดยพิจารณาค่าเฉลี่ย MAPE ในแต่ละกรณีภายใต้ขอบเขตการวิจัย

3.2 การกำหนดขนาดตัวอย่าง

เลือกข้อมูลของตัวแปรที่สนใจศึกษา y และตัวแปร x ที่มีความสัมพันธ์กับ y ซึ่งมีค่าสัมประสิทธิ์สหสัมพันธ์อยู่ในช่วง 0.50-0.99 เมื่อได้ข้อมูลตัวอย่างนำมาหาค่าความแปรปรวนของตัวประมาณค่ายอดรวมของประชากรดังแสดงในบทที่ 2 หัวข้อ (2.1) ถึง (2.14) กรณีใช้ค่าตัวแปร x แทนค่าวัดขนาด และพิจารณาว่าเมื่อเพิ่มขนาดตัวอย่าง ค่าRelative Efficiencyของแต่ละขนาดตัวอย่างลดลงมากน้อยเพียงใด จนกระทั่งเมื่อเพิ่มขนาดตัวอย่างระดับหนึ่ง ค่า Relative Efficiency นี้เพิ่มขึ้นเพียงเล็กน้อยเท่านั้น ก็แสดงว่าไม่จำเป็นต้องเพิ่มขนาดตัวอย่างให้มากกว่านี้แล้ว ซึ่งในที่นี้จะพิจารณาที่ขนาดตัวอย่าง 5 ระดับ คือ 2 , 3 , 4 , 6 และ 9

3.3 การคำนวณค่า MAPE ของตัวประมาณค่ายอดรวมประชากร

จากข้อมูลของตัวแปรที่สนใจศึกษา y และตัวแปร x ที่มีความสัมพันธ์กับ y ซึ่งมีค่าสัมประสิทธิ์สหสัมพันธ์อยู่ในช่วง 0.50-0.99

ใช้ค่าของตัวแปร x และค่าอันดับกำหนดความน่าจะเป็นที่กลุ่มตัวอย่างจะถูกเลือก โดยมีวิธีการให้ค่าอันดับดังนี้

1. กรณีที่ใช้แผนการเลือกตัวอย่างpps แบบไม่ใส่คืนของ Vasantha Kumar E., Srivenkatamana ,T. และ Srinath ,K.P. (1996)

1.1 ใช้ค่าตัวแปร x เป็นค่าวัดขนาด หรือเป็นตัวกำหนดความน่าจะเป็นที่กลุ่มในประชากร จะถูกเลือก ดังที่แสดงในบทที่ 2 หัวข้อ (2.1) และ (2.2)

1.2 ใช้ค่าอันดับที่จัดเรียงตามค่าตัวแปร x เป็นค่าวัดขนาดโดย จัดเรียงกลุ่มตามค่า x จากน้อยไปมาก แล้วให้ค่าอันดับกับกลุ่มประชากรมีค่าตั้งแต่ 1 จนถึง N เมื่อ N เป็นจำนวนกลุ่มในประชากร กรณีที่ x มีค่าซ้ำกัน จะใช้ค่าเฉลี่ยของค่าอันดับของกลุ่มในกลุ่มค่าเหล่านั้น คำนวณค่าความน่าจะเป็นที่กลุ่มในประชากร จะถูกเลือก ดังที่แสดงในบทที่ 2 หัวข้อ (2.1) ถึง (2.3)

1.3 ใช้ค่าอันดับที่ปรับเพื่อลดค่าความแตกต่างระหว่างค่าอันดับกับค่าของตัวแปร x ในเชิงปริมาณ เป็นค่าวัดขนาด โดยการหาค่าต่ำสุดและสูงสุดของข้อมูลตัวแปร x คำนวณ $\max(x_1, x_2, \ldots, x_N) = c^*\min(x_1, x_2, \ldots, x_N)$ เมื่อ c เป็นค่าคงที่ ปรับค่า c เป็นเลขจำนวน

