รายงานการวิจัย

การประมาณค่าผลรวมประชากรในการสำรวจตัวอย่าง เมื่อมีข้อมูลสูญหาย

Estimation of Population Total in Sample Survey with Missing Data

ชูเกียรติ โพนแก้ว
สาขาคณิตศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี
มหาวิทยาลัยราชภัฏเพชรบูรณ์

ประจำปีงบประมาณ 2560

รหัสโครงการ PCRU_2560_L026

รายงานวิจัยฉบับสมบูรณ์

การประมาณค่าผลรวมประชากรในการสำรวจตัวอย่างเมื่อมีข้อมูลสูญหาย

Estimation of Population Total in Sample Survey with Missing Data

ชูเกียรติ โพนแก้ว **สาขาคณิตศาสตร์**

คณะวิทยาศาสตร์และเทคโนโลยี

ทุนอุดหนุนโดย มหาวิทยาลัยราชภัฏเพชรบูรณ์

งบประมาณแผ่นดินที่พิจารณาโดยผ่านความเห็นชอบ

จากสำนักงานคณะกรรมการวิจัยแห่งชาติ

ประจำปีงบประมาณ 2560

ชื่องานวิจัย การประมาณค่าผลรวมประชากรในการสำรวจตัวอย่างเมื่อมีข้อมูลสูญหาย
ผู้วิจัย ชูเกียรติ โพนแก้ว
สาขาวิชา สาขาวิชาคณิตศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี
มหาวิทยาลยราชภัฏเพชรบูรณ์ ปีเสร็จวิจัย 2560

บทคัดย่อ

การวิจัยในครั้งนี้ศึกษาเกี่ยวกับการประมาณค่าผลรวมประชากรในการสำรวจตัวอย่าง เมื่อมีข้อมูลสูญหาย ที่เกิดจากการไม่ตอบเฉพาะบางคำถาม ภายใต้การสุ่มตัวอย่างแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มตัวอย่างโดยใช้ความน่าจะเป็นในการเลือกไม่เท่ากัน

ผู้วิจัยนำเสนอตัวประมาณค่าแบบผลรวมประชากร ทั้งตัวประมาณค่าแบบจุด และตัวประมาณค่าแบบช่วง ตัวประมาณค่าแบบจุดที่นำเสนอประกอบด้วยตัวประมาณค่าแบบเชิงเส้น และตัวประมาณค่าแบบไม่เป็นเชิงเส้น ส่วนตัวประมาณค่าแบบช่วงที่นำเสนอประกอบด้วย ช่วงความเชื่อมั่นแบบเชิงเส้น และช่วงความเชื่อมั่นแบบไม่เป็นเชิงเส้น การหาตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบจุดที่นำเสนอศึกษาภายใต้เฟรมเวิร์คแบบรีเวิร์ส และสัดส่วนตัวอย่างในแต่ละชั้นภูมิมีขนาดเล็ก

ผลการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าที่นำเสนอ สำหรับตัวประมาณค่าแบบจุดเมื่อพิจารณาจากค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ พบว่า ตัวประมาณค่าแบบไม่เป็นเชิงเส้นมีประสิทธิภาพดีกว่าตัวประมาณค่าแบบเป็นเชิงเส้น ส่วนตัวประมาณค่าแบบช่วงเมื่อพิจารณาจากค่าเฉลี่ยของจำนวนครั้งที่ช่วงความเชื่อมั่นจะครอบคลุมพารามิเตอร์ พบว่า ช่วงความเชื่อมั่นแบบไม่เป็นเชิงเส้น มีประสิทธิภาพดีกว่าช่วงความเชื่อมั่นแบบเชิงเส้น

คำสำคัญ : ผลรวมประชากร ข้อมูลสูญหาย ความแปรปรวน เฟรมเวิร์คแบบรีเวิร์ส

Research 's Title Estimation of Population in Sample Survey with Missing Data
Researcher Chugiat Ponkase
Major Mathematics
Phetchabun Rajabhat University **Year** 2017

Abstract

In this research paper, we propose population total estimators in the presence of item nonresponse under stratified sampling with unequal probability sampling without replacement in each stratum both of point estimator and interval estimator.

The point estimator is comprised of the linear estimator and nonlinear estimator. The interval estimator contains linear interval estimator and nonlinear interval estimator. The variance and associated estimator of the point estimators are investigated under reverse framework with sampling fraction in each stratum is negligible.

The simulation study show that the nonlinear estimator is efficiency than linear estimator and the nonlinear interval estimator is efficiency than linear interval estimator.

Keywords: Population total, Missing data, Variance, Reverse framework

กิตติกรรมประกาศ

ขอขอบคุณสถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏเพชรบูรณ์ ที่ให้ทุนสนับสนุนประจำปีงบประมาณ 2560

ขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.ชัยณรงค์ ขันผนึก ผู้ช่วยศาสตราจารย์อาดุลย์ จงรักษ์ และอาจารย์ ดร.เจษฎา โพนแก้ว ที่ได้ให้ความอนุเคราะห์เป็นผู้ทรงคุณวุฒิตรวจสอบเครื่องมือ และเนื้อหาในโครงการวิจัย

ท้ายที่สุดนี้ ขอขอบคุณผู้ให้ความช่วยเหลือและคำแนะนำจากผู้ที่เกี่ยวข้องทุกท่าน ตลอดจนบิดามารดาและสมาชิกในครอบครัวทุกคนที่คอยเป็นกำลังใจ ให้การสนับสนุนทุกอย่างที่เกี่ยวข้องกับงานวิจัยทั้งทางตรงและทางอ้อม จนงานวิจัยนี้สำเร็จลงได้

ชูเกียรติ โพนแก้ว

20 มีนาคม 2561

สารบัญ

		หน้า
บทคัดย่อภาษาไทย		ก
บทคัดย่อภาษาอังกฤษ		ข
กิตติกรรมประกาศ		ค
สารบัญ		ง
สารบัญตาราง		ฉ
สารบัญรูป		ช
บทที่ 1	บทนำ	1
1.1	ที่มาและความสำคัญของปัญหา	1
1.2	วัตถุประสงค์	3
1.3	สมมติฐานการวิจัย	3
1.4	ประโยชน์ที่คาดว่าจะได้รับ	3
1.5	ขอบเขตการวิจัย	3
1.6	นิยามศัพท์ที่เกี่ยวข้อง	4
1.7	กรอบแนวคิดในการวิจัย	4
บทที่ 2	เอกสารและงานวิจัยที่เกี่ยวข้อง	5
2.1	ข้อมูลสูญหาย	5
2.2	ตัวประมาณค่าผลรวมประชากร	7
2.3	การหาความแปรปรวนของตัวประมาณค่าแบบไม่เป็นเชิงเส้น	13
2.4	งานวิจัยที่เกี่ยวข้อง	14
บทที่ 3	วิธีดำเนินการวิจัย	17
3.1	สัญลักษณ์และข้อตกลงเบื้องต้น	17
3.2	การสุ่มตัวอย่างในแต่ละชั้นภูมิ	18
3.3	การประมาณค่าความแปรปวนภายใต้เฟรมเวิรค์แบบรีเวิร์ส	18
3.4	การนำเสนอตัวประมาณค่าผลรวมประชากรแบบจุด	19
3.5	นำเสนอตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร	19
3.6	นำเสนอช่วงความเชื่อมั่นของผลรวมประชากร	19
3.7	แสดงตัวอย่างเชิงตัวเลข	19

สารบัญ (ต่อ)

		หน้า
บทที่ 4	ผลการวิจัย	22
	4.1 ตัวประมาณแบบจุดของผลรวมประชากร	22
	4.2 ความแปรปรวนและตัวประมาณค่าความแปรปวนของตัวประมาณค่าแบบจุด	24
	4.3 ช่วงความเชื่อมั่นของผลรวมประชากร	30
	4.4 ผลการแสดงตัวอย่างเชิงตัวเลข	31
บทที่ 5	สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ	34
	5.1 สรุปผลการวิจัย	34
	5.2 การอภิปรายผล	36
	5.3 ข้อเสนอแนะ	37
บรรณานุกรม		38
ภาคผนวก		40
ภาคผนวก ก	รายชื่อผู้ทรงคุณวุฒิ	41
ภาคผนวก ข	การเผยแพร่งานวิจัย	43
ภาคผนวก ค	ประวัติผู้วิจัย	55

สารบัญตาราง

หน้า

ตารางที่

3.1 ขนาดตัวอย่างที่ใช้ในการศึกษา..... 20

4.1 ค่าพารามิเตอร์ของแต่ละชั้นภูมิ..... 31

4.2 ค่ารากที่สอง ของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ ของ $\hat{V}(\hat{Y}_{LMI})$ และ $\hat{V}(\hat{Y}_{NLMI})$ 32

4-3 ร้อยละของค่าเฉลี่ยของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมพารามิเตอร์..... 33

สารบัญรูป

หน้า

รูปที่

1.1 กรอบแนวคิดในการทำวิจัย..... 4

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ปัญหาเกี่ยวกับการประมาณค่าผลรวมประชากร (Population total) เป็นหนึ่งในปัญหาที่สำคัญในการศึกษาวิชาทางสถิติ การประมาณค่าผลรวมประชากรสามารถทำได้โดยใช้ข้อมูลของตัวอย่างที่ถูกสุ่มจากประชากร ตามข้อตกลงของแผนการสุ่มตัวอย่าง (Sampling design) และจำเป็นต้องใช้ข้อมูลของตัวอย่างที่มีความสมบูรณ์เพื่อใช้ในการหาตัวประมาณค่าที่ไม่เอนเอียง (Unbiased estimator) ณรงค์ โพธิ และ สมชาย ปราการเจริญ (2553) กล่าวว่า ตัวประมาณค่าผลรวมประชากร อาจจะไม่สามารถใช้งานได้อย่างเต็มประสิทธิภาพ เนื่องจากมีข้อมูลบางส่วนเกิดการสูญหาย (Missing data) ทำให้เกิดความไม่สมบูรณ์ของข้อมูล ซึ่งอาจเกิดจากขั้นตอนจัดเก็บ ปัญหาจากการป้อนข้อมูล หรือเกิดจากข้อจำกัดของเทคโนโลยีที่ใช้งาน ข้อมูลของตัวอย่างที่เกิดการสูญหายอาจส่งผลให้ตัวประมาณค่าผลรวมประชากรที่ได้มีความเอนเอียง (Bias estimator)ข้อมูลสูญหาย เป็นกรณีที่พบได้บ่อยในงานวิจัยทุกสาขา และนักวิจัยจำเป็นต้องพิจารณาถึงแนวทางที่เหมาะสมสำหรับใช้จัดการกับข้อมูลสูญหาย (ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และ สุคนธ์ ประสิทธิ์วัฒนเสรี, 2549) การสูญหายของข้อมูลอาจเกิดจากการไม่ตอบของหน่วยตัวอย่างบางหน่วย (Unit nonresponse) หรือเกิดจากการไม่ตอบเฉพาะบางคำถาม (Item nonresponse) เพียงออ ยีสา และกัลยา วานิชย์บัญชา (2552) กล่าวว่า การจัดการกับข้อมูลสูญหายมีหลายวิธี การเลือกใช้วิธีใดวิธีการหนึ่งขึ้นอยู่กับลักษณะข้อมูลสูญหายที่เกิดขึ้น สำหรับการจัดการกับข้อมูลที่สูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม นิยมใช้วิธีการแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ (Imputation) เช่น การประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย (Mean imputation) การประมาณข้อมูลสูญหายด้วยค่าอัตราส่วน (Ratio imputation) หรือ การประมาณข้อมูลสูญหายด้วยค่าถดถอย (Regression imputation) เป็นต้น

พิจารณาประชากร $\boldsymbol{U}=\{1,2,...,\boldsymbol{i},...,\boldsymbol{N}\}$ ขนาดจำกัด ขนาด $\boldsymbol{N}$ ตัวแปรที่สนใจศึกษาแทนด้วยตัวแปร $\boldsymbol{y}$ และพารามิเตอร์ที่ต้องการประมาณค่า คือ ผลรวมประชากร เขียนแทนด้วยสัญลักษณ์ $\boldsymbol{Y}=\sum_{\boldsymbol{i}\in\boldsymbol{U}}\boldsymbol{y_i}$ ซึ่ง $\boldsymbol{y_i}$ แทนค่าสังเกตของประชากรหน่วยที่ $\boldsymbol{i}$ เมื่อ $\boldsymbol{i}=1,2,...,\boldsymbol{N}$ ตัวอย่าง $\boldsymbol{s}$ ขนาด $\boldsymbol{n}$ ถูกสุ่มเลือกตามข้อมูลตกลงของแผนการสุ่มตัวอย่าง แทนด้วยสัญลักษณ์ $\boldsymbol{S}$ กรณีที่ไม่มีข้อมูลสูญหายการประมาณค่าผลรวมประชากร สามารถใช้ตัวประมาณค่าของ Horvitz and Thompson (1952) ตามที่แสดงในสมการที่ (1.1)

$$\hat{\boldsymbol{Y}}_{\boldsymbol{HT}}=\sum_{\boldsymbol{i}\in\boldsymbol{s}}\frac{\boldsymbol{y_i}}{\pi_{\boldsymbol{i}}} \qquad (1.1)$$

สำหรับค่าความแปรปรวนของ $\hat{Y}_{HT}$ เขียนแทนด้วย $V(\hat{Y}_{HT})$ ซึ่งกำหนดค่าตามที่แสดงในสมการที่ (1.2)

$$V(\hat{Y}_{HT}) = \sum_{i \in U}\left(\frac{1-\pi_i}{\pi_i}\right)y_i^2 + \sum_{i \in U}\sum_{j \neq i \in U}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)y_i y_j \qquad (1.2)$$

และตัวประมาณค่าของ $V(\hat{Y}_{HT})$ แทนด้วย $\hat{V}(\hat{Y}_{HT})$ ซึ่งกำหนดค่าตามที่แสดงในสมการ (1.3)

$$\hat{V}(\hat{Y}_{HT}) = \sum_{i \in s}\left(\frac{1-\pi_i}{\pi_i^2}\right)y_i^2 + \sum_{i \in s}\sum_{j \neq i \in s}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\right)y_i y_j \qquad (1.3)$$

เมื่อ π_i แทนความน่าจะเป็นที่หน่วยตัวอย่างที่ i ตกอยู่ในตัวอย่าง s ขนาด หรือ $\pi_i = P(i \in s)$ ส่วนค่า π_{ij} แทน ความน่าจะเป็นที่หน่วยตัวอย่างที่ i และ j ตกอยู่ในตัวอย่าง s ขนาด n หรือ $\pi_{ij} = P(i, j \in s)$ เมื่อ $i = 1, 2, ..., N$ และ $j = 1, 2, ..., N$ และได้ว่า $\hat{Y}_{HT}$ เป็นตัวประมาณค่าที่ไม่เอนเอียงของ Y เนื่องจาก $E_S(\hat{Y}_{HT}) = Y$

กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม และแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย Bethlehem (1988) กล่าวว่าการประมาณค่าผลรวมประชากรควรขึ้นอยู่กับเซตของข้อมูลที่ไม่สูญหายมากกว่าใช้เซตของตัวอย่างที่ถูกสุ่มมาทั้งหมด ดังนั้น Haziza (2010) ได้นำเสนอตัวประมาณค่าผลรวมประชากร ภายใต้การสุ่มตัวอย่างแบบง่ายและไม่ใส่คืน (Simple random sampling without replacement) ต่อมา ชูเกียรติ โพนแก้ว (2560) นำเสนอตัวประมาณค่าผลรวมประชากรภายใต้การสุ่มตัวอย่างโดยใช้ความน่าจะเป็นในการเลือกไม่เท่ากัน (Unequal probability sampling) และใช้แผนการสุ่มตัวอย่างของ Midzuno (1952) ในการสุ่มตัวอย่าง s ขนาด n

สำหรับการวิจัยในครั้งนี้ผู้วิจัยได้ขยายตัวประมาณค่าผลรวมประชากรที่ ชูเกียรติ โพนแก้ว (2560) นำเสนอ ไปยังแผนการเลือกตัวอย่างแบบชั้นภูมิ (Stratified sampling) และแต่ละชั้นภูมิสุ่มตัวอย่างโดยใช้แผนการสุ่มตัวอย่างของ Midzuno (Midzuno's scheme) จากนั้นนำเสนอตัวประมาณค่าผลรวมประชากรทั้งแบบจุด (Point estimation) และแบบช่วง (Interval estimation) สำหรับการประมาณค่าแบบช่วง ผู้วิจัยพิจารณาตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร ทั้งตัวประมาณค่าแบบเป็นเชิงเส้น และตัวประมาณค่าแบบไม่เป็นเชิงเส้น เกณฑ์การเปรียบเทียบประสิทธิภาพของการประมาณค่าแบบช่วง ผู้วิจัยพิจารณาจากค่าเฉลี่ยของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมพารามิเตอร์ ส่วนเกณฑ์การเปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร ผู้วิจัยเปรียบเทียบประสิทธิภาพจากค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ (Relative root mean square error)

1.2 วัตถุประสงค์

1.2.1 นำเสนอตัวประมาณค่าผลรวมประชากรแบบจุด ทั้งตัวประมาณค่าแบบเชิงเส้นและตัวประมาณค่าแบบไม่เป็นเชิงเส้น

1.2.2 นำเสนอตัวประมาณค่าความแปรปรวนแบบเป็นเชิงเส้น และตัวประมาณค่าความแปรปรวนแบบไม่เป็นเชิงเส้น

1.2.3 นำเสนอช่วงความเชื่อมั่นแบบเชิงเส้นและช่วงความเชื่อมั่นแบบไม่เป็นเชิงเส้น

1.2.4 เปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนแบบเชิงเส้น และตัวประมาณค่าความแปรปรวนแบบไม่เป็นเชิงเส้น โดยพิจารณาจากค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์

1.2.5 เปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นแบบเชิงเส้น และช่วงความเชื่อมั่นแบบไม่เป็นเชิงเส้น โดยพิจารณาจากร้อยละของค่าเฉลี่ยของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมพารามิเตอร์

1.3 สมมติฐานการวิจัย

1.3.1 ตัวประมาณค่าความแปรปรวนแบบไม่เป็นเชิงเส้นมีประสิทธิภาพดีกว่าตัวประมาณค่าความแปรปรวนแบบเชิงเส้น

1.3.2 ช่วงความเชื่อมั่นแบบไม่เป็นเชิงเส้น มีประสิทธิภาพดีกว่าช่วงความเชื่อมั่นแบบเชิงเส้น

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 เพื่อเป็นแนวทางในการเลือกวิธีการประมาณค่าข้อมูลสูญหาย ที่มีประสิทธิภาพและใกล้เคียงกับข้อมูลจริงมากที่สุด

1.4.2 เพื่อเป็นตัวช่วยในการวิเคราะห์ข้อมูลที่มีข้อมูลสูญหายให้สะดวก รวดเร็วและแม่นยำมากขึ้น

1.5 ขอบเขตของการวิจัย

1.5.1 แต่ละชั้นภูมิกำหนดให้ตัวอย่าง s_h ขนาด n_h เป็นตัวอย่างที่มีข้อมูลสูญหายเกิดขึ้นแบบสุ่มสมบูรณ์ (Missing completely at random) จากประชากร U_h ขนาด N_h ดังนั้น $p_{hi} = p_h$ ทุก $i = 1, 2, ..., N_h$ ค่า $E_R(r_{hi}) = p_h$ และ $V_R(r_{hi}) = p_h(1-p_h)$

1.5.2 แต่ละชั้นภูมิ กำหนดค่าความน่าจะเป็นที่ตอบสนองของหน่วยตัวอย่างที่ i เป็นอิสระกับหน่วยอื่น ๆ $p_{hij} = p(r_{hi} = 1, r_{hj} = 1) = p_{hi}p_{hj} = p_h p_h = p_h^2$

1.5.3 กำหนดให้การสูญหายของข้อมูลเกิดจากการไม่ตอบเฉพาะบางคำถาม และเกิดขึ้นเฉพาะตัวแปรที่สนใจศึกษา y

1.5.4 กำหนดให้ตัวแปรกำหนดขนาด (Size variable) k มีความสัมพันธ์กับตัวแปร y

1.5.5 ตัวแปร k ทราบค่าทุกค่า k_i เมื่อ $i \in U_h, h \in H$ และ $H = \{1, 2, ..., L\}$

1.6 นิยมศัพท์ที่เกี่ยวข้อง

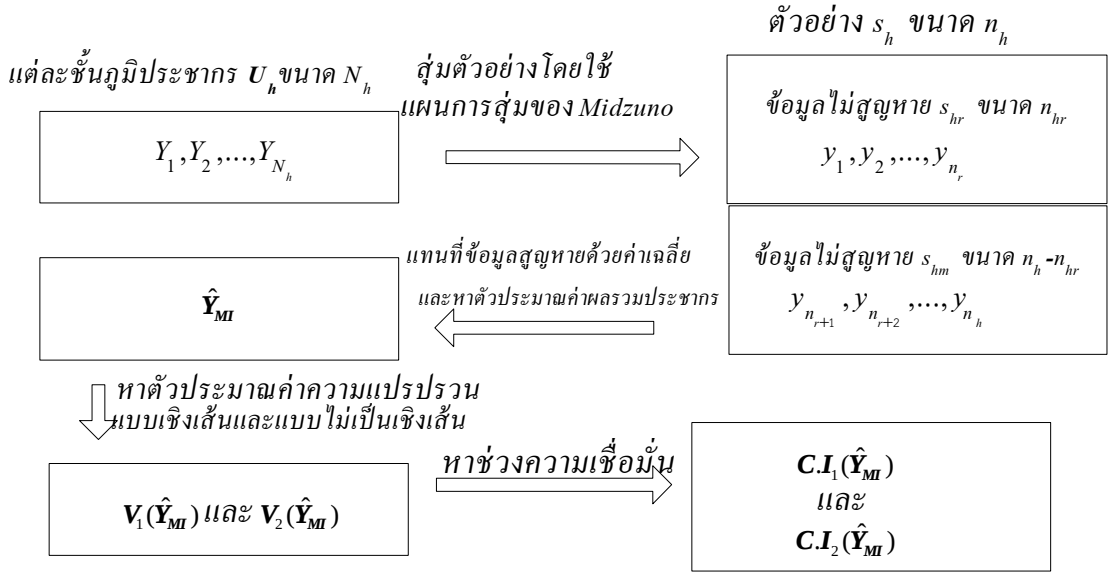
ตัวประมาณค่าความแปรปรวนแบบเชิงเส้น หมายถึง ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบเชิงเส้น

ตัวประมาณค่าความแปรปรวนแบบไม่เป็นเชิงเส้น หมายถึง ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น

ช่วงความเชื่อมั่นแบบเชิงเส้น หมายถึง ช่วงความเชื่อมั่นของผลรวมประชากรที่สร้างจากตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบเชิงเส้น

ช่วงความเชื่อมั่นแบบไม่เป็นเชิงเส้น หมายถึง ช่วงความเชื่อมั่นของผลรวมประชากรที่สร้างจากตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น

1.7 กรอบแนวคิดในการทำวิจัย



รูปที่ 1-1 กรอบแนวคิดในการวิจัย

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การศึกษาเกี่ยวกับการประมาณค่าผลรวมประชากร ในการสำรวจตัวอย่าง เมื่อมีข้อมูลสูญหาย ผู้วิจัยศึกษาทฤษฎีเอกสาร และงานวิจัยที่เกี่ยวข้องดังนี้

2.1 ข้อมูลสูญหาย

2.1.1 ความหมายของข้อมูลสูญหาย

ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และ สุคนธ์ ประสิทธิ์วัฒนเสรี (2549) กล่าวว่าข้อมูลสูญหาย คือค่าสังเกตที่ต้องการทราบค่าแต่ไม่สามารถทราบค่าได้ โดยที่ค่านั้นควรจะสามารถ ทราบค่าได้หากวิธีการที่ใช้ในการรวบรวมข้อมูลหรือในการวัดค่ามีประสิทธิภาพดีขึ้นหรือมีความเหมาะสมมากขึ้น พิจารณาจากตัวอย่างของการวัดระดับความดันเลือดของเด็กในชั้นเรียนหนึ่ง ซึ่งผลจากการตรวจวัดทำให้ทราบค่าความดันเลือดของเด็กทั้งหมดว่ามีค่าอยู่ในช่วง 90 – 180 mm Hg ยกเว้นในเด็กหนึ่งรายที่ป่วยจึงไม่ได้มาในวันที่ทำการตรวจวัดความดันเลือด ทำให้ไม่ทราบค่าระดับความดันเลือดของเด็กผู้นี้ ซึ่งเรียกได้ว่าเป็นค่าสูญหาย (Missing value) หากจะให้กล่าวถึงความสำคัญของข้อมูลสูญหายที่มีต่อการวิจัยควรต้องมุ่งไปยังประเด็นของผลกระทบต่องานวิจัยที่เกิดจากกรณีข้อมูล สูญหาย โดยข้อมูลสูญหายที่เกิดขึ้นอาจไม่ก่อผลกระทบใด ๆ ต่อผลการวิจัย หรืออาจก่อให้เกิดผลกระทบที่รุนแรงต่องานวิจัยก็ย่อมได้ ทั้งนี้สามารถจำแนกผลกระทบจากข้อมูลสูญหาย ได้ดังนี้

- ข้อมูลสูญหายสามารถทำให้เกิดการสูญเสียกำลังในการทดสอบ (Power of the test) เนื่องจากขนาดตัวอย่างที่ใช้ลดลงอันเป็นผลจากการตัดข้อมูลสูญหายออกจากการศึกษา ยกตัวอย่างเช่น ในตอนเริ่มต้นการศึกษา คำนวณขนาดตัวอย่างที่เหมาะสมในการศึกษาเท่ากับ 300 คน แต่ภายหลังจากการรวบรวมข้อมูลปรากฏว่า ผลจากข้อมูลสูญหายทำให้กลุ่มตัวอย่างที่มีข้อมูลสมบูรณ์เหลือเพียง 250 คน หากทำการวิเคราะห์ในกลุ่มตัวอย่างที่มีขนาดเล็กลงย่อมส่งผลต่อการสูญเสียระดับความเชื่อมั่น และการเพิ่มขึ้นของความแปรผันในการศึกษา

- ข้อมูลสูญหายอาจก่อให้เกิดความเอนเอียงของค่าประมาณ ยกตัวอย่างเช่น ในการศึกษาเกี่ยวกับลักษณะพฤติกรรมการดูแลสุขภาพ หากลักษณะข้อคำถามที่ใช้เป็นคำถามที่กระทบต่อความรู้สึกได้ง่าย (sensitive question) ไม่ว่าจะเป็นพฤติกรรมการเสพสารเสพติด หรือพฤติกรรมการมี เพศสัมพันธ์ หรือแม้แต่คำถามทั่ว ๆ ไปทางสังคม อย่างรายได้ต่อปี ซึ่งอาจไม่ได้รับความร่วมมือในการตอบคำถามจากหลาย ๆ คน ทำให้ข้อมูลที่ได้อาจไม่สามารถแทนลักษณะของประชากรได้ครบถ้วน ทั้งนี้เพราะผู้ที่ไม่ให้ความร่วมมือในการตอบคำถามเหล่านี้ ส่วนใหญ่อาจเป็นผู้ที่มีพฤติกรรมที่ไม่เป็นที่ยอมรับในสังคม หรือเป็นผู้ที่มีรายได้ต่ำมาก หรือในทางตรงกันข้ามอาจเป็นผู้ที่มีรายได้สูงมาก ทำให้ไม่ต้องการตอบคำถามเหล่านี้

- สืบเนื่องจากผลกระทบก่อนหน้านี้ ข้อมูลสูญหายก่อให้เกิดความยากลำบากในการตรวจสอบอิทธิพลของตัวแปรต่าง ๆ ยกตัวอย่างจากกรณีข้างต้น หากต้องการศึกษาความเกี่ยวข้องกันระหว่างรายได้กับพฤติกรรมการดูแลสุขภาพ หากเกิดปัญหาในลักษณะที่ผู้ที่มีรายได้น้อยหรือมากไม่ยอม

ให้คำตอบเกี่ยวกับรายได้ต่อปี จะทำให้ข้อมูลที่ใช้ในการวิเคราะห์มีเพียงกลุ่มผู้ที่มีรายได้ปานกลางหากผลการวิเคราะห์พบว่าไม่มีความเกี่ยวข้องกันระหว่างรายได้กับพฤติกรรมการดูแลสุขภาพผลสรุปที่ได้อาจผิดพลาด เพราะในความเป็นจริงอาจมีความสัมพันธ์กันก็เป็นไปได้ทั้งนี้จะเห็นได้ว่า ข้อมูลสูญหายสามารถส่งผลต่อการวิจัยทั้งในส่วนของการวิเคราะห์ และการสรุปผลตีความ โดยที่ระดับความรุนแรงของผลกระทบนี้ขึ้นอยู่กับองค์ประกอบจากหลายส่วน แต่ที่สำคัญก็คือ ขนาดของข้อมูลสูญหาย ประเภทของข้อมูลสูญหายที่เกิดขึ้น และวิธีการจัดการกับข้อมูลสูญหาย

