

การเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าเบี่ยงเบนมาตรฐานประชากร สำหรับข้อมูลที่มีการแจกแจงปกติและมีค่านอกเกณฑ์

นิธิภัทร กมลสุข*

สำนักการศึกษาทั่วไป สถาบันการจัดการปัญญาภิวัฒน์

Received: 7 June 2022

Revised: 13 December 2022

Accepted: 7 August 2023

บทคัดย่อ

งานวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าเบี่ยงเบนมาตรฐานของประชากร 4 วิธีคือ วิธีส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย วิธีพิสัยที่ถูกปรับค่า วิธีส่วนเบี่ยงเบนมาตรฐานที่ถูกปรับค่า และวิธีส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง สำหรับข้อมูลที่มีการแจกแจงปกติ เมื่อข้อมูลมีค่านอกเกณฑ์ โดยใช้ค่าความเอนเอียง สมบูรณ์และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดมาเป็นเกณฑ์พิจารณาภายใต้สถานการณ์จำลอง 90 สถานการณ์ ที่เป็นสถานการณ์ของข้อมูลที่มีการแจกแจงปกติ มีค่าเฉลี่ยเท่ากับ 30 และค่าเบี่ยงเบนมาตรฐานประชากร 5 จำนวน ได้แก่ 1, 5, 10, 15 และ 20 ขนาดตัวอย่าง 6 จำนวน ได้แก่ 10, 20, 30 50, 70 และ 100 ร้อยละของค่านอกเกณฑ์ 3 จำนวน ได้แก่ ร้อยละ 0, 10 และ 20 พบว่า เมื่อไม่มีค่านอกเกณฑ์ปนเปื้อนอยู่ วิธีส่วนเบี่ยงเบนมาตรฐานของตัวอย่างเป็นวิธีการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรที่มีประสิทธิภาพมากที่สุด ซึ่งจะมีประสิทธิภาพมากขึ้นถ้าขนาดตัวอย่างใหญ่ขึ้น แต่ประสิทธิภาพจะลดลงเมื่อค่าส่วนเบี่ยงเบนมาตรฐานเพิ่มมากขึ้น และพบว่าวิธีส่วนเบี่ยงเบนมาตรฐานที่ถูกปรับค่า มีประสิทธิภาพมากกว่าวิธีพิสัยที่ถูกปรับค่า ส่วนในกรณีที่ค่านอกเกณฑ์ร้อยละ 10 และ 20 พบว่าวิธีส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยมีประสิทธิภาพมากที่สุด

คำสำคัญ: ค่าเบี่ยงเบนมาตรฐานของประชากร ค่านอกเกณฑ์

* ผู้ประสานงานหลัก; อีเมล: nitipatkam@pim.ac.th

Efficiency Comparison of the Population Standard Deviation Estimation Methods for Data set with Normal Distribution and Containing Outlier

Nitipat Kamolsuk*

Office of General Education, Panyapiwat Institute of Management

Received: 7 June 2022

Revised: 13 December 2022

Accepted: 7 August 2023

Abstract

The objective of this research was to compare the efficiency of four population standard deviation estimation methods, mean absolute deviation, adjusted range, adjusted standard deviation and sample standard deviation – for a normal distribution when data set containing outlier when the lowest absolute bias and the lowest mean square error were used as criteria. Under 90 simulation scenarios that normal distribution with mean equals 30 and the population standard deviation equals 1, 5, 10, 15 and 20, sample size equals 10, 20, 30, 50, 70 and 100 with the percentage of outlier equals 0%, 10% and 20%. The results found that sample standard deviation was the most efficient when the data set not contain outlier and it was more efficient if the sample size was larger, but the efficiency would decrease when standard deviation was increase and found that adjusted standard deviation has more efficient than adjusted range in all situation which the data set not contain outlier and found that mean absolute deviation was the most efficient when the percentage of outlier equals 10% and 20%.

Keywords: Population Standard Deviation, Outlier

* Corresponding Author; E-mail: nitipatkam@pim.ac.th

ความเป็นมาและความสำคัญของปัญหา

การรวบรวมข้อมูล (Data Collection) บ่อยครั้งพบว่าข้อมูลที่รวบรวมมีการปนเปื้อน (Contaminated) ทำให้ข้อมูลมีค่าต่ำหรือสูงเกินจากข้อมูลส่วนใหญ่ โดยเรียกข้อมูลที่ปนเปื้อนนี้ว่า ค่านอกเกณฑ์ (Outliers) ค่าผิดปกติ (Abnormalities) หรือค่าความไม่สอดคล้อง (Discordant) ซึ่งการเกิดค่านอกเกณฑ์นี้เป็นการเบี่ยงเบน (Deviant) ของข้อมูล ส่งผลให้เกิดความผิดปกติในการทำเหมืองข้อมูล (Data Mining) หรือความผิดปกติจากผลการวิเคราะห์ข้อมูลทางสถิติได้ (Aggarwal, 2016) โดยส่วนใหญ่แล้วความผิดปกติที่ทำให้เกิดค่านอกเกณฑ์นั้นมาจากขั้นตอนการสร้างข้อมูล (Generate) การสังเกต (Observation) หรือการรวบรวม (Collection) ซึ่งเกิดขึ้นจากหลายสาเหตุ เช่น เกิดจากความคลาดเคลื่อนของการวัด (Measurement Error) ความคลาดเคลื่อนที่เกิดจากการบันทึกข้อมูล หรืออาจมาจากสาเหตุของสมมติฐานการแจกแจงของข้อมูลที่ไม่ตรงกับลักษณะการแจกแจงของข้อมูลที่แท้จริง (Chelishchev et al., 2018) โดยหากมีการนำข้อมูลที่มีค่านอกเกณฑ์มาใช้อาจทำให้ผลการวิเคราะห์ข้อมูลเกิดความผิดพลาด ที่แก้ปัญหามาโดยการตัดข้อมูลที่มีค่านอกเกณฑ์นี้ออกหรือปรับค่าให้เหมาะสม ขึ้นอยู่กับดุลยพินิจของผู้วิจัยหรือผู้ที่ต้องการนำผลการวิเคราะห์ข้อมูลนี้ไปใช้