เต็ม กำหนดให้ค่าอันดับของกลุ่มประชากรมีค่าต่ำสุดเป็น 1 และสูงสุดเป็น c และถ้าใช้ค่าสะสมของ $\sqrt{f}$ กำหนดขอบเขตของกลุ่ม ให้ค่าอันดับของหน่วย (cluster) ภายในกลุ่มจำนวน n_i หน่วย สำหรับ i = 1, 2 , . . . , c เป็นค่าเดียวกัน ดังนั้นแต่ละหน่วยภายในกลุ่มจะมีความน่าจะเป็นในการถูกเลือกเท่าๆกัน การคำนวณค่าความน่าจะเป็นที่หน่วยในประชากรจะถูกเลือก ต้องกระทำเช่นเดียวกับกรณีใช้ค่า x แทนค่าวัดขนาด

จากแต่ละกรณีของตัวแปรที่ใช้แทนค่าวัดขนาด เลือกตัวอย่างขนาดตามกำหนดข้างต้น เริ่มที่ขนาดตัวอย่าง2 โดยสุ่ม 2 หน่วยแรกด้วย pps แบบไม่ใส่คืน และสุ่มครั้งต่อๆไป แบบใส่คืน คำนวณค่าตัวประมาณค่ายอดรวมประชากรที่เป็นไปได้ ดังแสดงในบทที่ 2 หัวข้อ (2.4) และ (2.5) ชุดตัวอย่างที่เป็นไปได้เท่ากับ N!/(N-n)!*(n-2)! ชุดตัวอย่าง (เมื่อ N เป็นจำนวนกลุ่มในประชากร และ n เป็นจำนวนกลุ่มตัวอย่าง) พิจารณาว่าค่าต่างๆที่เป็นไปได้ของตัวประมาณ ($\hat{Y}_i$) ทั้งหมดแตกต่างจากค่ายอดรวมจริง(Y) เพียงไร ตัวประมาณใดมีค่าที่เป็นไปได้ต่างๆใกล้เคียงค่าจริงมากกว่า ก็ย่อมจะถูกต้องกว่าตัวประมาณอื่นที่มีค่าที่เป็นไปได้แตกต่างห่างออกไป เกณฑ์ในการวัดความถูกต้องอาจวัดได้ด้วยค่าเฉลี่ยเปอร์เซนต์ความคลาดเคลื่อนสัมบูรณ์ (MAPE)

$$MAPE = \frac{100}{m} * \sum_{i=1}^{m} \left| \frac{\hat{Y}_i - Y}{Y} \right|$$ เมื่อ m เป็นจำนวนชุดตัวอย่างที่เป็นไปได้

ถ้าขนาดตัวอย่างใหญ่มีชุดตัวอย่างที่เป็นไปได้จำนวนมากจะใช้วิธีสุ่มบางชุด แล้วคำนวณตัวประมาณค่ายอดรวมประชากร ภายใต้ข้อกำหนด

1. การแจกแจงของตัวประมาณค่ายอดรวมประชากรจากแต่ละชุดตัวอย่างที่เป็นไปได้ มีการแจกแจงแบบปกติ ทดสอบด้วยตัวสถิติ Kolmogorov-Smirnov
2. ค่าเฉลี่ยตัวประมาณค่ายอดรวมประชากรจากชุดตัวอย่างที่สุ่มมาได้ มีค่าเท่ากับค่ายอดรวมประชากรหรือค่าจริง ทดสอบด้วยค่าสถิติ t ที่ระดับนัยสำคัญ 0.05
3. ค่าความแปรปรวนของตัวประมาณค่ายอดรวมประชากรจากชุดตัวอย่างที่สุ่มมาต้องมีค่าต่ำกว่าค่าความแปรปรวนของตัวประมาณค่ายอดรวมประชากรของทุกชุดตัวอย่างที่เป็นไปได้