2.1.2 ประเภทของข้อมูลสูญหาย

ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และ สุคนธ์ ประสิทธิ์วัฒนเสรี (2549) กล่าวว่า การพิจารณาประเภทของข้อมูลสูญหายเป็นขั้นตอนที่สำคัญ ทั้งนี้เพราะหากสามารถทราบถึงลักษณะของข้อมูลสูญหายจะช่วยในการพิจารณาแนวทางสำหรับจัดการกับปัญหาความไม่สมบูรณ์ของข้อมูลได้อย่างเหมาะสม ซึ่งโดยทั่วไปมักจำแนกข้อมูลสูญหายออกเป็น 3 ประเภทด้วยกันคือ

(1) Missing completely at random (MCAR)

MCAR เป็นลักษณะของข้อมูลสูญหายที่เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด นั่นคือข้อมูลที่สูญหายเป็นอิสระจากตัวแปรต่าง ๆ สามารถทำการตรวจสอบลักษณะของข้อมูลสูญหายกลุ่มนี้โดยการแบ่งกลุ่มของค่าสังเกตเป็นกลุ่มข้อมูลปกติและข้อมูลสูญหาย ในกรณีนี้เมื่อทำการทดสอบจะไม่พบความแตกต่างอย่างมีนัยสำคัญระหว่างทั้งสองกลุ่มสำหรับตัวแปรต่าง ๆ ในฐานข้อมูลสำหรับสาเหตุที่ทำให้ข้อมูลเกิดการสูญหายมีอยู่หลากหลายเหตุผล อาจเกิดขึ้นเนื่องจากเครื่องมือเสีย อุปกรณ์เกิดข้อบกพร่อง สภาพอากาศเลวร้าย กลุ่มเป้าหมายที่ศึกษาล้มป่วย หรือการนำเข้าข้อมูลไม่ถูกต้องสำหรับข้อมูลสูญหายประเภทนี้จัดเป็นข้อมูลที่ก่อให้เกิดปัญหาน้อยที่สุด เพราะว่าข้อมูลสูญหายไม่มีความเกี่ยวข้องต่อผลลัพธ์ของข้อมูล เพราะฉะนั้นสามารถเลือกทำการวิเคราะห์ข้อมูลในส่วนที่สมบูรณ์ได้

(2) Missing at random (MAR)

MAR เป็นลักษณะของข้อมูลสูญหายซึ่งไม่ได้เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด แต่เกิดขึ้นอย่างสุ่มภายในบางส่วนหรือบางกลุ่มของค่าสังเกต นั่นคือค่าของข้อมูลสูญหายขึ้นอยู่กับตัวแปรตัวอื่น ๆ ในฐานข้อมูลซึ่งไม่ได้เป็นตัวแปรที่เกิดข้อมูลสูญหาย ยกตัวอย่างเช่น หากพบว่าเฉพาะกลุ่มผู้ได้รับการศึกษาน้อยที่ไม่ให้ความร่วมมือในการตอบข้อคำถามเกี่ยวกับทัศนคติในการเสพยาเสพติด ในลักษณะนี้สามารถกล่าวได้ว่าข้อมูลทัศนคติในการเสพยาเสพติดมีค่าสูญหายแบบ MAR ทั้งนี้เนื่องจากเป็นค่าสูญหายที่เกิดขึ้นเฉพาะในบางส่วนของตัวแปรระดับการศึกษาสำหรับข้อมูลสูญหายประเภทนี้ยังไม่ส่งผลกระทบรุนแรงเท่ากับข้อมูลสูญหายในประเภทสุดท้าย

(3) Not missing at random (NMAR)

NMAR เป็นลักษณะของข้อมูลสูญหายซึ่งไม่ได้เกิดขึ้นอย่างสุ่ม โดยค่าของข้อมูลสูญหายขึ้นอยู่กับค่าของข้อมูลสมบูรณ์ในตัวแปรเดียวกัน รวมถึงตัวแปรตัวอื่นด้วย เช่น หากข้อมูลสูญหายของระดับรายได้ขึ้นอยู่ระดับรายได้ในแต่ละช่วงอายุ ข้อมูลสูญหายที่เกิดขึ้นจัดอยู่ในประเภท NMAR หรือในบางกรณีค่าของข้อมูลสูญหายอาจไม่ขึ้นอยู่กับตัวแปรใด ๆ ในฐานข้อมูลเลย แต่ขึ้นอยู่กับตัวแปรอื่นที่ไม่ได้ถูกเก็บรวบรวมไว้ในการศึกษาครั้งนั้น เช่น ค่าน้ำหนักตัวที่ลดลงขึ้นอยู่กับน้ำหนักตัวตอนเริ่มต้น แต่เนื่องจากตัวแปรน้ำหนักตอนเริ่มต้นไม่ได้ถูกรวบรวมไว้ในฐานข้อมูล ดังนั้นค่าสูญหายของน้ำหนักตัวที่

ลดลงจึงขึ้นอยู่กับตัวแปรภายนอกฐานข้อมูลลักษณะข้อมูลสูญหายประเภทนี้จัดเป็นข้อมูลสูญหายที่สามารถส่งผลกระทบอย่างรุนแรงในการวิเคราะห์ข้อมูลในทางปฏิบัติ

2.2 ตัวประมาณค่าผลรวมประชากร

2.2.1 ตัวประมาณค่าแบบจุด

ปัญหาเกี่ยวกับการประมาณค่าผลรวมประชากร เป็นหนึ่งในปัญหาที่สำคัญในการศึกษาวิชาทางสถิติ การประมาณค่าผลรวมประชากรสามารถทำได้โดยใช้ข้อมูลของตัวอย่างที่ถูกสุ่มจากประชากรตามข้อตกลงของแผนการสุ่มตัวอย่าง (Sampling design) พิจารณาประชากร $\boldsymbol{U}=\{1,2,...,\boldsymbol{i},...,\boldsymbol{N}\}$ ขนาดจำกัด ขนาด $\boldsymbol{N}$ ตัวแปรที่สนใจศึกษาแทนด้วยตัวแปร $\boldsymbol{y}$ และพารามิเตอร์ที่ต้องการประมาณค่าคือผลรวมประชากร เขียนแทนด้วยสัญลักษณ์ $\boldsymbol{Y}=\sum_{\boldsymbol{i}\in\boldsymbol{U}}\boldsymbol{y}_{\boldsymbol{i}}$ ซึ่ง $\boldsymbol{y}_{\boldsymbol{i}}$ แทนค่าสังเกตของประชากรหน่วยที่ $\boldsymbol{i}$ เมื่อ $\boldsymbol{i}=1,2,...,\boldsymbol{N}$ ตัวอย่าง $\boldsymbol{s}$ ขนาด $\boldsymbol{n}$ ถูกสุ่มเลือกตามข้อมูลตกลงของแผนการสุ่มตัวอย่าง แทนด้วยสัญลักษณ์ $\boldsymbol{S}$ กรณีที่ไม่มีข้อมูลสูญหายการประมาณค่าผลรวมประชากร สามารถใช้ตัวประมาณค่าของ Horvitz and Thompson (1952) ตามที่แสดงในสมการที่ (2.1)

$$\hat{\boldsymbol{Y}}_{\boldsymbol{HT}}=\sum_{\boldsymbol{i}\in\boldsymbol{s}}\frac{\boldsymbol{y}_{\boldsymbol{i}}}{\pi_{\boldsymbol{i}}} \tag{2.1}$$

สำหรับค่าความแปรปรวนของ $\hat{\boldsymbol{Y}}_{\boldsymbol{HT}}$ เขียนแทนด้วย $\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{HT}})$ และกำหนดค่าตามที่แสดงในสมการที่ (2.2)

$$\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{HT}})=\sum_{\boldsymbol{i}\in\boldsymbol{U}}\left(\frac{1-\pi_{\boldsymbol{i}}}{\pi_{\boldsymbol{i}}}\right)\boldsymbol{y}_{\boldsymbol{i}}^{2}+\sum_{\boldsymbol{i}\in\boldsymbol{U}}\sum_{\boldsymbol{j}\neq\boldsymbol{i}\in\boldsymbol{U}}\left(\frac{\pi_{\boldsymbol{ij}}-\pi_{\boldsymbol{i}}\pi_{\boldsymbol{j}}}{\pi_{\boldsymbol{i}}\pi_{\boldsymbol{j}}}\right)\boldsymbol{y}_{\boldsymbol{i}}\boldsymbol{y}_{\boldsymbol{j}} \tag{2.2}$$

และตัวประมาณค่าของ $\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{HT}})$ แทนด้วย $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{HT}})$ และกำหนดค่าตามที่แสดงในสมการที่ (2.3)

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{HT}})=\sum_{\boldsymbol{i}\in\boldsymbol{s}}\left(\frac{1-\pi_{\boldsymbol{i}}}{\pi_{\boldsymbol{i}}^{2}}\right)\boldsymbol{y}_{\boldsymbol{i}}^{2}+\sum_{\boldsymbol{i}\in\boldsymbol{s}}\sum_{\boldsymbol{j}\neq\boldsymbol{i}\in\boldsymbol{s}}\left(\frac{\pi_{\boldsymbol{ij}}-\pi_{\boldsymbol{i}}\pi_{\boldsymbol{j}}}{\pi_{\boldsymbol{ij}}\pi_{\boldsymbol{i}}\pi_{\boldsymbol{j}}}\right)\boldsymbol{y}_{\boldsymbol{i}}\boldsymbol{y}_{\boldsymbol{j}} \tag{2.3}$$

เมื่อ $\pi_{\boldsymbol{i}}$ แทนความน่าจะเป็นที่หน่วยตัวอย่างที่ $\boldsymbol{i}$ ตกอยู่ในตัวอย่าง $\boldsymbol{s}$ ขนาด $\boldsymbol{n}$ หรือ $\pi_{\boldsymbol{i}}=\boldsymbol{P}(\boldsymbol{i}\in\boldsymbol{s})$ ส่วนค่า $\pi_{\boldsymbol{ij}}$ แทน ความน่าจะเป็นที่หน่วยตัวอย่างที่ $\boldsymbol{i}$ และ $\boldsymbol{j}$ ตกอยู่ในตัวอย่าง $\boldsymbol{s}$ ขนาด $\boldsymbol{n}$ หรือ $\pi_{\boldsymbol{ij}}=\boldsymbol{P}(\boldsymbol{i}\wedge\boldsymbol{j}\in\boldsymbol{s})$ เมื่อ $\boldsymbol{i}=1,2,...,\boldsymbol{N}$ และ $\boldsymbol{j}=1,2,...,\boldsymbol{N}$ และได้ว่า $\hat{\boldsymbol{Y}}_{\boldsymbol{HT}}$ เป็นตัวประมาณค่าที่ไม่เอนเอียงของ $\boldsymbol{Y}$ เนื่องจาก $\boldsymbol{E}_{\boldsymbol{S}}(\hat{\boldsymbol{Y}}_{\boldsymbol{HT}})=\boldsymbol{Y}$

บทตั้ง 2.1 *ภายใต้การสุ่มแบบง่าย* $\pi_{\boldsymbol{i}}=\frac{\boldsymbol{n}}{\boldsymbol{N}}$ *และ* $\pi_{\boldsymbol{ij}}=\frac{\boldsymbol{n}}{\boldsymbol{N}}\frac{\boldsymbol{n}-1}{\boldsymbol{N}-1}$

ทฤษฎีบท 2.1 *ภายใต้การสุ่มแบบง่าย*

1) $\hat{\boldsymbol{Y}}_{\boldsymbol{HT}}=\sum_{\boldsymbol{i}\in\boldsymbol{s}}\frac{\boldsymbol{y}_{\boldsymbol{i}}}{\pi_{\boldsymbol{i}}}=\frac{\boldsymbol{N}}{\boldsymbol{n}}\sum_{\boldsymbol{i}\in\boldsymbol{s}}\boldsymbol{y}_{\boldsymbol{i}}=\boldsymbol{N}\bar{\boldsymbol{y}}$

2) $\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{HT}})=\frac{\boldsymbol{N}^{2}}{\boldsymbol{n}}\left(1-\frac{\boldsymbol{n}}{\boldsymbol{N}}\right)\boldsymbol{S}^{2}$ *เมื่อ* $\boldsymbol{S}^{2}=\frac{1}{\boldsymbol{N}-1}\sum_{\boldsymbol{i}\in\boldsymbol{U}}(\boldsymbol{y}_{\boldsymbol{i}}-\bar{\boldsymbol{Y}})^{2}$

3) $\hat{V}(\hat{Y}_{HT}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\hat{S}^2$ *เมื่อ* $\hat{S}^2 = \frac{1}{n-1}\sum_{i\in s}(y_i-\bar{y})^2$ เมื่อ $\bar{y} = \frac{1}{n}\sum_{i\in s} y_i$

ตัวอย่าง 2.1 จากการสำรวจตัวอย่างสุ่มแบบง่ายขนาด 10 ครัวเรือน จากประชากรขนาด 100 สอบถามจำนวนคนในแต่ละครัวเรือน ได้ดังนี้

i	2	15	25	30	34	55	60	76	89	93
y_i	4	3	2	4	6	5	2	3	5	7

ที่มา: สุชาดา กีระนันท์ (2542)

จากข้างต้นจะได้ว่า

$$\sum_{i\in s} y_i = 4+3+\ldots+7 = 41$$

$$\bar{y} = n^{-1}\sum_{i\in s} y_i = 4.1$$

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i\in s}(y_i-\bar{y})^2 = \frac{1}{9}\left[(4-4.1)^2+\ldots+(7-4.1)^2\right] = 2.77$$

ดังนั้นตัวประมาณค่าผลรวมของจำนวนคนในแต่ครัวเรือน เมื่อหาตาม ทฤษฎีบท 2.1 ข้อ 1) ได้ว่า

$$\hat{Y}_{HT} = N\bar{y} = 4.1(100) = 410$$

ส่วนตัวประมาณค่าความแปรปรวนของ $\hat{Y}_{HT}$ เมื่อหาตาม ทฤษฎีบท 2.1 ข้อ 3) ได้ว่า

$$\hat{V}(\hat{Y}_{HT}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\hat{S}^2 = \frac{100^2}{10}\left(1-\frac{10}{100}\right)2.77 = 2493$$ ■

จากตัวอย่างที่ 2.1 แสดงวิธีการประมาณค่าผลรวมประชากร และการหาตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร ในกรณีที่มีข้อมูลสูญหาย ส่วนกรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม การประมาณค่าผลรวมประชากร โดยใช้ตัวประมาณค่าของ Horvitz and Thompson (1952) ไม่สามารถทำได้เนื่องจากข้อมูลของตัวอย่าง s ขนาด n บางหน่วยตัวอย่างสูญหาย กำหนดให้ z แทน เซตของตัวอย่างที่มีข้อมูลสูญหาย

$$z = \{y_1, y_2, \ldots, y_{n_r}, y_{n_{r+1}}, \ldots, y_n\} \qquad (2.4)$$

ค่าของ y_i เมื่อ $i = 1,2,\ldots,n_r$ สามารถสังเกตุค่าได้ ส่วน $i = n_{r+1}, n_{r+2}, \ldots, n$ ค่า y_i ไม่สามารถสังเกตค่าได้ (ข้อมูลสูญหาย) จากสมการ (2.4) เซตของตัวอย่าง z ขนาด n สามารถแบ่งได้เป็น เซตของตัวอย่างที่ไม่สูญหาย z_r ขนาด n_r และเซตของตัวอย่างที่สูญหาย z_r' ขนาด $n-n_r$ ซึ่ง

$$z = z_r \cup z_r' \qquad (2.5)$$

เมื่อ $\boldsymbol{z_r} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{\boldsymbol{n_r}}\}$ และ $\boldsymbol{z'_r} = \{\boldsymbol{y}_{\boldsymbol{n_{r+1}}}, \boldsymbol{y}_{\boldsymbol{n_{r+2}}}, ..., \boldsymbol{y_n}\}$ ดังนั้นการประมาณค่าผลรวมประชากร กรณีที่มีข้อมูลสูญหาย สามารถทำได้โดยใช้เซตของตัวอย่างที่ไม่สูญหาย $\boldsymbol{z_r} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{\boldsymbol{n_r}}\}$ แทนค่าในสมการ (2.1) กำหนดให้ $\hat{\boldsymbol{Y}}_{\boldsymbol{z_r}\pi}$ แทนตัวประมาณค่าผลรวมประชากร กรณีที่มีข้อมูลสูญหาย และกำหนดค่าตามที่แสดงในสมการ (2.6)

$$\hat{\boldsymbol{Y}}_{\boldsymbol{z_r}\pi} = \sum\nolimits_{\boldsymbol{z_r}} \frac{\boldsymbol{y_i}}{\boldsymbol{\pi_i}} \tag{2.6}$$

เมื่อ $\boldsymbol{z_r}$ กำหนดค่าตามสมการ (2.5) สำหรับความแปรปรวน และตัวประมาณค่าความแปรปรวนของ $\hat{\boldsymbol{Y}}_{\boldsymbol{z_r}\pi}$ สามารถหาได้โดยการแทนค่า $\boldsymbol{s}$ ด้วย $\boldsymbol{z_r}$ ในสมการที่ (2.2) และ (2.3) ตามลำดับ

ตัวประมาณค่าผลรวมประชากรตามที่แสดงในสมการ (2.6) เป็นตัวประมาณค่าที่ได้จากการตัดเซตข้อมูลที่สูญหายออกไป และนำเพียงข้อมูลที่ไม่สูญหายมาวิเคราะห์ ซึ่งวิธีการดังกล่าวเป็นวิธีการที่ง่าย แต่มีข้อเสียคือ เนื่องจากมีการตัดเซตข้อมูลที่สูญหายออกไป ทำให้ขนาดตัวอย่างลดลง ส่งผลให้อำนาจการทดสอบเชิงสถิติลดลง (ทัตตา หิรัญพต, 2554) นอกจากวิธีการตัดเซตข้อมูลที่สูญหายดังที่กล่าวข้างต้น นักสถิตินิยมใช้วิธีการแทนที่ข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ เช่นการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย ซึ่งมีขั้นตอนดังนี้

(1) หาค่าเฉลี่ยของเซตตัวอย่างที่ไม่สูญหาย เขียนแทนด้วย $\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{z_r}\pi}$ และกำหนดค่าตามที่แสดงในสมการ (2.7)

$$\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{z_r}\pi} = \sum_{\boldsymbol{i}\in\boldsymbol{s}} \frac{\boldsymbol{r_i}\boldsymbol{y_i}}{\boldsymbol{\pi_i}} \Big/ \sum_{\boldsymbol{i}\in\boldsymbol{s}} \frac{\boldsymbol{r_i}}{\boldsymbol{\pi_i}} \tag{2.7}$$

ค่า $\boldsymbol{r_i} = 1$ เมื่อ $\boldsymbol{y_i}$ ไม่สูญหาย และ $\boldsymbol{r_i} = 0$ เมื่อ $\boldsymbol{y_i}$ สูญหาย

(2) แทนที่ข้อมูลสูญหาย ในเซต $\boldsymbol{z'_r} = \{\boldsymbol{y}_{\boldsymbol{n_{r+1}}}, \boldsymbol{y}_{\boldsymbol{n_{r+2}}}, ..., \boldsymbol{y_n}\}$ ด้วย $\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{z_r}\pi}$ กำหนดให้ $\boldsymbol{z}^*$ แทนเซตของตัวอย่าง $\boldsymbol{s}$ ขนาด $\boldsymbol{n}$ หลังจากที่แทนที่ข้อมูลที่สูญหาย ในเซต $\boldsymbol{z'_r} = \{\boldsymbol{y}_{\boldsymbol{n_{r+1}}}, \boldsymbol{y}_{\boldsymbol{n_{r+2}}}, ..., \boldsymbol{y_n}\}$ ด้วย $\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{z_r}\pi}$ ดังนั้น

$$\boldsymbol{z}^* = \{\boldsymbol{y}_1^*, \boldsymbol{y}_2^*, ..., \boldsymbol{y}_{\boldsymbol{n}}^*\} \tag{2.8}$$

โดยที่ $\boldsymbol{y}_{\boldsymbol{i}}^* = \boldsymbol{y_i}$ เมื่อหน่วยตัวอย่างที่ $\boldsymbol{i}$ ไม่สูญหาย และ $\boldsymbol{y}_{\boldsymbol{i}}^* = \hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{z_r}\pi}$ เมื่อหน่วยตัวอย่างที่ $\boldsymbol{i}$ สูญหาย สำหรับตัวประมาณค่าผลรวมประชากร หลังจากที่แทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย เขียนแทนด้วยสัญลักษณ์ $\hat{\boldsymbol{Y}}_{\boldsymbol{z}^*\pi}$ กำหนดค่าตามที่แสดงในสมการ (2.9)

$$\hat{\boldsymbol{Y}}_{\boldsymbol{z}^*\pi} = \sum_{\boldsymbol{i}\in\boldsymbol{s}} \frac{\boldsymbol{y}_{\boldsymbol{i}}^*}{\boldsymbol{\pi_i}} \tag{2.9}$$

เมื่อ $\boldsymbol{y}_{\boldsymbol{i}}^*$ กำหนดค่าตามสมการ (2.8) สำหรับความแปรปรวน และตัวประมาณค่าความแปรปรวนของ $\hat{\boldsymbol{Y}}_{\boldsymbol{z}^*\pi}$ สามารถหาได้โดยการแทนค่า $\boldsymbol{y_i}$ ด้วย $\boldsymbol{y}_{\boldsymbol{i}}^*$ ในสมการที่ (2.2) และ (2.3) ตามลำดับ

ตัวอย่าง 2.2 จากตัวอย่าง 2.1 สมมติว่าหน่วยตัวอย่างที่ 15, 30, 76 และ 89 ไม่สามารถสังเกตค่าได้

ตามที่แสดงในตาราง

i	2	15	25	30	34	55	60	76	89	93
y_i	4	*	2	*	6	5	2	*	*	7

เมื่อ * แทนข้อมูลสูญหาย จากข้อมูลข้างต้นได้ว่า

$\boldsymbol{z}=\{4,2,6,5,2,7,*,*,*,*\}$

$\boldsymbol{z}_r=\{4,2,6,5,2,7\}$

$\boldsymbol{z}'_r=\{*,*,*,*\}$

(1) ใช้วิธีตัดข้อมูลสูญหาย

$$\bar{\boldsymbol{y}}=\frac{1}{6}(4+2+6+5+2+7)=4.33$$

$$\hat{\boldsymbol{S}}^2=\frac{1}{\boldsymbol{n}-1}\sum_{\boldsymbol{i}\in\boldsymbol{s}}(\boldsymbol{y}_i-\bar{\boldsymbol{y}})^2=\frac{1}{5}\left[(4-4.1)^2+\ldots+(7-4.1)^2\right]=4.27$$

ตัวประมาณค่าผลรวมประชากร เมื่อหาตามสูตรที่แสดงในสมการ (2.6) ได้ว่า

$$\hat{\boldsymbol{Y}}_{\boldsymbol{z}_r\pi}=\sum_{\boldsymbol{z}_r}\frac{\boldsymbol{y}_i}{\pi_i}=\frac{\boldsymbol{N}}{\boldsymbol{n}}\sum_{\boldsymbol{z}_r}\boldsymbol{y}_i$$

$$=\frac{100}{6}(4+2+6+5+2+7)=433.33$$

ส่วนตัวประมาณค่าความแปรปรวนของ $\hat{\boldsymbol{Y}}_{\boldsymbol{HT}}$ เมื่อหาตาม ทฤษฎีบท 2.1 ข้อ 3) ได้ว่า

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{HT}})=\frac{\boldsymbol{N}^2}{\boldsymbol{n}}\left(1-\frac{\boldsymbol{n}}{\boldsymbol{N}}\right)\hat{\boldsymbol{S}}^2=\frac{100^2}{6}\left(1-\frac{6}{100}\right)4.27=6689.67$$

(2) ใช้วิธีแทนที่ข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ

กรณีที่สุ่มตัวอย่างแบบง่าย $\pi_i=\frac{\boldsymbol{n}}{\boldsymbol{N}}$ ดังนั้นค่าเฉลี่ยของตัวอย่างที่ไม่สูญหายมีค่าเท่ากับ

$$\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{z}_r\pi}=\frac{\sum_{\boldsymbol{i}\in\boldsymbol{s}}\boldsymbol{r}_i\boldsymbol{y}_i/\pi_i}{\sum_{\boldsymbol{i}\in\boldsymbol{s}}\boldsymbol{r}_i/\pi_i}=\frac{\sum_{\boldsymbol{i}\in\boldsymbol{s}}\boldsymbol{r}_i\boldsymbol{y}_i}{\sum_{\boldsymbol{i}\in\boldsymbol{s}}\boldsymbol{r}_i}=\frac{1}{\boldsymbol{n}_r}\sum_{\boldsymbol{z}_r}\boldsymbol{y}_i$$

$$=\frac{1}{6}(4+2+6+5+2+7)=4.33$$

ดังนั้น

$$\boldsymbol{z}^*=\{4,2,6,5,2,7,4.33,4.33,4.33,4.33\}$$

$$\hat{\boldsymbol{Y}}_{\boldsymbol{z}^*\pi}=\sum_{\boldsymbol{i}\in\boldsymbol{s}}\frac{\boldsymbol{y}_i^*}{\pi_i}=\frac{100}{10}(4+2+\ldots+4.33)=433.20$$

$$\bar{\boldsymbol{y}}=\frac{1}{10}(4+2+\ldots+4.33)=4.33$$

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y})^2 = \frac{1}{9}\left[(4-4.33)^2 + \ldots + (4.33-4.33)^2\right] = 2.37$$

ส่วนตัวประมาณค่าความแปรปรวนของ $\hat{Y}_{z^*\pi}$ เมื่อหาตาม ทฤษฎีบท 2.1 ข้อ 3) ได้ว่า

$$\hat{V}(\hat{Y}_{z^*\pi}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\hat{S}^2 = \frac{100^2}{10}\left(1-\frac{10}{100}\right)2.37 = 2133$$

■

2.2.2 ตัวประมาณค่าแบบช่วง

หัวข้อที่แล้วผู้วิจัยได้กล่าวถึงตัวประมาณค่าแบบจุด ของผลรวมประชากร ซึ่งตัวประมาณค่า(ตัวสถิติ)ที่ได้อาจมีค่าเท่า หรือไม่เท่ากับ ผลรวมประชากร(พารามิเตอร์) หัวข้อนี้ผู้วิจัยจะได้กล่าวถึงตัวประมาณค่าแบบช่วง เพื่อใช้ในการประมาณค่าผลรวมประชากร เริ่มต้นผู้เขียนจะได้กล่าวถึง ทฤษฎีขีดจำกัดค่ากลาง (The central limit theorem) ดังนี้

ทฤษฎีบท 2.2 *กำหนดให้* $y_1, y_2, y_3, \ldots, y_n$ *เป็นตัวอย่างสุ่ม* s *ขนาด* n *ที่ถูกสุ่มจากประชากร* U *ที่มีการแจกแจงแบบใดๆ เมื่อ* n *มีขนาดใหญ่* $(n \geq 30)$ *ได้ว่า*

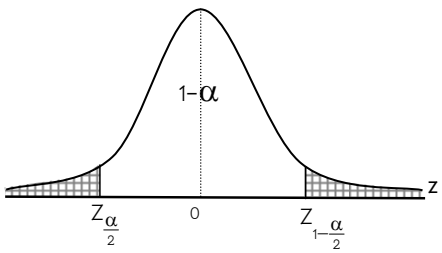
1) $Z = \dfrac{\hat{Y}_{HT} - Y}{\sqrt{V(\hat{Y}_{HT})}} \sim N(0,1)$

2) กรณีที่ไม่ทราบค่า $V(\hat{Y}_{HT})$ สามารถประมาณค่าด้วย $\hat{V}(\hat{Y}_{HT})$ และได้ว่า

$$Z = \frac{\hat{Y}_{HT} - Y}{\sqrt{\hat{V}(\hat{Y}_{HT})}} \sim appN(0,1)$$

เมื่อ $\hat{Y}_{HT}$ แทนตัวประมาณค่าผลรวมประชากร ส่วน $V(\hat{Y}_{HT})$ และ $\hat{V}(\hat{Y}_{HT})$ แทนความแปรปรวน และตัวประมาณค่าความแปรปรวนของ $\hat{Y}_{HT}$ ตามลำดับ

จาก ทฤษฎี 2.2 ข้อ 2) ได้ว่า $Z = \dfrac{\hat{Y}_{HT} - Y}{\sqrt{\hat{V}(\hat{Y}_{HT})}} \sim appN(0,1)$



$$P(Z_{\frac{\alpha}{2}} < Z < Z_{1-\frac{\alpha}{2}}) = 1-\alpha$$

$$P(Z_{\frac{\alpha}{2}} < \frac{\hat{Y}_{HT} - Y}{\sqrt{\hat{V}(\hat{Y}_{HT})}} < Z_{1-\frac{\alpha}{2}}) = 1-\alpha$$

$$P(-Z_{1-\frac{\alpha}{2}} < \frac{\hat{Y}_{HT} - Y}{\sqrt{\hat{V}(\hat{Y}_{HT})}} < Z_{1-\frac{\alpha}{2}}) = 1-\alpha$$

$$P(-Z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{Y}_{HT})} < \hat{Y}_{HT} - Y < Z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{Y}_{HT})}) = 1-\alpha$$

$$P(\hat{Y}_{HT} - Z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{Y}_{HT})} < Y < \hat{Y}_{HT} + Z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{Y}_{HT})}) = 1-\alpha$$