ทั้งนี้ ในการวิเคราะห์ข้อมูลหากต้องการทราบการกระจายของข้อมูลจากประชากรจะใช้วิธีการประมาณค่าได้หลายวิธี โดยวิธี Sample Standard Deviation (SD) เป็นวิธีหนึ่งที่นิยมใช้ แต่วิธีนี้จะทำให้ตัวประมาณค่า (Estimator) ขาดสมบัติความไม่เอนเอียง (Unbiased) และมีประสิทธิภาพต่ำเมื่อใช้กับตัวอย่างขนาดเล็ก (Montgomery, 2012) นอกจากนี้ยังมีวิธีการประมาณค่าเบี่ยงเบนมาตรฐานประชากรวิธีอื่นที่มีประสิทธิภาพ เช่น วิธี Mean Absolute Deviation (MAD) วิธี Adjusted Range (AR) วิธี Adjusted Standard Deviation (ASD) โดย Montgomery (2012) ได้เปรียบเทียบประสิทธิภาพตัวประมาณค่าวิธี AR และวิธี ASD ที่ใช้ค่าประสิทธิภาพสัมพัทธ์ (Relative Efficiency) เป็นเกณฑ์การเปรียบเทียบพบว่า เมื่อขนาดตัวอย่าง (n) เพิ่มขึ้น วิธี ASD จะมีประสิทธิภาพมากกว่าวิธี AR สอดคล้องกับผลงานวิจัยของ Boonyavaha et al. (2018) พบว่า วิธี ASD มีประสิทธิภาพมากกว่าวิธี AR เมื่อข้อมูลแจกแจงปกติ ในตัวอย่างขนาดกลางและเล็ก และไม่มีค่านอกเกณฑ์ปนเปื้อน โดยใช้ค่าความเอนเอียงสมบูรณ์ (Absolute Bias: ABS) เป็นเกณฑ์การพิจารณาที่กำหนดให้ข้อมูลมีค่านอกเกณฑ์ในระดับปานกลาง ตามเงื่อนไขของการตรวจสอบค่านอกเกณฑ์โดยใช้แผนภาพกล่อง (Box Diagram) และหนวด (Whisker) ที่เป็นค่าที่อยู่นอกช่วง ($Q_3 + 1.5(IQR), Q_3 + 3(IQR)$) เมื่อ Q_1 และ Q_3 แทนควอร์ไทล์ที่ 1 และ 3 ตามลำดับ และ IQR แทนพิสัยระหว่างควอร์ไทล์ (Interquartile Range) (Barnett and Lewis, 1995)

ดังนั้นจึงเป็นที่มาของการศึกษาวิจัยในครั้งนี้ว่า หากข้อมูลมีการแจกแจงปกติที่มีตัวอย่างขนาดกลางและขนาดเล็ก และมีค่านอกเกณฑ์ปนเปื้อนด้วยระดับต่าง ๆ ตามเงื่อนไขของการตรวจสอบค่านอกเกณฑ์ของ Junsawang et al. (2021) ที่เป็นค่าที่อยู่นอกช่วง

$$\left(Q_1 - 1.5 e^{\left(\frac{SIQR_L}{SIQR_U}\right)\delta} IQR, Q_3 + 1.5 e^{\left(\frac{SIQR_L}{SIQR_U}\right)\delta} IQR \right)$$

เมื่อ $SIQR_L = Q_2 - Q_1$ และ $SIQR_U = Q_3 - Q_2$ โดยที่ $\delta = \frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$ ที่มี Q_1 Q_2 และ Q_3 แทนควอร์ไทล์ที่ 1 ถึง 3 ตามลำดับ และ IQR แทนพิสัยระหว่าง ควอร์ไทล์ แล้วประสิทธิภาพตัวประมาณค่าเบี่ยงเบนมาตรฐานประชากรวิธี MAD วิธี AR และวิธี ASD วิธีใดมีมากที่สุด เมื่อใช้ค่าความเอนเอียงสมบูรณ์

(Absolute Bias: ABS) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) ต่ำสุดเป็นเกณฑ์การพิจารณา

วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าเบี่ยงเบนมาตรฐานประชากร สำหรับข้อมูลที่มีการแจกแจงปกติและมีค่านอกเกณฑ์ จำนวน 4 วิธี ได้แก่ วิธีส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute Deviation: MAD) วิธีพิสัยที่ถูกปรับค่า (Adjusted Range: AR) วิธีส่วนเบี่ยงเบนมาตรฐานที่ถูกปรับค่า (Adjusted Standard Deviation: ASD) และวิธีส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง (Sample Standard Deviation: SD) โดยใช้ค่าความเอนเอียงสัมบูรณ์ (Absolute Bias: ABS) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) ต่ำสุดมาเป็นเกณฑ์การเปรียบเทียบ

ขอบเขตการวิจัย

การเปรียบเทียบประสิทธิภาพตัวประมาณค่าเบี่ยงเบนมาตรฐานของประชากร 4 วิธี ได้แก่ วิธี SD วิธี MAD วิธี AR และวิธี ASD จากสถานการณ์จำลอง 90 สถานการณ์ ที่ใช้ ABS และ MSE ต่ำที่สุดมาเป็นเกณฑ์การพิจารณา อยู่ภายใต้ขอบเขตของการวิจัยดังนี้

1. ทำการศึกษาการจำลองสถานการณ์ด้วยวิธีมอนติคาร์โล (Monte Carlo Method) จากโปรแกรม Java โดยการทดลองซ้ำ จำนวน 2,000 ครั้งในแต่ละสถานการณ์
2. กำหนดขนาดตัวอย่าง (n) จำนวน 2 ขนาดคือ ตัวอย่างขนาดเล็กโดย $n = 10, 20$ และ ตัวอย่างขนาดกลางโดย $n = 50, 70$ และ 100
3. กำหนดข้อมูลที่มีการแจกแจงปกติ ด้วยค่าเฉลี่ย (μ) เท่ากับ 30 และค่าเบี่ยงเบนมาตรฐานของประชากร (σ) 5 จำนวน ได้แก่ 1, 5, 10, 15 และ 20
4. กำหนดร้อยละของค่านอกเกณฑ์ 3 ระดับ ได้แก่ ร้อยละ 0, 10 และ 20

