การเปรียบเทียบตัวประมาณค่าเฉลี่ยจากการประมาณค่าสูญหาย ภายใต้ความแตกต่างของรูปแบบการสูญหาย

The Comparison of Mean Estimator by Imputation Methods under the Different Mechanisms of Missing Data

คณิศา โชติจันทึก*

Kanisa Chodjuntug*

สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี

Department of Mathematics Statistics and Computer, Faculty of Science, Ubon Ratchathani University

Submitted 7/10/2022 ; Revised 18/11/2022 ; Accepted 5/12/2022

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหาย ได้แก่ วิธีของ Gira วิธีของ Singh และวิธีของ Aliyu ภายใต้รูปแบบการสูญหายของข้อมูลได้แก่ การสูญหายแบบสุ่มอย่างสมบูรณ์ การสูญหายแบบสุ่ม และการสูญหายแบบไม่สุ่ม เพื่อประมาณค่าเฉลี่ยในกรณีข้อมูลมีปัญหาการสูญหาย งานวิจัยนี้ทำการทดลองโดยใช้การจำลองข้อมูลทั้งหมด 10,000 ครั้ง กำหนดสถานการณ์ดังนี้ ขนาดตัวอย่างได้แก่ 30, 100, 400 ภายใต้ลักษณะการสูญหายของข้อมูลในรูปแบบต่าง ๆ ร้อยละข้อมูลสูญหาย ได้แก่ 5, 10, 15 เกณฑ์ที่ใช้ในการเปรียบเทียบคือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง และค่าร้อยละประสิทธิภาพสัมพัทธ์ จากผลการทดลองวิธีที่เหมาะสมในการประมาณค่าเฉลี่ยจากการประมาณค่าสูญหาย คือ วิธีของ Singh วิธีของ Aliyu และวิธีของ Gira ตามลำดับ

คำสำคัญ: ข้อมูลสูญหาย การประมาณค่าสูญหาย การจำลองข้อมูล ความคลาดเคลื่อนเฉลี่ยกำลังสอง

***ผู้ประสานงานหลัก (Corresponding Author)**

E-mail: kanisa.c@ubu.ac.th

Abstract

The purpose of this research was to compare the different imputation methods including Gira's method, Singh's method, and Aliyu's method under the missing data mechanisms. There were three types of missing data mechanisms for estimation of the mean values: missing completely at random, missing at random, and missing not at random. In this study, the data was conducted through 10,000 simulations. The sample sizes were set at 30, 100, and 400 for a variety of missing data scenarios. The percentages of missing data were 5, 10 and 15. The criteria used for comparison were the mean squared error and the percentage of relative efficiency. From the results, it was found that the appropriate methods for estimating the mean values were Singh's method, Aliyu's method, and Gira's method, respectively.

