

PROCEEDINGS

The 22nd Annual Meeting in Mathematics
June 2 - 4, 2017



การประชุมวิชาการทางคณิตศาสตร์
ประจำปี 2560 ครั้งที่ 22
2 - 4 มิถุนายน พ.ศ. 2560

โดย

ศูนย์ส่งเสริมและพัฒนาคุณภาพ
การเรียนรู้คณิตศาสตร์
โรงเรียนสาธิตมหาวิทยาลัยราชภัฏวชิรเวศน์
จังหวัดราชบุรี



TPI

การประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม

ชูเกียรติ โพนแก้ว *

หลักสูตรสาขาวิชาคณิตศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเพชรบูรณ์

บทคัดย่อ

บทความนี้มีวัตถุประสงค์เพื่อนำเสนอและเปรียบเทียบตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรกรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถามและแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรทั้งแบบที่เป็นและไม่เป็นเชิงเส้นถูกพิจารณาภายใต้การสุ่มตัวอย่างแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มตัวอย่างโดยใช้ความน่าจะเป็นในเลือกไม่เท่ากันตามแผนการสุ่มตัวอย่างของ Midzuno ผลการเปรียบเทียบประสิทธิภาพโดยพิจารณาจากราคที่สองของความคลาดเคลื่อนพบว่าตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบไม่เป็นเชิงเส้นมีประสิทธิภาพดีกว่าตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบเป็นเชิงเส้น

คำสำคัญ: ข้อมูลสูญหาย การประมาณค่าความแปรปรวน เทเลอร์แบบเชิงเส้น

1 บทนำ

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

* งานวิจัยเรื่องนี้ได้รับทุนสนับสนุนจากสถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏเพชรบูรณ์

¹ผู้แต่งหลัก

²ผู้พูด

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

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

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

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

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

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

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

กรณีที่ข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม และแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย [5] กล่าวว่า การประมาณค่าผลรวมประชากรควรขึ้นอยู่กับชุดของข้อมูลที่ไม่ว่างหาย มากกว่าใช้ชุดของตัวอย่างที่ถูกสุ่มมาทั้งหมด ดังนั้น [6] ได้นำเสนอตัวประมาณค่าผลรวมประชากร ภายใต้การสุ่มตัวอย่างแบบง่ายและไม่ใส่กลับ (Simple random sampling without replacement) ต่อมา [7] นำเสนอตัวประมาณค่าผลรวมประชากรภายใต้การ

สุ่มตัวอย่างโดยใช้ความน่าจะเป็นในการเลือกไม่เท่ากัน (Unequal probability sampling) และใช้แผนการสุ่มตัวอย่างของ [8] ในการสุ่มตัวอย่าง 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 นำเสนอตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร ทั้งแบบที่เป็นเชิงเส้นและไม่เป็นเชิงเส้น
- 2.3 เปรียบเทียบประสิทธิภาพระหว่างตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบที่เป็นเชิงเส้น กับแบบที่ไม่เป็นเชิงเส้น โดยพิจารณาจากค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์

3 สมมติฐานของการวิจัย

การสุ่มตัวอย่างกรณีที่มีข้อมูลสูญหายมีขนาดใหญ่ ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบเชิงเส้นจะมีประสิทธิภาพลดลง ดังนั้นการศึกษานี้ผู้วิจัยกำหนดสมมติฐานคือตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้นมีประสิทธิภาพดีกว่าตัวประมาณค่าแบบเชิงเส้น

4 วิธีดำเนินการวิจัย

4.1 สัญลักษณ์และข้อตกลงเบื้องต้น

พิจารณาประชากร $U = \{1, 2, \dots, i, \dots, N\}$ ขนาด N ถูกแบ่งออกเป็น L ชั้นภูมิ $U_1, U_2, \dots, U_h, \dots, U_L$ ขนาด $N_1, N_2, \dots, N_h, \dots, N_L$ ตามลำดับ กำหนดให้ $H = \{1, 2, \dots, h, \dots, L\}$ 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 แทนค่าคาดหวังและความแปรปรวนทั้งหมด E_s, V_s แทนค่าคาดหวัง และความแปรปรวนเทียบกับแผนการสุ่มตัวอย่างตามลำดับ