เอกสารและงานวิจัยที่เกี่ยวข้อง

ค่านอกเกณฑ์คือค่าสังเกตหรือข้อมูลที่แตกต่าง หรือเบี่ยงเบนไปจากค่าสังเกตอื่น ๆ มาก โดยค่านอกเกณฑ์อาจมีขนาดใหญ่หรือเล็กมากเมื่อเปรียบเทียบกับค่าสังเกตอื่นในชุดข้อมูล Beckman and Cook (1983) ได้แบ่งค่านอกเกณฑ์เป็น 2 ลักษณะ คือ 1) ค่านอกเกณฑ์ที่เป็น Discordant observations คือค่านอกเกณฑ์ที่มีค่าสูงหรือต่ำมาก หรือเป็นค่าที่เบี่ยงเบนไปจากค่าสังเกตส่วนใหญ่ และ 2) ค่านอกเกณฑ์ที่เป็น Contaminate observations เป็นค่านอกเกณฑ์ที่มีลักษณะการแจกแจงแตกต่างจากลักษณะการแจกแจงของประชากรที่สนใจศึกษา โดย Anscombe (1960) ได้สรุปสาเหตุสำคัญของการเกิดค่านอกเกณฑ์ จากความคลาดเคลื่อนที่เกิดขึ้น ประกอบด้วย (1) ความคลาดเคลื่อนจากความผันแปรของข้อมูลในประชากรที่ศึกษา (Inherent variability) ซึ่งเป็นความคลาดเคลื่อนที่ไม่ได้มีสาเหตุมาจากความผิดพลาดจากการวัดหรือการบันทึกข้อมูล ถือว่าเป็นความคลาดเคลื่อนที่เลี่ยงไม่ได้ เพราะไม่ว่าจะควบคุมการวัด หรือการบันทึกข้อมูลให้ดีเพียงใดความผันแปรในประชากรยังคงอยู่ นอกจากจะเปลี่ยนประชากร หรือวัตถุประสงค์ของการศึกษา (2) ความคลาดเคลื่อนที่เกิดจากการวัด (Measurement error) เป็นความคลาดเคลื่อน

จากการรวบรวมข้อมูลจากการวัดที่ผิดพลาดไม่ตรงกับสิ่งที่วัดได้จริง หรือการนำอุปกรณ์ เครื่องมือ เครื่องใช้คุณภาพต่ำมาวัดจึงทำให้ได้ผลการวัดหรือข้อมูลคลาดเคลื่อนได้ และ (3) ความคลาดเคลื่อนจากการปฏิบัติการ (Execution error) เป็นความคลาดเคลื่อนที่เกิดขึ้นจากความผิดพลาดของการบันทึกข้อมูล หรืออาจเกิดจากการเลือกตัวอย่างที่ไม่เหมาะสมกับลักษณะของประชากร สามารถลดความคลาดเคลื่อนนี้ลงได้ถ้าเพิ่มความระมัดระวังในการปฏิบัติการ

ทั้งนี้หากค่านอกเกณฑ์เป็นค่าสังเกตที่มีความสำคัญซึ่งมีอิทธิพลต่อค่าสังเกตอื่น ๆ หรือมีความน่าสนใจในตัวเองที่สามารถนำไปใช้เพื่อการวิเคราะห์ต่อเนื่อง การตัดค่าสังเกตส่วนนี้ออกอาจทำให้ขาดสารสนเทศบางอย่างไป แต่ถ้าค่านอกเกณะนั้นเกิดจากความคลาดเคลื่อนจากการวัด หรือเกิดจากความคลาดเคลื่อนจากการปฏิบัติ หรือเป็นหน่วยตัวอย่างที่ไม่ได้สุ่มมาจากประชากรเป้าหมาย สามารถตัดค่าสังเกตที่มีความผิดปกติออกไปได้ เพื่อให้ผลการวิเคราะห์ข้อมูลน่าเชื่อถือสามารถนำไปใช้ได้ ดังนั้นการตรวจสอบค่าสังเกตว่าเป็นค่านอกเกณฑ์หรือไม่ จึงถือเป็นเรื่องสำคัญที่ควรกระทำก่อนการวิเคราะห์ข้อมูลที่จะทำให้ผู้วิเคราะห์ข้อมูลเห็นลักษณะข้อมูลทั้งหมด และหาวิธีการแก้ไขปัญหาด้วยวิธีการอย่างหนึ่งอย่างใดให้เหมาะสม หากพบค่านอกเกณฑ์ที่ส่งผลหรือมีอิทธิพลต่อค่าสังเกตอื่น ๆ

นอกจากนี้หากพบค่านอกเกณฑ์ที่มีค่าความผิดปกติมาก จะเรียกค่านอกเกณฑ์ลักษณะนี้ว่า ค่าสุดขีด (Extreme value) ซึ่งสามารถสังเกตและหาวิธีการตรวจสอบได้ง่ายเพราะแสดงให้เห็นอย่างชัดเจน แต่ถ้าค่านอกเกณะนั้นมีค่าความผิดปกติเพียงเล็กน้อย (Mild value) ควรใช้วิธีการทางสถิติที่เหมาะสมมาตรวจสอบ โดยปัจจุบันมีวิธีการที่ใช้ตรวจสอบค่านอกเกณฑ์วิธีกราฟและวิธีการทดสอบสมมติฐาน ซึ่งการตรวจสอบค่านอกเกณฑ์วิธีกราฟเป็นการตรวจสอบที่ง่ายและสะดวก ทำได้โดยการตรวจสอบค่านอกเกณฑ์จากช่วงตัวเลขของแผนภาพกล่องซึ่งแนะนำโดย Tukey (1977) ซึ่งค่าสังเกตที่ตรวจสอบนั้นจะเป็นค่านอกเกณฑ์เมื่ออยู่นอกช่วง ($Q_3 + 1.5(IQR)$, $Q_3 + 3(IQR)$) เมื่อ Q_1 และ Q_3 แทนควอนไทล์ที่ 1 และ 3 ตามลำดับ และ IQR แทนพิสัยระหว่างควอร์ไทล์ (Interquartile Range) (Barnett and Lewis, 1995)

Boonyavaha et al. (2018) ได้เปรียบเทียบวิธีการประมาณค่าเบี่ยงเบนมาตรฐานประชากรโดยใช้เกณฑ์ของ Tukey (1977) ที่กำหนดให้ข้อมูลมีค่านอกเกณฑ์ในระดับปานกลาง ตามเงื่อนไขของการตรวจสอบค่านอกเกณฑ์โดยใช้แผนภาพกล่อง และหมวดที่เป็นค่าที่อยู่นอกช่วง ($Q_3 + 1.5(IQR)$, $Q_3 + 3(IQR)$) พบว่า วิธี ASD มีประสิทธิภาพมากกว่าวิธี AR เมื่อข้อมูลแจกแจงปกติในตัวอย่างขนาดกลางและเล็ก และไม่มีค่านอกเกณืที่ปนเปื้อน โดยใช้ ABS เป็นเกณฑ์การพิจารณา

Junsawang et al. (2021) ได้พัฒนาวิธีการตรวจสอบค่านอกเกณฑ์เรียกว่า แผนภาพกล่อง MH ซึ่งเป็นการพัฒนาจากแผนภาพกล่องของ Hubert and Vandervieren (2008) สำหรับข้อมูลที่แจกแจงแบบเบ้ (Skewed Distribution) พบว่า ผลการตรวจสอบค่านอกเกณฑ์มีประสิทธิภาพหลายกรณีภายใต้สถานการณ์จำลองและจากข้อมูลจริง

