

ตัวประมาณค่าอัตราส่วนใหม่ สำหรับประมาณค่าเฉลี่ยประชากร ภายใต้การสุ่มแบบชั้นภูมิและมีข้อมูลสูญหาย

New ratio estimators to estimated population mean under stratified sampling with missing data

ชูเกียรติ โพนแก้ว^{1*}

Chugiat Ponkaew^{1*}

บทคัดย่อ

ปกติตัวประมาณค่าอัตราส่วน นิยมนำมาใช้ประมาณค่าเฉลี่ยประชากรของตัวแปรที่สนใจศึกษา เมื่อทราบข้อมูลของตัวแปรช่วยที่มีความสัมพันธ์ทางบวกกับตัวแปรที่สนใจศึกษา ตัวประมาณค่าที่ผู้วิจัยนำเสนอพิจารณาภายใต้เงื่อนไขที่ตัวแปรที่สนใจศึกษาของแต่ละชั้นภูมิมีข้อมูลสูญหายที่เกิดจากการไม่ตอบของข้อมูลบางหน่วย และกลไกการไม่ตอบสนองแบบสุ่มสมบูรณ์ ผู้วิจัยใช้ค่ารากของค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนสัมพัทธ์ เพื่อเปรียบเทียบประสิทธิภาพของตัวประมาณที่นำเสนอ ผลการเปรียบเทียบพบว่าตัวประมาณค่าอัตราส่วนแบบใหม่ที่น่าสนใจมีประสิทธิภาพดีกว่าตัวประมาณค่าอัตราส่วนแบบตัดข้อมูลเนื่องจากตัวประมาณค่าอัตราส่วนแบบใหม่ถ่วงน้ำหนักข้อมูลที่สูญหายด้วยค่าความน่าจะเป็นที่จะตอบสนอง ทำให้ตัวประมาณดังกล่าวเป็นตัวประมาณค่าที่เกือบไม่เอนเอียงโดยประมาณ

คำสำคัญ: ข้อมูลสูญหาย ค่าเฉลี่ยประชากร สัดส่วนตัวอย่างมีขนาดเล็ก ตัวประมาณค่าอัตราส่วน การสุ่มแบบชั้นภูมิ กลไกการไม่ตอบสนอง

Abstract

The ratio estimator is commonly used to estimate the population mean of study variables. When knowing the information of the auxiliary variable which had a positive correlation with the study variable. The estimators proposed by the researcher were considered under the condition that the study variable in each stratum is non-response by unit non-response and missing completely at a random mechanism. The researcher used relative root mean square error to compare the efficiency of the proposed estimators. The results showed that the new ratio estimator is more efficient than the list wise deletion ratio estimator because the new ratio estimator is weighted by the probability of response then this estimator is an almost unbiased estimator.

Keywords: missing data, population mean, negligible sampling fraction, ratio estimator, stratified sampling, non-response mechanism

¹ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเพชรบูรณ์

¹ Faculty of Science and Technology, Petchabun Rajabhat University

* Corresponding author. E-mail: chugiat.pon@pcru.ac.th

Received: December 1, 2022; Revised: February 15, 2023; Accepted: March 20, 2023

บทนำ

ตัวประมาณค่าอัตราส่วน (ratio estimator) เป็นหนึ่งในตัวประมาณค่าที่นิยมใช้ประมาณค่าเฉลี่ยประชากรของตัวแปรที่สนใจศึกษา (study variable) เมื่อทราบข้อมูลของตัวแปรช่วย (auxiliary variable) ที่มีความสัมพันธ์ทางบวกกับตัวแปรที่สนใจศึกษา ตัวประมาณค่าอัตราส่วนเขียนอยู่ในรูปฟังก์ชันของอัตราส่วนระหว่างตัวประมาณค่าเฉลี่ยประชากรของตัวแปรที่สนใจศึกษากับตัวแปรช่วยคูณกับค่าเฉลี่ยประชากรของตัวแปรช่วย (Cochran, 1977) อย่างไรก็ตาม กรณีที่ตัวแปรที่สนใจศึกษาเกิดการสูญหายของข้อมูล ตัวประมาณค่าอัตราส่วนแบบปกติ ไม่สามารถประมาณค่าเฉลี่ยประชากรได้ (Ponkaew, & Lawson, 2018) ดังนั้นการวิจัยในครั้งนี้มุ่งเน้นศึกษาเกี่ยวกับประมาณค่าเฉลี่ยประชากรด้วยตัวประมาณค่าอัตราส่วนกรณีที่ตัวแปรที่สนใจศึกษาเกิดการสูญหายของข้อมูล

การทำวิจัยเชิงสำรวจเพื่อศึกษาข้อมูลของประชากร (population) กรณีที่ประชากรมีขนาดใหญ่จะส่งผลกระทบโดยตรงต่อผู้ศึกษาเนื่องจากต้องเก็บข้อมูลให้ครบทุกหน่วยของประชากร และนำค่าที่ได้มาคำนวณตามวิธีการทางสถิติ แต่ด้วยข้อจำกัดด้านเวลาและค่าใช้จ่าย (Dansawad, 2020) วิธีการดังกล่าวอาจมีความยุ่งยาก หรือไม่สามารถทำได้ในทางปฏิบัติเนื่องจากผู้ศึกษาไม่สามารถเก็บข้อมูลบางหน่วยของประชากรได้ ดังนั้นจำเป็นต้องประมาณค่าที่ต้องการศึกษาของประชากรจากข้อมูลเพียงบางส่วนที่เรียกว่าตัวอย่าง (sample) กำหนดให้ y แทนตัวแปรที่สนใจศึกษา $U = \{1, 2, \dots, N\}$ แทนเซตที่แสดงลำดับที่ของหน่วยประชากร ขนาด N และ y_i แทนค่าสังเกตของประชากรหน่วยที่ i เมื่อ $i \in U$ ค่าเฉลี่ยประชากร

ของ y คือ $\bar{y} = N^{-1} \sum_{i \in U} y_i$ ดังนั้นกรณีที่ N มีขนาดใหญ่ อาจไม่สามารถสังเกตค่า y_i ได้ครบทุกค่า ส่งผลให้ไม่สามารถคำนวณค่า $\bar{y}$ ได้ ทำให้ต้องประมาณค่าดังกล่าวจากข้อมูลของตัวอย่าง กำหนดให้ $s = \{1, 2, \dots, n\}$ แทนเซตที่แสดงลำดับที่ของหน่วยประชากรที่ถูกสุ่มมาเป็นตัวอย่างขนาด n ภายใต้การสุ่มแบบง่ายและไม่ใส่คืน (simple random sampling without replacement : SRSWOR) ค่าของ $\bar{y}$ สามารถประมาณได้จาก $\hat{\bar{y}} = n^{-1} \sum_{i \in s} y_i$ และ Horvitz, & Thompson (1952) ได้นำเสนอตัวประมาณค่าของ $\bar{y}$ ภายใต้การสุ่มแบบใช้ความน่าจะเป็นในการเลือกไม่เท่ากันและไม่ใส่คืน (unequal probability sampling without replacement : UPWOR) มีค่าเท่ากับ $\hat{\bar{y}}_{HT} = N^{-1} \sum_{i \in s} y_i \pi_i^{-1}$ เมื่อ $\pi_i = p(i \in s)$ และค่า π_i ขึ้นอยู่กับ ตัวแปรกำหนดขนาด (size variable) หรือตัวแปร k

กำหนดให้ x แทนตัวแปรช่วย (auxiliary variable) ที่มีความสัมพันธ์ทางบวกกับ y และทราบค่า $\bar{x} = N^{-1} \sum_{i \in U} x_i$ ค่า $\hat{\bar{y}} = n^{-1} \sum_{i \in s} y_i$ และ $\hat{\bar{x}} = n^{-1} \sum_{i \in s} x_i$ Cochran (1977) ได้นำเสนอตัวประมาณค่าอัตราส่วนสำหรับประมาณ $\bar{y}$ ดัง สมการ (1)

$$\hat{\bar{y}}_R = \left(\frac{\hat{\bar{y}}}{\hat{\bar{x}}} \right) \bar{x} \quad (1)$$

ลำดับถัดมา Sisodia, & Dwivedi (1981); Singh, & Kakran (1993); Upadhyaya, & Singh (1999) นำเสนอตัวประมาณค่าอัตราส่วนเมื่อทราบค่า