Keywords: missing data, imputation method, simulation, mean square error

บทนำ

ในงานวิจัยด้านต่าง ๆ นิยมใช้ข้อมูลที่ได้จากการสุ่มสำรวจตัวอย่างเนื่องจากข้อจำกัดเกี่ยวกับระยะเวลาหรืองบประมาณในการทำวิจัยเป็นต้น จึงจำเป็นต้องใช้ข้อมูลตัวอย่างเพื่อนำมาวิเคราะห์เพราะไม่สามารถเก็บรวมข้อมูลของประชากรหรือข้อมูลที่สนใจศึกษาได้ทั้งหมด แต่ข้อมูลที่เก็บรวบรวมมาจากกลุ่มตัวอย่างนั้นมักพบปัญหาข้อมูลสูญหาย อาจมีสาเหตุเนื่องจากความผิดพลาดของเครื่องมือที่ใช้ในการบันทึกข้อมูลหรือสาเหตุที่เกิดขึ้นจากตัวบุคคล ซึ่งปัญหาดังกล่าวเป็นเหตุการณ์ที่นักวิจัยไม่สามารถคาดการณ์ล่วงหน้าได้ ทำให้ข้อมูลที่ได้ไม่ครบถ้วนตามจำนวนที่ได้วางแผนไว้ ด้วยข้อจำกัดที่กล่าวมาแล้วข้างต้นในการทำวิจัยจึงไม่สามารถเก็บรวบรวมข้อมูลอีกครั้งเพื่อให้ได้ข้อมูลครบตามจำนวนที่วางแผนไว้ในงานวิจัย ดังนั้นข้อมูลที่มีการสูญหายจึงส่งผลกระทบต่อกระบวนการวิจัยสำหรับการวิเคราะห์ข้อมูลทางสถิติทำให้ผลที่ได้จากการวิเคราะห์ข้อมูลมีความผิดพลาดหรือความคลาดเคลื่อนเกิดขึ้นนำไปสู่ความไม่น่าเชื่อถือของผลสรุปที่ได้จากงานวิจัยนั้น [1, 2] เพื่อแก้ไขปัญหาดังกล่าวได้มีวิธีการที่ใช้จัดการกับข้อมูลที่มีค่าสูญหายโดยไม่จำเป็นต้องเก็บรวบรวมข้อมูลใหม่มีรายละเอียดดังนี้ วิธีการตัดกลุ่มข้อมูลสูญหายทิ้งไปซึ่งเป็นวิธีการที่ง่ายและรวดเร็ว อีกทั้งในโปรแกรมทางด้านสถิติส่วนมากจะตั้งค่าเป็นวิธีการเริ่มต้นในการจัดการข้อมูลสูญหายที่เกิดขึ้นกับข้อมูล แต่วิธีนี้มีผลทำให้ข้อมูลไม่ครบถ้วนตามจำนวนที่ได้วางแผนไว้ซึ่งส่งผลกระทบต่องานวิจัยดังที่ได้กล่าวมาแล้วในข้างต้น อย่างไรก็ตามในทางสถิติได้มีวิธีการแก้ไขปัญหาข้อมูลที่มีค่าสูญหายโดยไม่ต้องตัดข้อมูลทิ้งเพื่อยังคงรักษาจำนวนข้อมูลให้ครบถ้วนตามแผนที่ได้วางไว้ คือ วิธีการประมาณค่าพารามิเตอร์ และวิธีการประมาณค่าสูญหาย (imputation method) แต่สำหรับวิธีการประมาณค่าพารามิเตอร์เป็นวิธีที่ต้องอาศัยความรู้ทางด้านสถิติขั้นสูงในกระบวนทำงาน ซึ่งเป็นวิธีการที่มีความซับซ้อนทำให้ยากต่อการนำไปใช้ในทางปฏิบัติ ส่วนวิธีการประมาณค่าสูญหายเป็นการอาศัยข้อมูลที่เก็บรวบรวมได้มาใช้ในการคำนวณค่าสูญหายโดยไม่ต้องอาศัยหลักการทางด้านสถิติขั้นสูงจึงทำให้ง่ายต่อการนำไปใช้งานจริงในทางปฏิบัติ สำหรับวิธีที่นิยมใช้ในการประมาณค่าสูญหาย ได้แก่ วิธีการประมาณค่าสูญหายด้วยค่าเฉลี่ย เป็นการนำข้อมูลข้อมูลที่เก็นรวบรวมได้มาหาค่าเฉลี่ยและนำกลับไปแทนค่าข้อมูลที่สูญหายของข้อมูลที่สนใจศึกษา และวิธีการประมาณค่าสูญหายด้วยค่าอัตราส่วน เป็นการใช้ข้อมูลจากตัวแปรเสริม (auxiliary variable) ที่มีความสัมพันธ์กับข้อมูลที่สนใจศึกษามาช่วยในการประมาณค่าสูญหาย วิธีดังกล่าวนอกจากนำไปใช้ประมาณค่าสูญหายและยังสามารถพัฒนาไปสู่ตัวประมาณค่าสำหรับหาค่าเฉลี่ยของประชากรในกรณีข้อมูลสูญหายได้อีกด้วย ซึ่งเป็นที่ทราบกันดีว่าค่าเฉลี่ยเป็นการวิเคราะห์ข้อมูลพื้นฐานที่ใช้ในงานวิจัยด้านต่าง ๆ ดังนั้นจึงมีนักสถิติหลายท่านสนใจพัฒนาตัวประมาณค่าเฉลี่ยโดยใช้ข้อมูลของตัวแปรเสริมเพื่อเพิ่มประสิทธิภาพวิธีของตัวประมาณค่าเฉลี่ยได้แก่ วิธีของ Chakrabarty [3], วิธีของ Srivenkataramana & Tracy [4], วิธีของ Singh & Shukla [5], วิธีของ Bahl & Tuteja [6], วิธีของ Singh & Tailor [7], วิธีของ Kadilar & Cingi [8, 9], Upadhyaya [10], วิธีของ Yadav & Kadilar [11], วิธีของ Khan [12], วิธีของ Kadilar [13] แต่วิธีดังกล่าวไม่สามารถประมาณค่าเฉลี่ยในกรณีที่มีข้อมูลสูญหายได้ และในปัจจุบันได้มีการพัฒนาวิธีการประมาณค่าสูญหายโดยใช้แนวคิดการใช้ข้อมูลตัวแปรเสริมเพื่อเพิ่มประสิทธิภาพวิธีการประมาณค่าสูญหายเช่นเดียวกับวิธีต่าง ๆ ที่กล่าวมาแล้วข้างต้นและนำมาประยุกต์ใช้สู่ตัวประมาณค่าเฉลี่ยในกรณีที่ข้อมูลสูญหาย