วิธีดำเนินการวิจัย

1. สร้างข้อมูลเพื่อนำมาใช้ในการจำลองสถานการณ์ เป็นการสร้างข้อมูลแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 30 และค่าเบี่ยงเบนมาตรฐาน 5 จำนวน ได้แก่ 1, 5, 10, 15 และ 20
2. สร้างค่านอกเกณฑ์ร้อยละ 0, 10 และ 20 ตามเงื่อนไขของการตรวจสอบค่านอกเกณฑ์ของ Junsawang et al. (2021)

3. การตรวจสอบค่า outliers โดยวิธีกราฟซึ่งเป็นการตรวจสอบด้วยแผนภาพกล่อง (Box Plot) โดยค่าสังเกตที่มีค่า outliers ในระดับปานกลางทางด้านมาก จะมีค่ามากกว่า Q_3 ระหว่าง 1.5 ถึง 3 เท่าของ $e^{\left(\frac{SIQR_L}{SIQR_U}\right)\delta} IQR$

4. ให้ $x_1, x_2, x_3, \dots, x_n$ เป็นตัวอย่างสุ่มขนาด n เมื่อ $n = 10, 20, 30, 50, 70$ และ 100 จากประชากรที่มีการแจกแจงแบบปกติ ที่มีค่าเฉลี่ยเท่ากับ 30 และค่าเบี่ยงเบนมาตรฐานเท่ากับ 1, 5, 10, 15 และ 20

5. ประมาณค่าเบี่ยงเบนมาตรฐานของประชากรวิธี SD วิธี MAD วิธี AR และวิธี ASD โดยแต่ละวิธีมีสูตรที่ใช้ในการคำนวณ ดังนี้

5.1 วิธี SD หรือวิธีส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง แทนด้วย $\hat{\sigma}_{SD}$ มีสูตรที่ใช้คำนวณตามสมการที่ 1

$$\hat{\sigma}_{SD} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (1)$$

โดยที่ $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ แทนค่าเฉลี่ยของตัวอย่าง และ n แทนขนาดตัวอย่าง

5.2 วิธี MAD หรือวิธีค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย แทนด้วย $\hat{\sigma}_{MAD}$ มีสูตรที่ใช้คำนวณตามสมการที่ 2

$$\hat{\sigma}_{MAD} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} \quad (2)$$

โดยที่ $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ แทนค่าเฉลี่ยของตัวอย่าง และ n แทนขนาดตัวอย่าง

5.3 วิธี AR เป็นการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรโดยใช้พิสัยของตัวอย่างมาปรับค่าเพื่อให้ได้ตัวประมาณค่าที่ไม่เอนเอียง แทนด้วย $\hat{\sigma}_{AR}$ มีสูตรที่ใช้คำนวณตามสมการที่ 3

$$\hat{\sigma}_{AR} = \frac{R}{d_2} \quad (3)$$

โดยที่ $R = x_{max} - x_{min}$ แทนพิสัย (Range) ของตัวอย่างที่เป็นผลต่างของข้อมูลที่มีค่ามากที่สุด (x_{max}) กับข้อมูลที่มีค่าน้อยที่สุด (x_{min}) และ d_2 คือ ค่าคงที่ตามตารางที่ 1 ซึ่งขึ้นกับค่า n เมื่อ $n = 1, 2, 3, \dots, 50$ และ $d_2 = 3.4873 + 0.0250141(n) - 0.00009823(n^2)$ เมื่อ $n = 51, 52, 53, \dots, 100$

5.4 วิธี ASD เป็นการประมาณค่าเบี่ยงเบนมาตรฐานของประชากร โดยใช้ค่าเบี่ยงเบนมาตรฐานของตัวอย่างมาปรับค่าเพื่อให้ได้ตัวประมาณค่าที่ไม่เอนเอียง แทนด้วย $\hat{\sigma}_{ASD}$ มีสูตรที่ใช้คำนวณตามสมการที่ 4

$$\hat{\sigma}_{ASD} = \frac{\hat{\sigma}_{SD}}{c_4} \quad (4)$$

โดยที่ $\hat{\sigma}_{SD}$ เป็นค่าเบี่ยงเบนมาตรฐานของตัวอย่าง และ c_4 คือค่าคงที่ตามตารางที่ 1 ซึ่งขึ้นอยู่กับค่า n เมื่อ $n = 1, 2, 3, \dots, 25$ และ $c_4 = \sqrt{\frac{2}{n-1}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}$ เมื่อ $n = 26, 27, 28, \dots, 100$ เมื่อ $\Gamma(n)$ เป็นฟังก์ชันแกมมา (Gamma Function) ซึ่ง $\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx$

6. เปรียบเทียบประสิทธิภาพของตัวประมาณค่าพารามิเตอร์จากค่า ABS หรือ $\widehat{ABS}$ และ MSE หรือ $\widehat{MSE}$ ที่มีสูตรที่ใช้คำนวณ ตามสมการที่ 5 และ 6 ตามลำดับ

$$\widehat{ABS} = \left| \frac{\sum_{k=1}^m (\hat{\sigma})_k}{m} - \sigma \right| \quad (5)$$

$$\widehat{MSE} = \frac{1}{m} \sum_{k=1}^m [(\hat{\sigma})_k - \sigma] \quad (6)$$

เมื่อ $(\hat{\sigma})_k$ แทนตัวประมาณค่าเบี่ยงเบนมาตรฐานของประชากร (σ) ในแต่ละวิธีภายใต้สถานการณ์ต่าง ๆ จากการทำซ้ำในรอบที่ k และ m แทนจำนวนรอบที่ทำซ้ำในที่นี้กำหนดให้ $m = 2,000$

ตารางที่ 1 ค่าคงที่สำหรับคำนวณค่า $\hat{\sigma}_{AR}$ และ $\hat{\sigma}_{ASD}$ (Leon, 1960)

n	d_2	c_4	n	d_2
2	1.128	0.7979	26	3.964
3	1.693	0.8862	27	3.997
4	2.059	0.9213	28	4.027
5	2.326	0.9400	29	4.057
6	2.534	0.9515	30	4.086
7	2.704	0.9594	31	4.113
8	2.847	0.9650	32	4.139
9	2.970	0.9693	33	4.165
10	3.078	0.9727	34	4.189
11	3.173	0.9754	35	4.213
12	3.258	0.9776	36	4.236
13	3.336	0.9794	37	4.259
14	3.407	0.9810	38	4.280
15	3.472	0.9823	39	4.301
16	3.532	0.9835	40	4.322
17	3.588	0.9845	41	4.341
18	3.640	0.9854	42	4.361
19	3.689	0.9862	43	4.379
20	3.735	0.9869	44	4.398
21	3.778	0.9876	45	4.415
22	3.819	0.9882	46	4.433
23	3.858	0.9887	47	4.450
24	3.895	0.9892	48	4.466
25	3.931	0.9896	49	4.482
			50	4.498