สัมประสิทธิ์ความโด่ง (coefficient of kurtosis) และสัมประสิทธิ์การแปรผัน (coefficient of variance) ภายใต้การสุ่มเดียวกับ Cochran (1977) ภายใต้การสุ่มแบบ UPWOR และกำหนดให้ $\hat{X}_{HT} = N^{-1} \sum_{i \in s} x_i \pi_i^{-1}$ แทนตัวประมาณค่าของ $\bar{X}$ นักวิจัยสองท่าน ได้แก่ Bacanlı, & Kadilar (2008) ได้นำเสนอตัวประมาณค่าอัตราส่วนดัง สมการ (2)

$$\hat{Y}_{RHT} = \frac{\hat{Y}_{HT}}{\hat{X}_{HT}} \bar{X} = \frac{N^{-1} \sum_{i \in s} y_i \pi_i^{-1}}{N^{-1} \sum_{i \in s} x_i \pi_i^{-1}} \bar{X} \quad (2)$$

ภายใต้การสุ่มแบบชั้นภูมิและแต่ละชั้นภูมิสุ่มแบบ SRSWOR กำหนดให้ $W_h = \frac{N_h}{N}$, $\hat{Y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_{hi}$, $\hat{X}_h = \frac{1}{n_h} \sum_{i \in s_h} x_{hi}$ และ $\bar{X}_h = \frac{1}{N_h} \sum_{i \in U_h} x_{hi}$ Taylor, & Lone (2014) นำเสนอตัวประมาณค่าอัตราส่วนของตัวแปรที่สนใจศึกษา ดัง สมการ (3)

$$\hat{Y}_{RS} = \sum_{h \in H} W_h \left(\frac{\hat{Y}_h}{\hat{X}_h} \right) \bar{X}_h = \sum_{h \in H} \frac{N_h}{N} \left(\frac{\hat{Y}_h}{\hat{X}_h} \right) \bar{X}_h \quad (3)$$

นอกจากนี้ Taylor, & Lone (2014) ยังได้พัฒนาตัวประมาณค่าของ Sisodia, & Dwivedi (1981); Singh, & Kakran (1993); Upadhyaya, & Singh (1999) ไปยังแผนการสุ่มตัวอย่างเดียวกับตัวประมาณค่าที่แสดงในสมการ (3) ทั้งนี้ตัวประมาณค่าอัตราส่วน ตามที่แสดงในสมการ (1) ถึงสมการ (3)

อยู่ในรูปตัวประมาณค่าแบบไม่เป็นเชิงเส้น (nonlinear estimator) ดังนั้นการหาค่าคาดหวัง และค่าความแปรปรวน ต้องใช้เทอร์เลอร์แบบเชิงเส้น (Taylor linearization) ปรับตัวประมาณค่าอัตราส่วนให้อยู่ในรูปตัวประมาณค่าแบบเชิงเส้น และประมาณค่าคาดหวังและความแปรปรวนจากตัวประมาณค่าแบบเชิงเส้นดังกล่าว กรณีที่ใช้เทอร์เลอร์แบบเชิงเส้นอันดับที่หนึ่งสามารถแสดงได้ว่าตัวประมาณค่าอัตราส่วนเป็นตัวประมาณค่าเกือบจะไม่เอนเอียง (almost unbiased estimator) ของ $\bar{Y}$ ดังนั้นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณค่าอัตราส่วนมีค่าเท่ากับค่าความแปรปรวน โดยประมาณ

ตัวประมาณค่าอัตราส่วนอยู่ในรูปฟังก์ชันของตัวประมาณค่าเฉลี่ยประชากรของตัวแปรที่สนใจศึกษา ดังนั้นกรณีที่ตัวแปรที่สนใจศึกษา เกิดการสูญหายของข้อมูล กล่าวคือบางค่า y_i เมื่อ $i \in s$ ไม่สามารถสังเกตค่าได้ ส่งผลให้ไม่สามารถคำนวณ $\hat{Y}$ ค่า $\hat{Y}_{HT}$ หรือ $\hat{Y}_h$ ทำให้ไม่สามารถใช้ตัวประมาณค่าอัตราส่วนประมาณค่า $\bar{Y}$ ได้ อย่างไรก็ตามวิธีการประมาณค่าเฉลี่ยประชากร กรณีที่มีข้อมูลสูญหายสามารถใช้วิธีตัดข้อมูลในการประมาณค่า โดยวิธีดังกล่าวใช้ข้อมูล y_i เฉพาะ $i \in s_r$ เมื่อ s_r แทนเซตของตัวอย่างที่สามารถสังเกตค่าได้ วิธีการตัดข้อมูลเป็นวิธีที่ง่ายและสะดวกสำหรับใช้แก้ปัญหาข้อมูลสูญหาย แต่อาจส่งผลต่อประสิทธิภาพของตัวประมาณค่าในกรณีที่ข้อมูลสูญหายมีขนาดใหญ่ เนื่องจากขนาดของตัวอย่างที่ใช้ในการสร้างตัวประมาณค่ามีขนาดลดลง

ปกติการสูญหายของข้อมูลแบ่งออกเป็น การไม่ตอบของข้อมูลบางหน่วย (unit nonresponse) หรือการไม่ตอบเฉพาะบางคำถาม (item nonresponse)

กรณีที่เป็นการศึกษาประเภทแรก Samdal, & Lundstorm (2005) ได้กล่าวถึงตัวประมาณค่าผลรวม

ประชากร $\hat{Y}_{SL} = \sum_{i \in s} \frac{r_i Y_i}{\pi_i p_i}$ และตัวประมาณค่าเฉลี่ย

ประชากร คือ $\hat{\bar{Y}}_{SL} = N^{-1} \sum_{i \in s} \frac{r_i Y_i}{\pi_i p_i}$ โดยที่ค่า $r_i = 1$

เมื่อ y_i สามารถสังเกตค่าได้ และ $r_i = 0$ ในกรณีอื่น ๆ ส่วนค่า p_i แทนค่าความน่าจะเป็นที่จะตอบสนอง ซึ่งมีค่าขึ้นอยู่กับกลไกการไม่ตอบสนอง (nonresponse mechanism) ของข้อมูล ต่อมา

Ponkaew, & Lawson (2018) นำเสนอตัวประมาณค่าอัตราส่วน ภายใต้กลไกการไม่ตอบสนองแบบสุ่มสมบูรณ์ (missing complete at random: MCAR) ตัวประมาณค่าอัตราส่วนที่ Ponkaew, & Lawson

(2018) นำเสนอคือ $\hat{Y}_R^{(1)} = \left(\sum_{i \in s} \frac{r_i Y_i}{\pi_i p_i} / \sum_{i \in s} \frac{x_i}{\pi_i} \right) X$

เมื่อ $X = \sum_{i \in U} x_i$ นอกจากนี้ Ponkaew, & Lawson

(2018) ได้พิจารณาตัวประมาณค่าความแปรปรวนของ $\hat{Y}_R^{(1)}$ ภายใต้เฟรมเวิร์กแบบย้อนกลับ (reverse framework) และสัดส่วนตัวอย่างมีขนาดเล็ก (negligible sampling fraction)

ตามที่ผู้วิจัยได้กล่าวถึงตัวประมาณค่าอัตราส่วนข้างต้น กรณีที่เป็นการศึกษาแบบสุ่มแบบชั้นภูมิ ยังคงพิจารณาการสุ่มในแต่ละชั้นภูมิแบบ SRSWOR และพิจารณาเฉพาะกรณีที่ ตัวแปร y ไม่มีข้อมูลสูญหาย ดังนั้นการวิจัยในครั้งนี้ผู้วิจัยมุ่งศึกษาเกี่ยวกับการประมาณค่า $\bar{Y}$ ด้วยตัวประมาณค่าอัตราส่วนภายใต้การสุ่มแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มแบบ UPWOR นอกจากนี้แต่ละชั้นภูมิมีข้อมูลสูญหาย รวมถึงพิจารณา ค่าคาดหวังและตัวประมาณค่าความแปรปรวน

ของตัวประมาณค่าที่นำเสนอภายใต้เงื่อนไขเดียวกับ Ponkaew, & Lawson (2018)

วิธีการศึกษา

การวิจัยในครั้งนี้มุ่งศึกษาเกี่ยวกับการประมาณค่าเฉลี่ยประชากรของตัวแปร y ด้วยตัวประมาณอัตราส่วน ภายใต้การสุ่มแบบชั้นภูมิ และแต่ละชั้นภูมิสุ่มแบบ UPWOR และมีข้อมูลสูญหาย โดยมีขั้นตอนดังนี้

1. สัญลักษณ์และข้อตกลงเบื้องต้น

กำหนดให้ประชากร U ขนาด N ถูกแบ่งออกเป็น L ชั้นภูมิ และ $H = \{1, 2, \dots, L\}$ ผู้วิจัยกำหนดสัญลักษณ์แต่ละชั้นภูมิดังนี้

- U_h แทนประชากรขนาด N_h ของชั้นภูมิที่ h เมื่อ $h \in H$ โดยที่ $\bigcup_{h \in H} U_h = U$ และ $U_i \cap U_j = \phi$ เมื่อ $i \neq j \in H$

