

การประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากร ในกรณีที่ข้อมูลไม่ตอบสนอง เมื่อสัดส่วนตัวอย่างมีขนาดเล็ก

Estimating Variance for The Population Mean Estimator in The Presence of Nonresponse and Negligible Sampling Fraction

ชูเกียรติ โพนแก้ว^{1*} และ หยาดพิรุณ สุภารากุล²

Chugiat Ponkaew¹ and Yadpirun Supharakonsakun²

^{1,2} อาจารย์ หลักสูตรสาขาวิชาคณิตศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเพชรบูรณ์

*ผู้รับผิดชอบหลัก (corresponding author)

บทคัดย่อ

ข้อมูลไม่ตอบสนอง เป็นสาเหตุทำให้เกิดข้อมูลสูญหาย การแทนที่ข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ เป็นเทคนิคที่ใช้เกี่ยวกับการไม่ตอบสนองของข้อมูลที่เกิดจากการไม่ตอบเฉพาะบางคำถาม การวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อนำเสนอ และเปรียบเทียบประสิทธิภาพค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากรในกรณีที่ข้อมูลไม่ตอบสนอง และสัดส่วนตัวอย่างมีขนาดเล็ก ภายใต้เฟรมเวิร์คแบบรีเวิร์ส ผู้วิจัยกำหนดความน่าจะเป็นที่จะตอบสนอง เท่ากับ 0.3, 0.5, 0.7 และ 0.9 กำหนดขนาดตัวอย่างเท่ากับ 30 และ 45 จากนั้นทำการเลือกตัวอย่างสุ่มแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาดและใส่คืน ผลการเปรียบเทียบประสิทธิภาพ พบว่าค่าความคลาดเคลื่อน และค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากร ที่ได้จากวิธีของ Sampath มีค่าต่ำกว่าวิธีของ Horvitz-Thompson

คำสำคัญ : การไม่ตอบสนองที่เกิดจากการไม่ตอบเฉพาะบางคำถาม, สัดส่วนตัวอย่าง, ความน่าจะเป็นที่จะตอบสนอง, ข้อมูลสูญหาย

Abstract

Nonresponse data is caused of missing data. The Imputation technique is handled with the data which is item nonresponse. This research is aimed to purpose and compare estimating of variance for the population mean estimator in the presence of nonresponse and negligible sampling fraction under reverse framework. In experimentation, the sampling size is respectively defined as 30 and 45. Then, it was selected according to probability proportional to size with replacement sampling with probability of response is respectively defined as 0.3, 0.5, 0.7 and 0.9. Finally, we compare bias and root mean square error of propose estimator. Results show that bias and root mean square error

of estimator of variance of population mean estimator from Sampath method is lower than Horvitz and Thompson method.

Keywords : Item nonresponse, Sampling fraction, probability of response, Missing data

บทนำ

การทำงานวิจัยหรือการสำรวจสิ่งที่สนใจต่าง ๆ มักจะต้องเก็บรวบรวมข้อมูลแล้วนำมาวิเคราะห์ และประมวลผลให้ได้ข้อสรุปเพื่อนำมาตอบข้อสงสัย หรือแก้ไขปัญหาทางงานวิจัยหรือการสำรวจนั้น ๆ การเก็บรวบรวมข้อมูลจากการสำรวจ มักเกิดปัญหาข้อมูลไม่ตอบสนอง (Nonresponse) (ทัตตา หิรัญพต, 2554) ซึ่งเป็นสาเหตุทำให้เกิดการสูญหายของข้อมูล และส่งผลให้ตัวประมาณค่าของพารามิเตอร์ที่สนใจมีความเอนเอียง (Bias) นอกจากนี้ค่าความเอนเอียงของตัวประมาณค่าจะมีค่าเพิ่มขึ้น เมื่ออัตราส่วนของข้อมูลไม่ตอบสนองมีค่าเพิ่มขึ้น ปัญหาเกี่ยวกับการไม่ตอบสนองของข้อมูลถูกนำเสนอเป็นครั้งแรกโดย Hansen and Hurwitz (1946, 663-685) ขณะศึกษาเกี่ยวกับการสำรวจข้อมูลทางไปรษณีย์ (Mail survey) ต่อมา Armstrong and Overton (1997, 396-402) ศึกษาเกี่ยวกับความเอนเอียงของการสำรวจข้อมูลทางไปรษณีย์ โดยปกติข้อมูลไม่ตอบสนองจำแนกได้เป็น 2 ประเภทคือการไม่ตอบสนองสำหรับบางหน่วยตัวอย่าง (Unit nonresponse) และการไม่ตอบสนองที่เกิดจากการไม่ตอบเฉพาะบางคำถาม (Item nonresponse) วิธีจัดการกับข้อมูลที่ไม่ตอบสนองใช้วิธีการถ่วงน้ำหนัก (Weighting) สำหรับการไม่ตอบสนองสำหรับบางหน่วยตัวอย่าง ส่วนการไม่ตอบสนองที่เกิดจากการไม่ตอบเฉพาะบางคำถาม ใช้วิธีการแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ (Imputation)

การแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ อาศัยความสัมพันธ์ระหว่างข้อมูลเพื่อนำไปประมาณค่าข้อมูลที่สูญหาย การประมาณค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ มีหลายวิธี เช่นการประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย (Mean imputation หรือ MI) การประมาณข้อมูลสูญหายด้วยค่าฮอตเด็ค (Hot deck imputation หรือ HDI) การประมาณข้อมูลสูญหายด้วยค่าถดถอย (Regression imputation หรือ RI) (ทัตตา หิรัญพต, 2554)