คือ วิธีของ Gira [14], วิธีของ Singh [15] และวิธีของ Aliyu [16] จากการศึกษาพบว่าทั้ง 3 วิธีมีประสิทธิภาพดีกว่าวิธีการประมาณค่าสูญหายด้วยค่าเฉลี่ย และวิธีการประมาณค่าสูญหายด้วยค่าอัตราส่วน แต่อย่างไรก็ตามยังไม่มีงานวิจัยใดนำวิธีดังกล่าวมาเปรียบเทียบประสิทธิภาพกันและในแต่ละวิธีได้ศึกษาเฉพาะภายใต้รูปแบบการสูญหายของข้อมูลหายแบบสุ่มอย่างสมบูรณ์ (missing completely at random; MCAR) เท่านั้น ซึ่งในการสุ่มสำรวจตัวอย่างรูปแบบการสูญหายของข้อมูลไม่ได้มีแค่เฉพาะ MCAR แต่ยังมีรูปแบบอื่นอีกด้วย โดย Little & Rubin [17] ได้จำแนกรูปแบบการสูญหายของข้อมูล มีรายละเอียดดังนี้ 1) การสูญหายแบบสุ่มอย่างสมบูรณ์ (missing completely at random; MCAR) คือความน่าจะเป็นที่ข้อมูลสูญหายไม่ขึ้นกับค่าใด ๆ หรือไม่มีความสัมพันธ์กับข้อมูลอื่นเลย 2) การสูญหายแบบสุ่ม (missing at random; MAR) คือ ความน่าจะเป็นที่ข้อมูลสูญหายขึ้นอยู่กับข้อมูลอื่น ๆ หรือมีความสัมพันธ์กับข้อมูลอื่นที่ทราบค่า แต่ไม่ขึ้นกับค่าของข้อมูลที่สูญหาย 3) การสูญหายแบบไม่สุ่ม (missing not at random; MNAR) คือความน่าจะเป็นที่ข้อมูลสูญหายที่ตำแหน่งใด ๆ โดยขึ้นกับค่าของตำแหน่งนั้น ๆ

ดังนั้นงานวิจัยนี้จึงสนใจเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าข้อมูลสูญหายทั้ง 3 วิธี คือ วิธีของ Gira วิธีของ Singh และวิธีของ Aliyu ภายใต้รูปแบบข้อมูลสูญหายทั้งหมดได้แก่ MCAR MAR และ MNAR ซึ่งยังไม่มีผู้ใดศึกษา ในการประมาณค่าเฉลี่ย สำหรับเกณฑ์ที่ใช้ในการเปรียบเทียบคือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง และค่าร้อยละประสิทธิภาพสัมพัทธ์ อีกทั้งในงานวิจัยนี้ยังศึกษาอิทธิพลของร้อยละข้อมูลหายและขนาดตัวอย่างที่เปลี่ยนแปลงไปซึ่งอาจจะส่งผลต่อประสิทธิภาพของตัวประมาณค่าได้

วัตถุประสงค์

1. เปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายด้วยวิธีของ Gira วิธีของ Singh และวิธีของ Aliyu จากตัวประมาณค่าเฉลี่ยซึ่งได้มาจากวิธีของ Gira วิธีของ Singh และวิธีของ Aliyu ภายใต้ข้อมูลสูญหายทั้ง 3 แบบ ได้แก่ MCAR MAR และ MNAR

2. ศึกษาปัจจัยที่มีผลต่อประสิทธิภาพวิธีการประมาณค่าสูญหายจากการประมาณค่าเฉลี่ย ได้แก่ ร้อยละข้อมูลสูญหายและขนาดตัวอย่าง

วิธีดำเนินการวิจัย

กำหนดให้สัญลักษณ์ที่ใช้ในงานวิจัย แทนค่าต่าง ๆ ดังนี้

y_i คือ ค่าตัวแปรที่สนใจศึกษาลำดับที่ i โดย i = 1, 2, 3,...