ดังนั้นช่วงความเชื่อมั่น $(1-\alpha)100\%$ ของ Y มีค่าเป็น

$$\hat{Y}_{HT} - Z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{Y}_{HT})} < Y < \hat{Y}_{HT} + Z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{Y}_{HT})} \quad \blacksquare$$

ตัวอย่าง 2.3 จากตัวอย่าง SRS ขนาด $n=30$ จากเขตหนึ่งซึ่งมีจำนวนครัวเรือนทั้งหมด $N=14,800$ ได้จำนวนคนในแต่ละครัวเรือน ในตัวอย่างดังนี้

5	6	3	3	2	3	3	3	4	4
3	2	7	4	3	5	4	4	3	3
4	3	3	1	2	4	3	4	2	4

จากข้อมูลข้างต้น สามารถประมาณจำนวนคนทั้งหมด ตัวประมาณค่าความแปรปรวน และช่วงความเชื่อมั่นของจำนวนคนทั้งหมด ที่ช่วงความเชื่อมั่น 95% (สุชาดา กีระนันท์, 2542)
จากข้อมูลข้างต้นได้ว่า

$$\sum_{i\in s} y_i = 5+6+...+4 = 104$$

$$\bar{y} = \frac{1}{n}\sum_{i\in s} y_i = \frac{104}{30} = 3.47$$

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y})^2 = \frac{1}{29}\left[(5-3.47)^2 + ... + (4-3.47)^2\right] = 1.50$$

ดังนั้นตัวประมาณค่าผลรวมประชากร หรือ $\hat{Y}_{HT}$ มีค่าเป็น

$$\hat{Y}_{HT} = \sum_{i\in s}\frac{y_i}{\pi_i} = \frac{N}{n}\sum_{i\in s} y_i$$
$$= \frac{14800}{30}(104) = 51306.67$$

ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร $\hat{V}(\hat{Y}_{HT})$ มีค่าเท่ากับ

$$\hat{V}(\hat{Y}_{HT}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\hat{S}^2 = \frac{14800^2}{30}\left(1-\frac{30}{14800}\right)1.50 = 10,929,800$$

ช่วงความเชื่อมั่น 95% ของ ผลรวมประชากร มีค่าเท่ากับ

$$\hat{Y}_{HT} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{Y}_{HT})} = 51306.67 \pm 1.96\sqrt{10929800}$$

$$= 51306.67 \pm 6479.81$$
$$= (51306.67 - 6479.81, 51306.67 + 6479.81)$$
$$= (44826.86\ ,\ 57786.48)$$

ดังนั้นที่ช่วงความเชื่อมั่น 95% จำนวนคนทั้งหมดในเขตดังกล่าวมีค่าอยู่ระหว่าง 44826.86 คนถึง 57786.48 คน

■

2.3 การหาความแปรปรวนของตัวประมาณค่าแบบไม่เป็นเชิงเส้น

หัวข้อที่แล้วผู้วิจัยได้กล่าวถึงตัวประมาณค่าผลรวมประชากร ทั้งกรณีที่ไม่มีข้อมูลสูญหาย และมีข้อมูลสูญหาย ซึ่งตัวประมาณค่าที่นำกล่าวถึงจะถูกพิจารณาในรูปของตัวประมาณค่าแบบเชิงเส้น ดังนั้นการหาค่าความแปรปรวนสามารถทำได้โดยใช้สูตรของ Horvitz and Thompson (1952) ตามที่แสดงในสมการ (2.2) แต่ในบางครั้งตัวประมาณค่าผลรวมประชากรจะอยู่ในรูปของตัวประมาณค่าแบบไม่เป็นเชิงเส้น เช่นตัวประมาณค่าของ Hajek (1981) ซึ่งกำหนดค่าตามที่แสดงในสมการ (2.10)

$$\hat{\boldsymbol{Y}}_{\boldsymbol{Hajek}} = \boldsymbol{N}\frac{\sum_{\boldsymbol{i}\in\boldsymbol{s}}\frac{\boldsymbol{y_i}}{\pi_{\boldsymbol{i}}}}{\sum_{\boldsymbol{i}\in\boldsymbol{s}}\frac{1}{\pi_{\boldsymbol{i}}}} \qquad (2.10)$$

Särndal, Swenson, and Wretman (1992) กล่าวว่าในบางกรณี ตัวประมาณค่าของ Hajek (1971) จะเป็นตัวประมาณค่าที่ดีกว่าตัวประมาณค่าของ Horvitz and Thompson (1952) จากสมการ (2.10) จะเห็นว่าตัวประมาณค่าของ Hajek (1971) อยู่ในรูปของอัตราส่วน ดังนั้นตัวประมาณค่าของ Hajek (1971) เป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้น การหาค่าความแปรปรวนของตัวประมาณค่าแบบไม่เป็นเชิงเส้นมีหลายวิธีเช่น วิธีบูตสแตรป (Bootstrap Method) วิธีแจ็คไนฟ์ (Jackknife Method) หรือวิธีเทเลอร์แบบเชิงเส้น (Taylor linearization method) สำหรับงานวิจัยในครั้งนี้ผู้วิจัยใช้ วิธีเทเลอร์แบบเชิงเส้นซึ่งมีขั้นตอนดังนี้

(1) ปรับตัวประมาณค่าแบบไม่เป็นเชิงเส้น ให้เป็นตัวประมาณค่าแบบเชิงเส้น โดยใช้วิธีเทเลอร์แบบเชิงเส้น

กำหนดให้ $\hat{\boldsymbol{Y}}_{\boldsymbol{NL}}$ เป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้น และสามารถเขียนให้อยู่ในรูปของฟังก์ชันปรับเรียบได้ตามที่แสดงในสมการที่ (2.11)

$$\hat{\boldsymbol{Y}}_{\boldsymbol{NL}} = \psi(\hat{\mathrm{t}}) \qquad (2.11)$$

เมื่อ ψ แทนฟังก์ชันปรับเรียบ $\hat{\mathrm{t}} = \left[\hat{\boldsymbol{t}}_1\ \hat{\boldsymbol{t}}_2\ \hat{\boldsymbol{t}}_3\ \cdots\ \hat{\boldsymbol{t}}_{\boldsymbol{k}}\right]'$ แทนเวกเตอร์ของตัวแปรสุ่มขนาด $\boldsymbol{k}$ และ $\hat{\boldsymbol{t}}_{\boldsymbol{i}}$ เมื่อ $\boldsymbol{i} = 1,2,...,\boldsymbol{k}$ แทนฟังก์ชันของตัวแปรสุ่มที่อยู่ในรูปของตัวประมาณค่าแบบ Horvitz and Thompson (1952) กำหนดให้ $\mathrm{t} = \boldsymbol{E}(\hat{\mathrm{t}}) = \left[\boldsymbol{t}_1\ \boldsymbol{t}_2\ \boldsymbol{t}_3\ \cdots\ \boldsymbol{t}_{\boldsymbol{k}}\right]'$ แทนค่าคาดหมายของ $\hat{\mathrm{t}}$ และ $\boldsymbol{t}_{\boldsymbol{i}} = \boldsymbol{E}(\hat{\boldsymbol{t}}_{\boldsymbol{i}})$ เมื่อ $\boldsymbol{i} = 1,2,...,\boldsymbol{k}$ กำหนดให้ $\hat{\boldsymbol{Y}}_{\boldsymbol{NL,lin}}$ เป็นตัวประมาณค่าแบบเชิงเส้นของ $\hat{\boldsymbol{Y}}_{\boldsymbol{NL}}$ และมีค่าตามที่แสดงในสมการที่ 2.7

$$\hat{\boldsymbol{Y}}_{\boldsymbol{NL,lin}} \approx \psi(\mathrm{t}) + \sum_{\boldsymbol{i}=1}^{\boldsymbol{k}} \left.\frac{\partial\psi(\hat{\mathrm{t}})}{\partial\hat{\boldsymbol{t}}_{\boldsymbol{i}}}\right|_{\hat{\boldsymbol{t}}=\boldsymbol{t}} (\hat{\boldsymbol{t}}_{\boldsymbol{i}} - \boldsymbol{t}_{\boldsymbol{i}}) \qquad (2.12)$$

(2) หาค่าความแปรปรวนของ $\hat{\boldsymbol{Y}}_{\boldsymbol{NL}}$ ตามที่แสดงในสมการที่ (2.12)

$$\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{NL}}) \approx \boldsymbol{V}\left(\hat{\boldsymbol{Y}}_{\boldsymbol{NL,lin}}\right) \tag{2.13}$$

ตัวอย่างที่ 2.4 พิจารณาตัวอย่าง $\boldsymbol{s}$ ขนาด $\boldsymbol{n}$ ถูกสุ่มตามข้อตกลงของแผนการสุ่มตัวอย่าง กำหนดให้ $\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}} = \sum_{\boldsymbol{i}\in\boldsymbol{s}} \frac{\boldsymbol{y_i}}{\pi_{\boldsymbol{i}}} \Big/ \sum_{\boldsymbol{i}\in\boldsymbol{s}} \frac{1}{\pi_{\boldsymbol{i}}}$ แทนตัวประมาณค่าเฉลี่ยประชากร การหาค่าความแปรปรวนของ $\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}}$ เป็นดังนี้

(1) $\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}}$ สามารถเขียนให้อยู่ในรูปของฟังก์ชันปรับเรียบ ตามที่แสดงในสมการที่ (2.14)

$$\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}} = \sum_{\boldsymbol{i}\in\boldsymbol{s}} \frac{\boldsymbol{y_i}}{\pi_{\boldsymbol{i}}} \Big/ \sum_{\boldsymbol{i}\in\boldsymbol{s}} \frac{1}{\pi_{\boldsymbol{i}}} = \frac{\hat{\boldsymbol{t}}_1}{\hat{\boldsymbol{t}}_2} = \psi(\hat{\mathrm{t}}) \tag{2.14}$$

เมื่อ $\hat{\mathrm{t}} = \left[\hat{\boldsymbol{t}}_1 \;\; \hat{\boldsymbol{t}}_2\right]' = \left[\sum_{\boldsymbol{i}\in\boldsymbol{s}} \frac{\boldsymbol{y_i}}{\pi_{\boldsymbol{i}}} \;\; \sum_{\boldsymbol{i}\in\boldsymbol{s}} \frac{1}{\pi_{\boldsymbol{i}}}\right]'$ ค่าคาดหมายของ $\hat{\mathrm{t}}$ มีค่าเป็น $\mathrm{t} = \boldsymbol{E}(\hat{\mathrm{t}}) = \left[\boldsymbol{t}_1 \;\; \boldsymbol{t}_2\right]' = \left[\sum_{\boldsymbol{i}\in\boldsymbol{U}} \boldsymbol{y_i} \;\; \boldsymbol{N}\right]'$ เมื่อใช้วิธีเทเลอร์แบบเชิงเส้น ตัวประมาณแบบเชิงเส้นของ $\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}}$ เขียนแทนด้วย $\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek,lin}}$ มีค่าเป็น

$$\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek,lin}} \approx \boldsymbol{Cons}\tan\boldsymbol{t} + \frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{i}\in\boldsymbol{s}} \frac{1}{\pi_{\boldsymbol{i}}} (\boldsymbol{y_i} - \bar{\boldsymbol{Y}}) \tag{2.15}$$

(2) ความแปรปรวนของ $\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}}$ เขียนแทนด้วย $\boldsymbol{V}(\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}})$ กำหนดค่าตามที่แสดงในสมการ (2.16)

$$\begin{aligned}
\boldsymbol{V}(\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}}) &\approx \boldsymbol{V}\left(\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek,lin}}\right) \\
&= \boldsymbol{V}\left(\boldsymbol{Cons}\tan\boldsymbol{t} + \frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{i}\in\boldsymbol{s}} \frac{1}{\pi_{\boldsymbol{i}}} (\boldsymbol{y_i} - \bar{\boldsymbol{Y}})\right) = \frac{1}{\boldsymbol{N}_2} \boldsymbol{V}\left(\sum_{\boldsymbol{i}\in\boldsymbol{s}} \frac{1}{\pi_{\boldsymbol{i}}} (\boldsymbol{y_i} - \bar{\boldsymbol{Y}})\right) \\
&= \frac{1}{\boldsymbol{N}^2}\left[\sum_{\boldsymbol{i}\in\boldsymbol{U}} \left(\frac{1-\pi_{\boldsymbol{i}}}{\pi_{\boldsymbol{i}}}\right)(\boldsymbol{y_i} - \bar{\boldsymbol{Y}})^2 + \sum_{\boldsymbol{i}\in\boldsymbol{U}} \sum_{\boldsymbol{j}\neq\boldsymbol{i}\in\boldsymbol{U}} \left(\frac{\pi_{\boldsymbol{ij}} - \pi_{\boldsymbol{i}}\pi_{\boldsymbol{j}}}{\pi_{\boldsymbol{i}}\pi_{\boldsymbol{j}}}\right)(\boldsymbol{y_i} - \bar{\boldsymbol{Y}})(\boldsymbol{y_j} - \bar{\boldsymbol{Y}})\right]
\end{aligned} \tag{2.16}$$

ตัวประมาณของ $\boldsymbol{V}(\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}})$ เขียนแทนด้วย $\hat{\boldsymbol{V}}(\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}})$ และกำหนดค่าตามที่แสดงในสมการ (2.17)

$$\hat{\boldsymbol{V}}(\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}}) = \frac{1}{\boldsymbol{N}^2}\left[\sum_{\boldsymbol{i}\in\boldsymbol{s}} \left(\frac{1-\pi_{\boldsymbol{i}}}{\pi_{\boldsymbol{i}}^2}\right)(\boldsymbol{y_i} - \hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}})^2 + \sum_{\boldsymbol{i}\in\boldsymbol{s}} \sum_{\boldsymbol{j}\neq\boldsymbol{i}\in\boldsymbol{s}} \left(\frac{\pi_{\boldsymbol{ij}} - \pi_{\boldsymbol{i}}\pi_{\boldsymbol{j}}}{\pi_{\boldsymbol{ij}}\pi_{\boldsymbol{i}}\pi_{\boldsymbol{j}}}\right)(\boldsymbol{y_i} - \hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}})(\boldsymbol{y_j} - \hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Hajek}})\right] \tag{2.17}$$

2.4 งานวิจัยที่เกี่ยวข้อง

ทัตตา หิรัญพต (2554) ศึกษาเกี่ยวกับการเปรียบเทียบประสิทธิภาพระหว่างวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยและวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยแบบสโทแคสติก โดยกำหนดขนาดตัวอย่างเท่ากับ 50, 100, 200 และ 300 สุ่มโดยขั้นตอนวิธีของเบ็บบิงตัน (Bebbington algorithm) ซึ่งใช้วิธีการเลือกตัวอย่างสุ่มแบบง่ายและไม่ใส่คืน (Simple random sampling without replacement หรือ SRSWOR) และกำหนดเปอร์เซ็นต์ของข้อมูลสูญหายของตัวอย่าง เท่ากับ 5%, 10%, 15% และ 20% ของขนาดตัวอย่างผลการศึกษาการเปรียบเทียบประสิทธิภาพระหว่างวิธีการประมาณค่า

ข้อมูลสูญหายนั้น ได้วัดจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยของค่าประมาณของข้อมูลสูญหาย (MSE) โดยได้ข้อสรุปว่า วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอย (RI) มีประสิทธิภาพมากกว่าวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยแบบสโทแคสติก (SRI)

วริษฐา กณิกนันต์ และ อนุภาพ สมบูรณ์สวัสดี (2556) ศึกษาเกี่ยวกับ การเปรียบเทียบวิธีการประมาณสาหรับการวิเคราะห์การถดถอยเชิงเส้นพหุเมื่อตัวแปร งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณข้อมูลสูญหายสำหรับการวิเคราะห์การถดถอยเชิงเส้น พหุ เมื่อตัวแปรตามและตัวแปรอิสระมีการสูญหายแบบนอนอิกนอร์เรเบิล วิธีการประมาณค่าสูญหายที่ใช้ในงานวิจัยนี้คือ วิธี K-Nearest Neighbor (KNN) วิธี EM Algorithm (EM) และวิธี Predictive Mean Matching (PMM) ซึ่งข้อมูลที่ใช้ใน การศึกษาได้จากการจำลองโดยมีสัดส่วนของการสูญหาย 3 ระดับคือ 10% 20% และ 30% และมีระดับการสูญหายแบบ นอนอิกนอร์เรเบิล 3 ระดับคือ ไม่มี ปานกลาง และสูง โดยจะทาการเปรียบเทียบประสิทธิภาพของแต่ละวิธีการด้วยค่าเฉลี่ยใหญ่ขึ้น ii) วิธีการประมาณทุกวิธีจะแย่ลงเมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน สัดส่วนของการสูญ หาย และระดับการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มสูงขึ้น iii) วิธีการ KNN จะเป็นวิธีการประมาณที่ดีที่สุด เมื่อส่วน เบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีขนาดปานกลางและสูง (30 และ 90) iv) วิธีการ EM จะเป็นวิธีการที่ดีที่สุดเมื่อ ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนน้อย (10) ยกเว้นในกรณีที่ข้อมูลตัวแปรอิสระมีความแปรปรวนเท่ากัน

พัชรีพร สนรักษา และคณะ (2555) ศึกษาเกี่ยวกับการประมาณค่าเฉลี่ยประชากรในการสำรวจตัวอย่าง เมื่อมีข้อมูลสูญหาย วัตถุประสงค์ในการศึกษา คือ นำเสนอตัวประมาณค่าเฉลี่ยประชากรในการสำรวจตัวอย่างเมื่อมีข้อมูลสูญหายโดยการประยุกต์ใช้ตัวประมาณค่าเฉลี่ย $\boldsymbol{T_{d1}}$ ของ Singh และคณะ เมื่อเปรียบเทียบค่า the root mean square error (RMSE) ของตัวประมาณค่าที่นำเสนอและตัวประมาณค่า $\boldsymbol{T_{d1}}$ $\boldsymbol{T_{d2}}$ ตัวประมาณค่าเมื่อประมาณค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ย (mean imputation) ตัวประมาณค่าเมื่อประมาณค่าข้อมูลสูญหายด้วยค่าอัตราส่วน (ratio imputation) ตัวประมาณค่าเมื่อประมาณค่าข้อมูลสูญหายด้วยค่าคอมโพรไมซ (compromised imputation) และตัวประมาณค่าเมื่อประมาณค่าข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนัก (weighted nearest neighbor and regression imputation) (ภายใต้เงื่อนไขที่กำหนด) เราพบว่าตัวประมาณค่าที่นำเสนอมีประสิทธิภาพดีกว่าตัวประมาณค่า $\boldsymbol{T_{d1}}$ $\boldsymbol{T_{d2}}$ ตัวประมาณค่าเมื่อประมาณค่าข้อมูลสูญหายด้วยวิธีค่าเฉลี่ย (mean imputation) ตัวประมาณค่าเมื่อประมาณค่าข้อมูลสูญหายด้วยค่าอัตราส่วน (ratio imputation) ตัวประมาณค่าเมื่อประมาณค่าข้อมูลสูญหายด้วยค่าคอมโพรไมซ (compromised imputation) และตัวประมาณค่าเมื่อประมาณค่าข้อมูลสูญหายด้วยค่าถดถอยของค่าเข้าใกล้สุดแบบถ่วงน้ำหนัก (weighted nearest neighbor and regression imputation)

ภวัต ภักดิ์ศรานุวัต และ สอาด นิวิศพงศ์ (2553) ศึกษาเกี่ยวกับการหาช่วงความเชื่อมั่น และช่วงพยากรณ์สำหรับพารามิเตอร์ บางค่าเมื่อข้อมูลเกิดค่าสูญหาย วัตถุประสงค์ในการศึกษาคือ สร้างและเปรียบเทียบช่วงความเชื่อมั่น และช่วงพยากรณ์ของพารามิเตอร์บางค่า เมื่อมีข้อมูลสูญหาย เกิดขึ้น โดยช่วงความเชื่อมั่น และช่วงพยากรณ์ที่ทำการศึกษา คือช่วงความเชื่อมั่นของผลต่างเฉลี่ย 2 ประชากร ช่วงความเชื่อมั่นของความแปรปรวน 1 ประชากร และช่วงความเชื่อมั่น สำหรับค่าพยากรณ์ ผลการจำลองข้อมูลแสดงให้เห็นว่า เมื่อข้อมูลมีการแจกแจงไม่ปกติ และแทนที่ค่าสูญหายด้วยวิธีการแทนที่แบบสุ่ม ช่วง

ความเชื่อมั่นของผลต่างค่าเฉลี่ยแบบปรับปรุงโดยใช้ช่วงความเชื่อมั่นของ Miao-Chiou เป็นพื้นฐานที่มีประสิทธิภาพดีกว่าช่วงความเชื่อมั่นของ Welch-Satterthwaite และได้ว่าช่วงความเชื่อมั่นของความแปรปรวนแบบปรับปรุง โดยใช้ช่วงความเชื่อมั่นของ Bonett เป็นพื้นฐาน มีประสิทธิภาพดีกว่าช่วงความเชื่อมั่นที่ใช้ตัวสถิติ χ^2 และตัวสถิติ F ตามลำดับ และสำหรับช่วงความเชื่อมั่นสำหรับค่าพยากรณ์ ผลจากการจำลองแสดงให้เห็นว่าช่วงพยากรณ์ ที่ได้จากการแทนที่ค่าสูญหายโดยใช้วิธีการแทนที่แบบสุ่มมีประสิทธิภาพกว่าการแทนที่ค่าสูญหายโดยวิธีการแทนที่ด้วยค่าเฉลี่ย

ดวงพร หัชชะวณิช (2553) ศึกษาเกี่ยวกับการเปรียบเทียบค่าประมาณที่ปรับน้ำหนัก จากการไม่ได้รับคำตอบด้วยวิธีปรับค่าน้ำหนัก เมื่อทราบค่ายอดรวมของประชากร และวิธีจัดชั้นภูมิภายหลังการสุ่มตัวอย่าง วัตถุประสงค์ในการวิจัยคือเปรียบเทียบค่าประมาณที่ปรับค่าน้ำหนักจากการไม่ได้รับคำตอบด้วยวิธี ปรับค่าน้ำหนักเมื่อทราบค่ายอดรวมของประชากร และวิธีจัดชั้นภูมิภายหลังการสุ่มตัวอย่าง ผลการเปรียบเทียบค่าความเอนเอียงในการประมาณค่าเฉลี่ยของประชากร พบว่า วิธีทั้งสองให้ค่าประมาณ ที่มีความเอนเอียงไม่แตกต่างกัน และจากการเปรียบเทียบค่าความคลาดเคลื่อนมาตรฐานในการ ประมาณค่าเฉลี่ยของประชากร พบว่าในกรณีที่ประชากรมีการกระจายน้อยวิธีจัดชั้นภูมิภายหลัง การสุ่มตัวอย่างและวิธีปรับค่าน้ำหนักเมื่อทราบค่ายอดรวมของประชากร ให้ค่าความคลาดเคลื่อน มาตรฐานในการประมาณค่าเฉลี่ยของประชากรน้อยกว่ากรณีที่ไม่ได้มีการปรับค่าน้ำหนัก วิธีปรับ ค่าน้ำหนักเมื่อทราบค่ายอดรวมของประชากร จะให้ค่าประมาณที่มีความคลาดเคลื่อนมาตรฐานน้อย กว่าวิธีจัดชั้นภูมิภายหลังการสุ่มตัวอย่าง เมื่อประชากรมีการกระจายมากขึ้นสองวิธีนี้จะให้ค่าประมาณของค่าเฉลี่ยของประชากรที่มีความคลาดเคลื่อนมาตรฐานแตกต่างกันน้อยลง นอกจากนี้ ยังพบว่าเมื่อใช้ขนาดตัวอย่างมากขึ้นสองวิธีนี้จะให้ค่าประมาณที่มีความคลาดเคลื่อนมาตรฐานในการ ประมาณค่าที่ใกล้เคียงกับกรณีที่ไม่ได้ปรับค่าน้ำหนักมากขึ้น

Haziza (2010) นำเสนอวิธีการประมาณค่าความแปรปรวน ของตัวประมารค่าเฉลี่ยประชากร กรณีที่มีข้อมูลสูญหาย โดยใช้วิธีการสุ่มตัวอย่างซ้ำ ภายใต้เฟรมเวิรค์แบบรีเวิร์ส (Reverse framework) เฟรมเวิร์คดังกล่าวถูกนำเสนอเป็นครั้งแรก โดย Fay (1991) ซึ่งได้สลับลำดับระหว่างแผนการสุ่มตัวอย่างและกลไกการตอบสนองดังนี้ จากประชากร $U = \{1, 2, ..., i, ..., N\}$ ขนาด N หน่วย ถูกสุ่มออกเป็นเซตของประชากรที่ไม่สูญหาย U_r ขนาด N_r และเซตของประชากรที่สูญหาย U_m ขนาด N_m ตามข้อตกลงของกลไกการตอบสนอง ดังนั้นเมื่อกำหนด (U_r, U_m) ตัวอย่าง s ขนาด n จะประกอบด้วยหน่วยตัวอย่างทั้งที่สูญหาย และไม่สูญหาย ตามข้อตกลงของแผนการสุ่มตัวอย่าง Haziza (2010) กล่าวว่าการประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากร ภายใต้เฟรมเวิร์คแบบรีเวิร์ส นิยมใช้วิธีการสุ่มตัวอย่างซ้ำ (Resampling method) เช่นวิธีแจ็คไนฟ (Jackknifing) หรือวิธี บูทสแตรป (Bootstrap) มากกว่าวิธีเทเลอร์แบบเชิงเส้น (Taylor linearization) เนื่องจากวิธีการสุ่มตัวอย่างซ้ำไม่จำเป็นต้องหาค่าอนุพันธ์ย่อยของฟังก์ชันของ ตัวประมาณค่าเทียบกับตัวประมาณค่าแต่ละตัว ทั้งนี้ถ้าฟังก์ชันของตัวประมาณค่าประกอบด้วยตัวประมาณค่าเป็นจำนวนมาก การหาอนุพันธ์ย่อยเทียบกับตัวประมาณค่าดังกล่าวจะมีความยุ่งยากและซับซ้อน นอกจากนี้วิธีการสุ่มตัวอย่างซ้ำไม่ต้องการความน่าจะเป็นที่หน่วยตัวอย่างที่ i และ j ตกอยู่ในตัวอย่าง s ขนาด n (Second order inclusion probability) ซึ่งในบางแผนการสุ่มตัวอย่างการคำนวณค่าดังกล่าวค่อนข้างมีความยุ่งยาก

บทที่ 3
วิธีดำเนินงานวิจัย

ในการศึกษาเกี่ยวกับการประมาณค่าผลรวมประชากร ในการสำรวจตัวอย่าง เมื่อมีข้อมูลสูญหาย ผู้วิจัยได้ดำเนินการตามขั้นตอนการวิจัย ดังนี้

3.1 สัญลักษณ์และข้อตกลงเบื้องต้น

พิจารณาประชากร $\boldsymbol{U}=\{1,2,...,\boldsymbol{i},...,\boldsymbol{N}\}$ ขนาด $\boldsymbol{N}$ ถูกแบ่งออกเป็น $\boldsymbol{L}$ ชั้นภูมิ $\boldsymbol{U}_1,\boldsymbol{U}_2,...,\boldsymbol{U}_{\boldsymbol{h}},...,\boldsymbol{U}_{\boldsymbol{L}}$ ขนาด $\boldsymbol{N}_1,\boldsymbol{N}_2,...,\boldsymbol{N}_{\boldsymbol{h}},...,\boldsymbol{N}_{\boldsymbol{L}}$ ตามลำดับ กำหนดให้ $\boldsymbol{H}=\{1,2,...,\boldsymbol{h},...,\boldsymbol{L}\}$ สมมติว่าในแต่ละชั้นภูมิตัวอย่าง $\boldsymbol{s}_{\boldsymbol{h}}$ ขนาด $\boldsymbol{n}_{\boldsymbol{h}}$ ถูกสุ่มประชากร $\boldsymbol{U}_{\boldsymbol{h}}$ขนาด $\boldsymbol{N}_{\boldsymbol{h}}$ เมื่อ $\boldsymbol{h}\in\boldsymbol{H}$ ด้วยแผนการสุ่มตัวอย่างของ Midzuno (1952) และในแต่ละชั้นภูมิกำหนดให้สัดส่วนตัวอย่าง (Sampling fraction) หรือ $\boldsymbol{f}_{\boldsymbol{h}}=\frac{\boldsymbol{n}_{\boldsymbol{h}}}{\boldsymbol{N}_{\boldsymbol{h}}}$ มีขนาดเล็ก กำหนดให้ $\boldsymbol{y}$ แทนตัวแปรที่สนใจศึกษา และผลรวมของตัวแปรที่สนใจศึกษาแทนด้วยสัญลักษณ์ $\boldsymbol{Y}=\sum_{\boldsymbol{h}\in\boldsymbol{H}}\sum_{\boldsymbol{i}\in\boldsymbol{U}_{\boldsymbol{h}}}\boldsymbol{y}_{\boldsymbol{hi}}$