กำหนดให้แต่ละชั้นภูมิ h มีข้อมูลสูญหาย ให้ s_{hi} แทนเซตของตัวอย่างที่ไม่เกิดการสูญหาย ขนาด n_{hi} และ $s_{hmi} = s_h - s_{hi}$ ขนาด $n_{hmi} = n_h - n_{hi}$ แทนเซตของตัวอย่างที่เกิดการสูญหาย เมื่อ $h \in H$ กำหนดให้ตัวแปรกำหนดขนาด (Size variable) k มีความสัมพันธ์กับ y นอกจากนี้ ในแต่ละชั้นภูมิ h ตัวแปรกำหนดขนาด k ทราบค่าและมีค่าบวกทุก k_{hi} เมื่อ $i \in U_h$ และ $h \in H$ กำหนดให้การสูญหายของข้อมูลเกิดขึ้นเฉพาะตัวแปรที่สนใจศึกษา และเกิดขึ้นเฉพาะการไม่ตอบเฉพาะบางคำถาม กำหนดให้สัญลักษณ์ R แทนกลไกการไม่ตอบสนองของข้อมูล (Nonresponse mechanism) รวมถึงกำหนดให้ r_{hi} แทนตัวแปรบอกสภาพของการสูญหายของหน่วยตัวอย่างที่ 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 $R = (r_{h1}, r_{h2}, \dots, r_{hN_h})'$ แทนเวกเตอร์ของตัวแปรบอกสภาพ

ของการสูญหาย 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 เป็นอิสระกับหน่วยอื่นๆ 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 แทนค่าคาดหวังและความแปรปรวนเทียบกับกลไกการตอบสนอง 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] นำเสนอ ดังนี้ 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

$$\pi_{hi} = \frac{k_{hi}}{K_h} \frac{N_h - n_h}{N_h - 1} \frac{n_h - 1}{N_h - 1} \quad (4.1)$$

$$\pi_{hij} = \frac{k_{hi} \cdot k_{hj}}{K_h} \frac{N_h - n_h}{N_h - 1} \frac{n_h - 1}{N_h - 1} \frac{n_h - 2}{N_h - 1} \frac{n_h - 2}{N_h - 2} \quad (4.2)$$

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

4.3 นำเสนอตัวประมาณค่าผลรวมประชากร

กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถามและแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย ผู้วิจัยนำเสนอตัวประมาณค่าผลรวมประชากรในรูปแบบที่เป็นเชิงเส้นเขียนแทนด้วย $\hat{Y}_{LM}$ และรูปแบบที่ไม่เป็นเชิงเส้นเขียนแทนด้วย $\hat{Y}_{NLM}$

4.4 นำเสนอตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร

การหาตัวประมาณค่าความแปรปรวนของ $\hat{Y}_{LM}$ เขียนแทนด้วย $\hat{V}(\hat{Y}_{LM})$ สามารถหาได้จากตัวประมาณค่าของ [4] ตามที่แสดงในสมการ (1.3) ส่วนการหาตัวประมาณค่าความแปรปรวนของ $\hat{Y}_{NLM}$ เขียนแทนด้วย $\hat{V}(\hat{Y}_{NLM})$ จำเป็นต้องปรับ $\hat{Y}_{NLM}$ ให้เป็นตัวประมาณค่าแบบเชิงเส้น และหาตัวประมาณค่าความแปรปรวนภายใต้เฟรมเวิร์กแบบรีเวิร์ส

4.5 แสดงตัวอย่างเชิงตัวเลข

4.5.1 ประชากรที่ใช้ในการศึกษา

ประชากรที่ใช้ในการศึกษา เป็นข้อมูลที่ได้จากการจำลองสถานการณ์ โดยใช้เทคนิคมอนติคาร์โลซิโมเลชัน (Monte carlo simulation) ตามการศึกษามาก่อน [9] ซึ่งกำหนดข้อมูลของประชากร ตามที่แสดงในสมการ (4.3) ดังนี้