- y_{hi} แทน ค่าสังเกตของหน่วยตัวอย่างที่ i ในชั้นภูมิที่ h ของ y ค่าเฉลี่ยประชากร เขียนแทนด้วย $\bar{Y} = N^{-1} \sum_{h \in H} N_h \bar{Y}_h = N^{-1} \sum_{h \in H} W_h Y_h$

- แผนการสุ่มตัวอย่าง เป็นแบบชั้นภูมิและแต่ละชั้นภูมิสุ่มตัวอย่าง s_h ขนาด n_h ด้วย UPWOR

- $\pi_{hi} = P(i \in s_h)$ แทนความน่าจะเป็นที่หน่วยตัวอย่างที่ i ตกอยู่ในตัวอย่าง s_h และ $\pi_{hij} = P(i \wedge j \in s_h)$ แทนความน่าจะเป็นที่หน่วยตัวอย่างที่ i และ j ตกอยู่ในตัวอย่าง s_h

- การสูญหายของข้อมูลเกิดขึ้นเฉพาะตัวแปร y และเป็นแบบ MCAR กล่าวคือการสูญหายของข้อมูลไม่ขึ้นอยู่กับตัวแปร y และตัวแปร x ทำให้ค่า

ความน่าจะเป็นที่จะตอบสนองของแต่ละหน่วยตัวอย่างมีค่าเท่ากันทุกหน่วย

- เซต s_{hr} แทนเซตตัวอย่างที่สามารถสังเกตค่า y_{hi} ได้ แต่ s_{hm} แทนเซตที่ไม่สามารถสังเกตค่า y_{hi} เมื่อ $s_{hr} \cup s_{hm} = s_h$ และ $s_{hr} \cap s_{hm} = \phi$

- r_{hi} แทน ตัวแปรบอกรสภาพ (nonresponse indicator) ของ y_{hi} โดยที่ $r_{hi} = 1$ เมื่อ y_{hi} สามารถสังเกตค่าได้ แต่ $r_{hi} = 0$ เมื่อ y_{hi} สูญหาย กำหนดค่า $R_h = (r_{hi}, r_{hi}, \dots, r_{hN_h})'$ และ $R = (R_1, R_2, \dots, R_L)'$

- $p_{hi} = p(r_{hi} = 1)$ แทน ค่าความน่าจะเป็นที่จะตอบสนองของหน่วยตัวอย่างที่ i ในชั้นภูมิที่ h ภายใต้ MCAR ค่า $p_{hi} = p_h$ นอกจากนี้ค่า $p_{hij} = p(r_{hi} = 1 \wedge r_{hj} = 1) = p_h^2$

- $E_R(\cdot)$ และ $V_R(\cdot)$ แทนค่าคาดหวัง และความแปรปรวน เทียบกับกลไกการตอบสนอง ส่วน $E_S(\cdot)$ และ $V_S(\cdot)$ แทนค่าคาดหวัง และความแปรปรวน เทียบกับแผนการสุ่มตัวอย่าง

- กำหนดเงื่อนไข A: แต่ละชั้นภูมิเกิดการสูญหายเฉพาะตัวแปร y และกลไกการไม่ตอบสนองเป็นแบบ MCAR รวมถึงสัดส่วนตัวอย่างของแต่ละชั้นภูมิมีขนาดเล็ก

2. การหาค่าความแปรปรวนและตัวประมาณค่า

การหาค่าความแปรปรวนและตัวประมาณค่าความแปรปรวน ภายใต้แผนการสุ่มแบบ UPWOR โดยพิจารณาแยกเป็น กรณีที่ตัวแปรที่สนใจศึกษาไม่มีข้อมูลสูญหาย และกรณีเกิดการสูญหายของข้อมูล ดังนี้

2.1 ตัวแปรที่สนใจศึกษาไม่มีข้อมูลสูญหาย

กำหนดให้ตัวแปร z_{1i} แทนฟังก์ชันของ y_i หรือ $z_{1i} = f(y_i)$ และไม่มีการสูญหายของข้อมูล กำหนด $\hat{Y}_{HT}' = N^{-1} \sum_{i \in s} \frac{z_{1i}}{\pi_i}$ แทนตัวประมาณค่าเฉลี่ยประชากร กำหนดให้ $\hat{z}_{1i}$ แทนตัวประมาณค่าของ z_{1i} ค่า $D_i = (1 - \pi_i) \pi_i^{-1}$ และ $D_{ij} = (\pi_{ij} - \pi_i \pi_j) (\pi_i \pi_j)^{-1}$ เนื่องจากค่า $\hat{Y}_{HT}'$ อยู่ในรูปแบบตัวประมาณค่าของ Horvitz, & Thompson (1952) ดังนั้นความแปรปรวนของ $\hat{Y}_{HT}'$ มีค่าดัง สมการ (4)

$$v(\hat{Y}_{HT}') = \sum_{i \in U} D_i z_{1i}^2 + \sum_{i \in U} \sum_{j \in U \setminus \{i\}} D_{ij} z_{1i} z_{1j} \quad (4)$$

ตัวประมาณค่าของ $v(\hat{Y}_{HT}')$ มีค่าดังสมการ (5)

$$\hat{v}(\hat{Y}_{HT}') = \sum_{i \in s} \hat{D}_i \hat{z}_{1i}^2 + \sum_{i \in s} \sum_{j \in U \setminus \{i\}} \hat{D}_{ij} \hat{z}_{1i} \hat{z}_{1j} \quad (5)$$

2.2 ตัวแปรที่สนใจศึกษามีข้อมูลสูญหาย

กำหนดให้ตัวแปร z_{2i} แทนฟังก์ชันของ y_i เมื่อมีข้อมูลสูญหาย แบบ MCAR หรือ $z_{2i} = f(y_i)$

กำหนด $\hat{Y}_{HT}'' = N^{-1} \sum_{i \in s} \frac{z_{2i}}{\pi_i}$ แทนตัวประมาณค่าเฉลี่ยประชากร ภายใต้เฟรมเวิร์คแบบย้อนกลับค่าความแปรปรวนของ $\hat{Y}_{HT}''$ มีค่าดังสมการ (6)

$$v(\hat{Y}_{HT}'') = E_R V_S(\hat{Y}_{HT}'' | R) + V_R E_S(\hat{Y}_{HT}'' | R) \quad (6)$$

กรณีที่สัดส่วนตัวอย่างมีขนาดเล็ก ค่า $V(\hat{\bar{Y}}_{HT}'')$ ตัวประมาณค่าของ $V(\hat{\bar{Y}}_{HT}'')$ มีค่านอกจากนี้ตัว
สามารถประมาณได้ดังสมการ (7) นอกจากนี้ ประเมินค่าของ $V(\hat{\bar{Y}}_{HT}'')$ มีค่าดังสมการ (8)

$$V(\hat{\bar{Y}}_{HT}'') \cong E_R V_S(\hat{\bar{Y}}_{HT}'' | R) = \sum_{i \in U} D_i E_R(z_{1i}^2 | R) + \sum_{i \in U} \sum_{j \in U \setminus \{i\}} D_{ij} E_R(z_{1i} | R) E_R(z_{1j} | R) \quad (7)$$

$$\hat{V}(\hat{\bar{Y}}_{HT}'') \cong \hat{V}_S(\hat{\bar{Y}}_{HT}'' | R) = \sum_{i \in s} r_i p_i^{-1} \hat{D}_i \hat{z}_{2i}^2 + \sum_{i \in s} \sum_{j \in U \setminus \{i\}} r_i r_j (p_i p_j)^{-1} \hat{D}_{ij} \hat{z}_{2i} \hat{z}_{2j} \quad (8)$$

3. การนำเสนอตัวประมาณค่าอัตราส่วน และตัวประมาณค่าความแปรปรวน

ตัวประมาณค่าอัตราส่วนเพื่อประมาณ ค่าเฉลี่ยประชากร จำนวน 2 ตัว ภายใต้กลไกการไม่

ตอบสนองแบบ MCAR ประกอบด้วย ตัวประมาณค่า อัตราส่วนแบบตัดข้อมูล และตัวประมาณค่า อัตราส่วนแบบใหม่ ดัง (Figure 1)

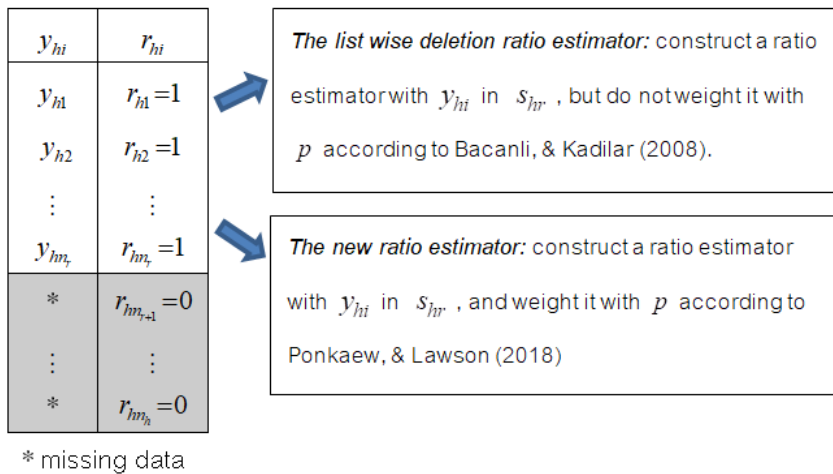


Figure 1 The concept of proposed estimators under MCAR mechanism.