ถ้าไม่เป็นไปตามข้อกำหนดอาจทำการสุ่มชุดตัวอย่างเพิ่ม หรือ สุ่มใหม่ทั้งหมด

คำนวณหาค่า APE ของแต่ละชุดตัวอย่างที่สุ่มมาได้ แล้วหาค่าเฉลี่ยของ APE จะได้ค่าประมาณของ MAPE

บันทึกเป็นค่า MAPE เพิ่มขนาดตัวอย่างเป็นขนาด 3 ทำการเลือกตัวอย่างเช่นเดียวกับที่กล่าวข้างต้นแล้วบันทึกค่า MAPE ของทั้ง 3 กรณีของตัวแปรที่ใช้แทนค่าวัดขนาด ทำเช่นเดียวกันนี้แต่เพิ่มขนาดตัวอย่างจนกระทั่งถึงขนาดตัวอย่าง 9

2. กรณีที่ใช้แผนการเลือกตัวอย่าง ppswor ของ Wright

จัดเรียงหน่วย (cluster) ตามค่า x จากน้อยไปมาก กำหนดจำนวนชั้นภูมิ ให้เท่ากับจำนวน ตัวอย่างที่ต้องการทั้งหมด จัดหน่วยเข้าชั้นภูมิ โดยมีวิธีการแบ่งชั้นภูมิ 2 วิธีคือ วิธีที่ 1 เป็นวิธีของ Wright ให้ N = 2kn เมื่อ k เป็นเลขจำนวนเต็มบวกชั้นภูมิที่ 1 มี 2k หน่วยที่ประกอบด้วย k หน่วยแรกและ k หน่วยสุดท้ายในอันดับของประชากร ชั้นภูมิที่ 2 ถูกจัดเช่นเดียวกันหลังจากตัดหน่วยที่ถูกจัดในชั้นภูมิที่ 1 แล้ว ทำเช่นนี้ต่อไปจนกระทั่งได้ n ชั้นภูมิ อีกวิธีหนึ่ง จะใช้ค่าสะสมของ $\sqrt{f}$ กำหนดขอบเขตของชั้นภูมิ ส่วนในกรณีที่ x มีค่าซ้ำกัน จะใช้ค่าเฉลี่ยของค่าอันดับของหน่วยในกลุ่มค่าเหล่านั้น จากนั้นเลือกตัวอย่างจากแต่ละชั้นภูมิ ชั้นภูมิละ 1 หน่วยด้วยความน่าจะเป็นที่เป็นสัดส่วนกับขนาด โดยเปรียบเทียบกรณีที่ใช้ค่าตัวแปร x กับกรณีที่ใช้ค่าอันดับที่จัดตามค่าของตัวแปร x เป็นค่าวัดขนาด

คำนวณค่าความน่าจะเป็นที่หน่วยในประชากรจะถูกเลือก และตัวประมาณค่ายอดรวมประชากรที่ได้จากแต่ละชุดตัวอย่าง ดังที่แสดงในบทที่ 2 หัวข้อ (2.16) ถึง (2.19) พิจารณาว่าค่าต่างๆที่เป็นไปได้ของตัวประมาณ ($\hat{Y}_i$) ทั้งหมดแตกต่างจากค่ายอดรวมจริง(Y) เพียงไรจากค่า MAPE บันทึกค่าMAPEในแต่ละขนาดตัวอย่างและตัวแปรที่ใช้แทนค่าวัดขนาด

ในกรณีที่ใช้วิธีของ Wright กำหนดขอบเขตของชั้นภูมิ และ $N \neq 2kn$ จะทำการตัดหน่วยบางหน่วยออก คือ หน่วยตัวอย่างที่อยู่ในอันดับกึ่งกลางเพื่อให้การกระจายของประชากรเปลี่ยนแปลงเพียงเล็กน้อย เช่น N 76 จังหวัด n = 3 ต้องตัดหน่วยออก 4 หน่วยคือ หน่วยที่ 37-40 เหลือ N = 72 และจะคิดค่ายอดรวมจริงของ 72 จังหวัดเท่านั้น