ในแต่ละชั้นภูมิ $\boldsymbol{h}$ กำหนดให้ $\pi_{\boldsymbol{hi}}$ แทนความน่าจะเป็นที่หน่วยตัวอย่างที่ $\boldsymbol{i}$ ตกอยู่ในหน่วยตัวอย่าง $\boldsymbol{s}_{\boldsymbol{h}}$ หรือ $\pi_{\boldsymbol{hi}}=\boldsymbol{P}(\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}})$ เมื่อ $\boldsymbol{i}\in\boldsymbol{U}_{\boldsymbol{h}}$ และ $\boldsymbol{h}\in\boldsymbol{H}$ ส่วนค่า $\pi_{\boldsymbol{hij}}$ แทนความน่าจะเป็นที่หน่วยตัวอย่างที่ $\boldsymbol{i}$ และ $\boldsymbol{j}$ ตกอยู่ในตัวอย่าง $\boldsymbol{s}_{\boldsymbol{h}}$ หรือ $\pi_{\boldsymbol{hij}}=\boldsymbol{P}(\boldsymbol{i}\wedge\boldsymbol{j}\in\boldsymbol{s}_{\boldsymbol{h}})$ เมื่อ $\boldsymbol{i}\wedge\boldsymbol{j}\in\boldsymbol{U}_{\boldsymbol{h}}$ และ $\boldsymbol{h}\in\boldsymbol{H}$ กำหนดให้ $\boldsymbol{E},\boldsymbol{V}$ แทนค่าคาดหมายและความแปรปรวนทั้งหมด $\boldsymbol{E}_{\boldsymbol{S}},\boldsymbol{V}_{\boldsymbol{S}}$ แทนค่าคาดหมาย และความแปรปรวนเทียบกับแผนการสุ่มตัวอย่างตามลำดับ

กำหนดให้แต่ละชั้นภูมิ $\boldsymbol{h}$ มีข้อมูลสูญหาย ให้ $\boldsymbol{s}_{\boldsymbol{hr}}$ แทนเซตของตัวอย่างที่ไม่เกิดการสูญหาย ขนาด $\boldsymbol{n}_{\boldsymbol{hr}}$ และ $\boldsymbol{s}_{\boldsymbol{hm}}=\boldsymbol{s}_{\boldsymbol{h}}-\boldsymbol{s}_{\boldsymbol{hr}}$ ขนาด $\boldsymbol{n}_{\boldsymbol{hm}}=\boldsymbol{n}_{\boldsymbol{h}}-\boldsymbol{n}_{\boldsymbol{hr}}$ แทนเซตของตัวอย่างที่เกิดการสูญหาย เมื่อ $\boldsymbol{h}\in\boldsymbol{H}$ กำหนดให้ตัวแปรกำหนดขนาด (Size variable) $\boldsymbol{k}$ มีความสัมพันธ์กับ $\boldsymbol{y}$ นอกจากนี้ ในแต่ละชั้นภูมิ $\boldsymbol{h}$ ตัวแปรกำหนดขนาด $\boldsymbol{k}$ ทราบค่า และมีค่าบวกทุก $\boldsymbol{k}_{\boldsymbol{hi}}$ เมื่อ $\boldsymbol{i}\in\boldsymbol{U}_{\boldsymbol{h}}$ และ $\boldsymbol{h}\in\boldsymbol{H}$ กำหนดให้การ สูญหายของข้อมูลเกิดขึ้นเฉพาะตัวแปรที่สนใจศึกษา และเกิดขึ้นเฉพาะการไม่ตอบเฉพาะบางคำถาม กำหนดให้สัญลักษณ์ $\boldsymbol{R}$ แทนกลไกการไม่ตอบสนองของข้อมูล (Nonresponse mechanism) รวมถึงกำหนดให้ $\boldsymbol{r}_{\boldsymbol{hi}}$ แทนตัวแปรบอกสภาพของการสูญหาย ของหน่วยตัวอย่างที่ $\boldsymbol{i}$ ในชั้นภูมิที่ $\boldsymbol{h}$ ซึ่ง $\boldsymbol{r}_{\boldsymbol{hi}}=1$ เมื่อ $\boldsymbol{y}_{\boldsymbol{hi}}$ ไม่เกิดการสูญหาย และ $\boldsymbol{r}_{\boldsymbol{hi}}=0$ เมื่อ $\boldsymbol{y}_{\boldsymbol{hi}}$ เกิดการสูญหาย เมื่อ $\boldsymbol{i}\in\boldsymbol{U}_{\boldsymbol{h}}$ และ $\boldsymbol{h}\in\boldsymbol{H}$ กำหนดให้ $\boldsymbol{R}=(\boldsymbol{r}_{11}\ \boldsymbol{r}_{12}\ \ldots\ \boldsymbol{r}_{\boldsymbol{hN}_{\boldsymbol{h}}})'$ แทนเวกเตอร์ของตัวแปรบอกสภาพของการสูญหาย $\boldsymbol{p}_{\boldsymbol{hi}}$ แทนความน่าจะเป็นที่จะตอบสนองของหน่วยตัวอย่างที่ $\boldsymbol{i}$ ในชั้นภูมิที่ $\boldsymbol{h}$ และกำหนดค่า $\boldsymbol{p}_{\boldsymbol{hi}}=\boldsymbol{P}(\boldsymbol{r}_{\boldsymbol{hi}}=1)$ สมมติว่า $\boldsymbol{p}_{\boldsymbol{hi}}=\boldsymbol{p}_{\boldsymbol{h}}$ ทุกค่า $\boldsymbol{i}$ เมื่อ $\boldsymbol{i}\in\boldsymbol{U}_{\boldsymbol{h}}$ และ $\boldsymbol{h}\in\boldsymbol{H}$ ในแต่ละชั้นภูมิกำหนดให้ความน่าจะเป็นที่จะไม่สูญหายของหน่วยตัวอย่างที่ $\boldsymbol{i}$ เป็นอิสระกับหน่วยอื่น ๆ หรือ $\boldsymbol{p}_{\boldsymbol{hij}}=\boldsymbol{P}(\boldsymbol{r}_{\boldsymbol{hi}}=1,\boldsymbol{r}_{\boldsymbol{hj}}=1)=\boldsymbol{p}_{\boldsymbol{h}}\boldsymbol{p}_{\boldsymbol{h}}=\boldsymbol{p}_{\boldsymbol{h}}^2$ เมื่อ $\boldsymbol{i},\boldsymbol{j}\in\boldsymbol{U}_{\boldsymbol{h}}$ และ $\boldsymbol{h}\in\boldsymbol{H}$

กำหนดให้ E_R, V_R แทนค่าคาดหมายและความแปรปรวนเทียบกับกลไกการตอบสนอง ดังนั้น $E_R(r_{hi}) = P(r_{hi} = 1) = p_h$ และ $V_R(r_{hi}) = p_h(1 - p_h)$ เมื่อ $i \wedge j \in U_h$ และ $h \in H$

3.2 การสุ่มตัวอย่างในแต่ละชั้นภูมิ

การวิจัยในครั้งนี้ผู้วิจัยสุ่มตัวอย่างในแต่ละชั้นภูมิ โดยใช้แผนการสุ่มตัวอย่างตามที่ Midzuno (1952) นำเสนอดังนี้ ขั้นแรกในแต่ละชั้นภูมิสุ่มตัวอย่างขนาดหนึ่งหน่วย จากประชากร U_h ขนาด N_h ด้วยแผนการสุ่มตัวอย่างแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาด (*Probability proportional to size*) จากนั้นสุ่มตัวอย่างขนาด $n_h - 1$ จากประชากรขนาด $N_h - 1$ โดยใช้แผนการสุ่มตัวอย่างแบบง่ายและไม่ใส่คืน กำหนดให้ $K_h = \sum_{i \in U_h} k_{hi}$ ซึ่งภายใต้แผนการสุ่มตัวอย่างดังกล่าวค่า π_{hi} และ π_{hij} กำหนดค่าตามที่แสดงในสมการที่ (3.1) และ (3.2) ตามลำดับ

$$\pi_{hi} = \frac{k_{hi}}{K_h}\frac{N_h - n_h}{N_h - 1} + \frac{n_h - 1}{N_h - 1} \tag{3.1}$$

$$\pi_{hij} = \frac{k_{hi} + k_{hj}}{K}\frac{N_h - n_h}{N_h - 1}\frac{n_h - 1}{N_h - 2} + \frac{n_h - 1}{N_h - 1}\frac{n_h - 2}{N_h - 2} \tag{3.2}$$

เมื่อ $i, j \in U_h$ และ $h \in H$ ค่า k_{hi} แทนค่าสังเกตุของตัวแปรกำหนดขนาด ของหน่วยตัวอย่างที่ i ในชั้นภูมิที่ h เมื่อ $h \in H$

3.3 การประมาณค่าความแปรปรวนภายใต้เฟรมเวิร์คแบบรีเวิร์ส

เฟรมเวิร์คแบบรีเวิร์สถูกนำเสนอเป็นครั้งแรกโดย Fay (1991) ที่ได้สลับลำดับระหว่าง แผนการสุ่มตัวอย่าง และการตอบสนอง (ศึกษารายละเอียดเพิ่มเติมได้จาก Shao and Steel (1999)) ภายใต้แผนการสุ่มตัวอย่างดังกล่าว ความแปรปรวนของตัวประมาณค่าผลรวม $\hat{Y}$ สามารถหาได้จากสมการ (3.3)

$$V(\hat{Y}) = E_R V_S(\hat{Y} \mid R) + V_R E_S(\hat{Y} \mid R) \tag{3.3}$$

เมื่อ $R = (r_{11}\ r_{12}\ \ldots\ r_{hN_h})'$ อย่างไรก็ตามกรณีที่สัดส่วนตัวอย่างในแต่ละชั้นภูมิ $f_h = \frac{n_h}{N_h}$ มีขนาดเล็กสามารถตัดเทอมที่ 2 ออกจากสมการ (3.3) ดังนั้นภายใต้เฟรมเวิร์คแบบรีเวิร์ส และสัดส่วนตัวอย่างในแต่ละชันภูมิมีขนาดเล็ก ค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร $\hat{Y}$ สามารถหาได้จาก

$$V(\hat{Y}) \approx E_R V_S(\hat{Y} \mid R) \tag{3.4}$$

สำหรับการประมาณ $E_R V_S(\hat{Y} \mid R)$ ตามที่แสดงในสมการ (3.4) จะคล้ายกับการประมาณค่า $V_S(\hat{Y} \mid R)$ กำหนดให้ $\hat{V}_S$ เป็นตัวประมาณค่าของ $V_S(\hat{Y} \mid R)$ ดังนั้น $E_S(\hat{V}_S \mid R) = V_S(\hat{Y} \mid R)$ และ $E\left[\hat{V}_S \mid R\right] = E_R\left[E_S(\hat{V}_S \mid R)\right] = E_R\left[V_S(\hat{Y} \mid R)\right]$ ดังนั้นค่า $V(\hat{Y})$ ตามที่แสดงในสมการที่ (3.4) สามารถประมาณค่าได้จาก

$$\hat{V}(\hat{Y}) = \hat{V}_S(\hat{Y} \mid R) \tag{3.5}$$

3.4 นำเสนอตัวประมาณค่าผลรวมประชากรแบบจุด

กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถามและแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย ผู้วิจัยนำเสนอตัวประมาณค่าผลรวมประชากรแบบจุด ทั้งตัวประมาณค่าผลรวมประชากรแบบเชิงเส้นเขียนแทนด้วย $\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}}$ และตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้นเขียนแทนด้วย $\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}$

3.5 นำเสนอตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร

การหาตัวประมาณค่าความแปรปรวนของ $\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}}$ เขียนแทนด้วย $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}})$ สามารถหาได้จากตัวประมาณค่าของ Horvitz and Thomson (1952) ตามที่แสดงในสมการ (1.3) ส่วนการหาตัวประมาณค่าความแปรปรวนของ $\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}$ เขียนแทนด้วย $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ จำเป็นต้องปรับ $\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}$ ให้เป็นตัวประมาณค่า แบบเชิงเส้น และหาตัวประมาณค่าความแปรปรวนภายใต้ เฟรมเวิร์คแบบรีเวิร์ส

3.6 นำเสนอช่วงความเชื่อมั่นของผลรวมประชากร

การสร้างช่วงความเชื่อมั่นของผลรวมประชากร ผู้วิจัยใช้ ทฤษฎีลิมิตเข้าสู่ส่วนกลางในการสร้างช่วงความเชื่อมั่น โดยกำหนดสัญลักษณ์ $\boldsymbol{C.I._1}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$ และ $\boldsymbol{C.I._2}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$ แทนช่วงความเชื่อมั่นแบบเชิงเส้นและช่วงความเชื่อมั่นแบบไม่เป็นเชิงเส้นตามลำดับ

3.7 แสดงตัวอย่างเชิงตัวเลข

3.7.1 ประชากรที่ใช้ในการศึกษา

ประชากรที่ใช้ในการศึกษา เป็นข้อมูลที่ได้จากการจำลองสถานการณ์ โดยใช้เทคนิคมอนติคาร์โลซิมูเลชั่น (Monte carlo simulation) ตามการศึกษาของ Cheng (2010) ซึ่งกำหนดข้อมูลของประชากร ตามที่แสดงในสมการ (3.3) ดังนี้

$$\boldsymbol{y_i} = \begin{cases} 5\boldsymbol{k_i^2} + 0.3\boldsymbol{k_i} + \boldsymbol{e_{1i}} & ,\boldsymbol{i} \in \boldsymbol{U}_1 \\ 5\boldsymbol{k_i^2} + 0.3\boldsymbol{k_i} + 0.8(\boldsymbol{k_i} - 2) + \boldsymbol{e_{2i}} & ,\boldsymbol{i} \in \boldsymbol{U}_2 \end{cases} \tag{3.3}$$

เมื่อ $\boldsymbol{x_i} \sim \boldsymbol{Gamma}(9,1.5)$ $\boldsymbol{e_{1i}} \sim \boldsymbol{N}(0,4)$ และ $\boldsymbol{e_{2i}} \sim \boldsymbol{N}(0,4)$ ส่วน $\boldsymbol{N}_1 = 1,500$ $\boldsymbol{N}_2 = 1,200$ ดังนั้น $\boldsymbol{N} = 2,700$

3.7.2 ตัวอย่างที่ใช้ในการศึกษา

จากประชากรในแต่ละชั้นภูมิ $\boldsymbol{U}_1$ และ $\boldsymbol{U}_2$ ขนาด $\boldsymbol{N}_1 = 1,500$ และ $\boldsymbol{N}_2 = 1,200$ ทำการสุ่มตัวอย่างขนาด $\boldsymbol{n}_1$ และ $\boldsymbol{n}_2$ โดยใช้แผนการสุ่มตัวอย่างของ Midzuno ทั้งหมด 8 ระดับ ตามที่แสดงในตารางที่ 3.1

ตารางที่ 3.1 ขนาดตัวอย่างที่ใช้ในการศึกษา

ชุดตัวอย่างที่	n_1	n_2	n
1	30	20	50
2	45	25	60
3	45	30	75
4	60	40	100
5	70	50	120
6	85	55	140
7	90	65	155
8	100	70	170

3.7.3 กำหนดความน่าจะเป็นที่จะตอบสนองและการสูญหายของข้อมูล

กำหนดความน่าจะเป็นที่จะตอบสนอง โดยชั้นภูมิที่ 1 กำหนดความน่าจะเป็นที่จะตอบสนองของแต่ละหน่วยตัวอย่างมีค่าเท่ากัน คือ $p_{1i} = p_1 = 0.6$; $i \in U_1$ และ กำหนดให้ความน่าจะเป็นที่จะตอบสนองของแต่ละหน่วยตัวอย่างในชั้นภูมิที่ 2 มีค่าเท่ากัน คือ $p_{2i} = p_2 = 0.8; i \in U_2$ จากนั้นกำหนดการสูญหายของข้อมูลในแต่ละชั้นภูมิ โดยพิจารณาจากตัวแปรสุ่ม A_{hi} ซึ่ง $A_{hi} \sim U(0,1)$ เมื่อ $i \in U_h$ และ $h = 1,2$ ถ้า $A_{hi} \le p_{hi}$ ค่า $r_{hi} = 1$ และ $r_{hi} = 0$ เมื่อ $A_{hi} > p$

3.7.4 เปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร

การวิจัยในครั้งนี้ใช้เทคนิคมอนติคาร์โลซิมูเลชั่นและกำหนดการทำซ้ำทั้งหมด 100,000 ครั้ง เพื่อเปรียบเทียบประสิทธิภาพของ $\hat{V}(\hat{Y}_{LMI})$ และ $\hat{V}(\hat{Y}_{NLMI})$ โดยพิจารณาจากค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ ตามที่แสดงในสมการ (3.4)

$$RMSE(\hat{V}) = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}[\hat{V}_m - V^*]^2}}{V^*}, \qquad (3.4)$$

เมื่อ $V^* = \frac{1}{M}\sum_{m=1}^{M}(\hat{Y}_m - \bar{Y}^*)^2$ ค่า $\bar{Y}^* = \frac{1}{M}\sum_{i=1}^{M}\hat{Y}_m$ และ $\hat{V}_m$ แทนตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร ของการสุ่มตัวอย่างครั้งที่ m เมื่อ $m = 1,2,...,M$ และ $M = 100,000$

สำหรับเกณฑ์การเปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร กำหนดตามนิยามที่ 3.1

นิยาม 3.1 *กำหนดให้* $\hat{V}_1$ *และ* $\hat{V}_2$ *เป็นตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร จะกล่าวว่า* $\hat{V}_1$ *เป็นตัวประมาณค่าที่มีประสิทธิภาพดีกว่า* $\hat{V}_2$ *ก็ต่อเมื่อ* $RMSE(\hat{V}_1) \le RMSE(\hat{V}_2)$

3.7.5 เปรียบเทียบประสิทธิภาพของตัวประมาณค่าผลรวมประชากรแบบช่วง

การเปรียบเทียบประสิทธิภาพของตัวประมาณค่าผลรวมประชากรแบบช่วง ผู้วิจัยพิจารณาจากค่าเฉลี่ยของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมพารามิเตอร์ (Average coverage percentage) กำหนดค่าตามที่แสดงในสมการที่ (3.5)

$$\boldsymbol{ACP}\left(\boldsymbol{C.I.}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})\right)=\sum_{\boldsymbol{m}=1}^{\boldsymbol{M}}\frac{\boldsymbol{a_i}}{\boldsymbol{M}} \tag{3.5}$$

เมื่อ

$$\boldsymbol{a_m}=\begin{cases}1 & ;\hat{\boldsymbol{Y}}_{\boldsymbol{L}}\le \boldsymbol{Y}\le \hat{\boldsymbol{Y}}_{\boldsymbol{U}}\\ 0 & ;\boldsymbol{Y}<\hat{\boldsymbol{Y}}_{\boldsymbol{L}}\vee \boldsymbol{Y}>\hat{\boldsymbol{Y}}_{\boldsymbol{U}}\end{cases}$$

$\hat{\boldsymbol{Y}}_{\boldsymbol{L}}=\hat{\boldsymbol{Y}}_{\boldsymbol{MI}}-\boldsymbol{z}_{\frac{\alpha}{2}}\sqrt{\hat{\boldsymbol{V}}_{\boldsymbol{m}}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})}$ และ $\hat{\boldsymbol{Y}}_{\boldsymbol{U}}=\hat{\boldsymbol{Y}}_{\boldsymbol{MI}}+\boldsymbol{z}_{\frac{\alpha}{2}}\sqrt{\hat{\boldsymbol{V}}_{\boldsymbol{m}}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})}$ ส่วน $\hat{\boldsymbol{V}}_{\boldsymbol{m}}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$ แทนค่า $\hat{\boldsymbol{V}}_{\boldsymbol{m}}(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}})$ หรือ $\hat{\boldsymbol{V}}_{\boldsymbol{m}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ ในรอบที่ $\boldsymbol{m}$ และ $\boldsymbol{m}=1,2,\ldots,\boldsymbol{M}$

สำหรับเกณฑ์การเปรียบเทียบประสิทธิภาพของตัวประมาณค่าผลรวมประชากรแบบช่วง กำหนดตามนิยามที่ 3.2

นิยาม 3.2 *กำหนดให้* $\boldsymbol{C.I.}_1(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$ *และ* $\boldsymbol{C.I.}_2(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$ *แทนตัวประมาณค่าผลรวมประชากรแบบช่วง จะกล่าวว่า* $\boldsymbol{C.I.}_1(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$ *เป็นตัวประมาณค่าที่มีประสิทธิภาพดีกว่า* $\boldsymbol{C.I.}_2(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$ *ก็ต่อเมื่อ* $\boldsymbol{ACP}\left(\boldsymbol{C.I.}_1(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})\right)\ge \boldsymbol{ACP}\left(\boldsymbol{C.I.}_2(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})\right)$

บทที่ 4
ผลการวิจัย

ผลการศึกษาเกี่ยวกับการประมาณค่าผลรวมประชากร ในการสำรวจตัวอย่าง เมื่อมีข้อมูลสูญหาย เป็นดังนี้

4.1 ตัวประมาณแบบจุดของผลรวมประชากร

กรณีที่ประชากร $\boldsymbol{U}$ ขนาด $\boldsymbol{N}$ ถูกแบ่งออกเป็น $\boldsymbol{H}$ ชั้นภูมิ $\boldsymbol{U}_1,\boldsymbol{U}_2,...,\boldsymbol{U}_h,...,\boldsymbol{U}_L$ ขนาด $\boldsymbol{N}_1,\boldsymbol{N}_2,...,\boldsymbol{N}_h,...,\boldsymbol{N}_L$ ตามลำดับ สมมติว่าในแต่ละชั้นภูมิ $\boldsymbol{h}$ ตัวอย่าง $\boldsymbol{s}_h$ ขนาด $\boldsymbol{n}_h$ ถูกสุ่มประชากร $\boldsymbol{U}_h$ ขนาด $\boldsymbol{N}_h$ เมื่อ $\boldsymbol{h}\in\boldsymbol{H}$ โดยใช้แผนการสุ่มตัวอย่างแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มตัวอย่างโดยใช้แผนการสุ่มตัวอย่างของ Midzuno กรณีที่ไม่มีข้อมูลสูญหาย การประมาณค่าผลรวมประชากร สามารถใช้ตัวประมาณค่าตามที่ Horvitz and Thompson (1952) นำเสนอดังนี้

$$\hat{\boldsymbol{Y}}=\sum_{\boldsymbol{h}\in\boldsymbol{H}}\sum_{\boldsymbol{i}\in\boldsymbol{s}_h}\frac{\boldsymbol{y}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}} \qquad (4.1)$$

เมื่อ $\pi_{\boldsymbol{hi}}$ กำหนดค่าตามสมการที่ (3.1) กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม ค่า $\boldsymbol{y}_{\boldsymbol{hi}}$ เมื่อ $\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{hm}}$ จะไม่สามารถสังเกตุค่าได้ ดังนั้นจำเป็นต้องแทนที่ข้อมูลสูญหายด้วยค่าประมาณค่าจากการคำนวณ ซึ่งการศึกษาในครั้งนี้ผู้วิจัยเลือกใช้การแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย เขียนแทนด้วยสัญลักษณ์ $\boldsymbol{y}^*_{\boldsymbol{hi}}$ ซึ่ง $\boldsymbol{y}^*_{\boldsymbol{hi}}=\sum_{\boldsymbol{i}\in\boldsymbol{s}_h}\frac{\boldsymbol{r}_{\boldsymbol{hi}}\boldsymbol{y}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}\Big/\sum_{\boldsymbol{i}\in\boldsymbol{s}_h}\frac{\boldsymbol{r}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}$ Deville and Särndal (1994) กล่าวว่าหลังจากประมาณค่าข้อมูลสูญหาย เซตของตัวอย่าง $\boldsymbol{s}_h$ ขนาด $\boldsymbol{n}_h$ ในแต่ละชั้นภูมิจะไม่มีข้อมูลสูญหาย เนื่องจากค่าข้อมูลสูญหายในเซต $\boldsymbol{s}_{\boldsymbol{hm}}=\boldsymbol{s}_h-\boldsymbol{s}_{\boldsymbol{hr}}$ ขนาด $\boldsymbol{n}_{\boldsymbol{hm}}=\boldsymbol{n}_h-\boldsymbol{n}_{\boldsymbol{hr}}$ ถูกแทนที่ด้วยค่า $\boldsymbol{y}^*_{\boldsymbol{hi}}$ เมื่อ $\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{hm}}$ กำหนดให้ $\boldsymbol{z}^*_{\boldsymbol{h}}=\{\boldsymbol{z}^*_{\boldsymbol{h}1},\boldsymbol{z}^*_{\boldsymbol{h}2},...,\boldsymbol{z}^*_{\boldsymbol{hn}_h}\}$ แทนเซตของข้อมูลที่สมบูรณ์ หลังจากที่แทนที่ข้อมูลสูญหายค่าเฉลี่ย ในชั้นภูมิที่ $\boldsymbol{h}$ เมื่อ $\boldsymbol{h}\in\boldsymbol{H}$ ซึ่ง $\boldsymbol{z}^*_{\boldsymbol{hi}}=\boldsymbol{y}_{\boldsymbol{hi}}$ เมื่อหน่วยตัวอย่างที่ $\boldsymbol{i}$ ไม่สูญหาย และ $\boldsymbol{z}^*_{\boldsymbol{hi}}=\boldsymbol{y}^*_{\boldsymbol{hi}}$ เมื่อหน่วยตัวอย่างที่ $\boldsymbol{i}$ สูญหาย ดังนั้นการประมาณค่าผลรวมประชากรสามารถทำได้โดยใช้ตัวประมาณค่าที่ Horvitz and Thompson (1952) นำเสนอ ตามที่แสดงในสมการ (4.2) และเรียกตัวประมาณค่าดังกล่าวว่าตัวประมาณค่าแบบเชิงเส้น

$$\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}}=\sum_{\boldsymbol{h}\in\boldsymbol{H}}\sum_{\boldsymbol{i}\in\boldsymbol{s}_h}\frac{\boldsymbol{z}^*_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}} \qquad (4.2)$$

เมื่อ $\pi_{\boldsymbol{hi}}\ ;\boldsymbol{h}\in\boldsymbol{H}$ กำหนดค่าตามสมการที่ (3.1) และจากสมการที่ (4.2) สามารถเขียนตัวประมาณค่าผลรวมประชากรให้อยู่ในรูปไม่เป็นเชิงเส้น และเรียกตัวประมาณค่าดังกล่าวว่าตัวประมาณค่าแบบไม่เป็นเชิงเส้น ตามที่แสดงในสมการ (4.3)

$$\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}=\sum_{\boldsymbol{h}\in\boldsymbol{H}}\left[\sum_{\boldsymbol{i}\in\boldsymbol{s}_h}\frac{\boldsymbol{r}_{\boldsymbol{hi}}\boldsymbol{y}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}+\sum_{\boldsymbol{i}\in\boldsymbol{s}_h}\frac{(1-\boldsymbol{r}_{\boldsymbol{hi}})\boldsymbol{y}^*_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}\right]$$

$$= \sum_{h \in H} \left[\frac{\sum_{i \in s_h} \frac{1}{\pi_{hi}} \sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi}}}{\sum_{i \in s_h} \frac{r_{hi}}{\pi_{hi}}} \right] \tag{4.3}$$

เมื่อ π_{hi} ; $h \in H$ กำหนดค่าตามที่แสดงในสมการที่ (3.1)

จากสมการที่ (4.2) และสมการที่ (4.3) ผู้วิจัยได้นำเสนอตัวประมาณค่าผลรวมประชากรแบบเชิงเส้น และตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น ดังนั้นตัวประมาณค่าแบบจุดของผลรวมประชากรที่ผู้วิจัยนำเสนอ กำหนดค่าตามที่แสดงในสมการ (4.4)

$$\hat{Y}_{MI} = \sum_{h \in H} \sum_{i \in s_h} \frac{z^*_{hi}}{\pi_{hi}} = \sum_{h \in H} \left[\frac{\sum_{i \in s_h} \frac{1}{\pi_{hi}} \sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi}}}{\sum_{i \in s_h} \frac{r_{hi}}{\pi_{hi}}} \right] \tag{4.4}$$

จากสมการที่ (4.4) จะได้ว่า $\hat{Y}_{MI}$ เป็นตัวประมาณค่าที่ไม่เอนเอียงโดยประมาณของ ของ Y ตามที่แสดงในทฤษฎี 4.1

ทฤษฎี 4.1 *กำหนดให้ $\hat{Y}_{MI}$ เป็นตัวประมาณค่าของผลรวมประชากร Y ซึ่ง $\hat{Y}_{MI}$ กำหนดค่าดังนี้*

$$\hat{Y}_{MI} = \sum_{h \in H} \sum_{i \in s_h} \frac{z^*_{hi}}{\pi_{hi}} = \sum_{h \in H} \left[\frac{\sum_{i \in s_h} \frac{1}{\pi_{hi}} \sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi}}}{\sum_{i \in s_h} \frac{r_{hi}}{\pi_{hi}}} \right]$$

ภายใต้เฟรมเวิร์คแบบรีเวิร์ส $\hat{Y}_{MI}$ จะเป็นตัวประมาณค่าที่ไม่เอนเอียงโดยประมาณของ Y เมื่อตัวอย่างเข้าใกล้อนันต์ $(n \to \infty)$