สำหรับรายละเอียดการสร้างตัวประมาณค่า อัตราส่วนแบบตัดข้อมูล และตัวประมาณค่าอัตราส่วน แบบใหม่ จาก (Figure 1) แสดงตามหัวข้อ 3.1 และ หัวข้อ 3.2 ตามลำดับ

3.1 ตัวประมาณค่าอัตราส่วนแบบตัดข้อมูล

ตัวประมาณค่าอัตราส่วนแบบตัดข้อมูล เขียนแทนด้วย $\hat{\bar{Y}}_R'$ หมายถึงตัวประมาณค่าอัตราส่วน ที่ใช้ตัวประมาณค่าของ Bacanli, & Kadilar (2008)

ประมาณค่าเฉลี่ยประชากรของตัวแปรที่สนใจศึกษาของแต่ละชั้นภูมิ แต่พิจารณาเฉพาะค่า y_{hi} ที่สามารถสังเกตค่าได้ กล่าวคือใช้เฉพาะตัวอย่างที่ i เมื่อ $i \in s_{hr}$ จากนั้นใช้ตัวประมาณค่าดังกล่าวปรับตัวประมาณค่าของ Taylor, & Lone (2014) สำหรับการหาตัวประมาณค่าความแปรปรวน เนื่องจาก $\hat{Y}_R'$ เป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้น ดังนั้นจำเป็นต้องใช้เทเลอร์แบบเชิงเส้น ปรับ $\hat{Y}_R'$ ให้เป็นตัวประมาณค่าแบบเชิงเส้น และหาค่า $V(\hat{Y}_R')$ และ $\hat{V}(\hat{Y}_R')$ ตามขั้นตอนที่อธิบายไว้ในหัวข้อ 2.1

3.2 ตัวประมาณค่าอัตราส่วนแบบใหม่

ตัวประมาณค่าอัตราส่วนแบบใหม่เขียนแทนด้วย $\hat{Y}_R''$ หมายถึงตัวประมาณค่าอัตราส่วนใช้ตัวประมาณค่าของ Ponkaew, & Lawson (2018) ประมาณค่าเฉลี่ยประชากรของตัวแปรที่สนใจศึกษาของแต่ละชั้นภูมิ แต่พิจารณาเฉพาะค่า y_{hi} ที่สามารถสังเกตค่าได้ กล่าวคือใช้เฉพาะตัวอย่างที่ i เมื่อ $i \in s_{hr}$ และมีการถ่วงน้ำหนักของแต่ละตัวอย่างที่ i เมื่อ $i \in s_{hr}$ ด้วยค่าความน่าจะเป็นที่จะตอบสนอง p จากนั้นใช้ตัวประมาณค่าดังกล่าวปรับตัวประมาณค่าของ Taylor, & Lone (2014) ส่วนการหาตัวประมาณค่าความแปรปรวนของ $\hat{Y}_R''$ หรือ $\hat{V}(\hat{Y}_R'')$ เนื่องจาก $\hat{Y}_R''$ เป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้น ดังนั้นจำเป็นต้องใช้เทเลอร์แบบเชิงเส้นปรับ $\hat{Y}_R''$ ให้เป็นตัวประมาณค่าแบบเชิงเส้น และหาค่า $V(\hat{Y}_R'')$ และ $\hat{V}(\hat{Y}_R'')$ ตามขั้นตอนที่อธิบายไว้ในหัวข้อ 2.2

4. การเปรียบเทียบประสิทธิภาพของตัวประมาณค่าที่นำเสนอ

การเปรียบเทียบประสิทธิภาพของตัวประมาณค่าอัตราส่วนแบบตัดข้อมูล $\hat{Y}_R'$ และตัวประมาณค่าอัตราส่วนใหม่ $\hat{Y}_R''$ กำหนดให้ $\hat{Y}_{R[m]}^k$ แทนค่าของตัวประมาณค่าอัตราส่วนในรอบที่ m เมื่อ $m=1,2,\dots,M$ และ $M=10,000$ การพิจารณาประสิทธิภาพของตัวประมาณค่าพิจารณา จากค่ารากของค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนสัมพัทธ์ (relative root mean square error : RRMSE) ดังสมการ (9)

$$RRMSE(\hat{Y}_R^k) = \frac{\sqrt{\frac{1}{M} \sum_{m=1}^M (\hat{Y}_{R[m]}^k - \bar{Y})^2}}{\bar{Y}} \quad (9)$$

สำหรับเกณฑ์ที่ใช้พิจารณาประสิทธิภาพของตัวประมาณค่า แสดงตามบทนิยามที่ 1

บทนิยาม 1 กำหนดให้ $\hat{Y}_A$ และ $\hat{Y}_B$ แทนตัวประมาณค่าเฉลี่ยของ $\bar{Y}$ นอกจากนี้กำหนดให้ $RRMSE(\hat{Y}_A)$ และ $RRMSE(\hat{Y}_B)$ แทน RRMSE ของ $\hat{Y}_A$ และ $\hat{Y}_B$ ตามลำดับ ถ้า $RRMSE(\hat{Y}_A) < RRMSE(\hat{Y}_B)$ และ $\hat{Y}_A$ จะมีประสิทธิภาพที่ดีกว่า $\hat{Y}_B$

ข้อมูลที่ใช้คำนวณค่าของตัวประมาณค่าอัตราส่วนผู้วิจัยใช้ข้อมูลข้าวโพดหวานของประเทศไทยในรอบปีการเพาะปลูก ปี 2562 และปี 2563 ที่จัดเก็บโดยสำนักงานเศรษฐกิจการเกษตร หน่วยประชากรที่ใช้คือจังหวัดที่ปลูกข้าวโพดหวาน ในประเทศไทยจำนวน 63 จังหวัด ดังนั้น $N = 63$ ชั้นภูมิที่ใช้แยกตามภาค ทั้งหมด 4 ชั้นภูมิ $H = \{1,2,3,4\}$

ประกอบด้วย ชั้นภูมิที่ 1 ภาคเหนือ ชั้นภูมิที่ 2 ภาคตะวันออกเฉียงเหนือ ชั้นภูมิที่ 3 ภาคกลาง และชั้นภูมิที่ 4 ภาคใต้ ทั้งนี้แต่ละชั้นภูมิ มีจำนวนประชากร คือ $N_1=16, N_2=19, N_3=13$ และ $N_4=14$ ตามลำดับ ตัวแปรที่ใช้ในการวิจัยประกอบด้วย ตัวแปรที่สนใจศึกษา y แทนผลผลิตข้าวโพดหวาน ในปี 2019 (หน่วย: ตัน) ตัวแปรช่วย x แทนเนื้อที่เพาะปลูก ในปี 2019 (หน่วย: ไร่) และตัวแปรกำหนดขนาด k แทนเนื้อที่เพาะปลูก ในปี 2018 (หน่วย: ไร่) สำหรับขั้นตอนการคำนวณเพื่อใช้ในการเปรียบเทียบเชิงตัวเลขมีดังนี้

ขั้นที่ 1 สุ่มตัวอย่างชั้นภูมิละ 40% ด้วยแผนการสุ่มตัวอย่างแบบ UPWOR

ขั้นที่ 2 กำหนดค่าของตัวแปรบอกสภาพ r_{hi} ภายใต้การสุ่มหายแบบ MCAR ในแต่ละชั้นภูมิ

ขั้นที่ 3 คำนวณค่า $\hat{Y}_R^m$ เมื่อ $m=1,2$ โดยที่ $\hat{Y}_R^1 = \hat{Y}_R'$ และ $\hat{Y}_R^2 = \hat{Y}_R''$