ผลการวิจัย

งานวิจัยนี้ได้เปรียบเทียบประสิทธิภาพวิธีการประมาณค่าเบี่ยงเบนมาตรฐานของประชากร สำหรับข้อมูลที่แจกแจงปกติ เมื่อข้อมูลมีค่านอกเกณฑ์ตามเงื่อนไขของการตรวจสอบค่านอกเกณฑ์ของ Junsawang et al. (2021) ทั้งหมด 4 วิธี ได้แก่ วิธีส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute Deviation: MAD) วิธีพิสัยที่ถูกปรับค่า (Adjusted Range: AR) วิธีส่วนเบี่ยงเบนมาตรฐานที่ถูกปรับค่า (Adjusted Standard Deviation: ASD) และวิธีส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง (Sample Standard Deviation: SD) โดยใช้ค่าความเอนเอียงสัมบูรณ์ (Absolute Bias: ABS) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) ต่ำสุดมาเป็นเกณฑ์การเปรียบเทียบภายใต้สถานการณ์จำลอง 90 สถานการณ์ ผลการเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรจากค่า ABS และ MSE เมื่อจำแนกตามร้อยละของค่านอกเกณฑ์ตั้งแต่ร้อยละ 0, 10 และ 20 แสดงได้ตามตารางที่ 2 ถึง 4 ตามลำดับดังนี้

ตารางที่ 2 ค่า ABS และ MSE ของวิธีประมาณค่าเบี่ยงเบนมาตรฐานของประชากร เมื่อไม่มีค่านอกเกณฑ์

<i>n</i>	σ	MAD		AR		ASD		SD	
		ABS	MSE	ABS	MSE	ABS	MSE	ABS	MSE
10	1	0.201502	0.062646	0.052268	0.049026	0.054078	0.036514	0.025301*	0.032421*
	5	1.174823	2.356909	0.039723	1.862444	0.123425	1.492634	0.013442*	1.427311*
	10	2.260362	8.275179	0.170957	5.547785	0.152379	5.24786	0.124781*	4.958839*
	15	2.657152	14.03503	0.288480	13.05941	0.717569	11.71952	0.183751*	10.68441*
	20	3.844282	33.21253	0.467768	21.86449	0.971260	19.95344	0.398744*	16.16658*
20	1	0.312613	0.114897	0.033471	0.051069	0.020641	0.031869	0.015042*	0.031745*
	5	0.993117	1.379923	0.147686	0.911122	0.076578	0.475432	0.010075*	0.457447*
	10	1.981525	6.588453	0.038007	4.512069	0.114178	3.868568	0.018318*	3.755514*
	15	3.309417	15.51159	1.043254	8.733755	0.80476	5.813431	0.616334*	5.683710*
	20	4.482087	28.27937	0.096081	17.94554	0.280467	12.0429	0.018712*	11.80776*
30	1	0.15687	0.032674	0.067711	0.02577	0.074840	0.01758	0.065617*	0.016080*
	5	1.052256	1.393684	0.115280	0.664717	0.068826	0.405273	0.025335*	0.394334*
	10	2.043747	5.213823	0.536362	2.516431	0.31349	1.65466	0.225000*	1.580417*
	15	2.43897	8.862411	1.145612	9.832943	0.917889	5.190306	0.781314*	4.883948*
	20	4.45125	23.06773	0.295782	18.29941	0.236148	7.261905	0.065107*	7.189425*
50	1	0.20721	0.054339	0.033048	0.02116	0.005549	0.013744	0.000431*	0.013574*
	5	1.137296	1.542947	0.405377	0.578064	0.221003	0.322684	0.196558*	0.315285*
	10	2.044377	5.030126	0.104102	2.969857	0.048757	1.240095	0.002144*	1.229878*
	15	3.259442	11.95164	0.512326	2.423303	0.287152	1.389017	0.211895*	1.364860*
	20	4.472556	22.90766	0.889105	6.720074	0.682313	4.388334	0.583502*	4.303491*
70	1	0.208251	0.047200	0.015301	0.009698	0.005551	0.005076	0.001941*	0.005067*
	5	0.96891	1.119261	0.081221	0.343326	0.085026	0.238182	0.066634*	0.233725*
	10	2.01331	4.593502	0.321739	1.56751	0.055226	0.852115	0.018856*	0.843290*
	15	2.640239	7.8665	0.47241	1.988288	0.528577	1.461678	0.418816*	1.396919*
	20	4.722874	23.80829	0.678648	6.757673	0.608509	3.029224	0.272420*	2.957627*
100	1	0.186442	0.037579	0.058099	0.016191	0.026988	0.004665	0.024398*	0.004512*
	5	1.158063	1.410282	0.269272	0.218343	0.164591	0.120287	0.152365*	0.116884*
	10	2.054227	4.436751	0.222418	1.790688	0.016772	0.260573	0.008491*	0.259053*
	15	3.156227	10.75941	0.131882	1.234199	0.055185	1.032138	0.017399*	1.029683*
	20	4.187752	18.66303	0.337925	4.482823	0.054671	1.722325	0.004242*	1.716619*

* หมายถึง ค่า ABS หรือ MSE ที่มีค่าน้อยที่สุดเมื่อเปรียบเทียบในสถานการณ์เดียวกัน

จากผลการวิจัยที่แสดงได้ตามตารางที่ 2 เมื่อไม่มีค่านอกเกณฑ์ปนเปื้อนอยู่พบว่า ค่า ABS และ MSE ของการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรวิธี SD ให้ค่าน้อยที่สุด แสดงว่าวิธี SD เป็นวิธีการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรที่มีประสิทธิภาพมากที่สุด เมื่อไม่มีค่านอกเกณฑ์ปนเปื้อนอยู่ ทั้งนี้ยังพบว่า ค่า ABS และ MSE ของการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรวิธี ASD มีค่าใกล้เคียงกับค่า ABS และ MSE ของ SD ทุกสถานการณ์ ดังนั้นหากเปรียบเทียบค่า ABS และ MSE ของการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรวิธี ASD กับวิธี AR พบว่าค่า ABS และ MSE ของวิธี ASD น้อยกว่าวิธี AR ทุกสถานการณ์ แสดงว่าวิธี ASD มีประสิทธิภาพมากกว่าวิธี AR เมื่อไม่มีค่านอกเกณฑ์ปนเปื้อนอยู่