พิจารณาประชากร $U = \{1, 2, \dots, i, \dots, N\}$ ขนาดจำกัด ขนาด N ตัวแปรที่สนใจศึกษาแทนด้วยตัวแปร y และพารามิเตอร์ที่ต้องการประมาณค่าคือค่าเฉลี่ยของประชากร (Population mean) เขียนแทนด้วยสัญลักษณ์ $\bar{Y} = N^{-1} \sum_{i \in U} y_i$ ซึ่ง y_i แทนค่าสังเกตของประชากรหน่วยที่ i เมื่อ $i = 1, 2, \dots, N$ ตัวอย่าง s ขนาด n ถูกสุ่มเลือกตามข้อมูลตกลงของแผนการสุ่มตัวอย่าง (Sampling design) แทนด้วยสัญลักษณ์ S โดยทั่วไปการประมาณค่าเฉลี่ยประชากร สามารถใช้ตัวประมาณค่าตามที่ Horvitz and Thompson (1952, 663-685) นำเสนอดังนี้

$$\hat{Y}_{HT} = N^{-1} \sum_{i \in S} d_i y_i \quad (1)$$

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

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

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

$$\hat{V}(\hat{Y}_{HT}) = N^{-2} \left[\sum_{i \in s} \left(\frac{1 - \pi_i}{\pi_i^2} \right) y_i^2 + \sum_{i \in s} \sum_{j \neq i \in s} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} \right) y_i y_j \right] \quad (3)$$

เมื่อ $d_i = \pi_i^{-1}$ แทนค่าถ่วงน้ำหนักของแผนการสุ่มตัวอย่าง (Design weight) และ 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 $\pi_{ij} = P(i, j \in s)$ เมื่อ $i = 1, 2, \dots, N$ และ $j = 1, 2, \dots, N$ และได้ว่า $\hat{Y}_{HT}$ เป็นตัวประมาณค่าที่ไม่เอนเอียงของ $\bar{Y}$ เนื่องจาก $E_s(\hat{Y}_{HT}) = \bar{Y}$

กรณีที่ข้อมูลไม่ตอบสนอง ตัวอย่าง s ขนาด n จำแนกได้เป็นตัวอย่างที่ตอบสนอง s_r ขนาด n_r และตัวอย่างที่ไม่ตอบสนอง s_m ขนาด $n - n_r$ ดังนั้นตัวประมาณค่าเฉลี่ยประชากร เมื่อใช้วิธีการแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณหาได้จาก

$$\hat{Y}_{HTI} = N^{-1} \sum_{i \in s} d_i y_i^* = N^{-1} \left[\sum_{i \in s} d_i r_i y_i + \sum_{i \in s} d_i (1 - r_i) y^* \right] \quad (4)$$

เมื่อ r_i แทนตัวแปรบอกสภาพของการตอบสนอง (Response indicator variable) ซึ่ง $r_i = 1$ เมื่อ y_i สามารถสังเกตค่าได้หรือ $i \in s_r$ และ $r_i = 0$ เมื่อ y_i สูญหาย หรือ $i \in s_m$ สำหรับค่า y_i^* กำหนดค่าดังนี้ $y_i^* = y_i$ ถ้า $i \in s_r$ และ $y_i^* = y^*$ ถ้า $i \in s_m$ และ y^* แทนค่าประมาณจากการคำนวณ (Imputed value) เพื่อใช้ในการแทนที่ข้อมูล y_i ที่สูญหาย สำหรับค่าความแปรปรวน และตัวประมาณค่าความแปรปรวนของ $\hat{Y}_{HTI}$ สามารถหาได้ โดยใช้สมการที่ (2) และสมการที่ (3) อย่างไรก็ตาม Deville and Särndal (1994, 381-394) กล่าวว่ากรณีที่ข้อมูลไม่ตอบสนองมีขนาดใหญ่ ความแปรปรวน และตัวประมาณของ $\hat{Y}_{HTI}$ ที่ได้จากสมการที่ (2) และสมการที่ (3) จะมีประสิทธิภาพลดลง ดังนั้น Haziza (2010) ได้นำเสนอตัวประมาณค่าเฉลี่ยประชากร โดยใช้วิธีการแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณดังนี้ จากสมการ (4) เมื่อแทนที่ข้อมูลสูญหายด้วยค่าเฉลี่ย หรือ $y^* = \sum_{i \in s} d_i r_i y_i / \sum_{i \in s} d_i r_i$ ได้ว่า

$$\hat{Y}_I = \frac{\sum_{i \in s} d_i \sum_{i \in s} d_i r_i y_i}{N \sum_{i \in s} d_i r_i} \quad (5)$$

Haziza (2010) กำหนดค่าความน่าจะเป็นที่ตอบสนองของหน่วยตัวอย่างที่ i เป็นอิสระกับหน่วยอื่น ด้วยความน่าจะเป็นคงที่ $p_i = p$ ซึ่งได้ว่า $E_R(r_i) = p$ และ $V_R(r_i) = p(1 - p)$ เมื่อ R แทนกลไกการไม่ตอบสนองของข้อมูล E_R และ V_R แทนค่าคาดหวังและความแปรปรวนเทียบกับกลไกการไม่ตอบสนอง ภายใต้การเลือกตัวอย่างสุ่มแบบง่ายและใส่คืน (Simple random sampling with replacement) ค่า $d_i = \pi_i^{-1} = N/n$ ทุกค่า $i = 1, 2, \dots, N$ ดังนั้น

$$\hat{Y}_I = \frac{\sum_{i \in s} d_i \sum_{i \in s} d_i r_i y_i}{N \sum_{i \in s} d_i r_i} = \frac{\sum_{i \in s} r_i y_i}{\sum_{i \in s} r_i} = \hat{Y}_r \quad (6)$$