ขั้นที่ 4 ทำซ้ำขั้นตอนที่ 1-3 จำนวน 10,000 รอบ

ขั้นที่ 5 คำนวณค่า $RRMSE(\hat{Y}_R^k)$ เมื่อ $m=1,2$

5. การประยุกต์ใช้ตัวประมาณค่ากับข้อมูลจริง

ผู้วิจัยใช้ตัวประมาณค่าอัตราส่วนใหม่ที่น่าสนใจ มาประยุกต์ใช้กับประมาณฝุ่นละอองที่มีขนาดเล็กกว่า 2.5 ไมครอน (ฝุ่น PM 2.5) ของกรุงเทพมหานคร ในระหว่างวันที่ 1-30 กันยายน พ.ศ. 2565 ที่จัดเก็บข้อมูล โดยกองจัดการคุณภาพอากาศและเสียงกรุงเทพมหานคร ตัวแปรที่ใช้ในการศึกษาประกอบด้วย ตัวแปรที่สนใจศึกษา y ได้แก่ ฝุ่น PM 2.5 (หน่วย: ไมโครกรัมต่อลูกบาศก์เมตร) ตัวแปรช่วย x ได้แก่ ฝุ่น PM 10 (หน่วย: ไมโครกรัมต่อลูกบาศก์เมตร) และตัวแปรกำหนดขนาด k ได้แก่อุณหภูมิ (หน่วย: องศาเซลเซียส) พื้นที่ที่ใช้ในการศึกษาแบ่งออกเป็น 3 ชั้นภูมิ ประกอบด้วย ชั้นภูมิที่ 1 เขตดินแดง ชั้นภูมิที่ 2 เขตพระโขนงและชั้นภูมิที่ 3 เขตดอนเมือง เนื่องจากข้อมูลของทั้ง 3 เขตที่ใช้ในการศึกษาสอดคล้องกับเงื่อนไขของผู้วิจัย กล่าวคือ การสูญหายของข้อมูลเกิดขึ้นเฉพาะตัวแปรที่สนใจศึกษา y ส่วนตัวแปรช่วย x และตัวแปรกำหนดขนาด k ไม่เกิดการสูญหายของข้อมูลค่าพารามิเตอร์หรือค่าสถิติที่เกี่ยวข้อง ดัง (Table 1)

Table 1 The statistic parameters of auxiliary and size variables.

h	P_h	N_h	n_h	$\bar{X}_h$	$\sum_{i \in U_h} k_{hi}$
1	0.70	30	10	53.23	844
2	0.70	30	10	27.57	1225
3	0.70	30	10	27.63	1029

จาก (Table 1) แสดงข้อมูลของฝุ่น PM 10 และอุณหภูมิ ที่เป็นตัวแปร x และ k ของกรุงเทพมหานคร แยกตามเขต เขตดินแดง เขตพระโขนงและ เขตดอนเมือง

ซึ่งแต่ละเขตเก็บบันทึกข้อมูลทั้งหมด 30 วัน หรือ $N_1=N_2=N_3=30$ และแต่ละเขตสุ่มตัวอย่างขนาด $n_1=n_2=n_3=10$ เมื่อพิจารณา ฝุ่น PM 10 ของ

แต่ละเขต พบว่า มีค่าเท่ากับ 53.23, 27.57 และ 27.63 ไมโครกรัมต่อลูกบาศก์เมตร ตามลำดับ ส่วนอุณหภูมิรวม พบว่าเขตดินแดง เขตพระโขนง และเขตดอนเมือง มีอุณหภูมิรวมเท่ากับ 844, 1225 และ 1029 องศาเซลเซียส ตามลำดับ

ผลการศึกษาและอภิปรายผล

1. ตัวประมาณค่าอัตราส่วนที่น่าสนใจ

1.1 ตัวประมาณค่าอัตราส่วนแบบตัดข้อมูล

ภายใต้การสุ่มแบบชั้นภูมิและแต่ละชั้นภูมิ สุ่มแบบ SRSWOR และไม่การสูญเสียหายของข้อมูล Tailor, & Lone (2014) นำเสนอตัวประมาณค่าอัตราส่วน $\hat{Y}_{RS} = \sum_{h \in H} w_h \hat{Y}_{RSTh}$ สำหรับประมาณค่า $\bar{Y}$ นอกจากนี้ Bacanli, & Kadilar (2008) ได้นำเสนอตัวประมาณค่าอัตราส่วน $\hat{Y}_{RHT} = \frac{\hat{Y}_{HT}}{\hat{X}_{HT}} \bar{X}$ ภายใต้การสุ่มแบบ UPWOR ดังนั้นผู้วิจัยปรับตัวประมาณค่าของ Bacanli, & Kadilar (2008) เพื่อใช้ประมาณค่าเฉลี่ยประชากรของชั้นภูมิที่ h หรือ $\bar{Y}_h$ เขียนแทนด้วย $\hat{Y}_{RHT}$ และมีค่าดังสมการ (10)

$$\hat{Y}_{RHT} = \frac{\hat{Y}_{HT}}{\hat{X}_{HT}} \bar{X} = \frac{\frac{1}{N_h} \sum_{i \in s_h} \frac{y_{hi}}{\pi_{hi}}}{\frac{1}{N_h} \sum_{i \in s_h} \frac{x_{hi}}{\pi_{hi}}} \bar{X} \quad (10)$$

ขั้นต่อไปแทนที่ค่า $\hat{Y}_{RSTh}$ ใน $\hat{Y}_{RS}$ ด้วย $\hat{Y}_{RHT}$ ทำให้ได้ตัวประมาณค่าอัตราส่วน กรณีที่ไม่มีข้อมูลสูญเสียหาย ภายใต้การสุ่มแบบชั้นภูมิและแต่ละชั้นภูมิ สุ่มแบบ UPWOR มีค่าดังสมการ (11)

$$\hat{Y}_{RSHT} = \sum_{h \in H} w_h \hat{Y}_{RHT} = \sum_{h \in H} w_h \frac{\hat{Y}_{HT}}{\hat{X}_{HT}} \bar{X}_h \quad (11)$$

ลำดับถัดไปเป็นการหาค่าความแปรปรวนและตัวประมาณค่าความแปรปรวนของ $\hat{Y}_{RSHT}$ เบื้องต้นผู้วิจัยกำหนดสัญลักษณ์ที่เกี่ยวข้องดังนี้ กำหนดให้ $e_{hi} = y_{hi} - R_h x_{hi}$ ค่า $R_h = \bar{Y}_h / \bar{X}_h$ ค่า $\hat{e}_{hi} = y_{hi} - \hat{R}_h x_{hi}$ และ $\hat{R}_h = \hat{Y}_{HT} / \hat{X}_{HT}$ ดังนั้น ค่าความแปรปรวนของ $\hat{Y}_{RSHT}$ คือ

$$v(\hat{Y}_{RSHT}) = \frac{1}{N^2} \sum_{h \in H} \left[\sum_{i \in U_h} D_{hi} e_{hi}^2 + \sum_{i \in U_h} \sum_{j \in U_h, j \neq i} D_{hij} e_{hi} e_{hj} \right]$$

ดังนั้นตัวประมาณค่าของ $v(\hat{Y}_{RSHT})$ มีค่าดังสมการ (12)

$$\hat{v}(\hat{Y}_{RSHT}) = \frac{1}{N^2} \sum_{h \in H} \left[\sum_{i \in s_h} \hat{D}_{hi} \hat{e}_{hi}^2 + \sum_{i \in s_h} \sum_{j \in U_h \setminus \{i\} \cap s_u} \hat{D}_{hij} \hat{e}_{hi} \hat{e}_{hj} \right] \quad (12)$$

ค่า $\hat{Y}_{HT}$ จากสมการ (11) อยู่ในรูปฟังก์ชัน y_{hi} ทุกค่า $i \in s_h$ กรณีที่มีข้อมูลสูญเสียหาย ค่าของ $\hat{Y}_{HT}$ เขียนอยู่ในฟังก์ชันของ y_{hi} เฉพาะ $i \in s_{hr}$ เขียนแทนด้วย $\hat{Y}_{HT}^*$ และมีค่าดังสมการ (13)

$$\hat{Y}_{HT}^* = \frac{1}{N_h} \sum_{i \in s_{hr}} \frac{y_{hi}}{\pi_{hi}} \quad (13)$$

ลำดับถัดไปแทนที่ค่า $\hat{Y}_{HTTh}$ ในสมการ (9) ด้วย $\hat{Y}_{HTTh}$ ดังนั้นตัวประมาณค่าอัตราส่วนแบบตัดข้อมูลมีค่าดังสมการ (14)

$$\hat{Y}_R' = \sum_{h \in H} W_h \frac{\hat{Y}_{HTTh}}{\hat{X}_{HTTh}} \bar{X}_h \quad (14)$$

สำหรับตัวประมาณค่าความแปรปรวนของ $\hat{Y}_R'$ ผู้วิจัยนำเสนอตามที่แสดงในทฤษฎีบทที่ 2