พิสูจน์ กำหนดให้ $\hat{Y}_{MI} = \sum_{h \in H} \left[\frac{\sum_{i \in s_h} \frac{1}{\pi_{hi}} \sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi}}}{\sum_{i \in s_h} \frac{r_{hi}}{\pi_{hi}}} \right]$ ภายใต้เฟรมเวิร์คแบบรีเวิร์ส ค่าคาดหมายของ $\hat{Y}_{MI}$ สามารถหาได้จาก

$$E\left(\hat{Y}_{MI}\right) = E_R E_S \left(\hat{Y}_{MI} \middle| R\right) = E_R E_S \left(\sum_{h \in H} \left[\frac{\sum_{i \in s_h} \frac{1}{\pi_{hi}} \sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi}}}{\sum_{i \in s_h} \frac{r_{hi}}{\pi_{hi}}} \right] \middle| R \right)$$

$$= E_R E_S \left(\sum_{h \in H} \left[\frac{\sum_{i \in U_h} \frac{I_{hi}}{\pi_{hi}} \sum_{i \in U_h} \frac{r_{hi} y_{hi} I_{hi}}{\pi_{hi}}}{\sum_{i \in sU} \frac{r_{hi} I_{hi}}{\pi_{hi}}} \right] \middle| R \right)$$

$$\approx \sum_{h\in H}\left[\frac{\sum_{i\in U_h}\frac{E_S(I_{hi})}{\pi_{hi}}\sum_{i\in U_h}\frac{E_R(r_{hi})y_{hi}E_S(I_{hi})}{\pi_{hi}}}{\sum_{i\in sU}\frac{E_R(r_{hi})E_S(I_{hi})}{\pi_{hi}}}\right]$$

$$=\sum_{h\in H}\left[\frac{\sum_{i\in U_h}\frac{\pi_{hi}}{\pi_{hi}}\sum_{i\in U_h}\frac{p_h y_{hi}\pi_{hi}}{\pi_{hi}}}{\sum_{i\in U_h}\frac{p_h\pi_{hi}}{\pi_{hi}}}\right]$$

$$=\sum_{h\in H}\left[\frac{\sum_{i\in U_h}1\, p_h\sum_{i\in U_h}y_{hi}}{p_h\sum_{i\in U_h}1}\right]$$

$$=\sum_{h\in H}\sum_{i\in U_h}y_{hi}$$

$$=Y$$

ดังนั้น $E\left(\hat{Y}_{MI}\right)\approx Y$ หรือ $\hat{Y}_{MI}=\sum_{h\in H}\left[\frac{\sum_{i\in s_h}\frac{1}{\pi_{hi}}\sum_{i\in s_h}\frac{r_{hi}y_{hi}}{\pi_{hi}}}{\sum_{i\in s_h}\frac{r_{hi}}{\pi_{hi}}}\right]$ เป็นตัวประมาณค่าที่ไม่เอนเอียงโดยประมาณของ Y นอกจากนี้ $\sum_{h\in H}\left[\frac{\sum_{i\in s_h}\frac{1}{\pi_{hi}}\sum_{i\in s_h}\frac{r_{hi}y_{hi}}{\pi_{hi}}}{\sum_{i\in s_h}\frac{r_{hi}}{\pi_{hi}}}\right]=\sum_{h\in H}\sum_{i\in s_h}\frac{z^*_{hi}}{\pi_{hi}}$ ทำให้ $\sum_{h\in H}\sum_{i\in s_h}\frac{z^*_{hi}}{\pi_{hi}}$ เป็นตัวประมาณค่าที่ไม่เอนเอียงโดยประมาณของ Y เมื่อตัวอย่างเข้าใกล้อนันต์

4.2 ค่าความแปรปรวนและตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบจุด

หัวข้อที่แล้วผู้วิจัยได้นำเสนอตัวประมาณค่าผลรวมประชากรแบบจุด กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม และแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย ซึ่งพิจารณาทั้งรูปแบบที่เป็นตัวประมาณค่าแบบเชิงเส้น และตัวประมาณค่าแบบไม่เป็นเชิงเส้น สำหรับหัวข้อนี้ผู้วิจัยจะได้ นำเสนอตัวประมาณค่าความแปรปรวนแบบเชิงเส้น และตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น

4.2.1 ตัวประมาณค่าความแปรปรวนแบบเป็นเชิงเส้น

จากสมการ (4.2) ตัวประมาณค่าผลรวมประชากรแบบเชิงเส้นมีค่าเป็น

$$\hat{Y}_{LMI}=\sum_{h\in H}\sum_{i\in s_h}\frac{z^*_{hi}}{\pi_{hi}} \qquad (4.5)$$

จากสมการ (4.5) เซตข้อมูล $z^*_h=\{z^*_{h1},z^*_{h2},\ldots,z^*_{hm_h}\}$ ในแต่ละชั้นภูมิเป็นเซตของข้อมูลที่ไม่เกิดการสูญหาย ดังนั้นค่าความแปรปรวนและตัวประมาณค่าความแปรปรวนของ $\hat{Y}_{LMI}$ มีค่าตามที่แสดงใน ทฤษฎีบทที่ 4.2

ทฤษฎี 4.2 *ภายใต้การสุ่มแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มตัวอย่าง* s_h *ขนาด* n_h *ด้วยแผนการสุ่มตัวอย่างของ Midzuno กำหนดให้* $z_h^* = \{z_{h1}^*, z_{h2}^*, ..., z_{hn_h}^*\}$ *แทนเซตของข้อมูลที่สมบูรณ์ หลังจากที่แทนที่ข้อมูลสูญหายค่าเฉลี่ย ในชั้นภูมิที่* h *เมื่อ* $h \in H$ *ซึ่ง* $z_{hi}^* = y_{hi}$ *เมื่อหน่วยตัวอย่างที่* i *ไม่ สูญหาย และ* $z_{hi}^* = y_{hi}^*$ *เมื่อหน่วยตัวอย่างที่* i *สูญหาย และตัวประมาณค่าผลรวมประชากรแบบเป็นเชิงเส้นกำหนดค่าตามที่แสดงในสมการ (4.5)*

(1) ความแปรปรวนของ $\hat{Y}_{LMI}$ เขียนแทนด้วย $V(\hat{Y}_{LMI})$ และกำหนดค่าตามที่แสดงในสมการ (4.6)

$$V(\hat{Y}_{LMI}) = \sum_{h\in H}\sum_{i\in U_h}\left(\frac{1-\pi_{hi}}{\pi_{hi}}\right)z_{hi}^{*2} + \sum_{h\in H}\sum_{i\in U_h}\sum_{j\neq i\in U_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hi}\pi_{hj}}\right)z_{hi}^* z_{hj}^* \qquad (4.6)$$

(2) ตัวประมาณค่าของ $V(\hat{Y}_{LMI})$ เขียนแทนด้วย $\hat{V}(\hat{Y}_{LMI})$ และกำหนดค่าตามที่แสดงในสมการที่ (4.7)

$$\hat{V}(\hat{Y}_{LMI}) = \sum_{h\in H}\sum_{i\in s_h}\left(\frac{1-\pi_{hi}}{\pi_{hi}^2}\right)z_{hi}^{*2} + \sum_{h\in H}\sum_{i\in s_h}\sum_{j\neq i\in s_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hij}\pi_{hi}\pi_{hj}}\right)z_{hi}^* z_{hj}^* \qquad (4.7)$$

(3) $\hat{V}(\hat{Y}_{LMI})$ เป็นตัวประมาณค่าที่ไม่เอนเอียงของ $V(\hat{Y}_{LMI})$ เมื่อขนาดตัวอย่างเข้าใกล้อนันต์

พิสูจน์

(1) จากสมการ (3.4) และ (4.5) ความแปรปรวนของ $\hat{Y}_{LMI}$ สามารถหาได้จาก

$$V(\hat{Y}_{LMI}) = V\left(\sum_{h\in H}\sum_{i\in s_h}\frac{z_{hi}^*}{\pi_{hi}}\right)$$

เนื่องจากการสุ่มตัวอย่างในแต่ละชั้นภูมิเป็นอิสระกัน ดังนั้น

$$V(\hat{Y}_{LMI}) = V\left(\sum_{h\in H}\sum_{i\in s_h}\frac{z_{hi}^*}{\pi_{hi}}\right) = \sum_{h\in H}V\left(\sum_{i\in s_h}\frac{z_{hi}^*}{\pi_{hi}}\right) \qquad (4.8)$$

จากสมการ (4.8) $\sum_{i\in s_h}\frac{z_{hi}^*}{\pi_{hi}}$ อยู่ในรูปตัวประมาณแบบ Horvitz และ Thompson (1952) ดังนั้น

$$V\left(\sum_{i\in s_h}\frac{z_{hi}^*}{\pi_{hi}}\right) = \sum_{i\in U_h}\left(\frac{1-\pi_{hi}}{\pi_{hi}}\right)z_{hi}^{*2} + \sum_{i\in U_h}\sum_{j\neq i\in U_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hi}\pi_{hj}}\right)z_{hi}^* z_{hj}^* \qquad (4.9)$$

แทนค่า (4.9) ใน (4.8) ได้ว่า

$$V(\hat{Y}_{LMI}) = \sum_{h\in H}\sum_{i\in U_h}\left(\frac{1-\pi_{hi}}{\pi_{hi}}\right)z_{hi}^{*2} + \sum_{h\in H}\sum_{i\in U_h}\sum_{j\neq i\in U_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hi}\pi_{hj}}\right)z_{hi}^* z_{hj}^*$$

(2) จากสมการที่ (3.5) และสมการที่ (4.5) จะได้ว่า

$$\hat{V}(\hat{Y}_{LMI}) = \hat{V}_S(\hat{Y}_{LMI} \mid R) = \hat{V}_S\left(\sum_{h\in H}\sum_{i\in s_h}\frac{z_{hi}^*}{\pi_{hi}}\middle| R\right) \qquad (4.10)$$

จากสมการ (4.8) $\sum_{i\in s_h}\frac{z^*_{hi}}{\pi_{hi}}$ อยู่ในรูปตัวประมาณแบบ Horvitz and Thompson (1952) ดังนั้นตัวประมาณค่าความแปรปรวนของ $\sum_{i\in s_h}\frac{z^*_{hi}}{\pi_{hi}}$ เทียบกับแผนการสุ่มตัวอย่าง เมื่อกำหนดค่า R มีค่าเป็น

$$\hat{V}(\hat{Y}_{LMI})=\sum_{h\in H}\sum_{i\in s_h}\left(\frac{1-\pi_{hi}}{\pi_{hi}^2}\right)z_{hi}^{*2}+\sum_{h\in H}\sum_{i\in s_h}\sum_{j\neq i\in s_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hij}\pi_{hi}\pi_{hj}}\right)z^*_{hi}z^*_{hj}$$

(3) จากสมการ (4.7) $\hat{V}(\hat{Y}_{LMI})$ มีค่าเป็น

$$\hat{V}(\hat{Y}_{LMI})=\sum_{h\in H}\sum_{i\in s_h}\left(\frac{1-\pi_{hi}}{\pi_{hi}^2}\right)z_{hi}^{*2}+\sum_{h\in H}\sum_{i\in s_h}\sum_{j\neq i\in s_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hij}\pi_{hi}\pi_{hj}}\right)z^*_{hi}z^*_{hj} \qquad (4.11)$$

จากสมการ (4.11) ค่าคาดหมายของ $\hat{V}(\hat{Y}_{LMI})$ มีค่าเป็น

$$\begin{aligned}
E\left[\hat{V}(\hat{Y}_{LMI})\right]&=E\left[\sum_{h\in H}\sum_{i\in s_h}\left(\frac{1-\pi_{hi}}{\pi_{hi}^2}\right)z_{hi}^{*2}+\sum_{h\in H}\sum_{i\in s_h}\sum_{j\neq i\in s_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hij}\pi_{hi}\pi_{hj}}\right)z^*_{hi}z^*_{hj}\right]\\
&=E\left[\sum_{h\in H}\sum_{i\in U_h}\left(\frac{1-\pi_{hi}}{\pi_{hi}^2}\right)z_{hi}^{*2}I_{hi}+\sum_{h\in H}\sum_{i\in U_h}\sum_{j\neq i\in U_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hij}\pi_{hi}\pi_{hj}}\right)z^*_{hi}z^*_{hj}I_{hi}I_{hj}\right]\\
&=\sum_{h\in H}\sum_{i\in U_h}\left(\frac{1-\pi_{hi}}{\pi_{hi}^2}\right)z_{hi}^{*2}E(I_{hi})+\sum_{h\in H}\sum_{i\in U_h}\sum_{j\neq i\in U_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hij}\pi_{hi}\pi_{hj}}\right)z^*_{hi}z^*_{hj}E(I_{hi}I_{hj})\\
&=\sum_{h\in H}\sum_{i\in U_h}\left(\frac{1-\pi_{hi}}{\pi_{hi}^2}\right)z_{hi}^{*2}\pi_{hi}+\sum_{h\in H}\sum_{i\in U_h}\sum_{j\neq i\in U_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hij}\pi_{hi}\pi_{hj}}\right)z^*_{hi}z^*_{hj}\pi_{hij}\\
&=\sum_{h\in H}\sum_{i\in U_h}\left(\frac{1-\pi_{hi}}{\pi_{hi}}\right)z_{hi}^{*2}+\sum_{h\in H}\sum_{i\in U_h}\sum_{j\neq i\in U_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hi}\pi_{hj}}\right)z^*_{hi}z^*_{hj}\\
&=V(\hat{Y}_{LMI})
\end{aligned}$$

4.2.2 ตัวประมาณค่าความแปรปรวนแบบไม่เป็นเชิงเส้น

จากสมการ (4.3) ตัวประมาณค่าแบบไม่เป็นเชิงเส้นมีค่าเป็น

$$\hat{Y}_{NLMI}=\sum_{h\in H}\left[\frac{\sum_{i\in s_h}\frac{1}{\pi_{hi}}\sum_{i\in s_h}\frac{r_{hi}y_{hi}}{\pi_{hi}}}{\sum_{i\in s_h}\frac{r_{hi}}{\pi_{hi}}}\right] \qquad (4.12)$$

สำหรับการหาค่าความแปรปรวนของ $\hat{Y}_{NLMI}$ ผู้วิจัยกำหนดเงื่อนไขเพิ่มเติมดังนี้

(1) ในแต่ละชั้นภูมิ $p_{hi}=p_h$ ทุกค่า $i\in U$

(2) สัดส่วนตัวอย่าง $f=\frac{n}{N}$ มีขนาดเล็ก

จากสมการ (3.4) และสมการ (4.12) จะได้ว่าค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้นมีค่าเป็น

$$\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}) \approx \boldsymbol{E}_{\boldsymbol{R}}\boldsymbol{V}_{\boldsymbol{S}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}} \mid \boldsymbol{R}) \tag{4.13}$$

ค่าความแปรปรวนและตัวประมาณค่าความแปรปรวนของ $\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}$ มีค่าตามที่แสดงใน ทฤษฎีบทที่ 4.3

ทฤษฎี 4.3 *กำหนดให้เงื่อนไข (1) และ (2) เป็นจริง ภายใต้การสุ่มแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มตัวอย่าง* $\boldsymbol{s}_{\boldsymbol{h}}$ *ขนาด* $\boldsymbol{n}_{\boldsymbol{h}}$ *ด้วยแผนการสุ่มตัวอย่างของ Midzuno ภายใต้เฟรมเวิร์คแบบรีเวิร์ส*

(1) ความแปรปรวนของ $\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}$ เขียนแทนด้วย $\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ และกำหนดค่าดังนี้

$$\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}) = \sum_{\boldsymbol{h}\in\boldsymbol{H}}\sum_{\boldsymbol{i}\in\boldsymbol{U}_{\boldsymbol{h}}}\frac{(1-\pi_{\boldsymbol{hi}})}{\pi_{\boldsymbol{hi}}}\boldsymbol{y}_{\boldsymbol{hi}}^{2} + \sum_{\boldsymbol{h}\in\boldsymbol{H}}\sum_{\boldsymbol{i}\in\boldsymbol{U}_{\boldsymbol{h}}}\sum_{\boldsymbol{j}\neq\boldsymbol{i}\in\boldsymbol{U}_{\boldsymbol{h}}}\frac{(\pi_{\boldsymbol{hij}}-\pi_{\boldsymbol{hi}}\pi_{\boldsymbol{hj}})}{\pi_{\boldsymbol{hi}}\pi_{\boldsymbol{hj}}}\boldsymbol{y}_{\boldsymbol{hi}}\boldsymbol{y}_{\boldsymbol{hj}} \tag{4.14}$$

(2) ตัวประมาณค่าของ $\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ เขียนแทนด้วย $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ และกำหนดค่าดังนี้

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}) = \sum_{\boldsymbol{h}\in\boldsymbol{H}}\frac{1}{\left(\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{r}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}\right)^{2}}\left[\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{(1-\pi_{\boldsymbol{hi}})}{\pi_{\boldsymbol{hi}}^{2}}\hat{\boldsymbol{z}}_{\boldsymbol{hi}}^{2} + \sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\sum_{\boldsymbol{j}\neq\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{(\pi_{\boldsymbol{hij}}-\pi_{\boldsymbol{hi}}\pi_{\boldsymbol{hj}})}{\pi_{\boldsymbol{hij}}\pi_{\boldsymbol{hi}}\pi_{\boldsymbol{hj}}}\hat{\boldsymbol{z}}_{\boldsymbol{hi}}\hat{\boldsymbol{z}}_{\boldsymbol{hj}}\right] \tag{4.15}$$

เมื่อ $\hat{\boldsymbol{z}}_{\boldsymbol{hi}} = \sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{r}_{\boldsymbol{hi}}\boldsymbol{y}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}} + \left(\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{1}{\pi_{\boldsymbol{hi}}}\right)\boldsymbol{r}_{\boldsymbol{hi}}(\boldsymbol{y}_{\boldsymbol{hi}} - \hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Rh}})$ ค่า $\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Rh}} = \frac{\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{r}_{\boldsymbol{hi}}\boldsymbol{y}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}}{\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{r}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}}$ และ $\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}};\boldsymbol{h}\in\boldsymbol{H}$

(3) $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ เป็นตัวประมาณค่าที่ไม่เอนเอียงเมื่อขนาดตัวอย่างเข้าใกล้อนันต์ ของ $\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$

พิสูจน์

(1) จากสมการ (4.13)

$$\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}) \approx \boldsymbol{E}_{\boldsymbol{R}}\boldsymbol{V}_{\boldsymbol{S}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}} \mid \boldsymbol{R}) \tag{4.16}$$

การหาค่า $\boldsymbol{V}_{\boldsymbol{S}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}} \mid \boldsymbol{R})$ ในสมการ (4.16) จากสมการ (4.12) จะเห็นว่า $\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}$ เป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้น ดังนั้นการหาค่าความแปรปรวน จำเป็นต้องเขียน $\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}$ ให้อยู่ในรูปของฟังก์ชันปรับเรียบ (Smooth function) เขียนแทนด้วยสัญลักษณ์ ψ จากนั้นใช้เทเลอร์แบบเชิงเส้น เพื่อปรับฟังก์ชันปรับเรียบของ $\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}$ ให้เป็นเชิงเส้นดังนี้ จากสมการที่ (4.12) สามารถเขียน

$$\begin{aligned}\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}} &= \sum_{\boldsymbol{h}\in\boldsymbol{H}}\left[\frac{\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{1}{\pi_{\boldsymbol{hi}}}\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{r}_{\boldsymbol{hi}}\boldsymbol{y}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}}{\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{r}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}}\right] \\ &= \sum_{\boldsymbol{h}\in\boldsymbol{H}}\frac{\hat{\boldsymbol{T}}_{\boldsymbol{h}1}\hat{\boldsymbol{T}}_{\boldsymbol{h}2}}{\hat{\boldsymbol{T}}_{\boldsymbol{h}3}} = \frac{\hat{\boldsymbol{T}}_{11}\hat{\boldsymbol{T}}_{12}}{\hat{\boldsymbol{T}}_{13}} + \frac{\hat{\boldsymbol{T}}_{21}\hat{\boldsymbol{T}}_{22}}{\hat{\boldsymbol{T}}_{23}} + \ldots + \frac{\hat{\boldsymbol{T}}_{\boldsymbol{H}1}\hat{\boldsymbol{T}}_{\boldsymbol{H}2}}{\hat{\boldsymbol{T}}_{\boldsymbol{H}3}} = \psi(\hat{\boldsymbol{T}})\end{aligned} \tag{4.17}$$

เมื่อ

$$\hat{\boldsymbol{T}} = \left[\sum_{i\in s_1} \frac{1}{\pi_{1i}} \;\; \sum_{i\in s_1} \frac{\boldsymbol{r}_{1i}\boldsymbol{y}_{1i}}{\pi_{1i}} \;\; \sum_{i\in s_1} \frac{\boldsymbol{r}_{1i}}{\pi_{1i}} \;\; \cdots \sum_{i\in s_H} \frac{1}{\pi_{Hi}} \;\; \sum_{i\in s_H} \frac{\boldsymbol{r}_{Hi}\boldsymbol{y}_{Hi}}{\pi_{Hi}} \;\; \sum_{i\in s_H} \frac{\boldsymbol{r}_{Hi}}{\pi_{Hi}} \right]' = \left[\hat{\boldsymbol{T}}_{11} \;\; \hat{\boldsymbol{T}}_{12} \;\; \hat{\boldsymbol{T}}_{13} \cdots \hat{\boldsymbol{T}}_{H1} \;\; \hat{\boldsymbol{T}}_{H2} \;\; \hat{\boldsymbol{T}}_{H3} \right]'$$

$$\boldsymbol{T} = \boldsymbol{E}_S(\hat{\boldsymbol{T}}) = \left[\boldsymbol{N}_1 \sum_{i\in U_1} \boldsymbol{r}_{1i}\boldsymbol{y}_{1i} \;\; \sum_{i\in U_1} \boldsymbol{r}_{1i} \;\; \cdots \boldsymbol{N}_h \sum_{i\in U_H} \boldsymbol{r}_{Hi}\boldsymbol{y}_{Hi} \;\; \sum_{i\in U_H} \boldsymbol{r}_{Hi} \right]' = \left[\boldsymbol{T}_{11} \;\; \boldsymbol{T}_{12} \;\; \boldsymbol{T}_{13} \;\; \cdots \boldsymbol{T}_{H1} \;\; \boldsymbol{T}_{H2} \;\; \boldsymbol{T}_{H3} \right]'$$

ดังนั้นเทเลอร์แบบเชิงเส้นของ $\hat{\boldsymbol{Y}}_{NLMI} = \psi(\hat{\boldsymbol{T}})$ รอบค่า $\boldsymbol{T}$ กำหนดค่าดังนี้

$$\hat{\boldsymbol{Y}}_{NLMI} = \psi(\hat{\boldsymbol{T}}) \approx \psi(\boldsymbol{T}) + \sum_{h\in H} \sum_{j=1}^{3H} \left. \frac{\partial \psi(\hat{\boldsymbol{T}})}{\partial \hat{\boldsymbol{T}}_{hj}} \right|_{\hat{\boldsymbol{T}}=\boldsymbol{T}} (\hat{\boldsymbol{T}}_{hj} - \boldsymbol{T}_{hj})$$

$$= \boldsymbol{Constant} + \sum_{h\in H} \frac{1}{\sum_{i\in U_h} \boldsymbol{r}_{hi}} \sum_{i\in s_h} \frac{\boldsymbol{z}_{hi}}{\pi_{hi}} \qquad (4.18)$$

เมื่อ $\boldsymbol{z}_{hi} = \sum_{i\in U_h} \boldsymbol{r}_{hi}\boldsymbol{y}_{hi} + \boldsymbol{N}_h \boldsymbol{r}_{hi}(\boldsymbol{y}_{hi} - \bar{\boldsymbol{Y}}_{Rh})$ และ $\bar{\boldsymbol{Y}}_{Rh} = \frac{\sum_{i\in U_h} \boldsymbol{r}_{hi}\boldsymbol{y}_{hi}}{\sum_{i\in U_h} \boldsymbol{r}_{hi}}$ และจากสมการที่ (4.18) สามารถหาค่าความแปรปวนของ $\hat{\boldsymbol{Y}}_{NLMI}$ เทียบกับแผนการสุ่มตัวอย่าง เมื่อกำหนด $\boldsymbol{R}$ หรือ $\boldsymbol{V}_S(\hat{\boldsymbol{Y}}_{NLMI} \mid \boldsymbol{R}) = \boldsymbol{V}_S(\psi(\hat{\boldsymbol{T}}) \mid \boldsymbol{R})$ ตามวิธีของ Horvitz and Thompson (1952) ดังนี้

$$\boldsymbol{V}_S(\hat{\boldsymbol{Y}}_{NLMI} \mid \boldsymbol{R}) = \boldsymbol{V}_S(\psi(\hat{\boldsymbol{T}}) \mid \mathrm{R}) = \boldsymbol{V}_S\left(\boldsymbol{Constant} + \sum_{h\in H} \frac{1}{\boldsymbol{T}_{h3}} \sum_{i\in s_h} \frac{\boldsymbol{z}_{hi}}{\pi_{hi}} \middle| \boldsymbol{R} \right)$$

$$= \boldsymbol{V}_S\left(\sum_{h\in H} \frac{1}{\boldsymbol{T}_{h3}} \sum_{i\in s_h} \frac{\boldsymbol{z}_{hi}}{\pi_{hi}} \middle| \mathrm{R} \right)$$

$$= \sum_{h\in H} \frac{1}{\left(\sum_{i\in U_h} \boldsymbol{r}_{hi} \right)^2} \left[\sum_{i\in U_h} \frac{(1-\pi_{hi})}{\pi_{hi}} \boldsymbol{z}_{hi}^2 + \sum_{i\in U_h} \sum_{j\neq i\in U_h} \frac{(\pi_{hij} - \pi_{hi}\pi_{hj})}{\pi_{hi}\pi_{hj}} \boldsymbol{z}_{hi}\boldsymbol{z}_{hj} \right] \qquad (4.19)$$

เมื่อ $\boldsymbol{z}_{hi}$ กำหนดค่าตามสมการที่ (4.18) ค่า π_{hi} และ π_{hij} กำหนดค่าตามสมการที่ (3.1) และสมการที่ (3.2) ตามลำดับ แทนค่า (4.19) ในสมการ (4.16) ได้ว่า

$$\boldsymbol{V}(\hat{\boldsymbol{Y}}_{NLMI}) \approx \boldsymbol{E}_R \boldsymbol{V}_S(\hat{\boldsymbol{Y}}_{NLMI} \mid \boldsymbol{R})$$

$$= \boldsymbol{E}_R \left[\sum_{h\in H} \frac{1}{\left(\sum_{i\in U_h} \boldsymbol{r}_{hi} \right)^2} \left[\sum_{i\in U_h} \frac{(1-\pi_{hi})}{\pi_{hi}} \boldsymbol{z}_{hi}^2 + \sum_{i\in U_h} \sum_{j\neq i\in U_h} \frac{(\pi_{hij} - \pi_{hi}\pi_{hj})}{\pi_{hi}\pi_{hj}} \boldsymbol{z}_{hi}\boldsymbol{z}_{hj} \right] \right]$$

$$\approx \sum_{h\in H} \frac{1}{\left(\sum_{i\in U_h} \boldsymbol{E}_R(\boldsymbol{r}_{hi}) \right)^2} \left[\sum_{i\in U_h} \frac{(1-\pi_{hi})}{\pi_{hi}} \boldsymbol{E}_R(\boldsymbol{z}_{hi}^2) + \sum_{i\in U_h} \sum_{j\neq i\in U_h} \frac{(\pi_{hij} - \pi_{hi}\pi_{hj})}{\pi_{hi}\pi_{hj}} \boldsymbol{E}_R(\boldsymbol{z}_{hi}) \boldsymbol{E}_R(\boldsymbol{z}_{hj}) \right]$$

$$\approx \sum_{h\in H}\frac{1}{\left(\sum_{i\in U_h} E_R(r_{hi})\right)^2}\left[\sum_{i\in U_h}\frac{(1-\pi_{hi})}{\pi_{hi}}\left(E_R(z_{hi})\right)^2+\sum_{i\in U_h}\sum_{j\neq i\in U_h}\frac{(\pi_{hij}-\pi_{hi}\pi_{hj})}{\pi_{hi}\pi_{hj}}E_R(z_{hi})E_R(z_{hj})\right]$$

$$=\sum_{h\in H}\frac{1}{\left(\sum_{i\in U_h} p_h\right)^2}\left[\sum_{i\in U_h}\frac{(1-\pi_{hi})}{\pi_{hi}}\left(p_hN_hy_{hi}\right)^2+\sum_{i\in U_h}\sum_{j\neq i\in U_h}\frac{(\pi_{hij}-\pi_{hi}\pi_{hj})}{\pi_{hi}\pi_{hj}}p_hN_hy_{hi}p_hN_hy_{hj}\right]$$

$$=\sum_{h\in H}\frac{1}{\left(N_hp_h\right)^2}\left[\sum_{i\in U_h}\frac{(1-\pi_{hi})}{\pi_{hi}}N_h^2p_h^2y_{hi}^2+\sum_{i\in U_h}\sum_{j\neq i\in U_h}\frac{(\pi_{hij}-\pi_{hi}\pi_{hj})}{\pi_{hi}\pi_{hj}}N_h^2p_h^2y_{hi}y_{hj}\right]$$

$$=\sum_{h\in H}\sum_{i\in U_h}\frac{(1-\pi_{hi})}{\pi_{hi}}y_{hi}^2+\sum_{h\in H}\sum_{i\in U_h}\sum_{j\neq i\in U_h}\frac{(\pi_{hij}-\pi_{hi}\pi_{hj})}{\pi_{hi}\pi_{hj}}y_{hi}y_{hj}$$