นอกจากนี้ภายใต้เฟรมเวิร์กแบบรีเวิร์ส (Reverse framework) และเลือกตัวอย่างสุ่มแบบง่ายและใส่คืน รวมถึงสัดส่วนตัวอย่างมีขนาดเล็ก (Sampling fraction is negligible) Haziza (2010) ได้นำเสนอตัวประมาณค่าความแปรปรวนของ $\hat{Y}_I$ ตามที่แสดงในสมการที่ (7) ดังนี้

$$\hat{V}(\hat{Y}_I) = s_r^2 / n_r \quad (7)$$

เมื่อ $s_r^2 = (n_r - 1)^{-1} \sum_{i \in S} ri(y_i - \hat{Y}_r)^2$ จากข้างต้นการประมาณค่าเฉลี่ยประชากรในกรณีที่มีข้อมูลไม่ตอบสนองที่เกิดจากการไม่ตอบเฉพาะบางคำถาม สามารถประมาณค่าข้อมูลที่สูญหายโดยใช้วิธีการแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณตามที่ Haziza (2010) เสนอ แต่เนื่องจาก Haziza (2010) ได้นำเสนอตัวประมาณค่าเฉลี่ยประชากร โดยใช้การประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย เฉพาะการเลือกตัวอย่างสุ่มแบบง่ายและใส่คืน ดังนั้นการศึกษาในครั้งนี้มุ่งเน้นที่จะประมาณค่าเฉลี่ยประชากร และประมาณค่าความแปรปรวนของตัวประมาณค่าดังกล่าว ในกรณีที่มีข้อมูลไม่ตอบสนองที่เกิดจากการไม่ตอบเฉพาะบางคำถาม ภายใต้เฟรมเวิร์คแบบรีเวิร์ส และสุ่มตัวอย่างเลือกแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาด และใส่คืน (Probability proportional to size with replacement) ซึ่งภายใต้แผนการสุ่มตัวอย่างดังกล่าวกรณีที่ไม่มีข้อมูลสูญหาย Sampath (2001) ได้นำเสนอตัวประมาณค่าเฉลี่ยประชากร ความแปรปรวน และตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากร นอกจากนี้ผู้วิจัยได้เปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากรที่นำเสนอที่ได้จากวิธีของ Sampath (2001) กับตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากรที่นำเสนอที่ได้จากวิธีของ Horvitz and Thompson (1952, 663-685) โดยพิจารณาจากค่าความคลาดเคลื่อน(Bias) และค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย(Root mean square error)

วัตถุประสงค์ของการวิจัย

สำหรับการวิจัยในครั้งนี้มีวัตถุประสงค์ เพื่อ

1. นำเสนอตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากร ในกรณีที่มีข้อมูลไม่ตอบสนอง โดยใช้การประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย
2. เปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากรที่นำเสนอที่ได้จากวิธีของ Sampath กับตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากรที่นำเสนอที่ได้จากวิธีของ Horvitz-Thompson

วรรณกรรมและทฤษฎีที่เกี่ยวข้อง

การประมาณค่าเฉลี่ยประชากร และประมาณค่าความแปรปรวนของตัวประมาณค่าดังกล่าว ในกรณีที่มีข้อมูลไม่ตอบสนองที่เกิดจากการไม่ตอบเฉพาะบางคำถามผู้วิจัยได้ทบทวนวรรณกรรมและทฤษฎีที่เกี่ยวข้องดังนี้

1. ประเภทของการสูญหายของข้อมูล

ข้อมูลไม่ตอบสนอง เป็นสาเหตุทำให้เกิดข้อมูลสูญหาย ซึ่ง Rubins (1976) ได้จำแนกประเภทของข้อมูลสูญหายออกเป็น 3 แบบดังนี้

- การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Complete at Random: MCAR) การสูญหายของข้อมูลด้วยวิธีการสุ่มอย่างสมบูรณ์เกิดขึ้นเมื่อความน่าจะเป็นของการสูญหายของข้อมูลไม่มีความสัมพันธ์กับค่าของ

ข้อมูลตัวอื่น ๆ ไม่ว่าจะเป็นข้อมูลที่เราพบค่า หรือข้อมูลที่เกิดการสูญหายด้วยกันก็ตาม นั่นคือความน่าจะเป็นที่จะเกิด การสูญหายของค่าข้อมูลในทุก ๆ ตำแหน่งมีค่าเท่ากัน

- การสูญหายแบบสุ่ม (Missing at Random: MAR) คือความน่าจะเป็นที่ข้อมูลสูญหาย ขึ้นอยู่ กับค่าของข้อมูลอื่น ๆ ที่เราพบค่าข้อมูล แต่ไม่ขึ้นกับค่าของข้อมูล

- การสูญหายแบบไม่สุ่ม (Not Missing at Random: NMAR) สาเหตุของการสูญหายไม่ สามารถบอกได้ และ สาเหตุของการสูญหายจะเกี่ยวข้องกับตัวแปรที่มีข้อมูลสูญหาย จะเรียกว่า เป็น Non-Ignorable ข้อมูลสูญหายสำหรับตัวแปร Y เป็น Non-Ignorable ถ้าความน่าจะเป็นของค่าสูญหายไปของ Y ไม่มีความสัมพันธ์กับค่า ของตัวแปรอื่น ๆ แต่จะมีความสัมพันธ์กับค่าของตัวเอง

2. เทเลอร์แบบเชิงเส้น (Taylor linearization)

การหาค่าคาดหวัง และความแปรปรวนของตัวประมาณค่าพารามิเตอร์ที่สนใจศึกษา กรณีที่ตัว ประมาณค่าเป็นแบบเชิงเส้น (Linear estimator) การหาค่าคาดหวัง และความแปรปรวนสามารถทำได้โดยตรง กรณีที่ ตัวประมาณค่าไม่เชิงเส้น (Nonlinear estimator) จำเป็นต้องใช้ เทเลอร์แบบเชิงเส้น (Taylor linearization) เพื่อ ปรับตัวประมาณค่าที่ไม่เป็นเชิงเส้น ให้เป็นเชิงเส้น จากนั้นหาค่าคาดหวังและความแปรปรวน เช่นเดียวกับตัวประมาณ ค่าแบบเชิงเส้น Haziza (2010) ได้นำเสนอขั้นตอนเทเลอร์แบบเชิงเส้น โดยเริ่มจากเขียนตัวประมาณค่าที่ไม่เป็นเชิงเส้น ให้อยู่ในรูปของฟังก์ชันปรับเรียบ (Smooth function) เขียนแทนด้วยสัญลักษณ์ $\psi(\hat{T})$ จากนั้นปรับตัวประมาณค่าที่ไม่ เป็นเชิงเส้นให้เป็นตัวประมาณค่าแบบเชิงเส้น โดยใช้เทเลอร์แบบเชิงเส้นอันดับที่หนึ่งดังนี้

$$\psi(\hat{T}) \approx \psi(T) + \sum_{j=1}^k \frac{\partial \psi(\hat{T})}{\partial \hat{T}_j} \bigg|_{\hat{T}=T} (\hat{T}_j - T_j) \quad (8)$$