ทฤษฎีบทที่ 2 ภายใต้แผนการสุ่มตัวอย่างแบบชั้นภูมิและแต่ละชั้นภูมิสุ่มแบบ UPWOR สมมติว่าแต่ละชั้นภูมิตัวแปรที่สนใจศึกษา เกิดการสูญหายของข้อมูลภายใต้การกลไกการไม่ตอบสนองแบบ MCAR และสัดส่วนตัวอย่างมีขนาดเล็ก กำหนดให้ $\hat{e}_{hi}' = y_{hi} - \hat{R}_h' x_{hi}$ ค่า $\hat{R}_h' = \hat{Y}_{HTTh} / \hat{X}_{HTTh}$ ค่า $\hat{Y}_{HTTh} = \frac{1}{N_h} \sum_{i \in s_{pr}} \frac{y_{hi}}{\pi_{hi}}$ และ $\hat{X}_{HTTh} = \frac{1}{N_h} \sum_{i \in s_{pr}} \frac{x_{hi}}{\pi_{hi}}$ ตัวประมาณค่าความแปรปรวนของ $\hat{Y}_R'$ คือ

$$\hat{V}(\hat{Y}_R') = \frac{1}{N^2} \sum_{h \in H} \left[\sum_{i \in s_{pr}} \hat{D}_{hi} \hat{e}_{hi}'^2 + \sum_{i \in s_{pr}} \sum_{j \in \{i\} \setminus s_{pr}} \hat{D}_{hij} \hat{e}_{hi}' \hat{e}_{hj}' \right]$$

1.2 ตัวประมาณค่าอัตราส่วนแบบใหม่

หัวข้อ 1.1 ผู้วิจัยได้นำเสนอตัวประมาณค่าอัตราส่วนแบบตัดข้อมูล รวมถึงนำเสนอตัวประมาณค่าความแปรปรวนของตัวประมาณค่าดังกล่าว ซึ่งตัวประมาณค่าที่นำเสนอมีฟังก์ชันการคำนวณ

คล้ายกับตัวประมาณค่าที่พิจารณาภายใต้เงื่อนไขที่ไม่มีการสูญหายของข้อมูลแต่ใช้ข้อมูลเฉพาะที่อยู่ในเซต s_{hr} แทนเซต s_h หัวข้อนี้ผู้วิจัยจะได้นำเสนอตัวประมาณค่าอัตราส่วนใหม่ โดยใช้ตัวประมาณค่าของ Ponkaew, & Lawson (2018) ปรับตัวประมาณค่าของ Tailor, & Lone (2014) ดังนั้น ตัวประมาณค่าของ Ponkaew, & Lawson (2018) สำหรับประมาณค่าเฉลี่ยประชากรของแต่ละชั้นภูมิ เขียนแทนด้วย

$$\hat{Y}_{PLh} \text{ และมีค่าเท่ากับ } \hat{Y}_{PLh} = \frac{1}{N_h} \sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi} p_h}$$

$$\hat{R}_h^{(1)} = \frac{\frac{1}{N_h} \sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi} p_h}}{\frac{1}{N_h} \sum_{i \in s_h} \frac{x_{hi}}{\pi_{hi}}} = \frac{\hat{Y}_{PLh}}{\hat{X}_{HTTh}}$$

ตัวประมาณค่าอัตราส่วนสำหรับประมาณค่า $\bar{Y}_h$ ที่พัฒนามาจาก Ponkaew, & Lawson (2018) คือ

$$\hat{Y}_{Rh}^{(1)} = \left(\frac{\frac{1}{N_h} \sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi} p_h}}{\frac{1}{N_h} \sum_{i \in s_h} \frac{x_{hi}}{\pi_{hi}}} \right) \bar{X}_h = \left(\frac{\hat{Y}_{PLh}}{\hat{X}_{HTTh}} \right) \bar{X}_h = \hat{R}_h^{(1)} \bar{X}_h$$

ขั้นตอนต่อไปแทนที่ค่า $\hat{Y}_{RSTh}$ ใน $\hat{Y}_{RS}$ ด้วย $\hat{Y}_{Rh}^{(1)}$ ทำให้ได้ตัวประมาณค่าอัตราส่วนใหม่ สำหรับประมาณค่าเฉลี่ยประชากร ภายใต้เงื่อนไข A และแผนการสุ่มตัวอย่างแบบชั้นภูมิและแต่ละชั้นภูมิสุ่มแบบ UPWOR มีค่าดังสมการ (15)

$$\hat{Y}_R'' = \sum_{h \in H} W_h \hat{R}_h^{(1)} \bar{X}_h = \sum_{h \in H} W_h \left(\frac{\hat{Y}_{PLh}}{\hat{X}_{HTTh}} \right) \bar{X}_h \quad (15)$$

ต่อไปผู้วิจัยจะได้แสดงว่า $\hat{Y}_R''$ เป็นตัวประมาณค่าที่เกือบไม่เอนเอียงของ $\bar{Y}$ ตามที่แสดงในทฤษฎีบทที่ 3

ทฤษฎีบทที่ 3 ภายใต้แผนการสุ่มตัวอย่างแบบชั้นภูมิและแต่ละชั้นภูมิสุ่มแบบ UPWOR สมมติว่าแต่ละชั้นภูมิตัวแปรที่สนใจศึกษา เกิดการสูญหาย

ของข้อมูลภายใต้การกลไกการไม่ตอบสนองแบบ MCAR ตัวประมาณค่า $\hat{Y}_R''$ เป็นตัวประมาณค่าที่เกือบไม่เอนเอียงของ $\bar{Y}$

พิสูจน์ กำหนดให้เงื่อนไข A เป็นจริง ค่าคาดหวังของ $\hat{Y}_R''$ คือ

$$E(\hat{Y}_R'') \cong E_{R'} E_S(\hat{Y}_R'' | R) = \sum_{h \in H} W_h \bar{X}_h E_{R'} E_S(\hat{R}_h^{(1)} | R) \quad (16)$$

ทั้งนี้ค่า $\hat{R}_h^{(1)}$ ในสมการ (16) เป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้น เมื่อใช้เทเลอร์แบบเชิงเส้น

อันดับที่หนึ่ง สามารถเขียน $\hat{R}_h^{(1)}$ ให้อยู่ในรูปแบบเชิงเส้น ดังสมการ (17)

$$\hat{R}_h^{(1)} \cong \tilde{R} + \frac{1}{\bar{X}_h}(\hat{t}_1 - t_1) - \frac{1}{\bar{X}_h} \tilde{R}(\hat{t}_2 - t_2) \quad (17)$$

$$\text{เมื่อ } \tilde{R} = \tilde{Y}_h / \bar{X}_h, \tilde{Y}_h = \frac{1}{N_h} \sum_{i \in U_h} \frac{r_{hi} y_{hi}}{p_h}, \hat{t}_1 = \frac{1}{N_h} \sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi} p_h}, \hat{t}_2 = \frac{1}{N_h} \sum_{i \in s_h} \frac{x_{hi}}{\pi_{hi}} \text{ และ } t_1 = \tilde{Y}_h, t_2 = \bar{X}_h$$

ถัดไปแทนค่าสมการ (17) ในสมการ (16) จะได้ $E(\hat{Y}_R'') \cong \sum_{h \in H} W_h \bar{X}_h \frac{1}{N_h} \frac{\bar{Y}_h}{\bar{X}_h} = \sum_{h \in H} W_h \bar{Y}_h = \bar{Y}$ นั่นคือ $E(\hat{Y}_R'') \cong \bar{Y}$ หรือกล่าวได้ว่า $\hat{Y}_R''$ เป็นตัวประมาณค่าที่เกือบจะไม่เอนเอียงของ $\bar{Y}$

สำหรับค่าความแปรปรวนและตัวประมาณค่าความแปรปรวนของตัวประมาณค่าอัตราส่วนใหม่

$\hat{Y}_R''$ ผู้วิจัยนำเสนอตามที่แสดงในทฤษฎีบทที่ 4

ทฤษฎีบทที่ 4 ภายใต้แผนการสุ่มตัวอย่างแบบชั้นภูมิและแต่ละชั้นภูมิสุ่มแบบ UPWOR สมมติว่าแต่ละชั้นภูมิตัวแปรที่สนใจศึกษา เกิดการสูญหายของข้อมูลภายใต้การกลไกการไม่ตอบสนองแบบ MCAR และสัดส่วนตัวอย่างมีขนาดเล็ก

1. ค่าความแปรปรวนของ $\hat{Y}_R''$ คือ

$$V(\hat{Y}_R'') = \frac{1}{N^2} \sum_{h \in H} \left[\sum_{i \in U_h} D_{hi} (e''_{hi})^2 + \sum_{i \in U_h} \sum_{j \in \{i\} \in U_h} D_{hij} e''_{hi} e''_{hj} \right] \text{ เมื่อ } e''_{hi} = e'_{hi} = y_{hi} - R'_h x_{hi}$$