ดังนั้น

$$V(\hat{Y}_{NLMI})\approx\sum_{h\in H}\sum_{i\in U_h}\frac{(1-\pi_{hi})}{\pi_{hi}}y_{hi}^2+\sum_{h\in H}\sum_{i\in U_h}\sum_{j\neq i\in U_h}\frac{(\pi_{hij}-\pi_{hi}\pi_{hj})}{\pi_{hi}\pi_{hj}}y_{hi}y_{hj} \qquad (4.20)$$

(2) จากสมการ (3.5) และสมการ (4.20) ตัวประมาณค่าของ $V(\hat{Y}_{NLMI})$ เขียนแทนด้วยสัญลักษณ์ $\hat{V}(\hat{Y}_{NLMI})$ มีค่าเป็น

$$\hat{V}(\hat{Y}_{NLMI})=\hat{V}_S(\hat{Y}_{NLMI}\mid R)$$

$$=\sum_{h\in H}\frac{1}{\left(\sum_{i\in s_h}\frac{r_{hi}y_{hi}}{\pi_{hi}}\right)^2}\left[\sum_{i\in s_h}\frac{(1-\pi_{hi})}{\pi_{hi}^2}\hat{z}_{hi}^2+\sum_{i\in s_h}\sum_{j\neq i\in s_h}\frac{(\pi_{hij}-\pi_{hi}\pi_{hj})}{\pi_{hij}\pi_{hi}\pi_{hj}}\hat{z}_{hi}\hat{z}_{hj}\right] \qquad (4.21)$$

เมื่อ $\hat{z}_{hi}=\sum_{i\in s_h}\frac{r_{hi}y_{hi}}{\pi_{hi}}+\sum_{i\in s_h}\frac{1}{\pi_{hi}}r_{hi}(y_{hi}-\hat{\bar{Y}}_{Rh})$ ค่า $\hat{\bar{Y}}_{Rh}=\frac{\sum_{i\in s_h}\frac{r_{hi}y_{hi}}{\pi_{hi}}}{\sum_{i\in s_h}\frac{r_{hi}}{\pi_{hi}}}$ และ $i\in s_h;h\in H$

(3) จากสมการ (4.21) ค่าคาดหมายของ $\hat{V}(\hat{Y}_{NLMI})$ ของ มีค่าเป็น

$$E\left[\hat{V}(\hat{Y}_{NLMI})\right]=E_RE_S\left[\hat{V}(\hat{Y}_{NLMI})\middle| R\right]$$

$$=E_RE_S\left[\sum_{h\in H}\frac{1}{\left(\sum_{i\in s_h}\frac{r_{hi}}{\pi_{hi}}\right)^2}\left[\sum_{i\in s_h}\frac{(1-\pi_{hi})}{\pi_{hi}^2}\hat{z}_{hi}^2+\sum_{i\in s_h}\sum_{j\neq i\in s_h}\frac{(\pi_{hij}-\pi_{hi}\pi_{hj})}{\pi_{hij}\pi_{hi}\pi_{hj}}\hat{z}_{hi}\hat{z}_{hj}\right]\middle| R\right]$$

$$= E_R E_S \left[\sum_{h\in H} \frac{1}{\left(\sum_{i\in U_h} \frac{r_{hi} I_{hi}}{\pi_{hi}} \right)^2} \left[\sum_{i\in U_h} \frac{(1-\pi_{hi}) I_{hi}}{\pi_{hi}^2} \hat{z}_{hi}^2 + \sum_{i\in U_h} \sum_{j\neq i\in U_h} \frac{(\pi_{hij} - \pi_{hi}\pi_{hj}) I_{hi} I_{hj}}{\pi_{hij}\pi_{hi}\pi_{hj}} \hat{z}_{hi}\hat{z}_{hj} \right] \middle| R \right]$$

$$\approx \sum_{h\in H} \frac{1}{\left(\sum_{i\in U_h} \frac{E_R(r_{hi}) E_S(I_{hi})}{\pi_{hi}} \right)^2} \left[\sum_{i\in U_h} \frac{(1-\pi_{hi}) E_S(I_{hi})}{\pi_{hi}^2} E_R E_S(\hat{z}_{hi}^2) \right.$$

$$\left. + \sum_{i\in U_h} \sum_{j\neq i\in U_h} \frac{(\pi_{hij} - \pi_{hi}\pi_{hj}) E_S(I_{hi} I_{hj})}{\pi_{hij}\pi_{hi}\pi_{hj}} E_R E_S(\hat{z}_{hi}) E_R E_S(\hat{z}_{hj}) \right]$$

$$\approx \sum_{h\in H} \frac{1}{\left(\sum_{i\in U_h} \frac{p_h \pi_{hi}}{\pi_{hi}} \right)^2} \left[\sum_{i\in U_h} \frac{(1-\pi_{hi})\pi_{hi}}{\pi_{hi}^2} \left(N_h p_h y_{hi} \right)^2 + \sum_{i\in U_h} \sum_{j\neq i\in U_h} \frac{(\pi_{hij} - \pi_{hi}\pi_{hj})\pi_{hij}}{\pi_{hij}\pi_{hi}\pi_{hj}} N_h p_h y_{hi} N_h p_h y_{hj} \right]$$

$$= \sum_{h\in H} \frac{1}{\left(N_h p_h \right)^2} \left[\sum_{i\in U_h} \frac{(1-\pi_{hi})}{\pi_{hi}} N_h^2 p_h^2 y_{hi}^2 + \sum_{i\in U_h} \sum_{j\neq i\in U_h} \frac{(\pi_{hij} - \pi_{hi}\pi_{hj})}{\pi_{hi}\pi_{hj}} N_h^2 p_h^2 y_{hi} y_{hj} \right]$$

$$= \sum_{h\in H} \left[\sum_{i\in U_h} \frac{(1-\pi_{hi})}{\pi_{hi}} y_{hi}^2 + \sum_{i\in U_h} \sum_{j\neq i\in U_h} \frac{(\pi_{hij} - \pi_{hi}\pi_{hj})}{\pi_{hi}\pi_{hj}} y_{hi} y_{hj} \right]$$

$$= \sum_{h\in H} \sum_{i\in U_h} \frac{(1-\pi_{hi})}{\pi_{hi}} y_{hi}^2 + \sum_{h\in H} \sum_{i\in U_h} \sum_{j\neq i\in U_h} \frac{(\pi_{hij} - \pi_{hi}\pi_{hj})}{\pi_{hi}\pi_{hj}} y_{hi} y_{hj} = V(\hat{Y}_{NLMI})$$

ดังนั้น $E\left[\hat{V}(\hat{Y}_{NLMI}) \right] = V(\hat{Y}_{NLMI})$

4.3 ช่วงความเชื่อมั่นของผลรวมประชากร

จากหัวข้อ 4.1 ผู้วิจัยได้กล่าวถึงตัวประมาณค่าแบบจุดของผลรวมประชากร ในกรณีที่มีข้อมูลสูญหาย ทั้งตัวประมาณผลรวมประชากรแบบเป็นเชิงเส้น และผลรวมประชากรแบบไม่เป็นเชิงเส้น ตามที่แสดงในสมการที่ (4.4) ซึ่งตัวประมาณค่าแบบจุดที่นำเสนอ เป็นตัวประมาณค่าที่ไม่เอนเอียงเมื่อขนาดตัวอย่างเข้าใกล้อนันต์ ตามที่แสดงในทฤษฎี 4.1 สำหรับในหัวข้อนี้ผู้วิจัยจะได้กล่าวถึง ช่วงความเชื่อมั่นของผลรวมประชากร โดยอาศัยทฤษฎีขีดจำกัดกลาง (The Central Limit Theorem)

ทฤษฎี 4.3 *กำหนดให้เงื่อนไข (1) และ (2) เป็นจริง ภายใต้การสุ่มแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มตัวอย่าง* s_h *ขนาด* n_h *ด้วยแผนการสุ่มตัวอย่างของ Midzuno ภายใต้เฟรมเวิร์คแบบรีเวิร์ส กำหนดค่า* $\hat{Y}_{MI}$ *ตามสมการ (4.4) กำหนดค่า* $\hat{V}(\hat{Y}_{LMI})$ *ตามสมการ (4.7) และ* $\hat{V}(\hat{Y}_{NLMI})$ *กำหนดค่าตามสมการ (4.15)*

(1) กำหนดให้ $\hat{V}(\hat{Y}_{MI}) = \hat{V}(\hat{Y}_{LMI})$ ดังนั้นช่วงความเชื่อมั่นแบบเชิงเส้นของผลรวมประชากร Y เขียนแทนด้วย $CI_1(\hat{Y}_{MI})$ ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$ Y มีค่าเป็น

$$\boldsymbol{C.I._1(\hat{Y}_{MI})=\left[\hat{Y}_{MI}-z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}\left(\hat{Y}_{LMI}\right)}\ ,\ \hat{Y}_{MI}+z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}\left(\hat{Y}_{LMI}\right)}\right]} \quad (4.22)$$

(2) กำหนดให้ $\boldsymbol{\hat{V}(\hat{Y}_{MI})=\hat{V}(\hat{Y}_{NLMI})}$ ดังนั้นช่วงความเชื่อมั่นแบบไม่เป็นเชิงเส้นของผลรวมประชากร $\boldsymbol{Y}$ เขียนแทนด้วย $\boldsymbol{C.I._2(\hat{Y}_{MI})}$ ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$ มีค่าเป็น

$$\boldsymbol{C.I._2(\hat{Y}_{MI})=\left[\hat{Y}_{MI}-z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}\left(\hat{Y}_{NLMI}\right)}\ ,\ \hat{Y}_{MI}+z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}\left(\hat{Y}_{NLMI}\right)}\right]} \quad (4.23)$$

พิสูจน์

(1) กำหนดค่า $\boldsymbol{\hat{Y}_{MI}}$ ตามสมการ (4.4) และ $\boldsymbol{\hat{V}(\hat{Y}_{MI})=\hat{V}(\hat{Y}_{LMI})}$ จากทฤษฎี 4.1 จะได้ว่า กรณีที่ $\boldsymbol{n\to\infty}$ $\boldsymbol{E\left[\hat{Y}_{MI}\right]=Y}$ โดยอาศัยทฤษฎีขีดจำกัดกลาง จะได้ว่า

$$\boldsymbol{Z=\frac{\hat{Y}_{MI}-E\left[\hat{Y}_{MI}\right]}{\sqrt{\hat{V}(\hat{Y}_{MI})}}=\frac{\hat{Y}_{MI}-E\left[\hat{Y}_{MI}\right]}{\sqrt{\hat{V}(\hat{Y}_{LMI})}}\sim appN(0,1)} \quad (4.24)$$

ดังนั้นช่วงความเชื่อมั่นของผลรวมประชากร $\boldsymbol{Y}$ ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$ มีค่าเป็น

$$\boldsymbol{C.I._1(\hat{Y}_{MI})=\left[\hat{Y}_{MI}-z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}\left(\hat{Y}_{LMI}\right)}\ ,\ \hat{Y}_{MI}+z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}\left(\hat{Y}_{LMI}\right)}\right]}$$

(2) พิสูจน์เช่นเดียวกับข้อ 1

4.4 ผลการแสดงตัวอย่างเชิงตัวเลข

4.4.1 ข้อมูลทั่วไปของประชากร

ค่าพารามิเตอร์ของประชากรทั้ง 2 ชั้นภูมิแสดงตารางที่ 4.1

ตารางที่ 4.1 ค่าพารามิเตอร์ ของแต่ละชั้นภูมิ

พารามิเตอร์	ชั้นภูมิที่ 1	ชั้นภูมิที่ 2
$\boldsymbol{Y_h}$	1549424.61	1205181.04
$\sigma^2_{Y_h}$	508641.40	467034.37
$\boldsymbol{X_h}$	20402.20	16019.47
$\sigma^2_{X_h}$	20.76	20.04
$\boldsymbol{N_h}$	1500	1200
$\sum \boldsymbol{r_{hi}}$	919	956

4.4.2 ผลการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร

เมื่อใช้เทคนิคมอนติคาร์โลซิมูเลชั่น และกำหนดการทำซ้ำทั้งหมด 100,000 ครั้ง เพื่อคำนวณค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ ของ $\hat{V}(\hat{Y}_{LMI})$ และ $\hat{V}(\hat{Y}_{NLMI})$ ผลการคำนวณแสดงตารางที่ 4.2

ตารางที่ 4.2 ค่ารากที่สอง ของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ ของ $\hat{V}(\hat{Y}_{LMI})$ และ $\hat{V}(\hat{Y}_{NLMI})$

ชุดตัวอย่างที่	n_1	n_2	n	$RMSE[\hat{V}(\hat{Y}_{LMI})]$	$RMSE[\hat{V}(\hat{Y}_{NLMI})]$
1	30	20	50	0.59	***0.52***
2	45	25	60	0.57	***0.44***
3	45	30	75	0.56	***0.43***
4	60	40	100	0.56	***0.37***
5	70	50	120	0.56	***0.34***
6	85	55	140	0.55	***0.30***
7	90	65	155	0.55	***0.30***
8	100	70	170	0.55	***0.28***

จากตางที่ 4.2 พบว่า ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ของตัวประมาณค่าความแปรปรวนแบบเชิงเส้น หรือ $RMSE[\hat{V}(\hat{Y}_{LMI})]$ มีค่าสูงสุดเท่ากับ 0.59 และต่ำสุดเท่ากับ 0.54 ที่ขนาดตัวอย่างเท่ากับ 50 และ 140 ตามลำดับ ส่วนค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ของตัวประมาณค่าความแปรปรวนแบบไม่เป็นเชิงเส้น หรือ $RMSE[\hat{V}(\hat{Y}_{NLMI})]$ มีค่าสูงสุดเท่ากับ 0.52 และต่ำสุดเท่ากับ 0.28 ที่ขนาดตัวอย่างเท่ากับ 50 และ 170 ตามลำดับ

เมื่อพิจารณาในแต่ละขนาดตัวอย่าง พบว่า ค่า $RMSE[\hat{V}(\hat{Y}_{NLMI})] \leq RMSE[\hat{V}(\hat{Y}_{LMI})]$ ทั้ง 8 ระดับของขนาดตัวอย่าง และเมื่อพิจารณาตามนิยามที่ 3.1 จะได้ว่า $\hat{V}(\hat{Y}_{NLMI})$ เป็นตัวประมาณค่าที่มีประสิทธิภาพที่ดีกว่า $\hat{V}(\hat{Y}_{LMI})$ ซึ่งสอดคล้องกับสมมติฐานการวิจัยที่ตั้งไว้

4.4.3 ผลการเปรียบเทียบประสิทธิภาพช่วงความเชื่อมั่นของผลรวมประชากร

หัวข้อนี้ผู้วิจัยจะได้กล่าวถึงผลการเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นของผลรวมประชากร โดยพิจารณาจากร้อยละของค่าเฉลี่ยของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมพารามิเตอร์ ตามที่แสดงในสมการที่ (3.5) ผลการคำนวณค่าร้อยละของค่าเฉลี่ยของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมพารามิเตอร์ เมื่อใช้เทคนิคมอนติคาร์โลซิมูเลชั่น และกำหนดการทำซ้ำทั้งหมด 100,000 ครั้ง ผลแสดงดังตารางที่ 4.3

ตารางที่ 4.3 ร้อยละของค่าเฉลี่ยของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมพารามิเตอร์

ชุดตัวอย่างที่	n_1	n_2	n	$C.I._1(\hat{Y}_{MI})$	$C.I._2(\hat{Y}_{MI})$
1	30	20	50	79.25	***92.73***
2	45	25	60	80.26	***93.43***
3	45	30	75	81.19	***93.71***
4	60	40	100	81.54	***93.88***
5	70	50	120	81.02	***94.42***
6	85	55	140	81.96	***94.50***
7	90	65	155	81.22	***94.24***
8	100	70	170	80.62	***94.21***

ผลจากตารางที่ 4.3 พบว่า ร้อยละของค่าเฉลี่ยของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมพารามิเตอร์ของช่วงความเชื่อมั่นแบบเชิงเส้น หรือ $C.I._1(\hat{Y}_{MI})$ มีค่าสูงสุดเท่ากับ ร้อยละ 81.96 และมีค่าต่ำสุด ร้อยละ 79.25 ที่ขนาดตัวอย่างเท่ากับ 140 และ 50 ตามลำดับ ส่วนร้อยละของค่าเฉลี่ยของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมพารามิเตอร์ของช่วงความเชื่อมั่นแบบไม่เป็นเชิงเส้น หรือ $C.I._2(\hat{Y}_{MI})$ มีค่าสูงสุดเท่ากับร้อยละ 94.21 และต่ำสุดร้อยละ 92.73

เมื่อพิจารณาในแต่ละขนาดตัวอย่าง พบว่า ค่า $C.I._1(\hat{Y}_{MI}) \leq C.I._2(\hat{Y}_{MI})$ ทั้ง 8 ระดับของขนาดตัวอย่าง และเมื่อพิจารณาตามนิยามที่ 3.2 จะได้ว่า $C.I._2(\hat{Y}_{MI})$ เป็นตัวประมาณค่าที่มีประสิทธิภาพที่ดีกว่า $C.I._1(\hat{Y}_{MI})$ ซึ่งสอดคล้องกับสมมติฐานการวิจัยที่ตั้งไว้

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

การศึกษาเกี่ยวกับการประมาณค่าผลรวมประชากรในการสำรวจตัวอย่าง เมื่อมีข้อมูลสูญหาย สามารถสรุปผลการวิจัย อภิปรายผล และมีข้อเสนอแนะดังนี้

5.1 สรุปผลการวิจัย

การวิจัยในครั้งนี้ มีวัตถุประสงค์เพื่อหาตัวประมาณค่าผลรวมประชากร ทั้งตัวประมาณแบบจุด และตัวประมาณค่าแบบช่วง กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม และแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย ภายใต้การสุ่มตัวอย่างโดยใช้ความน่าจะเป็นในการเลือกไม่เท่ากัน โดยผู้วิจัยเลือกใช้แผนการเลือกตัวอย่างแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มตัวอย่างโดยใช้แผนการสุ่มตัวอย่างของ Midzuno ผลการวิจัยเป็นดังนี้

5.1.1 ตัวประมาณค่าผลรวมประชากรแบบจุด

กรณีที่ประชากร $\boldsymbol{U}$ ขนาด $\boldsymbol{N}$ ถูกแบ่งออกเป็น $\boldsymbol{L}$ ชั้นภูมิ $\boldsymbol{U}_1,\boldsymbol{U}_2,...,\boldsymbol{U}_{\boldsymbol{h}},...,\boldsymbol{U}_{\boldsymbol{L}}$ ขนาด $\boldsymbol{N}_1,\boldsymbol{N}_2,...,\boldsymbol{N}_{\boldsymbol{h}},...,\boldsymbol{N}_{\boldsymbol{L}}$ ตามลำดับ สมมติว่าในแต่ละชั้นภูมิ $\boldsymbol{h}$ ตัวอย่าง $\boldsymbol{s}_{\boldsymbol{h}}$ ขนาด $\boldsymbol{n}_{\boldsymbol{h}}$ ถูกสุ่มประชากร $\boldsymbol{U}_{\boldsymbol{h}}$ ขนาด $\boldsymbol{N}_{\boldsymbol{h}}$ เมื่อ $\boldsymbol{h}\in\boldsymbol{H}$ เมื่อ $\boldsymbol{H}=\{1,2,3,...,\boldsymbol{L}\}$ โดยใช้แผนการสุ่มตัวอย่างแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มตัวอย่างตามแผนการสุ่มตัวอย่างของ Midzuno กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม และแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย ตัวประมาณค่าผลรวมประชากรแบบเชิงเส้นที่ผู้วิจัยนำเสนอแสดงดังสมการที่ (5.1)

$$\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}}=\sum_{\boldsymbol{h}\in\boldsymbol{H}}\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{z}^{*}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}} \tag{5.1}$$

เมื่อ $\boldsymbol{z}^{*}_{\boldsymbol{h}}=\{\boldsymbol{z}^{*}_{\boldsymbol{h}1},\boldsymbol{z}^{*}_{\boldsymbol{h}2},...,\boldsymbol{z}^{*}_{\boldsymbol{hn}_{\boldsymbol{h}}}\}$ แทนเซตของข้อมูลที่สมบูรณ์ หลังจากที่แทนที่ข้อมูลสูญหายค่าเฉลี่ย ในชั้นภูมิที่ $\boldsymbol{h}$ เมื่อ $\boldsymbol{h}\in\boldsymbol{H}$ ซึ่ง $\boldsymbol{z}^{*}_{\boldsymbol{hi}}=\boldsymbol{y}_{\boldsymbol{hi}}$ กรณีที่หน่วยตัวอย่างที่ $\boldsymbol{i}$ ไม่สูญหาย และ $\boldsymbol{z}^{*}_{\boldsymbol{hi}}=\boldsymbol{y}^{*}_{\boldsymbol{hi}}$ เมื่อหน่วยตัวอย่างที่ $\boldsymbol{i}$ สูญหาย และ $\boldsymbol{y}^{*}_{\boldsymbol{hi}}=\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{r}_{\boldsymbol{hi}}\boldsymbol{y}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}\Big/\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{r}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}$ และจากสมการที่ (5.1) สามารถปรับตัวประมาณค่าแบบเชิงเส้นให้เป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้นตามที่แสดงในสมการที่ (5.2)

$$\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}=\sum_{\boldsymbol{h}\in\boldsymbol{H}}\left[\frac{\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{1}{\pi_{\boldsymbol{hi}}}\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{r}_{\boldsymbol{hi}}\boldsymbol{y}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}}{\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{r}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}}\right] \tag{5.2}$$

5.1.2 ตัวประมาณค่าความแปรปรวนของตัวประมาณผลรวมประชากรค่าแบบจุด

ตัวประมาณค่าความแปรปรวนของตัวประมาณผลรวมประชากรแบบเป็นเชิงเส้นที่ผู้วิจัยนำเสนอแสดงดังสมการที่ (5.3) ส่วนตัวประมาณค่าความแปรปรวนของตัวประมาณผลรวมประชากรแบบเป็นไม่เป็นเชิงเส้นที่ผู้วิจัยนำเสนอแสดงดังสมการที่ (5.4)

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}})=\sum_{\boldsymbol{h}\in\boldsymbol{H}}\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\left(\frac{1-\pi_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}^{2}}\right)\boldsymbol{z}_{\boldsymbol{hi}}^{*2}+\sum_{\boldsymbol{h}\in\boldsymbol{H}}\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\sum_{\boldsymbol{j}\neq\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\left(\frac{\pi_{\boldsymbol{hij}}-\pi_{\boldsymbol{hi}}\pi_{\boldsymbol{hj}}}{\pi_{\boldsymbol{hij}}\pi_{\boldsymbol{hi}}\pi_{\boldsymbol{hj}}}\right)\boldsymbol{z}_{\boldsymbol{hi}}^{*}\boldsymbol{z}_{\boldsymbol{hj}}^{*} \tag{5.3}$$

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})=\sum_{\boldsymbol{h}\in\boldsymbol{H}}\frac{1}{\left(\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{r}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}\right)^{2}}\left[\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{(1-\pi_{\boldsymbol{hi}})}{\pi_{\boldsymbol{hi}}^{2}}\hat{\boldsymbol{z}}_{\boldsymbol{hi}}^{2}+\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\sum_{\boldsymbol{j}\neq\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{(\pi_{\boldsymbol{hij}}-\pi_{\boldsymbol{hi}}\pi_{\boldsymbol{hj}})}{\pi_{\boldsymbol{hij}}\pi_{\boldsymbol{hi}}\pi_{\boldsymbol{hj}}}\hat{\boldsymbol{z}}_{\boldsymbol{hi}}\hat{\boldsymbol{z}}_{\boldsymbol{hj}}\right] \tag{5.4}$$

เมื่อ $\hat{\boldsymbol{z}}_{\boldsymbol{hi}}=\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{\boldsymbol{r}_{\boldsymbol{hi}}\boldsymbol{y}_{\boldsymbol{hi}}}{\pi_{\boldsymbol{hi}}}+\left(\sum_{\boldsymbol{i}\in\boldsymbol{s}_{\boldsymbol{h}}}\frac{1}{\pi_{\boldsymbol{hi}}}\right)\boldsymbol{r}_{\boldsymbol{hi}}(\boldsymbol{y}_{\boldsymbol{hi}}-\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{Rh}})$ นอกจากนี้ $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}})$ และ $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ เป็นตัวประมาณค่าที่ไม่เอนเอียงโดยประมาณของ $\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}})$ และ $\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ ตามลำดับ เมื่อขนาดตัวอย่างเข้าใกล้อนันต์ ซึ่ง $\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}})$ กำหนดค่าตามที่แสดงในสมการที่ (4.14) และ $\boldsymbol{V}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ กำหนดค่าตามที่แสดงในสมการที่ (4.15) ผลการเปรียบเทียบประสิทธิภาพของ $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}})$ และ $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ เมื่อพิจารณาจากค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์พบว่า $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ มีประสิทธิภาพดีกว่า $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}})$

5.1.3 ตัวประมาณค่าผลรวมประชากรแบบช่วง

จากหัวข้อ 5.1.1 ผู้วิจัยได้กล่าวถึงตัวประมาณค่าแบบจุดของตัวประมาณค่าผลรวมประชากร ในกรณีที่มีข้อมูลสูญหายทั้งตัวประมาณค่าแบบเชิงเส้น และตัวประมาณค่าแบบไม่เป็นเชิงเส้น ส่วนหัวข้อ 5.1.2 ผู้วิจัยได้กล่าวถึงตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร ทั้งแบบเชิงเส้น และแบบที่ไม่เป็นเชิงเส้น เขียนแทนด้วย $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}})$ และ $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ ตามลำดับ โดยอาศัยทฤษฎีขีดจำกัดกลางสร้างตัวประมาณค่าผลรวมประชากรแบบช่วง ที่สร้างจาก $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}})$ และ $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ เขียนแทนด้วย $\boldsymbol{C}.\boldsymbol{I}._{1}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$ และ $\boldsymbol{C}.\boldsymbol{I}._{2}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$ และกำหนดค่าดังนี้

(1) ตัวประมาณค่าผลรวมประชากรแบบช่วงที่สร้างจาก $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}})$ ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$ มีค่าเป็น

$$\boldsymbol{C}.\boldsymbol{I}._{1}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})=\left[\hat{\boldsymbol{Y}}_{\boldsymbol{MI}}-\boldsymbol{z}_{1-\frac{\alpha}{2}}\sqrt{\hat{\boldsymbol{V}}\left(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}}\right)}\ ,\ \hat{\boldsymbol{Y}}_{\boldsymbol{MI}}+\boldsymbol{z}_{1-\frac{\alpha}{2}}\sqrt{\hat{\boldsymbol{V}}\left(\hat{\boldsymbol{Y}}_{\boldsymbol{LMI}}\right)}\right] \tag{5.5}$$

(2) ตัวประมาณค่าผลรวมประชากรแบบช่วงที่สร้างจาก $\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}})$ ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$ มีค่าเป็น

$$\boldsymbol{C}.\boldsymbol{I}._{2}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})=\left[\hat{\boldsymbol{Y}}_{\boldsymbol{MI}}-\boldsymbol{z}_{1-\frac{\alpha}{2}}\sqrt{\hat{\boldsymbol{V}}\left(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}\right)}\ ,\ \hat{\boldsymbol{Y}}_{\boldsymbol{MI}}+\boldsymbol{z}_{1-\frac{\alpha}{2}}\sqrt{\hat{\boldsymbol{V}}\left(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}\right)}\right] \tag{5.6}$$

เมื่อ $\hat{\boldsymbol{V}}\left(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}\right)$ และ $\hat{\boldsymbol{V}}\left(\hat{\boldsymbol{Y}}_{\boldsymbol{NLMI}}\right)$ กำหนดค่าตามที่แสดงในสมการที่ (5.3) และ (5.4) ตามลำดับ ผลการเปรียบเทียบประสิทธิภาพของ $\boldsymbol{C.I._1}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$ และ $\boldsymbol{C.I._2}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$ พบว่า $\boldsymbol{C.I._2}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$ มีประสิทธิภาพดีกว่า $\boldsymbol{C.I._1}(\hat{\boldsymbol{Y}}_{\boldsymbol{MI}})$

5.2 การอภิปรายผล

การวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อ นำเสนอตัวประมาณค่าผลรวมประชากรทั้งตัวประมาณค่าแบบจุดและแบบช่วง รวมถึงนำเสนอตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม ภายใต้การสุ่มตัวอย่างแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มตัวอย่างโดยใช้ความน่าจะเป็นไม่เท่ากัน นอกจากนี้ในแต่ละชั้นภูมิกำหนดให้สัดส่วนตัวอย่างมีขนาดเล็ก และความน่าจะเป็นที่จะตอบสนองมีค่าเท่ากันทุกหน่วยตัวอย่าง

ผลการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าผลรวมประชากร พบว่าตัวประมาณค่าความแปรปรวนแบบไม่เป็นเชิงเส้นมีค่ารากที่สองของความคลาดเคลื่อนสัมพัทธ์เฉลี่ยต่ำกว่า ตัวประมาณค่าความแปรปรวนแบบเป็นเชิงเส้น เนื่องจากตัวประมาณค่าความแปรปรวนแบบเชิงเส้น จะพิจารณาเซตตัวอย่างที่ถูกสุ่มจากประชากร เป็นเซตที่ไม่มีข้อมูลสูญหายเนื่องจากแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ยจากเซตตัวอย่างที่ไม่สูญหาย ดังนั้นการหาตัวประมาณค่าความแปรปรวน สามารถหาได้โดยตรงจากสูตรตามที่ Horvitz and Thompson (1952) นำเสนอ อย่างไรก็ตาม Deville and Särndal (1944) กล่าวว่ากรณีที่ข้อมูลสูญหายมีขนาดใหญ่ ตัวประมาณค่าผลรวมประชากรแบบเป็นเชิงเส้นจะเป็นตัวประมาณค่าที่เอนเอียง ส่งผลให้รากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์มีค่าเพิ่มขึ้น

สำหรับตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น จะพิจารณาเซตของตัวอย่างที่ถูกสุ่มจากประชากรที่มีข้อมูลสูญหาย ดังนั้นการหาค่าความแปรปรวนของตัวประมาณค่าดังกล่าว จะพิจารณาภายใต้เฟรมเวิร์คแบบรีเวิร์ส นอกจากนี้การหาค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น จำเป็นต้องปรับตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้นให้เป็นตัวประมาณค่าแบบเชิงเส้นโดยใช้เทเลอร์แบบเชิงเส้น จากนั้นหาตัวประมาณค่าความแปรปรวนตามสูตรที่ Haziza (2010) นำเสนอ อย่างไรก็ตามตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น จะเป็นตัวประมาณค่าที่เอนเอียงเช่นเดียวกับตัวประมาณค่าแบบเชิงเส้น แต่ค่าความเอนเอียงของตัวประมาณค่าแบบไม่เป็นเชิงเส้นจะมีค่าต่ำกว่าตัวประมาณค่าแบบเชิงเส้น และจะมีค่าเข้าใกล้ 0 เมื่อตัวอย่างมีขนาดใหญ่ ซึ่งจะส่งผลให้ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ของตัวประมาณค่าความแปรปรวนแบบไม่เป็นเชิงเส้น มีค่าต่ำกว่าตัวประมาณค่าความแปรปรวนแบบเชิงเส้น

การเปรียบเทียบประสิทธิภาพช่วงความเชื่อมั่นของผลรวมประชากรที่ระดับความเชื่อมั่น 95% เมื่อพิจารณาจากร้อยละของค่าเฉลี่ยของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมพารามิเตอร์ พบว่าช่วงความเชื่อมั่นแบบไม่เป็นเชิงเส้นมีประสิทธิดีกว่าช่วงความเชื่อมั่นแบบเชิงเส้น ซึ่ง Haziza(2010) กล่าวว่ากรณีที่ข้อมูลสูญหายมีขนาดใหญ่ ช่วงความเชื่อมั่นแบบเชิงเส้นจะมีประสิทธิภาพต่ำกว่าช่วงความเชื่อมั่นของข้อมูลระดับนามบัญญัติ ซึ่งจะส่งผลให้ประสิทธิภาพในการประมาณค่า มีค่าลดลง

5.3 ข้อเสนอแนะ

5.3.1 จากงานวิจัยผู้วิจัยได้กำหนดให้สัดส่วนตัวอย่างในแต่ละชั้นภูมิมีขนาดเล็ก ซึ่งจากผลการวิจัยพบว่ากรณีที่สัดส่วนตัวอย่างในแต่ละชั้นภูมิไม่มีขนาดเล็ก ร้อยละของค่าเฉลี่ยของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมพารามิเตอร์จะมีค่าลดลง ดังนั้นผู้ที่สนใจศึกษาอาจขยายผลการวิจัยไปยังกรณีที่สัดส่วนตัวอย่างในแต่ละชั้นภูมิไม่มีขนาดเล็ก

5.3.2 งานวิจัยในครั้งนี้ หาตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรโดยใช้สูตรของ Horvitz และ Thompson (1952) ซึ่งสูตรดังกล่าวต้องการค่า π_{hij} อย่างไรก็ตามในบางแผนการสุ่มตัวอย่าง การคำนวณค่า π_{hij} ค่อนข้างยุ่งยากหรือไม่สามารถคำนวณได้ ดังนั้นผู้ที่สนใจศึกษาอาจปรับตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร ที่ผู้วิจัยนำเสนอ โดยใช้วิธีของ Hájek (1964) หรือ Berger (2003) เป็นต้น

บรรณานุกรม

ชูเกียรติ โพนแก้ว. (2560). **การประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากรเมื่อมีข้อมูลสูญหาย และสุ่มตัวอย่างโดยใช้ความน่าจะเป็นในการเลือกไม่เท่ากัน.** วารสารวิจัยราชภัฏพระนคร สาขาวิทยาศาสตร์และเทคโนโลยี, หน้า 375-387.

ณรงค์ โพธิ และ สมชาย ปราการเจริญ.(2553). **กรอบแนวความคิดสำหรับการประมาณค่าสูญหายของข้อมูลด้วยเทคนิคการรู้จำรูปแบบ.** The 6TH National Conference on Computing and Information Technology. หน้า 686-687.

ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และ สุคนธ์ ประสิทธิ์วัฒนเสรี (2549). **ข้อมูลสูญหายและแนวทางการจัดการ.** Data Management & Biostatistics Journal. 4(3), หน้า 53 – 57.

พัชรีพร สนรักษา, พัชรี วงษ์เกษม, คณินทร์ ธีรภาพโอฬาร. (2553). **การประมาณค่าเฉลี่ยประชากรในการสำรวจตัวอย่างเมื่อมีข้อมูลสูญหลาย.**วารสารวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ, 28(1), หน้า 89-103.

เพียงออ ยีสา และกัลยา วานิชย์บัญชา. (2552). **การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้น.** Nation graduate research conference 12th. หน้า 560-562.

วริษฐา กณิกนันต์ และ อนุภาพ สมบูรณ์สวัสดี (2556). **การเปรียบเทียบวิธีการประมาณสาหรับการวิเคราะห์การถดถอยเชิงเส้นพหุเมื่อตัวแปรตามและตัวแปรอิสระมีการสูญหายแบบนอนอิกนอร์เรเบิล.** บทความวิจัย เสนอในการประชุมหาดใหญ่วิชาการ ครั้งที่ 4, หน้า 43 – 44.

สุชาดา กีระนันท์. (2542). **ทฤษฎีและวิธีการสำรวจตัวอย่าง.** กรุงเทพ: โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

ทัตตา หิรัญพต. (2554). **การเปรียบเทียบประสิทธิภาพระหว่างวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยและวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยแบบสโทแคสติก.** ปริญญาวิทยาศาสตรบัณฑิต ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลับูรพา.

ภวัต ภักดิ์ศรานุวัต และ สอาด นิวิศพงศ์. (2553). **ช่วงความเชื่อมั่น และช่วงพยากรณ์สำหรับพารามิเตอร์บางค่าเมื่อมีข้อมูลเกิดค่าสูญหาย.** วารสารบัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

ดวงพร หัชชะวณิช. (2553). **การเปรียบเทียบค่าประมาณที่ปรับน้ำหนักจากการไม่ได้รับคำตอบด้วยวิธีปรับค่าน้ำหนัก เมื่อทราบค่ายอดรวมของประชากร และวิธีจัดชั้นภูมิภายหลังการสุ่มตัวอย่าง.**วารสารวิชาการ มหาวิทยาลัยหอการค้าไทย. 30(2), หน้า 16-25.

Bethlehem, J.G. (1988). **Reduction of nonresponse bias through regression estimation. Journal of Official Statistics**, 4, pp. 251-260.

Deville JC, Särndal CE (1992). **Calibration estimators in survey sampling**. Journal of the American Statistical Association. 1992;87:376–382.

Fay, R.E. (1991). **A design-based perspective on missing data variance, Proceedings of the 1991 Annual Research Conference**, US Bureau of the Census. pp.429-440.

Hájek, J. (1971). **Foundations of Statistical Inference**. Toronto, Canada: Holt, Rinehart, Winston. Chap. Discussion of an essay on the logical foundations of survey sampling, part on by D. Basu.

Haziza, D.(2010).**Resampling methods for variance estimation in the presence of missing survey data**. Proceedings of the Annual Conference of the Italian Statistical Society.

Horvitz, D.G., and D.J. Thompson (1952). **A generalization of sampling without replacement from a finite universe.** Journal of the American Statistical Association, 47, pp. 663-685.

Midzuno, H. (1952). **On sampling system with probability proportional to sum of sizes**, Ann. Inst.Statist. Math. 3, pp 99-107.

Särndal,C.E., Swenson, B. and Wretman, J.H. (1992). **Model assisted survey sampling.** NewYork, Springer-Verlag.

ภาคผนวก

ภาคผนวก ก
รายชื่อผู้ทรงคุณวุฒิตรวจสอบเครื่องมือ และเนื้อหาในโครงการวิจัย

รายชื่อผู้ทรงคุณวุฒิตรวจสอบเครื่องมือ และเนื้อหาในโครงการวิจัย

1. ผู้ช่วยศาสตราจารย์ ดร.ชัยณรงค์ ขันผนึก
   ตำแหน่ง อาจารย์ประจำหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาคณิตศาสตร์
   คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเพชรบูรณ์
   จังหวัดเพชรบูรณ์
2. ผู้ช่วยศาสตราจารย์ อาดุลย์ จงรักษ์
   ตำแหน่ง อาจารย์ประจำหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาคณิตศาสตร์
   คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเพชรบูรณ์
   จังหวัดเพชรบูรณ์
3. อาจารย์ ดร.เจษฎา โพนแก้ว
   ตำแหน่ง อาจารย์ประจำหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
   คณะศิลปศาสตร์และวิทยาศาสตร์ มหาวิทยาลัยราชภัฏศรีสะเกษ
   จังหวัดศรีสะเกษ

ภาคผนวก ข
การเผยแพร่ผลงานวิจัย

The 22nd Annual Meeting in Mathematics (AMM 2017)
Department of Mathematics, Faculty of Science
Chiang Mai University, Chiang Mai, Thailand

การประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม

ชูเกียรติ โพนแก้ว[†] [‡]
หลักสูตรสาขาวิชาคณิตศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเพชรบูรณ์

บทคัดย่อ

บทความนี้มีวัตถุประสงค์เพื่อนำเสนอและเปรียบเทียบตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรกรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถามและแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรทั้งแบบที่เป็นและไม่เป็นเชิงเส้นถูกพิจารณาภายใต้การสุ่มตัวอย่างแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มตัวอย่างโดยใช้ความน่าจะเป็นในเลือกไม่เท่ากันตามแผนการสุ่มตัวอย่างของ Midzuno ผลการเปรียบเทียบประสิทธิภาพโดยพิจารณาจากรากที่สองของความคลาดเคลื่อนพบว่าตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบไม่เป็นเชิงเส้นมีประสิทธิภาพดีกว่าตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบเป็นเชิงเส้น

คำสำคัญ: ข้อมูลสูญหาย การประมาณค่าความแปรปรวน เทเลอร์แบบเชิงเส้น

1 บทนำ

ปัญหาเกี่ยวกับการประมาณค่าผลรวมประชากร (Population total) เป็นหนึ่งในปัญหาที่สำคัญในการศึกษาวิชาทางสถิติ การประมาณค่าผลรวมประชากรสามารถทำได้โดยใช้ข้อมูลของตัวอย่างที่ถูกสุ่มจากประชากร ตามข้อตกลงของแผนการสุ่มตัวอย่าง (Sampling design) และจำเป็นต้องใช้ข้อมูลของตัวอย่างที่มีความสมบูรณ์เพื่อใช้ในการหาตัวประมาณค่าที่ไม่เอนเอียง (Unbiased estimator)

*งานวิจัยเรื่องนี้ได้รับทุนสนับสนุนจากสถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏเพชรบูรณ์

[†]ผู้แต่งหลัก
[‡]ผู้พูด
อีเมล: chugiat.pon@pcru.ac.th

[1] กล่าวว่าตัวประมาณค่าผลรวมประชากร อาจจะไม่สามารถใช้งานได้อย่างเต็มประสิทธิภาพ เนื่องจากมีข้อมูลบางส่วนเกิดการสูญหาย (Missing data) ซึ่งอาจเกิดจากขั้นตอนจัดเก็บ ปัญหาจากการป้อนข้อมูล หรือเกิดจากข้อจำกัดของเทคโนโลยีที่ใช้งาน ข้อมูลของตัวอย่างที่เกิดการสูญหายอาจส่งผลให้ตัวประมาณค่าผลรวมประชากรที่ได้มีความเอนเอียง (Bias estimator)

ข้อมูลสูญหาย เป็นกรณีที่พบได้บ่อยในงานวิจัยทุกสาขา และนักวิจัยจำเป็นต้องพิจารณาถึงแนวทางที่เหมาะสมสำหรับใช้จัดการกับข้อมูลสูญหาย [2] การสูญหายของข้อมูลอาจเกิดจากการไม่ตอบของหน่วยตัวอย่างบางหน่วย (Unit nonresponse) หรือเกิดจากการไม่ตอบเฉพาะบางคำถาม (Item nonresponse) [3] กล่าวว่าการจัดการกับข้อมูลสูญหายมีหลายวิธี การเลือกใช้วิธีใดวิธีการหนึ่งขึ้นอยู่กับลักษณะข้อมูลสูญหายที่เกิดขึ้น สำหรับการจัดการกับข้อมูลที่สูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม นิยมใช้วิธีการแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ (Imputation) เช่น การประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย (Mean imputation) การประมาณข้อมูลสูญหายด้วยค่าอัตราส่วน (Ratio imputation) หรือ การประมาณข้อมูลสูญหายด้วยค่าถดถอย (Regression imputation) เป็นต้น

พิจารณาประชากร $U = \{1,2,...,i,...,N\}$ ขนาดจำกัด ขนาด N ตัวแปรที่สนใจศึกษาแทนด้วยตัวแปร y และพารามิเตอร์ที่ต้องการประมาณค่าคือผลรวมประชากร เขียนแทนด้วยสัญลักษณ์ $Y = \sum_{i\in U} y_i$ ซึ่ง y_i แทนค่าสังเกตของประชากรหน่วยที่ i เมื่อ $i = 1,2,...,N$ ตัวอย่าง s ขนาด n ถูกสุ่มจากประชากร U ขนาด N ตามข้อตกลงของแผนการสุ่มตัวอย่าง (Sampling design) กรณีที่ไม่มีข้อมูลสูญหาย [4] ได้นำเสนอตัวประมาณค่าผลรวมประชากร $\hat{Y}_{HT}$ ตามที่แสดงในสมการ (1.1)

$$\hat{Y}_{HT} = \sum_{i\in s} \frac{y_i}{\pi_i} \tag{1.1}$$

สำหรับค่าความแปรปรวนของ $\hat{Y}_{HT}$ เขียนแทนด้วย $V(\hat{Y}_{HT})$ และกำหนดค่าตามที่แสดงในสมการที่ (1.2)

$$V(\hat{Y}_{HT}) = \sum_{i\in U}\left(\frac{1-\pi_i}{\pi_i}\right)y_i^2 + \sum_{i\in U}\sum_{j\neq i\in U}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)y_iy_j \tag{1.2}$$

ส่วนตัวประมาณค่าของ $V(\hat{Y}_{HT})$ แทนด้วย $\hat{V}(\hat{Y}_{HT})$ ซึ่งกำหนดค่าตามที่แสดงในสมการที่ (1.3)

$$\hat{V}(\hat{Y}_{HT}) = \sum_{i\in s}\left(\frac{1-\pi_i}{\pi_i^2}\right)y_i^2 + \sum_{i\in s}\sum_{j\neq i\in s}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\right)y_iy_j \tag{1.3}$$

เมื่อ π_i แทนความน่าจะเป็นที่หน่วยตัวอย่างที่ i ตกอยู่ในตัวอย่าง s ขนาด n หรือ $\pi_i = P(i \in s)$ ส่วนค่า π_{ij} แทน ความน่าจะเป็นที่หน่วยตัวอย่างที่ i และ j ตกอยู่ในตัวอย่าง s ขนาด n หรือ $\pi_{ij} = P(i \wedge j \in s)$ เมื่อ $i; j = 1,2,...,N$ และได้ว่า $\hat{Y}_{HT}$ เป็นตัวประมาณค่าที่ไม่เอนเอียงของ Y เนื่องจาก $E_S(\hat{Y}_{HT}) = Y$

กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม และแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย [5] กล่าวว่าการประมาณค่าผลรวมประชากรควรขึ้นอยู่กับเซตของข้อมูลที่ไม่สูญหาย มากกว่าใช้เซตของตัวอย่างที่ถูกสุ่มมาทั้งหมด ดังนั้น [6] ได้นำเสนอตัวประมาณค่าผลรวมประชากร ภายใต้การสุ่มตัวอย่างแบบง่ายและไม่ใส่คืน (Simple random sampling without replacement) ต่อมา [7] นำเสนอตัวประมาณค่าผลรวมประชากรภายใต้การ

สุ่มตัวอย่างโดยใช้ความน่าจะเป็นในการเลือกไม่เท่ากัน (Unequal probability sampling) และใช้แผนการสุ่มตัวอย่างของ [8] ในการสุ่มตัวอย่าง s ขนาด n

สำหรับการวิจัยในครั้งนี้ผู้วิจัยได้ขยายตัวประมาณค่าผลรวมประชากรที่ [7] นำเสนอ ไปยังแผนการเลือกตัวอย่างแบบชั้นภูมิ (Stratified sampling) และแต่ละชั้นภูมิสุ่มตัวอย่างโดยใช้แผนการสุ่มตัวอย่างของ Midzuno (Midzuno's scheme) สำหรับการประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร ผู้วิจัยพิจารณาตัวประมาณค่าผลรวมประชากร ทั้งตัวประมาณค่าแบบเป็นเชิงเส้น และตัวประมาณค่าแบบไม่เป็นเชิงเส้น และเปรียบเทียบประสิทธิภาพจากค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ (Relative root mean square error)

2 วัตถุประสงค์ในการวิจัย

2.1 นำเสนอตัวประมาณค่าผลรวมประชากรเมื่อมีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถามและแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย ทั้งแบบที่เป็นเชิงเส้นและไม่เป็นเชิงเส้น

2.2 นำเสนอตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรทั้งแบบที่เป็นเชิงเส้นและไม่เป็นเชิงเส้น

2.3 เปรียบเทียบประสิทธิภาพระหว่างตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบที่เป็นเชิงเส้น กับแบบที่ไม่เป็นเชิงเส้นโดยพิจารณาจากค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์

3 สมมติฐานของการวิจัย

การสุ่มตัวอย่างกรณีที่ข้อมูลสูญหายมีขนาดใหญ่ ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบเชิงเส้นจะมีประสิทธิภาพลดลง ดังนั้นการศึกษาในครั้งนี้ผู้วิจัยกำหนดสมมติฐานคือตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้นมีประสิทธิภาพดีกว่าตัวประมาณค่าแบบเชิงเส้น

4 วิธีดำเนินการวิจัย

4.1 สัญลักษณ์และข้อตกลงเบื้องต้น

พิจารณาประชากร $U = \{1,2,...,i,...,N\}$ ขนาด N ถูกแบ่งออกเป็น L ชั้นภูมิ $U_1, U_2, ..., U_h, ..., U_L$ ขนาด $N_1, N_2, ..., N_h, ..., N_L$ ตามลำดับ กำหนดให้ $H = \{1,2,...,h,...,L\}$ สมมติว่าในแต่ละชั้นภูมิตัวอย่าง s_h ขนาด n_h ถูกสุ่มประชากร U_h ขนาด N_h เมื่อ $h \in H$ ด้วยแผนการสุ่มตัวอย่างของ [8] ในแต่ละชั้นภูมิกำหนดให้สัดส่วนตัวอย่าง (Sampling fraction) มีขนาดเล็ก กำหนดให้ y แทนตัวแปรที่สนใจศึกษา และผลรวมของตัวแปรที่สนใจศึกษาแทนด้วยสัญลักษณ์ $Y = \sum_{h \in H} \sum_{i \in U_h} y_{hi}$

ในแต่ละชั้นภูมิ h กำหนดให้ π_{hi} แทนความน่าจะเป็นที่หน่วยตัวอย่างที่ i ตกอยู่ในหน่วยตัวอย่าง s_h หรือ $\pi_{hi} = P(i \in s_h)$ เมื่อ $i \in U_h$ และ $h \in H$ ส่วนค่า π_{hij} แทนความน่าจะเป็นที่หน่วยตัวอย่างที่ i และ j ตกอยู่ในตัวอย่าง s_h หรือ $\pi_{hij} = P(i,j \in s_h)$ เมื่อ $i,j \in U_h$ และ $h \in H$ กำหนดให้ E,V แทนค่าคาดหมายและความแปรปรวนทั้งหมด E_S, V_S แทนค่าคาดหมาย และความแปรปรวนเทียบกับแผนการสุ่มตัวอย่างตามลำดับ

กำหนดให้แต่ละชั้นภูมิ h มีข้อมูลสูญหาย ให้ s_{hr} แทนเซตของตัวอย่างที่ไม่เกิดการสูญหาย ขนาด n_{hr} และ $s_{hm} = s_h - s_{hr}$ ขนาด $n_{hm} = n_h - n_{hr}$ แทนเซตของตัวอย่างที่เกิดการสูญหาย เมื่อ $h \in H$ กำหนดให้ตัวแปรกำหนดขนาด (Size variable) k มีความสัมพันธ์กับ y นอกจากนี้ ในแต่ละชั้นภูมิ h ตัวแปรกำหนดขนาด k ทราบค่า และมีค่าบวกทุก k_{hi} เมื่อ $i \in U_h$ และ $h \in H$ กำหนดให้การสูญหายของข้อมูลเกิดขึ้นเฉพาะตัวแปรที่สนใจศึกษา และเกิดขึ้นเฉพาะการไม่ตอบเฉพาะบางคำถาม กำหนดให้สัญลักษณ์ R แทนกลไกการไม่ตอบสนองของข้อมูล (Nonresponse mechanism) รวมถึงกำหนดให้ r_{hi} แทนตัวแปรบอกสภาพของการสูญหาย ของหน่วยตัวอย่างที่ i ในชั้นภูมิที่ h ซึ่ง $r_{hi} = 1$ เมื่อ y_{hi} ไม่เกิดการสูญหาย และ $r_{hi} = 0$ เมื่อ y_{hi} เกิดการสูญหาย เมื่อ $i \in U_h$ และ $h \in H$ กำหนดให้ $\boldsymbol{R} = (r_{11} \quad r_{12} \quad \dots \quad r_{hN_h})'$ แทนเวกเตอร์ของตัวแปรบอกสภาพของการสูญหาย p_{hi} แทนความน่าจะเป็นที่จะตอบสนองของหน่วยตัวอย่างที่ i ในชั้นภูมิที่ h และกำหนดค่า $p_{hi} = P(r_{hi} = 1)$ สมมติว่า $p_{hi} = p_h$ ทุกค่า i เมื่อ $i \in U_h$ และ $h \in H$ ในแต่ละชั้นภูมิกำหนดให้ความน่าจะเป็นที่จะไม่สูญหายของหน่วยตัวอย่างที่ i เป็นอิสระกับหน่วยอื่น ๆ หรือ $p_{hij} = P(r_{hi} = 1, r_{hj} = 1) = p_h p_h = p_h^2$ เมื่อ $i,j \in U_h$ และ $h \in H$ กำหนดให้ E_R, V_R แทนค่าคาดหมายและความแปรปรวนเทียบกับกลไกการตอบสนอง ดังนั้น $E_R(r_{hi}) = P(r_{hi} = 1) = p_h$ และ $V_R(r_{hi}) = p_h(1 - p_h)$ เมื่อ $i,j \in U_h$ และ $h \in H$

4.2 การสุ่มตัวอย่างในแต่ละชั้นภูมิ

การวิจัยในครั้งนี้ผู้วิจัยสุ่มตัวอย่างในแต่ละชั้นภูมิ โดยใช้แผนการสุ่มตัวอย่างตามที่ [8] นำเสนอดังนี้ ขั้นแรกในแต่ละชั้นภูมิสุ่มตัวอย่างขนาดหนึ่งหน่วย จากประชากร U_h ขนาด N_h ด้วยแผนการสุ่มตัวอย่างแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาด (probability proportional to size) จากนั้นสุ่มตัวอย่างขนาด $n_h - 1$ จากประชากรขนาด $N_h - 1$ โดยใช้แผนการสุ่มตัวอย่างแบบง่ายและไม่ใส่คืน กำหนดให้ $K_h = \sum_{i \in U_h} k_{hi}$ ซึ่งภายใต้แผนการสุ่มตัวอย่างดังกล่าวค่า π_{hi} และ π_{hij} กำหนดค่าตามที่แสดงในสมการที่ (4.1) และ (4.2) ตามลำดับ

$$\pi_{hi} = \frac{k_{hi}}{K_h}\frac{N_h - n_h}{N_h - 1} + \frac{n_h - 1}{N_h - 1} \tag{4.1}$$

$$\pi_{hij} = \frac{k_{hi} + k_{hj}}{K}\frac{N_h - n_h}{N_h - 1}\frac{n_h - 1}{N_h - 2} + \frac{n_h - 1}{N_h - 1}\frac{n_h - 2}{N_h - 2} \tag{4.2}$$

เมื่อ $i,j \in U_h$ และ $h \in H$ ค่า k_{hi} แทนค่าสังเกตุของตัวแปรกำหนดขนาด ของหน่วยตัวอย่างที่ i ในชั้นภูมิที่ h เมื่อ $h \in H$

4.3 นำเสนอตัวประมาณค่าผลรวมประชากร

กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถามและแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย ผู้วิจัยนำเสนอตัวประมาณค่าผลรวมประชากรในรูปแบบที่เป็นเชิงเส้นเขียนแทนด้วย $\hat{Y}_{LMI}$ และรูปแบบที่ไม่เป็นเชิงเส้นเขียนแทนด้วย $\hat{Y}_{NLMI}$

4.4 นำเสนอตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร

การหาตัวประมาณค่าความแปรปรวนของ $\hat{Y}_{LMI}$ เขียนแทนด้วย $\hat{V}(\hat{Y}_{LMI})$ สามารถหาได้จากตัวประมาณค่าของ [4] ตามที่แสดงในสมการ (1.3) ส่วนการหาตัวประมาณค่าความแปรปรวนของ $\hat{Y}_{NLMI}$ เขียนแทนด้วย $\hat{V}(\hat{Y}_{NLMI})$ จำเป็นต้องปรับ $\hat{Y}_{NLMI}$ ให้เป็นตัวประมาณค่าแบบเชิงเส้น และหาตัวประมาณค่าความแปรปรวนภายใต้เฟรมเวิร์คแบบรีเวิร์ส

4.5 แสดงตัวอย่างเชิงตัวเลข

4.5.1 ประชากรที่ใช้ในการศึกษา

ประชากรที่ใช้ในการศึกษา เป็นข้อมูลที่ได้จากการจำลองสถานการณ์ โดยใช้เทคนิคมอนติคาร์โลซิมูเลชั่น (Monte carlo simulation) ตามการศึกษาของ [9] ซึ่งกำหนดข้อมูลของประชากร ตามที่แสดงในสมการ (4.3) ดังนี้

$$y_i = \begin{cases} 5k_i^2 + 0.3k_i + e_{1i} & , i \in U_1 \\ 5k_i^2 + 0.3k_i + 0.8(k_i - 2) + e_{2i} & , i \in U_2 \end{cases} \tag{4.3}$$

เมื่อ $k_i \sim Gamma(9,1.5)$ $e_{1i} \sim N(0,4)$ และ $e_{2i} \sim N(0,4)$ ส่วน $N_1 = 1,500$ $N_2 = 1,200$ ดังนั้น $N = 2,700$

4.5.2 ตัวอย่างที่ใช้ในการศึกษา

จากประชากรในแต่ละชั้นภูมิ U_1 และ U_2 ขนาด $N_1 = 1,500$ และ $N_2 = 1,200$ ทำการสุ่มตัวอย่างขนาด n_1 และ n_2 โดยใช้แผนการสุ่มตัวอย่างของ Midzuno ทั้งหมด 8 ระดับ ตามที่แสดงในตารางที่ 1

ตารางที่ 1 ขนาดตัวอย่างที่ใช้ในการศึกษา

ชุดตัวอย่างที่	n_1	n_2	n
1	30	20	50
2	45	25	60
3	45	30	75
4	60	40	100
5	70	50	120
6	85	55	140
7	90	65	155
8	100	70	170