ทั้งนี้เมื่อเปรียบเทียบค่า MSE ของวิธี SD จากค่า σ เดียวกันพบว่า MSE จะมีแนวโน้มลดลงถ้าขนาดตัวอย่างใหญ่ขึ้น แต่เมื่อเปรียบเทียบค่า MSE จากขนาดตัวอย่างเดียวกันกลับพบว่า MSE มีแนวโน้มเพิ่มขึ้นเมื่อค่า σ มากขึ้น

ตารางที่ 3 ค่า ABS และ MSE ของวิธีประมาณค่าเบี่ยงเบนมาตรฐานของประชากร เมื่อมีค่านอกเกณฑ์ร้อยละ 10

n	σ	MAD		AR		ASD		SD	
		ABS	MSE	ABS	MSE	ABS	MSE	ABS	MSE
10	1	0.444397*	2.127662*	1.749383	3.174160	1.477691	2.330928	1.410050	0.267907
	5	0.988626*	1.111642*	2.577096	6.807672	2.724237	7.561919	2.513365	6.449893
	10	0.642392*	0.840333*	2.773871	9.330638	2.414861	6.508522	2.075935	4.950018
	15	1.549462*	3.155401*	8.584624	74.72702	7.140208	51.57871	6.535781	43.28046
	20	2.771389*	9.571256*	4.901272	25.31493	3.629561	13.87569	2.984474	9.775137
20	1	0.089515*	0.028835*	0.767979	0.590156	0.652089	0.425693	0.630446	0.397923
	5	0.707295*	0.721337*	2.31273	5.359051	2.383519	5.733604	2.286795	5.280507
	10	0.279923*	0.312861*	2.330623	5.550265	2.388828	5.926786	2.226535	5.172008
	15	2.457240*	6.947215*	9.745333	94.97799	10.12713	102.7246	9.797967	96.16163
	20	1.580740*	3.640673*	7.261596	54.35934	7.553497	58.22723	7.192546	52.87413
30	1	0.303192*	0.093850*	0.756287	0.574815	0.740172	0.549285	0.725242	0.527381
	5	0.526209*	0.343166*	2.191995	4.867157	2.349837	5.615005	2.286775	5.32102
	10	0.406183*	0.238097*	2.897679	8.396545	3.570791	12.79308	3.454353	11.97436
	15	1.699725*	3.283421*	5.637142	35.11158	6.696445	45.38420	6.510289	42.91644
	20	4.090650*	17.60610*	15.01807	228.8105	12.99938	171.0805	12.71625	163.7637
50	1	0.381946*	0.146364*	1.334045	1.779947	1.083726	1.174899	1.073122	1.152023
	5	0.651423*	0.432787*	2.49141	6.207126	2.639583	6.972845	2.600705	6.769059
	10	1.314825*	1.749506*	4.232156	17.91114	4.714714	22.24527	4.639831	21.54460
	15	1.164829*	1.534439*	4.686284	22.7673	5.790229	33.67447	5.684428	32.45894
	20	0.381977*	0.291108*	0.814319	0.663115	5.040211	25.50171	4.912782	24.23241
70	1	0.614337*	0.377736*	1.688567	2.851257	1.639981	2.689609	1.630432	2.65838
	5	0.006257*	0.004235*	1.556912	2.423975	1.781566	3.174763	1.757037	3.08796
	10	3.373077*	11.39291*	6.314935	40.12765	8.705506	75.80060	8.637848	74.62708
	15	3.149500*	10.02646*	5.208692	27.13047	8.571543	73.67908	8.486285	72.22325
	20	2.266628*	5.220576*	6.192145	38.34266	9.262085	85.84158	9.156244	83.89177
100	1	0.661654*	0.437955*	2.108481	4.450268	1.867976	3.489469	1.860743	3.462498
	5	0.603495*	0.371946*	2.532627	6.414197	2.669802	7.146637	2.650458	7.043632
	10	1.626508*	2.666338*	4.190658	17.56162	5.395988	29.13933	5.357159	28.72168
	15	0.770300*	0.593362*	0.967492	0.982274	3.023111	9.231047	2.977657	8.957824
	20	1.061042*	1.306670*	7.021769	49.30524	8.510116	72.51162	8.438213	71.29254

* หมายถึง ค่า ABS หรือ MSE ที่มีค่าน้อยที่สุดเมื่อเปรียบเทียบในสถานการณ์เดียวกัน

ตารางที่ 4 ค่า ABS และ MSE ของวิธีการประมาณค่าเบี่ยงเบนมาตรฐานของประชากร เมื่อมีค่านอกเกณฑ์ร้อยละ 20

n	σ	MAD		AR		ASD		SD	
		ABS	MSE	ABS	MSE	ABS	MSE	ABS	MSE
10	1	1.343864*	2.089195*	2.058948	4.360126	2.386109	5.982085	2.293669	5.533942
	5	1.874668*	3.664232*	3.176143	10.64950	3.768203	14.62657	3.528831	12.85686
	10	0.465381*	0.768361*	2.797927	9.909936	3.549755	13.64744	3.179847	11.10173
	15	4.556804*	23.28023*	8.807850	78.62740	11.33656	129.9771	10.61757	114.1137
	20	1.094687*	3.322667*	4.914545	26.15912	7.406625	56.97665	6.658424	46.33907
20	1	0.981474*	1.109212*	1.297091	1.683094	1.527898	2.335578	1.494782	2.235451
	5	2.345927*	6.736973*	2.814325	7.938792	3.960050	15.77026	3.842673	14.85210
	10	2.267185*	6.621373*	2.641098	7.246952	4.193006	18.42974	4.007078	16.88303
	15	10.17109*	105.1563*	10.25401	112.6014	16.42048	270.0728	16.00887	256.7131
	20	6.490676*	45.64218*	7.371689	57.23700	11.01266	123.7604	10.60640	114.9127
30	1	0.980055*	0.962652*	1.502738	2.258222	1.661890	2.763972	1.639051	2.688546
	5	1.880359*	3.704061*	2.821344	8.062617	3.724777	14.07320	3.649918	13.51774
	10	3.076199*	9.583497*	3.631894	13.19065	6.409009	41.15816	6.268220	39.37192
	15	4.277095*	18.63701*	5.604291	36.89972	9.137256	84.07990	8.930159	80.32810
	20	8.846079*	79.41772*	14.99285	230.1684	16.89849	288.4307	16.58190	277.7821
50	1	1.653218*	2.733545*	2.436165	5.935262	2.711194	7.351424	2.692308	7.249363
	5	2.835454*	8.049767*	3.603016	12.98172	5.077016	25.78039	5.025735	25.26226
	10	0.199407*	15.35231*	1.132545	27.42976	2.570234	57.36437	2.480819	56.01801
	15	4.454730*	19.96952*	5.049724	27.53721	9.482801	90.07315	9.358208	87.72417
	20	1.925924*	3.709182*	4.395854	19.51267	9.169411	84.17417	9.020968	81.47296
70	1	2.648619*	7.015227*	3.160095	9.986198	4.028343	16.22765	4.010155	16.08145
	5	2.346084*	5.511671*	2.991009	8.972757	4.507709	20.32689	4.473320	20.01799
	10	7.629458*	58.48252*	9.184180	84.36446	14.71317	216.4873	14.62378	213.8649
	15	6.680220*	44.62534*	7.716670	59.82913	13.23612	175.4930	13.13399	172.7977
	20	7.663673*	58.73189*	8.310151	69.10354	15.43290	238.2288	15.30474	234.2891
100	1	3.315719*	10.99406*	4.072133	16.58570	5.085361	25.86108	5.070014	25.70522
	5	3.464956*	12.02048*	4.377454	19.24284	6.166283	38.04117	6.138122	37.69457
	10	6.188130*	38.29295*	6.327860	40.05336	10.47644	109.7674	10.42480	108.6880
	15	2.603490*	6.844426*	2.767771	7.660556	6.681489	44.69195	6.626808	43.9640
	20	5.775657*	33.49227*	9.019240	81.34669	13.93637	194.3164	13.85078	191.9377