x_i คือ ค่าตัวแปรเสริมลำดับที่ i

n คือ ขนาดตัวอย่าง

r คือ จำนวนข้อมูลที่เก็บรวบรวมได้ ($r<n$)

n-r คือ จำนวนข้อมูลสูญหาย

R คือ เซตของข้อมูลที่เก็บรวบรวมได้

R^c คือ เซตของข้อมูลสูญหาย

$\bar{Y}$ คือ ค่าเฉลี่ยของประชากรตัวแปรที่สนใจศึกษา

$\bar{X}$ คือ ค่าเฉลี่ยของประชากรตัวแปรเสริม

S_{xy} คือ ความแปรร่วมระหว่างตัวแปรเสริมและตัวแปรที่สนใจศึกษา

S_x^2 คือ ความแปรปรวนของตัวแปรเสริม

S_y^2 คือ ความแปรปรวนของตัวแปรที่สนใจศึกษา

และ ρ คือ สัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรเสริมและตัวแปรที่สนใจศึกษา

ในงานวิจัยนี้ใช้การจำลองข้อมูลเพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย โดยใช้โปรแกรม R จำลองข้อมูลซึ่งประกอบด้วยตัวแปร Y แทนตัวแปรที่สนใจศึกษากับตัวแปร X แทนตัวแปรเสริม โดยมีจำนวนประชากรทั้งหมด N ตัว กำหนดให้มีการแจกแจงปกติ ภายใต้สถานการณ์ต่าง ๆ โดยมีรายละเอียดดังนี้

1) กำหนดให้ค่าพารามิเตอร์ ได้แก่ N = 3,000, $\bar{Y}$ = 200, $\bar{X}$ = 100, S_y = 50, S_x = 15, ρ = 0.9
2) กำหนดขนาดตัวอย่าง (n) ได้แก่ 30, 100, 400 เพื่อใช้เป็นกลุ่มตัวอย่างขนาดเล็ก ปานกลาง และใหญ่ ตามลำดับ โดยสุ่มจากประชากรด้วยวิธีการเลือกตัวอย่างสุ่มแบบง่ายแบบไม่ใส่คืน
3) กำหนดร้อยละข้อมูลสูญหาย ได้แก่ 5, 10, 15 ของขนาดตัวอย่าง เพื่อใช้เป็นระดับการสูญหายของข้อมูลที่มีจำนวนน้อย ปานกลาง และมาก ตามลำดับ
4) จากข้อ 3) จำลองลักษณะการสูญหายของข้อมูลทั้ง 3 แบบ ได้แก่ MCAR MAR และ MNAR
5) คำนวณตัวประมาณค่าเฉลี่ยที่ได้จากวิธีการประมาณค่าสูญหายแต่ละวิธี มีรายละเอียดของวิธีการประมาณค่าสูญหายดังนี้

5.1) วิธีของ Gira

$$y_{.i}=\begin{cases} y_i & ;i\in R \\ \bar{y}_r\left[n\left(\alpha_1-\bar{x}_r\right)/\left(\alpha_1-\bar{X}\right)-r\right]\left(x_i/\sum_{i\in R^c}x_i\right) & ;i\in R^c \end{cases} \quad (1)$$

เมื่อ $\alpha_1=\left[\bar{X}\left(R-B\right)\right]/B$, $R=\bar{Y}/\bar{X}$, $B=S_{xy}/S_x^2$

ตัวประมาณค่าเฉลี่ยที่ได้จากสมการที่ 1 คือ

$$\bar{y}_1=\bar{y}_r\left(\alpha_1-\bar{x}_r\right)/\left(\alpha_1-\bar{X}\right) \quad (2)$$

5.2) วิธีของ Singh

$$y_{.i}=\begin{cases} y_i & ;i\epsilon R \\ \left[\bar{y}_r/(n-r)\right](n\exp(\alpha_2 K)-r) & ;i\epsilon R^c \end{cases} \quad (3)$$

เมื่อ $K=\left(\bar{X}^{1/h}-\bar{x}_r^{1/h}\right)/\left(\bar{X}^{1/h}+(a-1)\bar{x}_r^{1/h}\right)$

นอกจากนี้ Singh *et al*. [15] ได้เสนอให้ $\alpha_2=ah\rho\left(C_y/C_x\right)$ โดยที่ a=1, h=1

ตัวประมาณค่าเฉลี่ยที่ได้จากสมการที่ 3 คือ

$$\bar{y}_2=\bar{y}_r\exp\left(\alpha_2\left[\left(\bar{X}^{1/h}-\bar{x}_r^{1/h}\right)/\left(\bar{X}^{1/h}+(a-1)\bar{x}_r^{1/h}\right)\right]\right) \quad (4)$$

5.3) วิธีของ Aliyu

$$y_{.i}=\begin{cases} y_i & ;i\epsilon R \\ \left[1/(n-r)\right]\left(nT-r\bar{y}_r\right) & ;i\epsilon R^c \end{cases} \quad (5)$$

เมื่อ $T=\left(\alpha_3\bar{y}_r+(1-\alpha_3)\bar{y}_r\left(\bar{x}_r/\bar{X}\right)\right)\exp\left((\bar{X}-\bar{x}_r)/(\bar{X}+\bar{x}_r)\right)$,