เมื่อ $\hat{T} = [\hat{T}_1 \ \hat{T}_2 \ \dots \ \hat{T}_k]'$ กำหนดให้ $T = E_s(\hat{T}) = [T_1 \ T_2 \ \dots \ T_k]'$ และ $\frac{\partial \psi(\hat{T})}{\partial \hat{T}_j}$ แทนอนุพันธ์

ย่อย (Partial derivative) ของ $\psi(\hat{T})$ เทียบกับ $\hat{T}_j$ เมื่อ $j=1,2,\dots,k$

3. ตัวประมาณค่าเฉลี่ยประชากร ภายใต้การเลือกตัวอย่างสุ่มแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับ ขนาด และใส่คีน

กำหนดให้ y แทนตัวแปรที่สนใจศึกษา N แทนขนาดของประชากรขนาดจำกัด $U = \{1, 2, \dots, i, \dots, N\}$ $\bar{Y} = \sum_{i \in U} y_i / N$ แทนค่าเฉลี่ยของประชากร P_i แทนความน่าจะเป็นที่หน่วยตัวอย่างที่ i จะถูกเลือก ภายใต้การเลือกตัวอย่างสุ่มแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาด และใส่คีน ตัวประมาณค่าที่ไม่เอนเอียง ของค่าเฉลี่ยประชากรที่ Sampath (2001) นำเสนอเป็นดังนี้

$$\hat{Y}_{pps} = \frac{1}{nN} \sum_{i \in s} \frac{y_i}{P_i} \quad (9)$$

ความแปรปรวนของ $\hat{Y}_{pps}$ ที่ Sampath (2001) นำเสนอเป็นดังนี้

$$V(\hat{Y}_{pps}) = \frac{1}{nN^2} \sum_{i \in U} \left[\frac{y_i}{P_i} - \bar{Y} \right]^2 P_i \quad (10)$$

และตัวประมาณค่าความแปรปรวนของ $\hat{Y}_{pps}$ เป็นดังนี้

$$\hat{V}(\hat{Y}_{pps}) = \frac{1}{n(n-1)N^2} \sum_{i \in u} \left[\frac{y_i}{P_i} - \hat{Y} \right]^2 P_i \quad (11)$$

4. วรรณกรรมที่เกี่ยวข้องวิธีการแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ

Nangsue (2014) กล่าวว่าวิธีการแทนที่ค่าข้อมูลสูญหายด้วยค่าประมาณจากการคำนวณ หมายถึงการแทนที่ข้อมูลที่สูญหาย ด้วยตัวประมาณค่าที่สามารถคำนวณค่าได้ วิธีที่นิยมสำหรับการแทนที่ข้อมูลสูญหาย โดยใช้ค่าเดียว (Single imputation) ได้แก่ การประมาณข้อมูลสูญหายด้วยค่าเฉลี่ย การประมาณข้อมูลสูญหายด้วยค่าฮอท หรือ การประมาณข้อมูลสูญหายด้วยค่าถดถอย เป็นต้น

หัตถา หิรัญพต (2554) ศึกษาเกี่ยวกับการเปรียบเทียบประสิทธิภาพระหว่างวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยและวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยแบบสโตนแคสติก โดยกำหนดขนาดตัวอย่างเท่ากับ 50, 100, 200 และ 300 สุ่มโดยขั้นตอนวิธีของเบบบิงตัน (Bebbington algorithm) ซึ่งใช้วิธีการเลือกตัวอย่างสุ่มแบบง่ายและไม่ใส่คืน (Simple random sampling without replacement หรือ SRSWOR) และกำหนดเปอร์เซ็นต์ของข้อมูลสูญหายของตัวอย่าง เท่ากับ 5%, 10%, 15% และ 20% ของขนาดตัวอย่าง ผลการศึกษาการเปรียบเทียบประสิทธิภาพระหว่างวิธีการประมาณค่าข้อมูลสูญหายนั้น ได้วัดจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยของค่าประมาณของข้อมูลสูญหาย (MSE) โดยได้ข้อสรุปว่า วิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอย (RI) มีประสิทธิภาพมากกว่าวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยแบบสโตนแคสติก (SRI)

วิธีดำเนินการวิจัย

1. สัญลักษณ์และข้อตกลง

กำหนดให้ y แทนตัวแปรที่สนใจศึกษา N แทนขนาดของประชากรขนาดจำกัด $U = \{1, 2, \dots, i, \dots, N\}$ และ $\bar{Y} = N^{-1} \sum_{i \in U} y_i$ แทนค่าเฉลี่ยของประชากร กำหนดให้ s แทนเซตของตัวอย่างที่ถูกสุ่มขนาด n ตามข้อตกลงของแผนการเลือกตัวอย่างแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาด และใส่คืน กำหนดให้ s_r แทนเซตของตัวอย่างที่ตอบสนองขนาด n_r

กำหนดให้ E_S, V_S แทนค่าคาดหวังและความแปรปรวน เทียบกับแผนการสุ่มตัวอย่าง ในกรณีที่ข้อมูลไม่ตอบสนอง กำหนดให้สัญลักษณ์ R แทนกลไกการไม่ตอบสนองของข้อมูล (Nonresponse mechanism) r_i แทนตัวแปรบอกรสภาพ ของการตอบสนอง (Response indicator variable) กำหนดค่าโดย

$$r_i = \begin{cases} 1, & \text{เมื่อ } y_i \text{ สังเกตค่าได้} \\ 0, & \text{เมื่อ } y_i \text{ สูญหาย} \end{cases} \quad (12)$$