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

เมื่อ $k_i \sim \text{Gamma}(9, 1.5)$ $e_{1i} \sim N(0, 4)$ 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 $p_1 = p_1 = 0.6 : i \in U_1$ และ กำหนดให้ความน่าจะเป็นที่จะตอบสนองของแต่ละหน่วยตัวอย่างในชั้นภูมิที่ 2 มีค่าเท่ากัน คือ $p_2 = p_2 = 0.8 : i \in U_2$ จากนั้นกำหนดการสูญหายของ

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

4.5.4 เปรียบเทียบประสิทธิภาพของตัวประมาณค่า

การวิจัยในครั้งนี้ใช้เทคนิคมอนติคาร์โลซิมูเลชันและกำหนดการทำซ้ำทั้งหมด 100,000 ครั้ง เพื่อเปรียบเทียบประสิทธิภาพของ $\hat{V}(\hat{Y}_{LM})$ และ $\hat{V}(\hat{Y}_{NLM})$ โดยพิจารณาจากค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ตามที่แสดงในสมการ (4.4)

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

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

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

5 ผลวิจัย

5.1 ตัวประมาณค่าผลรวมประชากร

กรณีที่ประชากร U ขนาด N ถูกแบ่งออกเป็น H ชั้นภูมิ $U_1, U_2, \dots, U_h, \dots, U_H$ ขนาด $N_1, N_2, \dots, N_h, \dots, N_H$ 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

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

เมื่อ π_{hi} กำหนดค่าตามสมการที่ (4.1) กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม ค่า y_{hi} เมื่อ $i \in s_{h_{mi}}$ จะไม่สามารถสังเกตค่าได้ ดังนั้นจึงเป็นต้องแทนที่ข้อมูลสูญหายด้วยค่าประมาณค่าจากการคำนวณ ซึ่งการศึกษาในครั้งนี้ผู้วิจัยเลือกใช้การแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย เขียนแทนด้วยสัญลักษณ์ y_{hi}^* ซึ่ง $y_{hi}^* = \sum_{i \in s_{h_{mi}}} \frac{y_{hi}}{\pi_{hi}} / \sum_{i \in s_{h_{mi}}} \frac{1}{\pi_{hi}}$ [11] กล่าวว่าหลังจากประมาณค่าข้อมูลสูญหายเซตของตัวอย่าง $s_{h_{mi}}$ ขนาด $n_{h_{mi}}$ ในแต่ละชั้นภูมิจะไม่มีข้อมูลสูญหาย เนื่องจากค่าข้อมูลสูญหายในเซตของข้อมูลที่เกิดการสูญหาย $s_{h_{mi}} = s_{h_{mi}} - s_{h_{mi}}$ ขนาด $n_{h_{mi}} = n_h - n_{h_{mi}}$ ถูกแทนที่ด้วยค่า y_{hi}^* เมื่อ $i \in s_{h_{mi}}$ กำหนดให้ $z_{hi}^* = \{z_{hi1}^*, z_{hi2}^*, \dots, z_{hin_{hi}}^*\}$ 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

$$\hat{Y}_{LMF} = \sum_{h \in H} \sum_{i \in \mathcal{N}_h} \frac{z_{hi}^*}{\pi_{hi}} \quad (5.2)$$

เมื่อ $\pi_{hi} : h \in H$ กำหนดค่าตามสมการที่ (4.1) และจากสมการที่ (5.2) สามารถเขียนตัวประมาณค่าผลรวมประชากรให้อยู่ในรูปไม่เป็นเชิงเส้น และเรียกตัวประมาณค่าดังกล่าวว่าตัวประมาณค่าแบบไม่เป็นเชิงเส้นตามที่แสดงในสมการ (5.3)

$$\begin{aligned} \hat{Y}_{NLMF} &= \sum_{h \in H} \left[\sum_{i \in \mathcal{N}_h} \frac{r_{hi} y_{hi}}{\pi_{hi}} + \sum_{i \in \mathcal{N}_h} \frac{(1 - r_{hi}) y_{hi}^*}{\pi_{hi}} \right] \\ &= \sum_{h \in H} \left[\frac{\sum_{i \in \mathcal{N}_h} \frac{1}{\pi_{hi}} \sum_{i \in \mathcal{N}_h} \frac{r_{hi} y_{hi}}{\pi_{hi}}}{\sum_{i \in \mathcal{N}_h} \frac{r_{hi}}{\pi_{hi}}} \right] \end{aligned} \quad (5.3)$$