2. ตัวประมาณค่าของ $\hat{Y}_R''$ คือ

$$\hat{V}(\hat{Y}_R'') = \frac{1}{N^2} \sum_{h \in H} \left[\sum_{i \in s_h} r_{hi} p_h^{-1} \hat{D}_{hi} (\hat{e}_{hi}'')^2 + \sum_{i \in s_h} \sum_{j \in s_h, j \neq i} r_{hi} r_{hj} (p_h^{-2}) \hat{D}_{hj} \hat{e}_{hi}'' \hat{e}_{hj}'' \right] \text{ เมื่อ } \hat{e}_{hi}'' = y_{hi} - \hat{R}_h^{(1)} x_{hi}$$

พิสูจน์ กำหนดให้เงื่อนไข A เป็นจริงภายใต้ และใช้วิธีตาม สมการ (6) ถึงสมการ (8) ค่าความแปรปรวนเฟรมเวิร์คแบบย้อนกลับ และสัดส่วนตัวอย่างมีขนาดเล็ก ของ $\hat{Y}_R''$ คือ

$$V(\hat{Y}_R'') \cong E_R V_S \left(\hat{Y}_R'' \mid R \right) = E_R V_S \left(\sum_{h \in H} W_h \hat{R}_h^{(1)} \bar{X}_h \mid R \right) = E_R \left(\sum_{h \in H} W_h^2 \bar{X}_h^2 V_S(\hat{R}_h^{(1)} \mid R) \right) \quad (18)$$

ดังนั้น

$$V(\hat{Y}_R'') \cong E_R \left(\sum_{h \in H} W_h^2 \bar{X}_h^2 V_S(\hat{R}_h^{(1)} \mid R) \right)$$

ตามที่ได้กล่าวในทฤษฎีบทที่ 3 ตัวประมาณค่า $\hat{R}_h^{(1)}$ เป็นตัวประมาณค่าแบบไม่เป็นเชิงเส้น เมื่อใช้เทเลอร์แบบเชิงเส้นอันดับหนึ่งปรับค่า $\hat{R}_h^{(1)}$ มีค่า

ดังสมการ (17) และจากสมการดังกล่าว สามารถเขียน $\hat{R}_h^{(1)}$ ได้ดังนี้

$$\hat{R}_h^{(1)} \cong \tilde{R} + \frac{1}{N_h \bar{X}_h} \sum_{i \in s_h} \frac{\tilde{e}_{hi}''}{\pi_{hi}} \text{ เมื่อ } \tilde{e}_{hi}'' = \frac{r_{hi} y_{hi}}{\pi_{hi} p_h} - \tilde{R} x_{hi} \quad (19)$$

ลำดับถัดไปแทนค่าสมการ (19) ในสมการ (18) จะได้ค่า $V(\hat{Y}_R'')$ ตามที่แสดงในทฤษฎีบทที่ 3 ข้อ 1 สำหรับการพิสูจน์ค่า $\hat{V}(\hat{Y}_R'')$ ใช้แนวคิดเดียวกับ $\hat{V}(\hat{Y}_{HT})$ ดังสมการ (5)

ทฤษฎีบทที่ 5 ภายใต้แผนการสุ่มตัวอย่างแบบชั้นภูมิและแต่ละชั้นภูมิสุ่มแบบ UPWOR สมมติว่าแต่ละชั้นภูมิตัวแปรที่สนใจศึกษา เกิดการสูญหายของข้อมูลภายใต้การกลไกการไม่ตอบสนอง

แบบ MCAR กำหนดให้ $\hat{Y}_{HTh} = \frac{1}{N_h} \sum_{i \in s_{hr}} \frac{y_{hi}}{\pi_{hi}}$ และ $\hat{Y}_{PLh} = \frac{1}{N_h} \sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi} p_h}$ แทนตัวประมาณค่าเฉลี่ยประชากรของชั้นภูมิที่ h สำหรับตัวประมาณค่าอัตราส่วนแบบตัดข้อมูลและตัวประมาณค่าอัตราส่วนแบบใหม่ มีค่าเท่ากับ

$$\hat{Y}_R' = \sum_{h \in H} W_h \frac{\hat{Y}_{HTh}}{\hat{X}_{HTh}} \bar{X}_h \text{ และ}$$

$$\hat{Y}_R'' = \sum_{h \in H} w_h \left(\frac{\hat{Y}_{PLh}}{\hat{X}_{HTh}} \right) \bar{X}_h \text{ ตามลำดับ}$$

$$(1) \hat{Y}_{PLh} < \hat{Y}_{HTh} \text{ สำหรับทุก } H = \{1, 2, \dots, L\}$$

$$(2) \hat{Y}_R'' < \hat{Y}_R'$$

พิสูจน์ (1) จะแสดงว่า $\hat{Y}_{PLh} < \hat{Y}_{HTh}$ พิจารณา

ค่า $\hat{Y}_{PLh} = \frac{1}{N_h} \sum_{i \in s_h} \frac{r_{hi} y_{hi}}{\pi_{hi} p_h}$ สามารถเขียนให้อยู่ในรูปของเซตที่สามารถสังเกตค่าได้ ดังนี้

$$\hat{Y}_{PLh} = \frac{1}{N_h} \sum_{i \in s_{hr}} \frac{y_{hi}}{\pi_{hi} p_h} \text{ เนื่องจาก } y_{hi} > 0, \\ 0 < \pi_{hi} < 1 \text{ และ } 0 < p_h < 1 \text{ ดังนั้น}$$

$$\frac{y_{hi}}{\pi_{hi} p_h} < \frac{y_{hi}}{\pi_{hi}} \\ \frac{1}{N_h} \sum_{i \in s_{hr}} \frac{r_{hi} y_{hi}}{\pi_{hi} p_h} < \frac{1}{N_h} \sum_{i \in s_{hr}} \frac{r_{hi} y_{hi}}{\pi_{hi}} \\ \hat{Y}_{PLh} < \hat{Y}_{HTh}$$

ดังนั้น ในแต่ละชั้นภูมิสามารถสรุปได้ว่า

$$\hat{Y}_{PLh} < \hat{Y}_{HTh}$$

(2) จะแสดงว่า $\hat{Y}_R'' < \hat{Y}_R'$ จากข้อ (1) แต่ละชั้นภูมิ

$$\hat{Y}_{PLh} < \hat{Y}_{HTh} \text{ ดังนั้น}$$

$$\sum_{h \in H} w_h \left(\frac{\hat{Y}_{PLh}}{\hat{X}_{HTh}} \right) \bar{X}_h < \sum_{h \in H} w_h \frac{\hat{Y}_{HTh}}{\hat{X}_{HTh}} \bar{X} \text{ และ}$$

สามารถสรุปได้ว่า $\hat{Y}_R'' < \hat{Y}_R'$

2. ผลการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าที่นำเสนอ

ผลการวิเคราะห์ค่ารากของค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนสัมพัทธ์ของตัวประมาณค่า ดัง (Table 2) และจาก (Table 2) พบว่าทุกระดับของค่าความน่าจะเป็นที่จะตอบสนอง ตัวประมาณค่าอัตราส่วนแบบใหม่มีค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนสัมพัทธ์ต่ำกว่าตัวประมาณค่าอัตราส่วนแบบตัดข้อมูล โดยอาศัยบทนิยามที่ 1 สามารถสรุปได้ว่าตัวประมาณค่าอัตราส่วนแบบใหม่ มีประสิทธิภาพดีกว่าตัวประมาณค่าอัตราส่วนแบบตัดข้อมูล เมื่อพิจารณาค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนสัมพัทธ์ของตัวประมาณค่าอัตราส่วนใหม่และตัวประมาณค่าแบบตัดข้อมูล ในแต่ละระดับของอัตราการตอบสนอง (response rate) พบว่าค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนสัมพัทธ์ของตัวประมาณค่าอัตราส่วนใหม่จะมีต่ำกว่าค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนสัมพัทธ์ของตัวประมาณค่าแบบตัดข้อมูล ทุกระดับของอัตราการตอบสนอง นอกจากนี้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนสัมพัทธ์ของตัวประมาณค่าทั้ง 2 ตัว มีค่าลดลงเมื่ออัตราการตอบสนองเพิ่มขึ้น หรือกล่าวได้ว่ากรณีที่อัตราการตอบสนองในแต่ละชั้นภูมิมีค่าต่ำ ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนสัมพัทธ์ของตัวประมาณค่าจะมีขนาดใหญ่สาเหตุเนื่องมาจากในแต่ละชั้นภูมิถ้าอัตราการตอบสนองมีขนาดต่ำจะทำให้ขนาดของข้อมูลที่สามารถสังเกตได้ในตัวอย่าง s_h มีขนาดเล็ก ซึ่งจะส่งผลโดยตรงต่อประสิทธิภาพของตัวประมาณค่า นอกจากนี้เมื่อพิจารณาในแต่ละระดับของอัตราการตอบสนอง พบว่าที่ระดับ $p_1 = p_2 = p_3 = p_4 = 0.7$ ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนสัมพัทธ์ของตัวประมาณค่าอัตราส่วนใหม่และตัวประมาณค่าแบบตัดข้อมูล มีค่าต่ำกว่าระดับอื่น ๆ