กำหนดให้ $\mathbf{R} = (r_1 \ r_2 \ \dots \ r_N)'$ แทนเวกเตอร์ของตัวแปรบอกรสภาพของการตอบสนอง กำหนดให้ p_i แทนความน่าจะเป็นที่จะตอบสนอง ซึ่งกำหนดค่าโดย $p_i = P(r_i = 1)$ สมมติว่า $p_i = p$ ทุกค่า $i \in U$ กำหนดให้ E_R, V_R แทนค่าคาดหวังและความแปรปรวนเทียบกับการตอบสนอง ดังนั้น $E_R(r_i) = P(r_i = 1) = p$ และ $V_R(r_i) = p(1-p)$ รวมถึงกำหนดให้สัดส่วนของตัวอย่าง (sampling fraction) หรือค่า n/N มีขนาดเล็ก

2. นำเสนอตัวประมาณค่าเฉลี่ยประชากร ในกรณีที่ข้อมูลไม่ตอบสนอง ภายใต้การสุ่มตัวอย่างเลือกแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาด และใส่คืน

3. หาค่าความแปรปรวน และตัวประมาณค่าความแปรปรวน ของตัวประมาณค่าเฉลี่ยประชากรภายใต้การสุ่มตัวอย่างเลือกแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาด และใส่คืน

ภายใต้เฟรมเวิร์คแบบรีเวิร์สความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากรสามารถหาได้จากสูตรที่ Shao and Steel (1999, 254-265) เสนอดังนี้

$$V(\hat{Y}) = E_R V_S(\hat{Y} | \mathbf{R}) + V_R E_S(\hat{Y} | \mathbf{R}) \quad (13)$$

เมื่อ $\hat{Y}$ แทนตัวประมาณค่าเฉลี่ยของประชากร กรณีที่สัดส่วนของตัวอย่างมีขนาดเล็ก Haziza (2010) กล่าวว่าความแปรปรวนของ $\hat{Y}$ สามารถหาได้จาก

$$V(\hat{Y}) \approx E_R V_S(\hat{Y} | \mathbf{R}) \quad (14)$$

กรณีที่สัดส่วนตัวอย่างมีขนาดเล็ก Haziza (2010) กล่าวว่าหาตัวประมาณค่าของ $V(\hat{Y})$ คล้ายกับการหาตัวประมาณค่าของ $V_S(\hat{Y} | \mathbf{R})$ เมื่อกำหนดเวกเตอร์ของตัวแปรบอกสภาพของการตอบสนอง หรือเมื่อกำหนด $\mathbf{R}$

กำหนดให้ $\hat{V}_S$ ซึ่ง $E_S(\hat{V}_S) = V_S(\hat{Y} | \mathbf{R})$ และ $E_R[E_S(\hat{V}_S)] = E_R[V_S(\hat{Y} | \mathbf{R})]$ ดังนั้นตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยของประชากร หาได้จาก

$$\hat{V}(\hat{Y}) \approx \hat{V}_S(\hat{Y} | \mathbf{R}) \quad (15)$$

4. เปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากรที่นำเสนอที่ได้จากวิธีของ Sampath (2001) กับตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากรที่นำเสนอที่ได้จากวิธีของ Horvitz and Thompson (1952, 663-685) โดยใช้ข้อมูลที่ได้จากการศึกษาของ Lohr (1999) ชื่อไฟล์ **agstrat.dat** ซึ่งประกอบด้วยประชากรขนาด $N = 299$ และตัวแปรที่สนใจศึกษาได้แก่ $y = ACRES92$ แทนจำนวนพื้นที่ที่ใช้ในการทำฟาร์มในปี 1992 ส่วนตัวแปรกำหนดขนาดได้แก่ $k = FARMS92$ แทนจำนวนฟาร์มในปี 1992 กำหนดให้ความน่าจะเป็นที่จะตอบสนองมีค่าคงที่เท่ากับ p ทั้งหมด 4 ระดับประกอบด้วย $p = 0.3, 0.5, 0.7$ และ 0.9 จากนั้นกำหนดข้อมูลสุ่มในแต่ละระดับ p โดยพิจารณาจากตัวแปรสุ่ม A_i ซึ่ง $A_i \sim U(0,1)$ เมื่อ $i = 1, 2, \dots, N$ ถ้า $A_i \leq p$ ค่า $r_i = 1$ และ $r_i = 0$ ในกรณีอื่นๆ จากนั้นสุ่มตัวอย่างเลือกแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาดและใส่คืน $n = 30, 45$ จำนวน 10,000 ครั้ง ($M = 100,000$) ในแต่ละระดับของความน่าจะเป็นที่จะตอบสนอง และขนาดตัวอย่าง จากนั้นทำการเปรียบเทียบประสิทธิภาพโดยใช้ค่าความคลาดเคลื่อน และค่ารากที่สองของค่าความคลาดเคลื่อน กำลังสองเฉลี่ย ตามที่ Berger (2003) นำเสนอดังนี้

- ค่าความคลาดเคลื่อน เขียนแทนด้วย $RB(\hat{\sigma}^2)$ ซึ่งกำหนดค่าดังนี้

$$RB(\hat{\sigma}^2) = \frac{1}{M} \sum_{m=1}^M \hat{\sigma}_m^2 - \sigma^2 \quad (16)$$

- ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเขียนแทนด้วย $RMSE(\hat{\sigma}^2)$ ซึ่งกำหนดค่าดังนี้

$$RMSE(\hat{\sigma}^2) = \sqrt{\frac{1}{M} \sum_{m=1}^M (\hat{\sigma}_m^2 - \sigma^2)^2} \quad (17)$$

เมื่อ $\sigma^2 = M^{-1} \sum_{m=1}^M (\hat{Y}_m - \bar{Y}^*)^2$, $\bar{Y}^* = M^{-1} \sum_{m=1}^M \hat{Y}_m$ และ $\hat{Y}_m$ ตัวประมาณค่าเฉลี่ยประชากรในรอบที่ m