$\alpha_3=0.5+\rho\left(C_y/C_x\right)$

ตัวประมาณค่าเฉลี่ยที่ได้จากสมการที่ 5 คือ

$$\bar{y}_3=\left(\alpha_3\bar{y}_r+(1-\alpha_3)\bar{y}_r\left(\bar{x}_r/\bar{X}\right)\right)\exp\left((\bar{X}-\bar{x}_r)/(\bar{X}+\bar{x}_r)\right) \quad (6)$$

6) ในแต่ละสถานการณ์ทำซ้ำทั้งหมด 10,000 ครั้ง

7) คำนวณค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (mean square error: MSE) และค่าร้อยละประสิทธิภาพสัมพัทธ์ (percentage of relative efficiency: PRE) ได้ดังนี้

$$MSE\left(\bar{y}_M\right)=\frac{\sum_{l=1}^{10,000}\left(\left(\bar{y}_M\right)_l-\bar{Y}\right)^2}{10,000}$$

กำหนดให้ $\bar{y}_M$ คือตัวประมาณค่าเฉลี่ยใด ๆ; M=1, 2, 3 และในงานวิจัยนี้ให้ค่าร้อยละประสิทธิภาพสัมพัทธ์ของ $\bar{y}_M$ เทียบกับ $\bar{y}_1$

$$PRE\left(\bar{y}_M\right)=\frac{MSE\left(\bar{y}_1\right)}{MSE\left(\bar{y}_M\right)}\times 100$$

8) สรุปผลการวิจัย

การเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายจะพิจารณา MSE และ PRE ของตัวประมาณค่าเฉลี่ยที่คำนวณได้จากวิธีต่าง ๆ หากตัวประมาณค่าเฉลี่ยใดมีค่า MSE ต่ำที่สุด และมีค่า PRE สูงที่สุด แสดงว่าวิธีที่ใช้ในการประมาณค่าเฉลี่ยนั้นมีประสิทธิภาพดีที่สุด หรือวิธีดังกล่าวเหมาะสมสำหรับการนำไปใช้การประมาณค่าเฉลี่ยเมื่อข้อมูลสูญหาย

ผลการวิจัย

จากการจำลองข้อมูลสามารถคำนวณค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (MSE) และค่าร้อยละประสิทธิภาพสัมพัทธ์ (PRE) ของตัวประมาณค่าที่ได้จากวิธีต่าง ๆ ภายใต้สถานการณ์รูปแบบการสูญหายของข้อมูล คือ MCAR, MAR และ MNAR มีร้อยละข้อมูลสูญหายคือ 5, 10 และ 15 สำหรับขนาดตัวอย่างคือ 30, 100 และ 400 ได้ดังตารางที่ 1

ตารางที่ 1 ค่า MSE และ PRE ของตัวประมาณค่าเฉลี่ย

ร้อยละ	ขนาดตัวอย่าง	รูปแบบ	ความคลาดเคลื่อนเฉลี่ยกำลังสอง			ร้อยละประสิทธิภาพสัมพัทธ์		
ข้อมูลสูญหาย			วิธีที่ 1	วิธีที่ 2	วิธีที่ 3	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
5	30	MCAR	80.4882	16.0733	16.1219	100.00	500.76	499.25
		MAR	71.7576	15.4128	15.5182	100.00	465.57	462.41
		MNAR	73.4147	16.2373	16.2135	100.00	452.14	452.80
	100	MCAR	23.4195	4.5848	4.5911	100.00	510.80	510.11
		MAR	21.1596	4.4431	4.4610	100.00	476.23	474.33
		MNAR	22.1841	5.1366	5.1133	100.00	431.88	433.85
	400	MCAR	5.4709	1.0324	1.0327	100.00	529.91	529.75
		MAR	4.9426	0.9909	0.9923	100.00	498.79	498.11
		MNAR	5.8669	1.8732	1.8492	100.00	313.20	317.26
10	30	MCAR	82.5217	16.7722	16.8180	100.00	492.02	490.68
		MAR	69.3303	15.9368	16.1084	100.00	435.03	430.40
		MNAR	72.0434	17.3874	17.3009	100.00	414.34	416.42
	100	MCAR	23.4954	4.9125	4.9199	100.00	478.28	477.56
		MAR	19.7269	4.6680	4.7405	100.00	422.60	416.14
		MNAR	21.8346	6.6883	6.5871	100.00	326.46	331.48
	400	MCAR	5.4298	1.0911	1.0916	100.00	497.63	497.39
		MAR	4.5618	1.0481	1.0764	100.00	435.25	423.81
		MNAR	6.7993	3.5024	3.3851	100.00	194.13	200.86