เมื่อ $\pi_{hi} : h \in H$ กำหนดค่าตามที่แสดงในสมการที่ (4.1)

5.2 ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากร

หัวข้อที่แล้วผู้วิจัยได้นำเสนอตัวประมาณค่าผลรวมประชากร กรณีที่มีข้อมูลสูญหายที่เกิดจากการไม่ตอบเฉพาะบางคำถาม และแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย ซึ่งพิจารณาทั้งรูปแบบที่เป็นตัวประมาณค่าแบบเชิงเส้น และตัวประมาณค่าแบบไม่เป็นเชิงเส้น สำหรับหัวข้อนี้ผู้วิจัยจะได้ นำเสนอวิธีการประมาณค่าความแปรปรวนของตัวประมาณค่าผลรวมประชากรทั้งแบบที่เป็นและไม่เป็นเชิงเส้นดังนี้

5.2.1 ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบเชิงเส้น

จากสมการที่ (11) เซตข้อมูล $z_{hi}^* = \{z_{hi1}^*, z_{hi2}^*, \dots, z_{hin_{hi}}^*\}$ ในแต่ละชั้นภูมิเป็นเซตของข้อมูลที่ไม่เกิดการสูญหาย และ $\hat{Y}_{LMF}$ อยู่ในรูปตัวประมาณค่าของ [4] ดังนั้นค่าความแปรปรวนของ $\hat{Y}_{LMF}$ สามารถหาค่าได้จาก

$$V(\hat{Y}_{LMF}) = \sum_{h \in H} \sum_{i \in \mathcal{N}_h} \left(\frac{1 - \pi_{hi}}{\pi_{hi}} \right) z_{hi}^{*2} + \sum_{h \in H} \sum_{i \in \mathcal{N}_h} \sum_{j \in \mathcal{N}_h} \left(\frac{\pi_{hij} - \pi_{hi} \pi_{hj}}{\pi_{hi} \pi_{hj}} \right) z_{hi}^* z_{hj}^* \quad (5.4)$$

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

$$\hat{V}(\hat{Y}_{LMF}) = \sum_{h \in H} \sum_{i \in \mathcal{N}_h} \left(\frac{1 - \pi_{hi}}{\pi_{hi}^2} \right) z_{hi}^{*2} + \sum_{h \in H} \sum_{i \in \mathcal{N}_h} \sum_{j \in \mathcal{N}_h} \left(\frac{\pi_{hij} - \pi_{hi} \pi_{hj}}{\pi_{hi} \pi_{hj} \pi_{hj}} \right) z_{hi}^* z_{hj}^* \quad (5.5)$$

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

5.2.2 ตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบไม่เป็นเชิงเส้น

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

$$V(\hat{Y}_{NLMF}) = E_R V_S(\hat{Y}_{NLMF} | R) \quad (5.6)$$

เมื่อ $\mathbf{R} = (r_{11} \ r_{12} \ \dots \ r_{HS})'$ แทนเวกเตอร์ของตัวแปรบอกสภาพของการตอบสนอง [12] กล่าวว่าการประมาณค่า $E_R V_S(\hat{Y}_{NLM} | \mathbf{R})$ คล้ายกับการประมาณค่า $V_S(\hat{Y}_{NLM} | \mathbf{R})$ เนื่องจากค่า r_{hi} เมื่อ $i \in U_h$ และ $h \in H$ ถูกมองว่าเป็นค่าเฉพาะของหน่วยตัวอย่างที่ i และทราบค่า สำหรับการหาค่า $V_S(\hat{Y}_{NLM} | \mathbf{R})$ เนื่องจากตัวประมาณผลรวมประชากร $\hat{Y}_{NLM}$ ตามที่แสดงในสมการ (5.3) เป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้น ดังนั้นจำเป็นต้องเขียน $\hat{Y}_{NLM}$ ให้อยู่ในรูปของฟังก์ชันปรีเรียบ (Smooth function) และเขียนแทนด้วยสัญลักษณ์ ψ จากนั้นใช้เทเลอร์แบบเชิงเส้น เพื่อปรับฟังก์ชันปรีเรียบของ $\hat{Y}_{NLM}$ ที่ไม่เป็นเชิงเส้น ให้เป็นเชิงเส้นดังนี้ จากสมการที่ (5.3) สามารถเขียน