ดังนั้นชุดข้อมูล 1 ชุด จะบันทึกค่า MAPE ไว้ 35 กรณี และทำการวิเคราะห์ความแปรปรวนในแต่ละระดับปัจจัยด้วยตัวสถิติ F โดยจะต้องทดสอบว่าข้อมูลทั้งหมดเป็นไปตามข้อตกลงเบื้องต้นของการวิเคราะห์ความแปรปรวนหรือไม่ คือ ค่าสังเกตหรือค่า MAPE แต่ละค่าเป็นอิสระกัน มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเท่ากับ μ_i และมีความแปรปรวนในแต่ละระดับปัจจัย (k ระดับ) เท่ากันคือ $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots\ldots = \sigma_k^2$ ถ้าไม่เป็นไปตามข้อตกลง จะทำการเลือกชุดข้อมูลเพิ่ม และทุกชุดข้อมูลจะทำการเลือกตัวอย่างและหาตัวประมาณค่ายอดรวมประชากรเช่นเดียวกับที่กล่าวมาข้างต้น

3.4 การเปรียบเทียบประสิทธิภาพของตัวประมาณค่ายอดรวมประชากรโดยพิจารณาค่าเฉลี่ย MAPE ในแต่ละกรณีภายใต้ขอบเขตการวิจัย

การหาข้อสรุปเกี่ยวกับความแตกต่างของค่าเฉลี่ย MAPE ของตัวประมาณค่ายอดรวมประชากร เพื่อศึกษาว่าค่าอันดับที่จัดเรียงตามค่าตัวแปร X และค่าอันดับที่ปรับใหม่ สามารถใช้แทนค่าวัดขนาดได้ดีเพียงไรเมื่อเทียบกับตัวแปรปริมาณหรือตัวแปร x ที่มีความสัมพันธ์กับ y ภายใต้แผนการเลือกตัวอย่างแบบ pps แบบไม่ใส่คืน และเปรียบเทียบประสิทธิภาพของตัวประมาณค่ายอดรวมประชากรที่ได้จากแต่ละแผนการเลือกตัวอย่าง ที่ระดับนัยสำคัญ 0.05 จึงได้ทำการวิเคราะห์ความแปรปรวนทดสอบค่าเฉลี่ย MAPE จำแนกตามปัจจัยต่างๆ ดังนี้

ปัจจัยที่ 1 คือ สัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรที่สนใจ(y) กับตัวแปร x ซึ่งมี 5 ระดับปัจจัยคือ ระดับ 0.50-0.59 , 0.60-0.69, 0.70-0.79, 0.80-0.89 และ 0.90-0.99

ปัจจัยที่ 2 คือ แผนการเลือกตัวอย่าง

1 แผนการเลือกตัวอย่างppsแบบไม่ใส่คืนของ Vasantha kumar E., Srivenkataramana,T. และ Srinath ,K.P. (1996)

2 แผนการเลือกตัวอย่าง pps แบบไม่ใส่คืน และวิธีการกำหนดขอบเขตของชั้นภูมิของ Wright (1990)

3 แผนการเลือกตัวอย่าง pps แบบไม่ใส่คืน ของ Wright และวิธีการกำหนดขอบเขตของชั้นภูมิด้วยวิธีค่าสะสมของค่ารากที่สองของความถี่ ($\sqrt{f}$) โดยอาศัยค่าของตัวแปร x

ปัจจัยที่ 3 คือ ขนาดตัวอย่าง มี 5 ระดับ คือ 2, 3, 4, 6 และ 9

ปัจจัยที่ 4 คือ ตัวแปรที่ใช้แทนค่าวัดขนาด แบ่งเป็น 3 กลุ่มคือ

1 ค่าของตัวแปร(x)ที่มีความสัมพันธ์กับตัวแปรที่สนใจ (y)