* หมายถึง ค่า ABS หรือ MSE ที่มีค่าน้อยที่สุดเมื่อเปรียบเทียบในสถานการณ์เดียวกัน

จากผลการวิจัยที่แสดงได้ตามตารางที่ 3 และ 4 เมื่อมีค่านอกเกณฑ์ปนเปื้อนอยู่ร้อยละ 10 และ 20 ตามลำดับพบว่า ค่า ABS และ MSE ของการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรวิธี MAD ให้ค่าน้อยที่สุดทุกสถานการณ์ และมีความแตกต่างจากวิธีอื่น ๆ ค่อนข้างมาก แสดงว่าวิธี MAD เป็นวิธีการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรที่มีประสิทธิภาพมากที่สุดเมื่อมีค่านอกเกณฑ์ปนเปื้อนอยู่ร้อยละ 10 และ 20

เมื่อพิจารณาข้อมูลขนาดเล็กที่รวบรวมโดย Christos (2020) ดังนี้

12.29	8.88	7.27	12.51	11.33	12.04	13.94	7.61	12.57	10.07
12.22	9.42	7.84	10.72	9.75	8.86	9.88	11.39	9.98	10.09
10.86	9.78	8.1	11.98	7.71	10.62	12.03	18	11.13	20

จากการตรวจสอบค่านอกเกณฑ์ตามเงื่อนไขของ Junsawang et al. (2021) พบว่ามีค่านอกเกณฑ์ 2 จำนวน คือ 18 และ 20 โดยคำนวณค่าประมาณของส่วนเบี่ยงเบนมาตรฐานประชากรพบว่า MAD มีค่าน้อยที่สุดเท่ากับ 2.32 เมื่อเปรียบเทียบกับค่า AR เท่ากับ 3.11 ค่า ASD เท่ากับ 2.79 และค่า SD เท่ากับ 2.77

นอกจากนี้เมื่อพิจารณาข้อมูลขนาดกลางรวบรวมโดย NIST/SEMATECH (2012) ดังนี้

30	171	184	201	212	250	265	270	272	289
305	306	322	322	336	346	351	370	390	404
409	411	436	437	439	441	444	448	451	453
470	480	482	487	494	495	499	503	514	521
522	527	548	550	559	560	570	572	574	578
585	592	592	607	616	618	621	629	637	638
640	656	668	707	709	719	737	739	752	758
766	792	792	794	802	818	830	832	843	858
860	869	918	925	953	991	1000	1005	1068	1441

เมื่อใช้เงื่อนไขของ Junsawang et al. (2021) พบว่ามีค่านอกเกณฑ์ 1 จำนวน คือ 1441 โดยคำนวณค่าประมาณส่วนเบี่ยงเบนมาตรฐานของประชากร พบว่า MAD มีค่าน้อยที่สุดเท่ากับ 166.72 เมื่อเปรียบเทียบกับค่า AR เท่ากับ 281.84 ค่า ASD เท่ากับ 238.20 และค่า SD เท่ากับ 237.60

ทั้งนี้เมื่อเปรียบเทียบค่า MSE จากค่า σ เดียวกันแล้ว พบว่า MSE ของ MAD จากตารางที่ 3 และ 4 ไม่ได้มีแนวโน้มลดลงถ้าขนาดตัวอย่างใหญ่ขึ้นเช่นเดียวกับ MSE ของ SD จากตารางที่ 2 และไม่ได้มีแนวโน้มเพิ่มขึ้นเมื่อค่า σ มากขึ้นเช่นเดียวกัน

สรุปผลการวิจัย

งานวิจัยนี้ได้เปรียบเทียบประสิทธิภาพวิธีการประมาณค่าเบี่ยงเบนมาตรฐานของประชากร 4 วิธี ได้แก่ วิธีส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute Deviation: MAD) วิธีพิสัยที่ถูกปรับค่า (Adjusted Range: AR) วิธีส่วนเบี่ยงเบนมาตรฐานที่ถูกปรับค่า (Adjusted Standard Deviation: ASD) และวิธีส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง (Sample Standard Deviation: SD) โดยใช้ค่าความเอนเอียงสัมบูรณ์ (Absolute Bias: ABS) สำหรับข้อมูลที่แจกแจงปกติ เมื่อข้อมูลมีค่านอกเกณฑ์ตามเงื่อนไขของการตรวจสอบค่านอกเกณฑ์ของ Junsawang et al. (2021) คือค่าที่อยู่นอกช่วง

$$\left(Q_1 - 1.5 e^{\left(\frac{SIQR_L}{SIQR_U}\right)\delta} IQR, Q_3 + 1.5 e^{\left(\frac{SIQR_L}{SIQR_U}\right)\delta} IQR \right)$$