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

เมื่อ

$$\hat{\mathbf{T}} = \left[\sum_{i \in u_1} \frac{1}{\pi_{1i}} \sum_{i \in u_1} r_{1i} y_{1i} \sum_{i \in u_1} \frac{r_{1i}}{\pi_{1i}} \dots \sum_{i \in u_H} \frac{1}{\pi_{Hi}} \sum_{i \in u_H} r_{Hi} y_{Hi} \sum_{i \in u_H} \frac{r_{Hi}}{\pi_{Hi}} \right]' = [\hat{T}_{11} \ \hat{T}_{12} \ \hat{T}_{13} \ \dots \ \hat{T}_{H1} \ \hat{T}_{H2} \ \hat{T}_{H3}]'$$

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

ดังนั้นเทเลอร์แบบเชิงเส้น $\hat{Y}_{NLM} = \psi(\hat{\mathbf{T}})$ รอบค่า $\mathbf{T}$ กำหนดค่าดังนี้

$$\hat{Y}_{NLM} = \psi(\hat{\mathbf{T}}) = \psi(\mathbf{T}) + \sum_{h \in H} \sum_{i=1}^{3H} \left. \frac{\partial \psi(\hat{\mathbf{T}})}{\partial \hat{T}_{hi}} \right|_{\hat{\mathbf{T}}=\mathbf{T}} (\hat{T}_{hi} - T_{hi})$$

$$= Constant + \sum_{h \in H} \frac{1}{T_{h3}} \sum_{i \in u_h} \frac{z_{hi}}{\pi_{hi}} \quad (5.8)$$

เมื่อ $z_{hi} = T_{h2} - T_{h1} r_{hi}(y_{hi} - \bar{Y}_{Hh})$ ถ้า $T_{h2} = \sum_{i \in U_h} r_{hi} y_{hi}$ ถ้า $T_{h3} = \sum_{i \in U_h} r_{hi}$ และ $\bar{Y}_{Hh} = \frac{\sum_{i \in U_h} r_{hi} y_{hi}}{\sum_{i \in U_h} r_{hi}}$ และจากสมการที่ (5.8)

สามารถหาค่าความแปรปรวนของ $\hat{Y}_{NLM}$ เทียบกับแผนการสุ่มตัวอย่าง เมื่อกำหนด $\mathbf{R}$ หรือ $V_S(\hat{Y}_{NLM} | \mathbf{R}) = V_S(\psi(\hat{\mathbf{T}}) | \mathbf{R})$ ตามวิธีที่ [4] นำเสนอดังนี้

$$V_S(\hat{Y}_{NLM} | \mathbf{R}) = V_S(\psi(\hat{\mathbf{T}}) | \mathbf{R})$$

$$= V_S \left(Constant + \sum_{h \in H} \frac{1}{T_{h3}} \sum_{i \in u_h} \frac{z_{hi}}{\pi_{hi}} \middle| \mathbf{R} \right)$$

$$= \sum_{h \in H} \frac{1}{T_{h3}^2} \left[\sum_{i \in U_h} \frac{(1 - \pi_{hi})}{\pi_{hi}} z_{hi}^2 + \sum_{i \in U_h} \sum_{j \in U_h} \frac{(\pi_{hi} - \pi_{hi} \pi_{hj})}{\pi_{hi} \pi_{hj}} z_{hi} z_{hj} \right] \quad (5.9)$$