ตารางที่ 1 ค่า MSE และ PRE ของตัวประมาณค่าเฉลี่ย (ต่อ)

ร้อยละ ข้อมูลสูญหาย	ขนาดตัวอย่าง	รูปแบบ	ความคลาดเคลื่อนเฉลี่ยกำลังสอง			ร้อยละประสิทธิภาพสัมพัทธ์		
			วิธีที่ 1	วิธีที่ 2	วิธีที่ 3	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
15	30	MCAR	81.9480	17.1633	17.2025	100.00	477.46	476.37
		MAR	66.2559	16.0951	16.3800	100.00	411.65	404.49
		MNAR	69.2515	18.4353	18.2857	100.00	375.65	378.72
	100	MCAR	24.1001	5.2541	5.2620	100.00	458.69	458.00
		MAR	19.7028	5.0246	5.2059	100.00	392.13	378.47
		MNAR	21.7601	8.3617	8.1021	100.00	260.24	268.57
	400	MCAR	5.4576	1.1690	1.1695	100.00	466.85	466.68
		MAR	4.9454	1.2829	1.3955	100.00	385.50	354.37
		MNAR	7.3621	5.1161	4.8235	100.00	143.90	152.63

จากตารางที่ 1 เมื่อพิจารณาภายใต้รูปแบบการสูญหายข้อมูล MCAR และ MAR พบว่าวิธีที่ 2 มีค่า PRE สูงที่สุดและมีค่า MSE ต่ำที่สุด แสดงว่าวิธีที่ 2 มีประสิทธิภาพดีที่สุด และภายใต้รูปแบบการสูญหายข้อมูล MNAR พบว่า วิธีที่ 3 มีค่า PRE สูงที่สุดและมีค่า MSE ต่ำที่สุด แสดงว่าวิธีที่ 3 มีประสิทธิภาพดีที่สุด

เมื่อพิจารณา MSE ในแต่ละวิธีภายใต้ขนาดตัวอย่างและร้อยละข้อมูลสูญหายที่ใช้ในการทดลอง เพื่อวิเคราะห์อิทธิพลของรูปแบบการสูญหายข้อมูลที่มีต่อค่า MSE มีรายละเอียดดังนี้

วิธีที่ 1 ภายใต้ขนาดตัวอย่าง 30, 100 และข้อมูลสูญหายร้อยละ 5, 10, 15 พบว่า รูปแบบการสูญหายข้อมูลที่มีค่า MSE ต่ำที่สุดคือ MAR รองลงมาคือ MNAR และ MCAR ตามลำดับ

วิธีที่ 1 ภายใต้ขนาดตัวอย่าง 400 และข้อมูลสูญหายร้อยละ 5, 10, 15 พบว่า รูปแบบการสูญหายข้อมูลที่มีค่า MSE ต่ำที่สุดคือ MAR รองลงมาคือ MCAR และ MNAR ตามลำดับ

วิธีที่ 2 และ 3 ภายใต้ขนาดตัวอย่าง 30, 100, 400 และข้อมูลสูญหายร้อยละ 5, 10, 15 พบว่า รูปแบบการสูญหายข้อมูลที่มีค่า MSE ต่ำที่สุดคือ MAR รองลงมาคือ MCAR และ MNAR ตามลำดับ ยกเว้น ขนาดตัวอย่าง 400 และข้อมูลสูญหายร้อยละ 15 มีค่า MSE ต่ำที่สุด คือ MCAR, MAR และ MNAR ตามลำดับ

การอภิปรายผลการวิจัย

งานวิจัยนี้ได้ประมาณค่าเฉลี่ยจากวิธีการประมาณค่าสูญหายจากวิธีที่ 1 คือ วิธีของ Gira วิธีที่ 2 คือ วิธีของ Singh และวิธีที่ 3 คือ วิธีของ Aliyu ภายใต้รูปแบบการสูญหายข้อมูล ได้แก่ การสูญหายแบบสุ่มอย่างสมบูรณ์ (MCAR) การสูญหายแบบสุ่ม (MAR) และการสูญหายแบบไม่สุ่ม (MNAR) จากผลการทดลองแสดงให้เห็นว่ารูปแบบการสูญหายที่แตกต่างกันมีผลต่อประสิทธิภาพของการประมาณค่าเฉลี่ยที่ได้จากวิธีการประมาณค่าสูญหายในแต่ละวิธีและมีความสอดคล้องกับงานวิจัยเรื่อง ความถูกต้องในการแทนค่าข้อมูลสูญหายในการจำแนกประเภทกรณีข้อมูลสองกลุ่ม โดยพบว่ารูปแบบการสูญหายมีผลต่อประสิทธิภาพของวิธีการประมาณค่า