ผลการวิจัย

1. ตัวประมาณค่าเฉลี่ยประชากร

พิจารณาตัวอย่าง s ขนาด n ถูกสุ่มจากประชากร U ขนาด N โดยใช้การเลือกตัวอย่างสุ่มแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาดและใส่คืน กำหนดให้ y_i แทนค่าสังเกตของหน่วยตัวอย่างที่ i ของตัวแปร y กำหนดให้ k_i แทนค่าสังเกตของหน่วยตัวอย่างที่ i ของตัวแปร k และ P_i แทนความน่าจะเป็นที่หน่วยตัวอย่างที่ i จะถูกเลือก ซึ่ง $P_i = k_i / \sum_{i \in U} k_i$ ภายใต้การเลือกตัวอย่างสุ่มแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาดและใส่คืนค่า $\pi_i = 1 - (1 - P_i)^n$ และ $\pi_{ij} = \pi_i + \pi_j - [1 - (1 - P_i - P_j)^n]$ ค่าเฉลี่ยประชากรในกรณีที่มีข้อมูลไม่ตอบสนองสามารถหาได้จากสมการที่ (4) ดังนี้

$$\hat{Y}_{HTI} = N^{-1} \sum_{i \in s} d_i y_i^* \quad (18)$$

$$\text{เมื่อ } y_i^* = y_i \text{ ถ้า } i \in s_r \text{ และ } y_i^* = y^* \text{ ถ้า } i \in s_m \text{ และ } y^* = \sum_{i \in s} d_i r_i y_i / \sum_{i \in s} d_i r_i$$

และจากสมการที่ (4) สามารถเขียนได้เป็น

$$\hat{Y}_{IPPS} = \frac{\sum_{i \in s} d_i \sum_{i \in s} d_i r_i y_i}{N \sum_{i \in s} d_i r_i} \quad (19)$$

2. ค่าความแปรปรวน และตัวประมาณค่าความแปรปรวน ของตัวประมาณค่าเฉลี่ยประชากร

เมื่อกำหนดเงื่อนไขตามหัวข้อที่ 1 และจากสมการที่ (19) สามารถหาค่าความแปรปรวนของ $\hat{Y}_{HTI}$ ตามวิธีที่ Horvitz and Thompson (1952, 663-685) นำเสนอดังนี้

$$V(\hat{Y}_{HTI}) = N^{-2} \left[\sum_{i \in U} \left(\frac{1 - \pi_i}{\pi_i} \right) y_i^{*2} + \sum_{i \in U} \sum_{j \neq i \in U} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_i^* y_j^* \right] \quad (20)$$

และตัวประมาณค่าของ $V(\hat{Y}_{HTI})$ แทนด้วย $\hat{V}(\hat{Y}_{HTI})$ ซึ่งกำหนดค่าดังนี้

$$\hat{V}(\hat{Y}_{HTI}) = N^{-2} \left[\sum_{i \in s} \left(\frac{1 - \pi_i}{\pi_i^2} \right) y_i^{*2} + \sum_{i \in s} \sum_{j \neq i \in s} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} \right) y_i^* y_j^* \right] \quad (21)$$

จากสมการที่ (19) สามารถหาค่าความแปรปรวนของ $\hat{Y}_{IPPS}$ ตามวิธีของ Sampath (2001) ได้ดังนี้

$$\begin{aligned} V(\hat{Y}_{IPPS}) &= E_R V_S \left[\frac{\sum_{i \in s} d_i \sum_{i \in s} d_i r_i y_i}{N \sum_{i \in s} d_i r_i} \mid \mathbf{R} \right] \\ &= \frac{1}{N^2} E_R V_S \left[\sum_{i \in s} d_i \frac{\sum_{i \in s} d_i r_i y_i}{\sum_{i \in s} d_i r_i} \mid \mathbf{R} \right] = \frac{1}{N^2} E_R V_S [\hat{Y}_{IPPS} \mid \mathbf{R}] \end{aligned} \quad (22)$$

เมื่อ $\hat{Y}_{IPPS} = \sum_{i \in s} d_i \frac{\sum_{i \in s} d_i r_i y_i}{\sum_{i \in s} d_i r_i}$ พิจารณาค่า $V_S [\hat{Y}_{IPPS} \mid \mathbf{R}]$ ตามที่แสดงในสมการที่ (22) เนื่องจาก

$\hat{Y}_{IPPS}$ เป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้น (Nonlinear estimator) จำเป็นต้องใช้เทเลอร์แบบเชิงเส้น

ในการปรับตัวประมาณค่าที่ไม่เป็นเชิงเส้นให้เป็นเชิงเส้นดังนี้ จากสมการที่ (22) สามารถเขียน

$$\hat{Y}_{IPPS} = \sum_{i \in S} d_i \frac{\sum_{i \in S} d_i r_i y_i}{\sum_{i \in S} d_i r_i} = \frac{\hat{T}_1 \hat{T}_2}{\hat{T}_3} = \psi(\hat{T}) \text{ ดังนั้น}$$

$$\hat{Y}_{IPPS} = \psi(\hat{T}) = \frac{\hat{T}_1 \hat{T}_2}{\hat{T}_3} \quad (23)$$

$$\text{เมื่อ } \hat{T} = \left[\sum_{i \in S} d_i \quad \sum_{i \in S} d_i r_i y_i \quad \sum_{i \in S} d_i r_i \right]' = [\hat{T}_1 \quad \hat{T}_2 \quad \hat{T}_3]'$$

$$\text{กำหนดให้ } T = E_S(\hat{T}) = \left[N \quad \sum_{i \in U} r_i y_i \quad \sum_{i \in U} r_i \right]' = [T_1 \quad T_2 \quad T_3]' \text{ เทเลอร์แบบเชิงเส้นอันดับหนึ่งของ}$$

$\hat{Y}_{IPPS} = \psi(\hat{T})$ รอบค่า T กำหนดค่าดังนี้

$$\psi(\hat{T}) \approx \psi(T) + \sum_{j=1}^3 \frac{\partial \psi(\hat{T})}{\partial \hat{T}_j} \bigg|_{\hat{T}=T} (\hat{T}_j - T_j) = \frac{1}{T_3} \sum_{i \in S} d_i [T_2 + T_1 r_i (y_i - \bar{Y}_R)] = \frac{1}{T_3} \sum_{i \in S} z_i \quad (24)$$

$$\text{เมื่อ } z_i = d_i [T_2 + T_1 r_i (y_i - \bar{Y}_R)] \text{ และ } \bar{Y}_R = T_2 / T_3$$