Table 2 The relative root mean square error of the estimators.

response rate				relative root mean square error	
p_1	p_2	p_3	p_4	$RRMSE(\hat{Y}_R')$	$RRMSE(\hat{Y}_R'')$
0.5	0.5	0.5	0.5	0.51	0.32
0.5	0.7	0.5	0.5	0.71	0.54
0.5	0.5	0.7	0.5	0.46	0.32
0.5	0.5	0.5	0.7	0.45	0.34
0.5	0.5	0.7	0.7	0.73	0.56
0.5	0.7	0.5	0.7	0.47	0.32
0.7	0.5	0.7	0.5	0.33	0.22
0.7	0.7	0.5	0.5	0.46	0.27
0.5	0.7	0.7	0.7	0.53	0.35
0.7	0.5	0.7	0.7	0.34	0.21
0.7	0.7	0.5	0.7	0.42	0.25
0.7	0.7	0.7	0.5	0.32	0.19
0.7	0.7	0.7	0.7	0.29	0.19

3. ผลการประยุกต์ใช้ตัวประมาณค่ากับข้อมูลจริง

การประมาณค่าเฉลี่ยของฝุ่น PM 2.5 ในเขตดินแดง พระโขนง และเขตดอนเมือง กรุงเทพมหานคร ดัง (Table 3)

Table 3 The estimated of PM 2.5.

estimators	estimated of PM 2.5
$\hat{Y}_R'$	13.28
$\hat{Y}_R''$	18.97

ผลการประมาณค่าจาก (Table 3) พบว่า ปริมาณฝุ่นเฉลี่ยที่ได้จาก ตัวประมาณค่าแบบตัดข้อมูล และตัวประมาณค่าอัตราส่วนใหม่ มีค่าเท่ากับ 13.28 และ 18.97 ไมโครกรัมต่อลูกบาศก์เมตร ตามลำดับ ซึ่งผล

การประมาณค่าสอดคล้องกับทฤษฎีบทที่ 5 ข้อ (2) ที่ว่า $\hat{Y}_R' < \hat{Y}_R''$ หรือตัวประมาณค่าอัตราส่วนแบบตัดข้อมูลมีค่าน้อยกว่าตัวประมาณค่าอัตราส่วนแบบใหม่

สรุป

ตัวประมาณค่าเฉลี่ยประชากรของตัวแปรที่สนใจศึกษาเมื่อมีข้อมูลสูญหายที่ผู้วิจัยนำเสนอประกอบด้วยตัวประมาณค่าอัตราส่วนแบบตัดข้อมูล และตัวประมาณค่าอัตราส่วนแบบใหม่ ตัวประมาณค่าทั้งสองตัว ศึกษาภายใต้การสุ่มแบบชั้นภูมิและแต่ละชั้นภูมิสุ่มแบบใช้ความน่าจะเป็นในการเลือกไม่เท่ากันและไม่ใส่คืน สำหรับการสูญหายของข้อมูลในแต่ละชั้นภูมิเกิดขึ้นเฉพาะตัวแปรที่สนใจศึกษาภายใต้กลไกการไม่ตอบสนองแบบสุ่มสมบูรณ์ ทำให้ความน่าจะเป็นที่แต่ละหน่วยตัวอย่างจะตอบสนองมี

ค่าเท่ากันในแต่ละชั้นภูมิ นอกจากนี้ในแต่ละชั้นภูมิ ต้องทราบค่าเฉลี่ยประชากรของตัวแปรช่วยในการสร้างตัวประมาณค่าอัตราส่วนของแต่ละชั้นภูมิ นอกจากนี้ทราบผลรวมของตัวแปรกำหนดขนาดเพื่อใช้คำนวณค่า π_i และ π_{ji} ตัวประมาณค่าอัตราส่วนแบบตัดข้อมูลใช้ตัวประมาณค่าของ Bacanli, & Kadilar (2008) ประเมินค่าเฉลี่ยประชากรของตัวแปรที่สนใจศึกษาของแต่ละชั้นภูมิ แต่พิจารณาเฉพาะค่า y_{hi} ที่สามารถสังเกตค่าได้ จากนั้นใช้ตัวประมาณค่าดังกล่าวปรับตัวประมาณค่าของ Tailor, & Lone (2014) ส่วนตัวประมาณค่าอัตราส่วนแบบใหม่ใช้ตัวประมาณค่าของ Ponkaew, & Lawson (2018) ประเมินค่าเฉลี่ยประชากรของตัวแปรที่สนใจศึกษาของแต่ละชั้นภูมิ แต่พิจารณาเฉพาะค่า y_{hi} ที่สามารถสังเกตค่าได้ แต่มีการถ่วงน้ำหนักของแต่ละตัวอย่างที่ i ที่ใช้สร้างตัวประมาณค่า ด้วยค่าความน่าจะเป็นที่จะตอบสนอง p ผลการเปรียบเทียบประสิทธิภาพของตัวประมาณค่า พบว่าตัวประมาณค่าอัตราส่วนแบบใหม่ เป็นตัวประมาณค่าที่มีประสิทธิภาพดีกว่า ตัวประมาณค่าอัตราส่วนแบบตัดข้อมูล เนื่องจากตัวประมาณค่าอัตราส่วนแบบใหม่มีค่ารากเฉลี่ยกำลังสองของความคลาดเคลื่อนสัมพัทธ์ค่าต่ำกว่าตัวประมาณค่าอัตราส่วนแบบตัดข้อมูล อย่างไรก็ตามการวิจัยในครั้งนี้ศึกษาภายใต้กลไกการไม่ตอบสนองแบบ MCAR ที่การสูญหายของข้อมูลไม่มีความสัมพันธ์กับตัวแปรที่สนใจศึกษาหรือตัวแปรอื่น ๆ ซึ่งพบยากในทางปฏิบัติ ดังนั้นควรขยายตัวประมาณค่าที่ผู้วิจัยนำเสนอไปยังกลไกการไม่ตอบสนองแบบ MAR หรือ NMAR ซึ่งพบได้จริงในทางปฏิบัติมากกว่ากลไกการไม่ตอบสนองแบบ MCAR

นอกจากนี้ควรขยายไปยังกรณีที่มีการสูญหายของข้อมูล เกิดขึ้นทั้งแปรที่สนใจศึกษาและตัวแปรช่วย เพื่อเพิ่มขอบเขตในการประมาณค่า

คำขอบคุณ

งานวิจัยในครั้งนี้ได้รับทุนอุดหนุนจากมหาวิทยาลัยราชภัฏเพชรบูรณ์ ทุนอุดหนุนงานวิจัยพัฒนาวิทยาศาสตร์วิจัยและนวัตกรรม ประจำปีงบประมาณ พ.ศ. 2564 สัญญาเลขที่ FRB640052/07

เอกสารอ้างอิง

- Bacanli, S., & Kadilar, C. (2008). Ratio estimators with unequal probability designs. *Pakistan Journal of Statistics*, 24(3), 167-172.
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.
- Dansawad, N. (2020). Ratio-cum-product type of exponential estimator for the population mean in simple random sampling using the information of auxiliary variable. *Burapha Science Journal*, 25(2), 563-577. (in Thai)
- Horvitz, D. F., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685.
- Ponkaew, C., & Lawson, N. (2018). A new ratio estimator for population total in the presence of nonresponse under unequal probability sampling without replacement. *Thai Journal of Mathematics*, 2018(Special issue: ACFPTO2018), 417-429.
- Samdal, C. E., & Lundstorm, S. (2005). *Estimation in surveys with nonresponse*. New York: John Wiley & Sons.

- Singh, H. P., & Kakran, M. S. (1993). A modified ratio estimator using known coefficient of kurtosis of an auxiliary character. (Unpublished manuscript).
- Sisodia, B. V. S., & Dwivedi, V. K. (1981). A modified ratio estimator using coefficient of variation of auxiliary variable. *Journal-Indian Society of Agricultural Statistics*, 33, 13-18.
- Tailor, R., & Lone, H. A. (2014). Separate ratio-type estimators of population mean in stratified random sampling, *Journal of Modern Applied Statistical Method*, 13(1), 223-233.
- Upadhyaya, L. N., & Singh, H. P. (1999). Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal*, 41(5), 627-636.