2 ค่าอันดับของหน่วยประชากรที่จัดตามค่าตัวแปร x

3 ค่าอันดับที่ปรับใหม่ ซึ่งจะพิจารณาเฉพาะในแผนการเลือกตัวอย่างที่1 เนื่องจากแผนการเลือกตัวอย่างที่ 2 และ 3 กำหนดจำนวนชั้นภูมิเท่ากับขนาดตัวอย่าง แต่แผนการเลือกตัวอย่างที่ 1 กำหนดจำนวนชั้นภูมิเท่ากับ จำนวนเท่าของความแตกต่างระหว่างค่าสูงสุดและค่าต่ำสุดของค่าตัวแปร x

พิจารณาทั้งในลักษณะของปัจจัยเดี่ยว , และผลกระทบร่วมของสองปัจจัย , สามปัจจัย และ สี่ปัจจัย และทำการทดสอบ multiple comparison หรือทดสอบความแตกต่างของค่าเฉลี่ยที่ละคู่โดยใช้ตัวสถิติ Tukey's studentized range test เมื่อปฏิเสธสมมติฐานว่าค่าเฉลี่ย MAPEในแต่ละระดับปัจจัยไม่แตกต่างกัน การวิเคราะห์เหล่านี้จะทำให้ทราบด้วยว่าที่ระดับนัยสำคัญ 0.05

แต่ละระดับความสัมพันธ์ระหว่างตัวแปรที่สนใจ (y) กับตัวแปร x หรือที่แต่ละขนาดตัวอย่าง แผนการเลือกตัวอย่างใดหรือการใช้ตัวแปรใดแทนค่าวัดขนาด จะให้ตัวประมาณ ($\hat{Y}$) ที่มีประสิทธิภาพมากที่สุด หรือมาก น้อยกว่ากันเท่าใด เป็นต้น

ส่วนในกรณีที่การวิเคราะห์ความแปรปรวนของประชากรมากกว่า 2 กลุ่มด้วยสถิติทดสอบเอฟที่เสนอโดย Brown & Forsythe (1974)เมื่อประชากรแต่ละกลุ่มมีความแปรปรวนไม่เท่ากัน

$$BF = \frac{\sum_{i}^{c} n_i(\bar{x}_i - \bar{x})^2}{\sum_{i}^{c}(1-n_i/N)s_i^2}$$

โดยที่ x_{ij} แทนค่า MAPE ที่ j ในระดับปัจจัยที่ i ($j=1,2,\ldots,n_i$),($i=1,2,\ldots,c$)

$$N = \sum_{i}^{c} n_i$$

$$\bar{x} = \frac{\sum_{j}^{c} n_i \bar{x}_i}{N}$$

$$\bar{x}_i = \frac{\sum_{j}^{n_i} x_{ij}}{n_i}$$

$$s_i^2 = \frac{\sum_{j}^{n_i}(x_{ij} - \bar{x}_i)^2}{(n_i - 1)}$$

ภายใต้สมมติฐานหลัก ($H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_c$) ที่เป็นจริง แล้วสถิติทดสอบแบบ Brown&Forsythe จะมีลักษณะการแจกแจกแบบเอฟโดยประมาณ ที่องศาความเป็นอิสระ (c-1,f*) โดยที่

$$f^* = [\sum_i^c v_i^2/(n_i-1)]^{-1}$$

$$v_i = \frac{(1-n_i/N)s_i^2}{\sum_i^c (1-n_i/N)s_i^2}$$

เกณฑ์การตัดสินใจจะปฏิเสธ H_0 เมื่อค่า BF มากกว่า $F_{\alpha,(c-1,f^*)}$ โดยที่ $F_{\alpha,(c-1,f^*)}$ คือค่าวิกฤตที่เปิดจากตาราง F ที่ระดับนัยสำคัญ α และองศาความเป็นอิสระ (c-1,f*) ตามลำดับ