และจากสมการที่ (24) สามารถหาค่า $V_S[\hat{Y}_{IPPS} | R] = V_S[\psi(\hat{T}) | R]$ ตามที่วิธีของ Sampath (2001) ดังนี้

$$\begin{aligned} V_S[\hat{Y}_{IPPS} | R] &= V_S[\psi(\hat{T}) | R] = V_S\left[\frac{1}{T_3} \sum_{i \in S} z_i | R\right] = \frac{1}{T_3^2} V_S\left[\sum_{i \in S} z_i\right] \\ &= \frac{1}{T_3^2} n \left[\sum_{i \in U} z_i^2 P_i - \frac{1}{n} \left(\sum_{i \in U} z_i P_i \right)^2 \right] \end{aligned} \quad (25)$$

แทนค่า $V_S[\hat{Y}_{IPPS} | R]$ ในสมการที่ (22) ได้ว่า

$$\begin{aligned} V(\hat{Y}_{IPPS}) &= \frac{1}{N^2} E_R V_S[\hat{Y}_{IPPS} | R] = \frac{1}{N^2} E_R \left[\frac{1}{T_3^2} n \left(\sum_{i \in U} z_i^2 P_i - \frac{1}{n} \left(\sum_{i \in U} z_i P_i \right)^2 \right) \right] \\ &= \frac{n}{N^2} \sum_{i \in U} d_i^2 P_i^2 y_{1i}^* - \frac{1}{N^2} \left(\sum_{i \in U} d_i y_{2i}^* \right)^2 \end{aligned} \quad (26)$$

$$\text{เมื่อ } y_{1i}^* = 3\bar{Y}^2 - 2y_i \bar{Y} + p^{-1}(y_i - \bar{Y})^2 \text{ และ } y_{2i}^* = \bar{Y} + y_i P_i - \bar{Y} P_i$$

และสามารถประมาณค่า $V(\hat{Y}_{IPPS})$ ด้วย $\hat{V}(\hat{Y}_{IPPS})$ ซึ่งกำหนดค่าดังนี้

$$\hat{V}(\hat{Y}_{IPPS}) = \frac{1}{N^2} \frac{1}{\hat{T}_3^2} \frac{n}{n-1} \left[\sum_{i \in S} \hat{z}_i^2 - \frac{1}{n} \left(\sum_{i \in S} \hat{z}_i \right)^2 \right] \quad (27)$$

$$\text{เมื่อ } \hat{z}_i = \hat{T}_2 + \hat{T}_1 r_i (y_i - \hat{Y}_R) \text{ และ } \hat{Y}_R = \hat{T}_2 / \hat{T}_3$$

3. ผลการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากร

สำหรับหัวข้อนี้ผู้เขียนจะได้กล่าวถึงผลการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากร ที่ได้จากวิธีของ Horvitz and Thompson (1952, 663-685) เขียนแทนด้วย $\hat{V}(\hat{Y}_{HT})$ และวิธีของ Sampath (2001) เขียนแทนด้วย $\hat{V}(\hat{Y}_{IPPS})$ โดยพิจารณาจากค่าความคลาดเคลื่อน และค่า

รากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ตามที่ Berger (2003) เสนอผลการเปรียบเทียบ แสดงดังตารางที่ 1

ตารางที่ 1 ผลการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวน

n	p	Estimators	RB($\hat{\sigma}^2$)	RMSE($\hat{\sigma}^2$)	n	p	Estimators	RB($\hat{\sigma}^2$)	RMSE($\hat{\sigma}^2$)
30	0.3	$\hat{V}(\hat{Y}_{IPPS})$	62,723,584	89,450,479	45	0.3	$\hat{V}(\hat{Y}_{IPPS})$	96,016,955	151,459,492
		$\hat{V}(\hat{Y}_{HT})$	10,607,373,867	127,209,288,706			$\hat{V}(\hat{Y}_{HT})$	12,719,500,751	126,118,558,931
	0.5	$\hat{V}(\hat{Y}_{IPPS})$	65,905,583	87,651,458		0.5	$\hat{V}(\hat{Y}_{IPPS})$	90,670,873	107,265,198
		$\hat{V}(\hat{Y}_{HT})$	10,940,447,254	128,322,779,253			$\hat{V}(\hat{Y}_{HT})$	11,931,462,700	113,757,557,150
	0.7	$\hat{V}(\hat{Y}_{IPPS})$	57,744,631	71,141,806		0.7	$\hat{V}(\hat{Y}_{IPPS})$	81,293,920	94,308,617
		$\hat{V}(\hat{Y}_{HT})$	11,763,116,239	129,622,608,322			$\hat{V}(\hat{Y}_{HT})$	12,608,505,040	117,997,206,247
	0.9	$\hat{V}(\hat{Y}_{IPPS})$	36,028,104	40,356,297		0.9	$\hat{V}(\hat{Y}_{IPPS})$	53,917,994	60,767,702
		$\hat{V}(\hat{Y}_{HT})$	10,393,967,347	120,190,730,802			$\hat{V}(\hat{Y}_{HT})$	15,653,852,241	141,070,741,278

ผลการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากร ตามที่แสดงในตารางที่ 1 พบว่า ในแต่ละระดับของขนาดตัวอย่างและค่าความน่าจะเป็นที่จะตอบสนอง ค่าความคลาดเคลื่อนและค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากร ที่ได้จากวิธีของ Sampath (2001) จะมีค่าต่ำกว่าวิธีของ Horvitz and Thompson (1952, 663-685)