สูญหาย [18] นอกจากนี้เมื่อศึกษาปัจจัยที่มีผลต่อประสิทธิภาพของวิธีการประมาณค่าสูญหายภายใต้รูปแบบการสูญหายของข้อมูลที่แตกต่างกัน ได้แก่ ร้อยละข้อมูลสูญหายและขนาดตัวอย่าง พบว่าในทุกรูปแบบการสูญหายของข้อมูลเมื่อร้อยละข้อมูลสูญหายเพิ่มขึ้นมีผลทำให้ค่า MSE เพิ่มขึ้น แสดงว่าประสิทธิภาพของวิธีการประมาณค่าสูญหายมีประสิทธิภาพลดลงเมื่อร้อยละการสูญหายเพิ่มขึ้น และเมื่อขนาดตัวอย่างเพิ่มขึ้นมีผลทำให้ค่า MSE ลดลง แสดงว่าประสิทธิภาพของวิธีการประมาณค่าสูญหายเพิ่มขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้นซึ่งสอดคล้องกับงานวิจัยของ Chodjuntug และ Lawson [19] โดยพบว่าขนาดตัวอย่างและข้อมูลหายมีผลต่อประสิทธิภาพของการประมาณค่าเฉลี่ยที่ได้จากวิธีการประมาณค่าสูญหาย [19] และสอดคล้องกับงานวิจัยของ Chodjuntug และ Lawson [20] โดยพบว่าอิทธิพลของขนาดตัวอย่างและร้อยละการสูญหาย มีผลต่อการประมาณสูญหายสำหรับการประมาณค่าเฉลี่ย [20] ผลการทดลองที่ได้จากงานวิจัยนี้ แสดงให้เห็นว่าควรมีการคำนึงถึงรูปแบบข้อมูลสูญหายก่อนการนำไปประยุกต์ใช้ในงานวิจัยเพื่อเลือกวิธีการประมาณค่าสูญหายที่เหมาะสม โดยรูปแบบข้อมูลสูญหายที่แตกต่างกันมีผลต่อประสิทธิภาพของวิธีการประมาณค่าสูญหาย

สรุปผลการวิจัย

งานวิจัยนี้เป็นการเปรียบเทียบวิธีการประมาณค่าสูญหายภายใต้รูปแบบการสูญหายของข้อมูลที่แตกต่างกันสำหรับการประมาณค่าเฉลี่ย จากผลการทดลองพบว่า ภายใต้การสูญหายแบบไม่สุ่มทำให้วิธีการประมาณค่าสูญหายสำหรับการประมาณค่าเฉลี่ยในทุกวิธีมีประสิทธิภาพดีที่สุด อย่างไรก็ตามเมื่อพิจารณารูปแบบการสูญหายทั้งหมดพบว่า วิธีของ Singh และวิธีของ Aliyu ดีกว่าวิธีของ Gira หากพิจารณาจากรูปแบบในการคำนวณพบว่า วิธีของ Singh และวิธีของ Aliyu จะมีความซับซ้อนมากกว่าวิธีของ Gira แต่วิธีของ Singh และวิธีของ Aliyu มีประสิทธิภาพที่ดีกว่าอย่างชัดเจนเมื่อพิจารณาจากค่า MSE และ PRE ที่ได้จากผลการทดลอง ซึ่งค่าที่ได้มีความแตกต่างจากวิธีของ Gira อย่างเห็นได้ชัด และทั้งสองวิธีมีประสิทธิภาพใกล้เคียงกันโดยการพิจารณาจากค่า MSE และ PRE ซึ่งมีค่าใกล้เคียงกัน นอกจากนี้ขนาดตัวอย่างและปริมาณจำนวนข้อมูลที่สูญหาย ซึ่งในงานวิจัยนี้ได้แสดงอยู่ในรูปแบบร้อยละข้อมูลสูญหายเป็นปัจจัยที่มีผลต่อประสิทธิภาพของวิธีการประมาณค่าสูญหายเช่นกัน โดยประสิทธิภาพของวิธีการประมาณค่าสูญหายมีประสิทธิภาพลดลงเมื่อร้อยละการสูญหายเพิ่มขึ้นและขนาดตัวอย่างลดลง ดังนั้นการเลือกวิธีการประมาณค่าสูญหายสำหรับการประยุกต์ใช้ในงานวิจัยควรมีการพิจารณาความเหมาะสมของรูปแบบข้อมูลสูญหาย เพื่อประสิทธิภาพที่ดีของผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลเมื่อเกิดปัญหาข้อมูลมีค่าสูญหายซึ่งจะนำไปสู่การใช้ประโยชน์สูงสุดจากการทำวิจัย อย่างไรก็ตามงานวิจัยนี้ศึกษาเฉพาะข้อมูลสูญหายกรณีข้อมูลถูกเลือกสุ่มแบบง่าย ซึ่งในอนาคตสามารถต่อยอดจากงานวิจัยครั้งนี้โดยศึกษาในกรณีข้อมูลที่ถูกเลือกสุ่มมีความซับซ้อนมากยิ่งขึ้น เช่น แบบเป็นระบบ แบบแบ่งชั้นภูมิ แบบแบ่งกลุ่ม เป็นต้น