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

$$\hat{V}(\hat{Y}_{NLM}) = \hat{V}_s(\hat{Y}_{NLM}) - \sum_{h \in H} \frac{1}{\hat{T}_{h,3}} \left[\sum_{i \in s_h} \frac{(1 - \pi_{hi})}{\pi_{hi}} \hat{z}_{hi}^2 + \sum_{i \in s_h} \sum_{j \in (s_h - s_i)} \frac{(\pi_{h(ij)} - \pi_{hi} \pi_{hj})}{\pi_{h(ij)} \pi_{hi} \pi_{hj}} \hat{z}_{hi} \hat{z}_{hj} \right] \quad (5.10)$$

$$\text{เมื่อ } \hat{z}_{hi} = \hat{T}_{h,2} \cdot \hat{T}_{hi} r_{hi} (y_{hi} - \hat{Y}_{Rhi}) \text{ ถ้า } \hat{T}_{h,2} = \sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi}} \text{ ถ้า } \hat{T}_{h,3} = \sum_{i \in s_h} \frac{r_{hi}}{\pi_{hi}} \text{ และ } \hat{Y}_{Rhi} = \frac{\sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi}}}{\sum_{i \in s_h} \frac{r_{hi}}{\pi_{hi}}} \text{ และ } i \in s_h; h \in H$$

5.2.3 ผลการเปรียบเทียบประสิทธิภาพ

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

ตารางที่ 2 ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยของ $\hat{V}(\hat{Y}_{LM})$ และ $\hat{V}(\hat{Y}_{NLM})$

ชุดตัวอย่างที่	n_1	n_2	n	$MSE[\hat{V}(\hat{Y}_{LM})]$	$MSE[\hat{V}(\hat{Y}_{NLM})]$
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] กล่าวว่าการมีข้อมูลที่สูญหายมีขนาดใหญ่ ตัวประมาณค่าผลรวมประชากรแบบเชิงเส้นจะเป็นตัวประมาณค่าที่เอนเอียง ซึ่งจะส่งผลให้ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์มีค่าเพิ่มขึ้น

สำหรับตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น จะพิจารณาเซตของตัวอย่างที่ถูกสุ่มจากประชากรมีข้อมูลสูญหาย ดังนั้นการประมาณค่าความแปรปรวนของตัวประมาณค่าดังกล่าวจะพิจารณาภายใต้เฟรมเวิร์กแบบบริวิวิธ และสามารถหาได้ตามสูตรที่ [12] นำเสนอ การหาความแปรปรวนของตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น จำเป็นต้องปรับให้เป็นเชิงเส้นโดยใช้เทเลอร์แบบเชิงเส้น จากนั้นหาตัวประมาณค่าความแปรปรวนตามสูตรที่ [4] นำเสนอ ตัวประมาณค่าผลรวมประชากรแบบไม่เป็นเชิงเส้น จะเป็นตัวประมาณค่าที่เอนเอียงเช่นเดียวกับตัวประมาณค่าแบบเชิงเส้น แต่ค่าความเอนเอียงของตัวประมาณค่าแบบไม่เป็นเชิงเส้นจะมีค่าต่ำกว่าตัวประมาณค่าแบบเชิงเส้นและจะมีค่าเข้าใกล้ 0 เมื่อตัวอย่างมีขนาดใหญ่ ซึ่งจะส่งผลให้ค่ารากที่สองของความคลาดเคลื่อนเฉลี่ยสัมพัทธ์ของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบไม่เป็นเชิงเส้น มีค่าต่ำกว่าตัวประมาณค่าความแปรปรวนของตัวประมาณค่าแบบเชิงเส้น

กิตติกรรมประกาศ งานวิจัยนี้ได้รับการสนับสนุนจาก สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏเพชรบูรณ์

References

- [1] ณรงค์ โปธิ และ สมชาย ปราการเจริญ (2010). กรอบแนวความคิดสำหรับการประมาณค่าสูญหายของข้อมูลด้วยเทคนิคการรู้จำรูปแบบ. The 6TH National Conference on Computing and Information Technology. 686-687.
- [2] ปิยะภรณ์ ประสิทธิ์วัฒนศิริ และ สุคนธ์ ประสิทธิ์วัฒนศิริ (2006). ข้อมูลสูญหายและแนวทางการจัดการ. Data Management & Biostatistics Journal, 4(3):53 – 57.
- [3] เพ็ญอ้อ 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 มหาวิทยาลัยราชภัฏอุดรดิตถ์. 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] พิชรีพร สันรักษา, พิชรี วังเกษม, คณินทร์ ชีรภาพโอฬาร. (2010). การประมาณค่าเฉลี่ยประชากรในการสำรวจตัวอย่างเมื่อมีข้อมูลสูญหาย.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.