สรุปและวิจารณ์ผล

การวิจัยในครั้งนี้ศึกษาเกี่ยวกับการประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากร ในกรณีที่ข้อมูลไม่ตอบสนองที่เกิดจากการไม่ตอบเฉพาะบางคำถาม เมื่อสัดส่วนตัวอย่างมีขนาดเล็ก และศึกษาภายใต้เฟรมเวิร์คแบบรีเวิร์ส สำหรับตัวประมาณค่าเฉลี่ยประชากร ผู้วิจัยใช้วิธีการแทนที่ค่าข้อมูลสูญหายด้วยค่าเฉลี่ย ซึ่งนำเสนอโดย Haziza (2010) และประยุกต์ใช้ในการเลือกตัวอย่างสุ่มแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาด เนื่องจากตัวประมาณค่าเฉลี่ยประชากรที่ Haziza (2010) นำเสนอเป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้น ดังนั้นการหาค่าความ

แปรปรวนจำเป็นต้องใช้เทอร์แบบเชิงเส้น เพื่อปรับตัวประมาณค่าที่ไม่เป็นเชิงเส้น ให้เป็นเชิงเส้น จากนั้นหาค่าความแปรปรวนโดยใช้วิธีของ Horvitz and Thompson (1952, 663-685) และวิธีของ Sampath (2001)

ผู้วิจัยเปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนที่นำเสนอโดยพิจารณาจากค่าความคลาดเคลื่อน และค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ตามที่ Berger (2003) แนะนำ โดยใช้ข้อมูลที่ได้จากการศึกษาของ Lohr (1999) ชื่อไฟล์ *agstrat.dat* ซึ่งประกอบด้วยประชากรขนาด $N = 299$ และตัวแปรที่สนใจศึกษาได้แก่ $y = ACRES92$ ส่วนตัวแปรกำหนดขนาดได้แก่ $k = FARMS92$ กำหนดให้ความน่าจะเป็นที่จะตอบสนองมีค่าคงที่เท่ากับ p ทั้งหมด 4 ระดับประกอบด้วย $p = 0.3, 0.5, 0.7$ และ 0.9 จากนั้นกำหนดข้อมูลสุ่มในแต่ละระดับ p โดยพิจารณาจากตัวแปรสุ่ม A_i ซึ่ง $A_i \sim U(0,1)$ เมื่อ $i = 1, 2, \dots, N$ ถ้า $A_i \leq p$ ค่า $r_i = 1$ และ $r_i = 0$ ในกรณีอื่นๆ จากนั้นสุ่มตัวอย่างเลือกแบบกำหนดความน่าจะเป็นให้เป็นสัดส่วนกับขนาดและใส่คืน $n = 30, 45$ จำนวน 100,000 ครั้ง ($M = 100,000$) ในแต่ละระดับของความน่าจะเป็นที่จะตอบสนอง และขนาดตัวอย่าง ผลการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากร พบว่าในแต่ละระดับของขนาดตัวอย่างและค่าความน่าจะเป็นที่จะตอบสนอง ค่าความคลาดเคลื่อนและค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ของตัวประมาณค่าความแปรปรวนของตัวประมาณค่าเฉลี่ยประชากร ที่ได้จากวิธีของ Sampath (2001) มีค่าต่ำกว่าวิธีของ Horvitz and Thompson (1952, 663-685)

ข้อเสนอแนะ

ผู้สนใจศึกษาสามารถขยายวิธีการที่ผู้เขียนนำเสนอ ไปยังแผนการเลือกตัวอย่างสุ่มแบบอื่น ๆ เช่น การสุ่มแบบชั้นภูมิ การสุ่มแบบเป็นระบบ หรือ การสุ่มแบบกลุ่ม สำหรับตัวประมาณค่าเฉลี่ยประชากรที่ไม่เป็นเชิงเส้นที่ผู้เขียนนำเสนอ ในกรณีที่ต้องการพิจารณาค่าความเอนเอียงของตัวประมาณค่า สามารถทำได้โดยขยายเทอร์แบบเชิงเส้นอันดับที่หนึ่ง เป็นเทอร์แบบเชิงเส้นอันดับที่สอง (Second-order Taylor linearization)

กิตติกรรมประกาศ

การวิจัยในครั้งนี้สำเร็จลงได้ด้วยความรู้และความช่วยเหลืออย่างสูงยิ่งจาก อาจารย์ดร.เจษฎา โพนแก้ว ที่ได้ให้คำปรึกษาแนะนำ และตรวจสอบ แก้ไข ข้อบกพร่อง ทุกขั้นตอนของการวิจัย รวมถึงคำปรึกษาเกี่ยวกับการเขียนบทความภาษาอังกฤษ

เอกสารอ้างอิง

ทัตดา หิรัญพต. (2554) การเปรียบเทียบประสิทธิภาพระหว่างวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอย และวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยแบบสโทแคสติก. ปรินญาณิพนธ์วิทยาศาสตร์มหาบัณฑิต สาขาวิชาสถิติ คณะวิทยาศาสตร์, มหาวิทยาลัยบูรพา.

Armstrong, J.S., Overton, T.S. (1977, August). Estimating nonresponse bias in mail surveys. Journal of Marketing Research. 14, pp 396-402.

- Berger, Yves G. (2003) *A simple variance estimator for unequal probability sampling without replacement*. Southampton, GB, University of Southampton.
- Deville, J.-C., and Särndal, C.-E. (1944, December). **Variance Estimation for the Regression Imputed HorvitzThompson Estimator**. Journal of Official Statistics. 10(4), pp381-394.
- Hansen, M.H. and Hurwitz, W.N. (1946, December). **The problem of non-response in sample surveys**. J. Am. Stat. 41(236), pp 517–529.
- 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, December). **A generalization of sampling without replacement from a finite universe**. Journal of the American Statistical Association. 47(260), pp 663-685.
- Lohr, S. L. (1999). Sampling: Design and Analysis. Pacific Grove, CA: Duxbury Press.
- Nangsue, Nuanpan. (2014) Adjusting for nonresponse in the analysis and estimation of sample survey data for cluster designs. Ph.D. Thesis of University of Southampton, Social Sciences.
- Rubins,D. (1976). **Inference and missing data**. Biometrika.
- Sampath, S. (2001). **Sampling theory and methods**. Norosa publishing, India.
- Shao, J. and Steel, P. (1999). **Variance estimation for survey data with composite imputation and nonnegligible sampling fractions**. Journal of the American Statistical Association. 9, pp254–65.