เอกสารอ้างอิง

[1] Gibson, N. M., & Olejnik, S. (2003). Treatment of missing data at the second level of hierarchical linear models. *Educational and Psychological Measurement*, 63(2), 204-238.

[2] วราฤทธิ์ พานิชกิจโกศลกุล. (2552). การจำลองแบบมอนติคาร์โลสำหรับประมาณค่าความแปรปรวนของการแจกแจงอินเวอร์สเกาส์เซียน เมื่อข้อมูลมีค่าสูญหาย, *ว.การวิจัยกาสะลองคำ*, 3(1), 14-23.

[3] Chakrabarty, R. P. (1979). Some ratio estimators. *Journal of the Indian Society of Agricultural Statistics*, 31(1), 49-57.

[4] Srivenkataramana T., & Tracy D. S. (1980). An alternative to ratio method in sample survey. *Annals of the Institute of Statistical Mathematics*, 32, 111-120.

[5] Singh, V. K., & Shukla, D. (1987). One parameter family of factor-type ratio estimation. *Metron*, 45, 273-283.

[6] Bahl, S., & Tuteja, R. K. (1991). Ratio and product type exponential estimator. *Information and Optimization Sciences*, 12(1), 159-163.

[7] Singh, H. P., & Tailor, R. (2003). Use of known correlation coefficient in estimating the finite population means. *Statistics in Transition*, 6(4), 555-560.

[8] Kadilar, C., & Cingi, H. (2006a). An improvement in estimating the population mean by using the correlation coefficient. *Hacettepe Journal of Mathematics and Statistics*, 35(1), 103-109.

[9] Kadilar, C., & Cingi, H. (2006b). Improvement in estimating the population mean in simple random sampling. *Applied Mathematics Letters*, 19, 75-79.

[10] Upadhyaya, L. N., Singh, H. P., Chatterjee, S., & Yadav, R. (2011). Improved ratio and product exponential type estimators. *Journal of Statistical Theory and Practice*, 5(2), 285-302.

[11] Yadav, S. K., & Kadilar, C. (2013). Efficient family of exponential estimators for the population mean. *Hacettepe Journal of Mathematics and Statistics*, 42(6), 671-677.

[12] Khan, S. A., Ali, H., Manzoor, S., & Alamgir, K. (2015). A class of transformed efficient ratio estimators of finite population mean. *Pakistan Journal of Statistics*, 31(4), 353-362.

[13] Kadilar, G. O. (2016). A new exponential type estimator for the population mean in simple random sampling. *Journal of Modern Applied Statistical Methods*, 15(2), 207-214.

[14] Gira, A. A. (2015). Estimation of population mean with a new imputation method. *Applied Mathematical Sciences*, 9(34), 1663-1672.

[15] Singh, R., Verma, H. K., & Sharma, P. (2016). Estimation of population mean using exponential type imputation technique for missing observations. *Journal of Modern Applied Statistical Methods*, 15(1), 358-372.

[16] Aliyu, Y. H., Adewara, A. A., Audu, A., Abidoye, O. A., Sulaiman, I., & Aliyu, M. B. (2022). Modified compromized type method of imputation for estimating population mean. *Journal of Scientific Research*, 66(1), 404-410.

[17] Little, R., & Rubin, D. (2002). *Statistical analysis with missing data*. New York: Wiley.

[18] Vongprasert, J. (2021). Missing Data Imputation Based on Accuracy of Binary Classification. *The Journal of King Mongkut's University of Technology North Bangkok*, 31(1), 90-98.

[19] Chodjuntug, K., & Lawson, N. (2020). A chain ratio exponential-type compromised imputation for mean estimation: case study on ozone pollution in Saraburi, Thailand. *Journal of Probability and Statistics*, 1, 1-6.

[20] Chodjuntug, K., & Lawson, N. (2022). Imputation for estimating the population mean in the presence of nonresponse, with application to fine particle density in Bangkok. *Mathematical Population Studies*, [Online], available from : https://doi.org/10.1080/08898480.2021.1997466