4.5.3 กำหนดความน่าจะเป็นที่จะตอบสนองและการสูญหายของข้อมูล

กำหนดความน่าจะเป็นที่จะตอบสนอง โดยชั้นภูมิที่ 1 กำหนดความน่าจะเป็นที่จะตอบสนองของแต่ละหน่วยตัวอย่างมีค่าเท่ากัน คือ $p_{1i} = p_1 = 0.6$; $i \in U_1$ และ กำหนดให้ความน่าจะเป็นที่จะตอบสนองของแต่ละหน่วยตัวอย่างในชั้นภูมิที่ 2 มีค่าเท่ากัน คือ $p_{2i} = p_2 = 0.8$; $i \in U_2$ จากนั้นกำหนดการสูญหายของ

ข้อมูลในแต่ละชั้นภูมิ โดยพิจารณาจากตัวแปรสุ่ม A_{hi} ซึ่ง $A_{hi} \sim U(0,1)$ เมื่อ $i \in U_h$ และ $h = 1,2$ ถ้า $A_{hi} \leq p_{hi}$ ค่า $r_{hi} = 1$ และ $r_{hi} = 0$ เมื่อ $A_{hi} > p$

4.5.4 เปรียบเทียบประสิทธิภาพของตัวประมาณค่า

การวิจัยในครั้งนี้ใช้เทคนิคมอนติคาร์โลซิมูเลชั่นและกำหนดการทำซ้ำทั้งหมด 100,000 ครั้ง เพื่อเปรียบเทียบประสิทธิภาพของ $\hat{V}(\hat{Y}_{LMI})$ และ $\hat{V}(\hat{Y}_{NLMI})$ โดยพิจารณาจากค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ตามที่แสดงในสมการ (4.4)

$$RMSE(\hat{V}) = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}[\hat{V}_m - V^*]^2}}{V^*}, \tag{4.4}$$

เมื่อ $V^* = \frac{1}{M}\sum_{m=1}^{M}(\hat{Y}_m - \bar{Y}^*)^2$ ค่า $\bar{Y}^* = \frac{1}{M}\sum_{i=1}^{M}\hat{Y}_m$ และ $\hat{V}_m$ แทนตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร ของการสุ่มตัวอย่างครั้งที่ m เมื่อ $m = 1,2,...,M$ และ $M = 100,000$

สำหรับเกณฑ์การเปรียบเทียบประสิทธิภาพของตัวประมาณค่า [10] กล่าวว่าตัวประมาณค่า $\hat{V}_1$ เป็นตัวประมาณค่าที่มีประสิทธิภาพดีกว่า $\hat{V}_2$ ก็ต่อเมื่อ $RMSE(\hat{V}_1) \leq RMSE(\hat{V}_2)$

5 ผลวิจัย

5.1 ตัวประมาณค่าผลรวมประชากร

กรณีที่ประชากร U ขนาด N ถูกแบ่งออกเป็น H ชั้นภูมิ $U_1,U_2,...,U_h,...,U_L$ ขนาด $N_1,N_2,...,N_h,...,N_L$ ตามลำดับ สมมติว่าในแต่ละชั้นภูมิ h ตัวอย่าง s_h ขนาด n_h ถูกสุ่มประชากร U_h ขนาด N_h เมื่อ $h \in H$ โดยใช้แผนการสุ่มตัวอย่างแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มตัวอย่างตามแผนการสุ่มตัวอย่างของ Midzuno กรณีที่ไม่มีข้อมูลสูญหาย การประมาณค่าผลรวมประชากร สามารถใช้ตัวประมาณค่าตามที่ [4] นำเสนอดังนี้

$$\hat{Y} = \sum_{h\in H}\sum_{i\in s_h}\frac{y_{hi}}{\pi_{hi}} \tag{5.1}$$

เมื่อ π_{hi} กำหนดค่าตามสมการที่ (4.1) กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม ค่า y_{hi} เมื่อ $i \in s_{hm}$ จะไม่สามารถสังเกตุค่าได้ ดังนั้นจำเป็นต้องแทนที่ข้อมูลสูญหายด้วยค่าประมาณค่าจากการคำนวณ ซึ่งการศึกษาในครั้งนี้ผู้วิจัยเลือกใช้การแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย เขียนแทนด้วยสัญลักษณ์ y_{hi}^* ซึ่ง $y_{hi}^* = \sum_{i\in s_h}\frac{r_{hi}y_{hi}}{\pi_{hi}} \Big/ \sum_{i\in s_h}\frac{r_{hi}}{\pi_{hi}}$ [11] กล่าวว่าหลังจากประมาณค่าข้อมูลสูญหายเซตของตัวอย่าง s_h ขนาด n_h ในแต่ละชั้นภูมิจะไม่มีข้อมูลสูญหาย เนื่องจากค่าข้อมูลสูญหายในเซตของข้อมูลที่เกิดการสูญหาย $s_{hm} = s_h - s_{hr}$ ขนาด $n_{hm} = n_h - n_{hr}$ ถูกแทนที่ด้วยค่า y_{hi}^* เมื่อ $i \in s_{hm}$ กำหนดให้ $z_h^* = \{z_{h1}^*, z_{h2}^*, ..., z_{hn_h}^*\}$ แทนเซตของข้อมูลที่สมบูรณ์หลังจากที่แทนที่ข้อมูลสูญหายค่าเฉลี่ย ในชั้นภูมิที่ h เมื่อ $h \in H$ ซึ่ง $z_{hi}^* = y_{hi}$ เมื่อหน่วยตัวอย่างที่ i ไม่สูญหาย และ $z_{hi}^* = y_{hi}^*$ เมื่อหน่วยตัวอย่างที่ i สูญหาย ดังนั้นการประมาณค่าผลรวมประชากรสามารถทำได้โดยใช้ตัวประมาณค่าที่ [4] นำเสนอ ตามที่แสดงในสมการ (5.2) และเรียกตัวประมาณค่าดังกล่าวว่าตัวประมาณค่าแบบเชิงเส้น

$$\hat{Y}_{LMI} = \sum_{h\in H}\sum_{i\in s_h}\frac{z^*_{hi}}{\pi_{hi}} \tag{5.2}$$

เมื่อ π_{hi} ; $h \in H$ กำหนดค่าตามสมการที่ (4.1) และจากสมการที่ (5.2) สามารถเขียนตัวประมาณค่าผลรวมประชากรให้อยู่ในรูปไม่เป็นเชิงเส้น และเรียกตัวประมาณค่าดังกล่าวว่าตัวประมาณค่าแบบไม่เป็นเชิงเส้น ตามที่แสดงในสมการ (5.3)

$$\begin{aligned}\hat{Y}_{NLMI} &= \sum_{h\in H}\left[\sum_{i\in s_h}\frac{r_{hi}y_{hi}}{\pi_{hi}} + \sum_{i\in s_h}\frac{(1-r_{hi})y^*_{hi}}{\pi_{hi}}\right] \\ &= \sum_{h\in H}\left[\frac{\sum_{i\in s_h}\frac{1}{\pi_{hi}}\sum_{i\in s_h}\frac{r_{hi}y_{hi}}{\pi_{hi}}}{\sum_{i\in s_h}\frac{r_{hi}}{\pi_{hi}}}\right]\end{aligned} \tag{5.3}$$

เมื่อ π_{hi} ; $h \in H$ กำหนดค่าตามที่แสดงในสมการที่ (4.1)

5.2 ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร

หัวข้อที่แล้วผู้วิจัยได้นำเสนอตัวประมาณค่าผลรวมประชากร กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม และแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย ซึ่งพิจารณาทั้งรูปแบบที่เป็นตัวประมาณค่าแบบเชิงเส้น และตัวประมาณค่าแบบไม่เป็นเชิงเส้น สำหรับหัวข้อนี้ผู้วิจัยจะได้ นำเสนอวิธีการประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรทั้งแบบที่เป็นและไม่เป็นเชิงเส้นดังนี้

5.2.1 ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบเชิงเส้น

จากสมการที่ (11) เซตข้อมูล $z^*_h = \{z^*_{h1}, z^*_{h2}, ..., z^*_{hn_h}\}$ ในแต่ละชั้นภูมิเป็นเซตของข้อมูลที่ไม่เกิดการสูญหาย และ $\hat{Y}_{LMI}$ อยู่ในรูปตัวประมาณค่าของ [4] ดังนั้นค่าความแปรปรวนของ $\hat{Y}_{LMI}$ สามารถหาค่าได้จาก

$$V(\hat{Y}_{LMI}) = \sum_{h\in H}\sum_{i\in U_h}\left(\frac{1-\pi_{hi}}{\pi_{hi}}\right)z^{*2}_{hi} + \sum_{h\in H}\sum_{i\in U_h}\sum_{j\neq i\in U_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hi}\pi_{hj}}\right)z^*_{hi}z^*_{hj} \tag{5.4}$$

ส่วนตัวประมาณค่าของ $V(\hat{Y}_{LMI})$ เขียนแทนด้วยสัญลักษณ์ $\hat{V}(\hat{Y}_{LMI})$ และกำหนดค่าดังนี้

$$\hat{V}(\hat{Y}_{LMI}) = \sum_{h\in H}\sum_{i\in s_h}\left(\frac{1-\pi_{hi}}{\pi^2_{hi}}\right)z^{*2}_{hi} + \sum_{h\in H}\sum_{i\in s_h}\sum_{j\neq i\in s_h}\left(\frac{\pi_{hij}-\pi_{hi}\pi_{hj}}{\pi_{hij}\pi_{hi}\pi_{hj}}\right)z^*_{hi}z^*_{hj} \tag{5.5}$$

เมื่อ π_{hi} และ π_{hij} กำหนดค่าตามที่แสดงในสมการ (4.1) และสมการ (4.2) ตามลำดับ

5.2.2 ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบไม่เป็นเชิงเส้น

[11] กล่าวว่ากรณีที่ข้อมูลสูญหายมีขนาดใหญ่ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร ตามที่แสดงในสมการที่ (5.5) จะมีประสิทธิภาพลดลง ดังนั้นกรณีที่สัดส่วนตัวอย่างในแต่ละชั้นภูมิมีขนาดเล็ก [12] ได้นำเสนอวิธีการประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร เมื่อพิจารณาตัวประมาณค่าผลรวมประชากรเป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้น ซึ่งสูตรที่ [12] นำเสนอเป็นดังนี้

$$V(\hat{Y}_{NLMI}) \approx E_R V_S(\hat{Y}_{NLMI} \mid \boldsymbol{R}) \tag{5.6}$$

เมื่อ $\boldsymbol{R} = (r_{11} \quad r_{12} \quad \ldots \quad r_{HN})'$ แทนเวกเตอร์ของตัวแปรบอกสภาพของการตอบสนอง [12] กล่าวว่า การประมาณค่า $E_R V_S(\hat{Y}_{NLMI} \mid \boldsymbol{R})$ คล้ายกับการประมาณค่า $V_S(\hat{Y}_{NLMI} \mid \boldsymbol{R})$ เนื่องจากค่า r_{hi} เมื่อ $i \in U_h$ และ $h \in H$ ถูกมองว่าเป็นค่าเฉพาะของหน่วยตัวอย่างที่ i และทราบค่า สำหรับการหาค่า $V_S(\hat{Y}_{NLMI} \mid \boldsymbol{R})$ เนื่องจากตัวประมาณผลรวมประชากร $\hat{Y}_{NLMI}$ ตามที่แสดงในสมการ (5.3) เป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้น ดังนั้นจำเป็นต้องเขียน $\hat{Y}_{NLMI}$ ให้อยู่ในรูปของฟังก์ชันปรับเรียบ (Smooth function) และเขียนแทนด้วยสัญลักษณ์ ψ จากนั้นใช้เทเลอร์แบบเชิงเส้น เพื่อปรับฟังก์ชันปรับเรียบของ $\hat{Y}_{LMI}$ ที่ไม่เป็นเชิงเส้น ให้เป็นเชิงเส้นดังนี้ จากสมการที่ (5.3) สามารถเขียน

$$\hat{Y}_{NLMI} = \sum_{h\in H}\left[\frac{\sum_{i\in s_h}\frac{1}{\pi_{hi}}\sum_{i\in s_h}\frac{r_{hi}y_{hi}}{\pi_{hi}}}{\sum_{i\in s_h}\frac{r_{hi}}{\pi_{hi}}}\right]$$
$$= \sum_{h\in H}\frac{\hat{T}_{h1}\hat{T}_{h2}}{\hat{T}_{h3}} = \frac{\hat{T}_{11}\hat{T}_{12}}{\hat{T}_{13}} + \frac{\hat{T}_{21}\hat{T}_{22}}{\hat{T}_{23}} + \ldots + \frac{\hat{T}_{H1}\hat{T}_{H2}}{\hat{T}_{H3}} = \psi(\hat{\mathrm{T}}) \qquad (5.7)$$

เมื่อ

$$\hat{\mathrm{T}} = \left[\sum_{i\in s_1}\frac{1}{\pi_{1i}} \quad \sum_{i\in s_1}\frac{r_{1i}y_{1i}}{\pi_{1i}} \quad \sum_{i\in s_1}\frac{r_{1i}}{\pi_{1i}} \quad \cdots \sum_{i\in s_H}\frac{1}{\pi_{Hi}} \quad \sum_{i\in s_H}\frac{r_{Hi}y_{Hi}}{\pi_{Hi}} \quad \sum_{i\in s_H}\frac{r_{Hi}}{\pi_{Hi}}\right]' = \left[\hat{T}_{11} \quad \hat{T}_{12} \quad \hat{T}_{13}\cdots \quad \hat{T}_{H1} \quad \hat{T}_{H2} \quad \hat{T}_{H3}\right]'$$

$$\mathrm{T} = E_S(\hat{\mathrm{T}}) = \left[N_1 \quad \sum_{i\in U_1} r_{1i}y_{1i} \quad \sum_{i\in U_1} r_{1i} \quad \cdots N_h \quad \sum_{i\in U_H} r_{Hi}y_{Hi} \quad \sum_{i\in U_H} r_{Hi}\right]' = \left[T_{11} \quad T_{12} \quad T_{13} \quad \cdots T_{H1} \quad T_{H2} \quad T_{H3}\right]'$$

ดังนั้นเทเลอร์แบบเชิงเส้น $\hat{Y}_{NLMI} = \psi(\hat{\mathrm{T}})$ รอบค่า T กำหนดค่าดังนี้

$$\hat{Y}_{NLMI} = \psi(\hat{\mathrm{T}}) \approx \psi(\mathrm{T}) + \sum_{h\in H}\sum_{j=1}^{3H}\left.\frac{\partial\psi(\hat{\mathrm{T}})}{\partial\hat{T}_{hj}}\right|_{\hat{\mathrm{T}}=\mathrm{T}}(\hat{T}_{hj} - T_{hj})$$
$$= Constant + \sum_{h\in H}\frac{1}{T_{h3}}\sum_{i\in s_h}\frac{z_{hi}}{\pi_{hi}} \qquad (5.8)$$

เมื่อ $z_{hi} = T_{h2} + T_{h1}r_{hi}(y_{hi} - \bar{Y}_{Rh})$ ค่า $T_{h2} = \sum_{i\in U_h} r_{hi}y_{hi}$ ค่า $T_{h3} = \sum_{i\in U_h} r_{hi}$ และ $\bar{Y}_{Rh} = \frac{\sum_{i\in U_h} r_{hi}y_{hi}}{\sum_{i\in U_h} r_{hi}}$ และจากสมการที่ (5.8) สามารถหาค่าความแปรปวนของ $\hat{Y}_{NLMI}$ เทียบกับแผนการสุ่มตัวอย่าง เมื่อกำหนด R หรือ $V_S(\hat{Y}_{NLMI} \mid \mathrm{R}) = V_S(\psi(\hat{\mathrm{T}}) \mid \mathrm{R})$ ตามวิธีที่ [4] นำเสนอดังนี้

$$V_S(\hat{Y}_{NLMI} \mid \mathrm{R}) = V_S(\psi(\hat{\mathrm{T}}) \mid \boldsymbol{R})$$
$$= V_S\left(Constant + \sum_{h\in H}\frac{1}{T_{h3}}\sum_{i\in s_h}\frac{z_{hi}}{\pi_{hi}}\middle|\boldsymbol{R}\right)$$
$$= \sum_{h\in H}\frac{1}{T_{h3}^2}\left[\sum_{i\in U_h}\frac{(1-\pi_{hi})}{\pi_{hi}}z_{hi}^2 + \sum_{i\in U_h}\sum_{j\neq i\in U_h}\frac{(\pi_{hij} - \pi_{hi}\pi_{hj})}{\pi_{hi}\pi_{hj}}z_{hi}z_{hj}\right] \qquad (5.9)$$

เมื่อ z_{hi} กำหนดค่าตามสมการที่ (5.8) ค่า π_{hi} และ π_{hij} กำหนดค่าตามสมการที่ (4.1) และสมการที่ (4.2) ตามลำดับ ส่วนค่า $\hat{V}$ มีค่าตามที่แสดงในสมการที่ (5.20)

$$\hat{V}(\hat{Y}_{NLMI}) = \hat{V}_S(\hat{Y}_{NLMI}) = \sum_{h\in H}\frac{1}{\hat{T}_{h3}^2}\left[\sum_{i\in s_h}\frac{(1-\pi_{hi})}{\pi_{hi}^2}\hat{z}_{hi}^2 + \sum_{i\in s_h}\sum_{j\neq i\in s_h}\frac{(\pi_{hij}-\pi_{hi}\pi_{hj})}{\pi_{hij}\pi_{hi}\pi_{hj}}\hat{z}_{hi}\hat{z}_{hj}\right] \qquad (5.10)$$

เมื่อ $\hat{z}_{hi} = \hat{T}_{h2} + \hat{T}_{h1}r_{hi}(y_{hi} - \hat{\bar{Y}}_{Rh})$ ค่า $\hat{T}_{h2} = \sum_{i\in s_h}\frac{r_{hi}y_{hi}}{\pi_{hi}}$ ค่า $\hat{T}_{h3} = \sum_{i\in s_1}\frac{r_{1i}}{\pi_{1i}}$ และ $\hat{\bar{Y}}_{Rh} = \frac{\sum_{i\in s_h}\frac{r_{hi}y_{hi}}{\pi_{hi}}}{\sum_{i\in s_h}\frac{r_{hi}}{\pi_{hi}}}$ และ $i \in s_h; h \in H$

5.2.3 ผลการเปรียบเทียบประสิทธิภาพ

การเปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบเชิงเส้นเขียนแทนด้วยสัญลักษณ์ $\hat{V}(\hat{Y}_{LMI})$ และตัวประมาณค่าแบบไม่เป็นเชิงเส้นเขียนแทนด้วยสัญลักษณ์ $\hat{V}(\hat{Y}_{NLMI})$ ผู้วิจัยใช้เทคนิคมอนติคาร์โลซิมูเลชั่น ตามการศึกษาของ [9] ซึ่งพิจารณาประชากร U ขนาด N ออกเป็น 2 ชั้นภูมิ ($H = 2$) และเปรียบเทียบประสิทธิภาพของ $\hat{V}(\hat{Y}_{LMI})$ และ $\hat{V}(\hat{Y}_{NLMI})$ จากค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ ผลการเปรียบเทียบแสดงค่าตามตารางที่ 2

ตารางที่ 2 ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยของ $\hat{V}(\hat{Y}_{LMI})$ และ $\hat{V}(\hat{Y}_{NLMI})$

ชุดตัวอย่างที่	n_1	n_2	n	$MSE[\hat{V}(\hat{Y}_{LMI})]$	$MSE[\hat{V}(\hat{Y}_{NLMI})]$
1	30	20	50	0.5748	***0.5364***
2	45	25	60	0.5748	***0.4488***
3	45	30	75	0.5690	***0.4276***
4	60	40	100	0.5588	***0.3738***
5	70	50	120	0.5556	***0.3397***
6	85	55	140	0.5489	***0.3021***
7	90	65	155	0.5494	***0.2928***
8	100	70	170	0.5474	***0.2808***

6 สรุปผลการวิจัยและอภิปรายผล

การศึกษาในครั้งนี้ มีวัตถุประสงค์เพื่อหาตัวประมาณค่าผลรวมประชากร และหาตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม และแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย ภายใต้การสุ่มตัวอย่างโดยใช้ความน่าจะเป็นในการเลือกไม่เท่ากัน โดยผู้วิจัยเลือกใช้แผนการเลือกตัวอย่างแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มตัวอย่างโดยใช้แผนการสุ่มตัวอย่างของ Midzuno การหาตัวประมาณค่าความแปรปรวนของประมาณค่าผลรวมประชากรผู้วิจัยพิจารณาตัวประมาณค่าผลรวมประชากรทั้งแบบที่เป็นเชิงเส้น และไม่เป็นเชิงเส้น สำหรับตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้นผู้วิจัยปรับให้เป็นเชิงเส้นโดยใช้เทเลอร์แบบเชิงเส้นอันดับที่หนึ่ง จากนั้นหาตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร ตามสูตรที่ [4] นำเสนอ

การเปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบที่ไม่เป็นเชิงเส้น กับแบบที่เป็นเชิงเส้นผู้วิจัยใช้เทคนิคมอนติคาร์โลซิมูเลชัน ทำซ้ำทั้งหมด 100,000 ครั้ง โดยแบ่งประชากรออกเป็น 2 ชั้นภูมิ และกำหนดความน่าจะเป็นที่จะตอบสนองของแต่ละหน่วยตัวอย่างในชั้นภูมิที่ 1

เท่ากับ 0.6 และชั้นภูมิที่ 2 เท่ากับ 0.8 นอกจากนี้ในแต่ละชั้นภูมิผู้วิจัยสุ่มตัวอย่างทั้งหมด 8 ระดับ ตามแผนการสุ่มตัวอย่างของ Midzuno ผลการเปรียบเทียบประสิทธิภาพพบว่าตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้นมีค่ารากที่สองของความคลาดเคลื่อนสัมพัทธ์เฉลี่ยต่ำกว่าตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบเชิงเส้น เนื่องจากการหาตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบเชิงเส้น จะพิจารณาเซตตัวอย่างที่ถูกสุ่มจากประชากร เป็นเซตที่ไม่มีข้อมูลสูญหายเนื่องจากแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ยจากเซตตัวอย่างที่ไม่สูญหาย ดังนั้นการหาตัวประมาณค่าความแปรปรวน สามารถหาได้จากสูตรตามที่ [4] นำเสนอ [11] กล่าวว่ากรณีที่ข้อมูลสูญหายมีขนาดใหญ่ ตัวประมาณค่าผลรวมประชากรแบบเชิงเส้นจะเป็นตัวประมาณค่าที่เอนเอียง ซึ่งจะส่งผลให้ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์มีค่าเพิ่มขึ้น

สำหรับตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น จะพิจารณาเซตของตัวอย่างที่ถูกสุ่มจากประชากรมีข้อมูลสูญหาย ดังนั้นการประมาณค่าความแปรปรวนของตัวประมาณค่าดังกล่าวจะพิจารณาภายใต้เฟรมเวิร์คแบบรีเวิร์ส และสามารถหาค่าได้ตามสูตรที่ [12] นำเสนอ การหาค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น จำเป็นต้องปรับให้เป็นเชิงเส้นโดยใช้เทเลอร์แบบเชิงเส้น จากนั้นหาตัวประมาณค่าความแปรปรวนตามสูตรที่ [4] นำเสนอ ตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น จะเป็นตัวประมาณค่าที่เอนเอียงเช่นเดียวกับตัวประมาณค่าแบบเชิงเส้น แต่ค่าความเอนเอียงของตัวประมาณค่าแบบไม่เป็นเชิงเส้นจะมีค่าต่ำกว่าตัวประมาณค่าแบบเชิงเส้นและจะมีค่าเข้าใกล้ 0 เมื่อตัวอย่างมีขนาดใหญ่ ซึ่งจะส่งผลให้ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบไม่เป็นเชิงเส้น มีค่าต่ำกว่าตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบเชิงเส้น

กิตติกรรมประกาศ งานวิจัยนี้ได้รับการสนับสนุนจาก สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏเพชรบูรณ์

References

[1] ณรงค์ โพธิ และ สมชาย ปราการเจริญ (2010). ***กรอบแนวความคิดสำหรับการประมาณค่าสูญหายของข้อมูลด้วยเทคนิคการรู้จำรูปแบบ.*** The 6TH National Conference on Computing and Information Technology. 686-687.

[2] ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และ สุคนธ์ ประสิทธิ์วัฒนเสรี (2006). ***ข้อมูลสูญหายและแนวทางการจัดการ***. Data Management & Biostatistics Journal, 4(3):53 – 57.

[3] เพียงออ ยีสา และกัลยา วานิชย์บัญชา. (2009*). การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้น*. Nation graduate research conference 12th, 560-562.

[4] Horvitz, D.G., and D.J. Thompson (1952). *A generalization of sampling without replacement from a finite universe***.** Journal of the American Statistical Association, 47:663-685.

[5] Bethlehem, J.G. (1988). *Reduction of nonresponse bias through regression estimation. Journal of Official Statistics,* 4:251-260.

[6] Haziza, D.(2010).*Resampling methods for variance estimation in the presence of missing survey data.* Proceedings of the Annual Conference of the Italian Statistical Society.

[7] ชูเกียรติ โพนแก้ว (2559). *การประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรในการสำรวจตัวอย่างเมื่อมีข้อมูลสูญหาย.*วารสารวิชาการ มหาวิทยาลัยราชภัฏอุตรดิตถ์. 11(2): 68-80.

[8] Midzuno, H. (1952). *On sampling system with probability proportional to sum of sizes,* Ann. Inst.Statist. Math, 3:99-107.

[9] Cheng, Y., Slud, E., Hogue, C. (2010). *Variance Estimation for Decision-based Estimators with Application to the Annual Survey of Public Employment* , JSM Proceedings.

[10] พัชรีพร สนรักษา, พัชรี วงษ์เกษม, คณินทร์ ธีรภาพโอฬาร. (2010). *การประมาณค่าเฉลี่ยประชากรในการสำรวจตัวอย่างเมื่อมีข้อมูลสูญหลาย.*วารสารวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ, 2010 28(1): 89-103.

[11] Deville, J.-C., and Särndal, C.-E. (1944). *Variance Estimation for the Regression Imputed HorvitzThompson Estimator.* Journal of Official Statistics, 10(4):381-394.

[12] Horvitz, D.G., and D.J. Thompson (1952). *A generalization of sampling without replacement from a finite universe.* Journal of the American Statistical Association, 47:663-685.

ภาคผนวก ค
ประวัติผู้วิจัย

ประวัติผู้วิจัย

1. ชื่อ - นามสกุล (ภาษาไทย) นายชูเกียรติ โพนแก้ว
   ชื่อ - นามสกุล (ภาษาอังกฤษ) Mr.Chugiat Ponkaew
2. เลขหมายบัตรประจำตัวประชาชน 3330101443210
3. ตำแหน่งปัจจุบัน อาจารย์
   เงินเดือน 28,0600 บาท
   เวลาที่ใช้ทำวิจัย 15 ชั่วโมง : สัปดาห์
4. หน่วยงานและสถานที่อยู่ที่ติดต่อได้สะดวก พร้อมหมายเลขโทรศัพท์ โทรสาร และไปรษณีย์อิเล็กทรอนิกส์ (e-mail)
   สาขาวิชาคณิตศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเพชรบูรณ์
   83 หมู่ 11 ถนนสระบุรี-หล่มสัก ตำบลสะเดียง อำเภอเมือง จังหวัดเพชรบูรณ์
   67000 โทร. 056-717100 ต่อ 1407
   e-mail : chugiat.pon@pcru.ac.th
5. ประวัติการศึกษา
   พ.ศ. 2547 วิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์ มหาวิทยาลัยขอนแก่น
   พ.ศ. 2544 วิทยาศาสตรบัณฑิต สาขาวิชาคณิตศาสตร์ มหาวิทยาลัยบูรพา
6. สาขาวิชาการที่มีความชำนาญพิเศษ (แตกต่างจากวุฒิการศึกษา) ระบุสาขาวิชาการการศึกษา
7. ประสบการณ์ที่เกี่ยวข้องกับการบริหารงานวิจัยทั้งภายในและภายนอกประเทศ โดยระบุสถานภาพในการทำการวิจัยว่าเป็นผู้อำนวยการแผนงานวิจัย หัวหน้าโครงการวิจัย หรือผู้ร่วมวิจัยในแต่ละผลงานวิจัย
   7.1 ผู้อำนวยการแผนงานวิจัย : -
   7.2 หัวหน้าโครงการวิจัย : เป็นหัวหน้าโครงการวิจัยดังรายละเอียดในข้อ 7.3
   7.3 งานวิจัยที่ทำเสร็จแล้ว : ชื่อผลงานวิจัย ปีที่พิมพ์ การเผยแพร่ และแหล่งทุน (อาจมากกว่า 1 เรื่อง)
   - การศึกษารูปแบบการเรียนรู้และผลสัมฤทธิ์ทางการเรียนของนักศึกษาที่เรียนแบบร่วมมือด้วยวิธีแบ่งกลุ่มผลสัมฤทธิ์ วิชาคณิตศาสตร์เบื้องต้น ประจำปีงบประมาณ พ.ศ. 2552
     สถานะโครงการ : เสร็จสิ้นโครงการ
   - ปัจจัยที่มีผลต่อ ต้นทุนการผลิต ผลผลิต และผลตอบแทน จากการปลูกข้าวโพดเลี้ยงสัตว์ของเกษตรกรในเขตจังหวัดเพชรบูรณ์ ประจำปีงบประมาณ พ.ศ. 2554
     สถานะโครงการ : เสร็จสิ้นโครงการ