เมื่อ $SIQR_L = Q_2 - Q_1$ และ $SIQR_U = Q_3 - Q_2$ โดยที่ $\delta = \frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$ ที่มี Q_1 Q_2 และ Q_3 แทนควอนไทล์ที่ 1 ถึง 3 ตามลำดับ และ IQR แทนพิสัยระหว่างควอร์ไทล์ เมื่อใช้ค่าความเอนเอียงสัมบูรณ์ (ABS) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ต่ำสุดมาเป็นเกณฑ์การพิจารณา ภายใต้สถานการณ์จำลอง 90 สถานการณ์ ที่เป็นสถานการณ์ของข้อมูลที่แจกแจงปกติ มีค่าเฉลี่ย (μ) เท่ากับ 30 และค่าเบี่ยงเบนมาตรฐานประชากร (σ) 5 จำนวน ได้แก่ 1, 5, 10, 15 และ 20 ขนาดตัวอย่าง (n) 6 จำนวน ได้แก่ 10, 20, 30, 50, 70 และ 100 ร้อยละของค่าผิดปกติ 3 จำนวน ได้แก่ ร้อยละ 0, 10 และ 20 พบว่า เมื่อไม่มีค่าผิดปกติปนเปื้อนอยู่ วิธี SD เป็นวิธีการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรที่มีประสิทธิภาพมากที่สุด ซึ่งมีประสิทธิภาพมากขึ้นถ้าขนาดตัวอย่างใหญ่ขึ้น แต่ประสิทธิภาพจะลดลงเมื่อค่า σ เพิ่มมากขึ้น และพบว่าวิธี ASD มีประสิทธิภาพมากกว่าวิธี AR และสำหรับกรณีที่มีค่าผิดปกติร้อยละ 10 และ 20 พบว่าวิธี MAD มีประสิทธิภาพมากที่สุด

อภิปรายผล

จากผลการวิจัยที่พบว่า เมื่อไม่มีค่าผิดปกติปนเปื้อนอยู่ ค่า ABS และ MSE ของการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรวิธี SD ให้ค่าน้อยที่สุด ซึ่งแสดงว่าวิธี SD เป็นวิธีการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรที่มีประสิทธิภาพมากที่สุด เมื่อไม่มีค่าผิดปกติปนเปื้อนอยู่ แต่หากมีค่าผิดปกติปนเปื้อนอยู่วิธี SD จะไม่ใช่วิธีประมาณค่าเบี่ยงเบนมาตรฐานของประชากรที่ดีที่สุดแต่วิธี MAD จะเป็นวิธีวิธีการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรที่มีประสิทธิภาพมากที่สุด สอดคล้องกับ Leys et al. (2013) ที่ได้อธิบายว่า หากมีค่าผิดปกติเกิดขึ้นกับข้อมูลที่แจกแจงปกติไม่ควรใช้การวัดแนวโน้มเข้าสู่กลางจากค่าเฉลี่ย และวัดการกระจายของข้อมูลจากค่าเบี่ยงเบนมาตรฐาน เพราะค่าทั้งสองนี้มีผลต่อข้อมูลที่เป็นค่าผิดปกติอย่างมาก (Strongly Impact of Outlier) แต่ควรใช้การหาค่าสัมบูรณ์ (Absolute) ที่เบี่ยงเบนออกจากค่ามัธยฐาน ซึ่งในการวิจัยนี้ใช้การหาค่า MAD ซึ่งเป็นการหาค่าสัมบูรณ์จากการเบี่ยงเบนของข้อมูลกับค่าเฉลี่ย ภายใต้ค่าผิดปกติตามเงื่อนไขของการตรวจสอบค่าผิดปกติของ Junsawang et al. (2021) ที่แสดงได้อย่างชัดเจนว่ามีประสิทธิภาพมากกว่า SD แม้จะเป็นการวัดการเบี่ยงเบนของข้อมูลจากค่าเฉลี่ย

ทั้งนี้จากงานวิจัยพบว่า หากข้อมูลแจกแจงปกติและไม่มีค่าผิดปกติปนเปื้อนอยู่ ค่า ABS และ MSE ของการประมาณค่าเบี่ยงเบนมาตรฐานของประชากรวิธี ASD มีค่าใกล้เคียงกับค่า ABS และ MSE ของ SD ซึ่งเป็นค่าน้อยกว่าค่า ABS และ MSE วิธี AR ทุกสถานการณ์ แสดงว่าวิธี ASD มีประสิทธิภาพมากกว่าวิธี AR สอดคล้องกับงานวิจัยของ Boonyavaha et al. (2018) ที่พบว่า วิธี ASD มีประสิทธิภาพมากกว่าวิธี AR เมื่อข้อมูลแจกแจงปกติในตัวอย่างขนาดกลางและเล็ก และไม่มีค่าผิดปกติปนเปื้อน โดยใช้ค่าความเอนเอียงสัมบูรณ์ (Absolute Bias: ABS) มาเป็นเกณฑ์การพิจารณา ที่กำหนดให้ข้อมูลมีค่าผิดปกติในระดับปานกลาง ตามเงื่อนไขของการตรวจสอบค่าผิดปกติโดยใช้แผนภาพกล่องและหนวด ที่เป็นค่าที่อยู่นอกช่วง ($Q_3 + 1.5(IQR)$, $Q_3 + 3(IQR)$) เมื่อ Q_1 และ Q_3 แทนควอนไทล์ที่ 1 และ 3 ตามลำดับ และ IQR แทนพิสัยระหว่างควอร์ไทล์ (Interquartile Range)

ทั้งนี้เมื่อเปรียบเทียบค่า MSE ของวิธี SD จากค่า σ เดียวกันในกรณีที่ไม่มีค่านอกเกณฑ์ พบว่า MSE จะมีแนวโน้มลดลงถ้าขนาดตัวอย่างใหญ่ขึ้น ซึ่งสอดคล้องกับงานวิจัยของ Montgomery (2012) ว่าวิธี SD จะมีประสิทธิภาพต่ำเมื่อใช้กับตัวอย่างขนาดเล็ก

ข้อเสนอแนะ

1. ข้อเสนอแนะในการนำผลวิจัยไปใช้

การวิจัยนี้เป็นการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าเบี่ยงเบนมาตรฐาน สำหรับข้อมูลขนาดกลางและขนาดเล็กที่แจกแจงปกติโดยวิธีกราฟ ซึ่งเป็นการตรวจสอบโดยใช้แผนภาพกล่องตามเงื่อนไขการตรวจสอบค่านอกเกณฑ์วิธีของ Junsawang et al. (2021) ซึ่งเป็นค่าสังเกตที่มีค่านอกเกณฑ์ในระดับปานกลาง ซึ่งค่าสังเกตนั้นจะเป็นค่านอกเกณฑ์ถ้ามีค่ามากกว่า Q_3 ระหว่าง 1.5 ถึง 3 เท่าของ $e^{(\frac{SIQR_L}{SIQR_U})\delta} IQR$ ดังนั้น หากมีการนำไปใช้ต้องระบุเกณฑ์ที่นำมาตรวจสอบ ซึ่งผลที่ได้อาจมีความแตกต่างจากการใช้เกณฑ์ที่อยู่นอกช่วง $(Q_3 + 1.5(IQR), Q_3 + 3(IQR))$ ของ Barnett and Lewis (1995) ที่ใช้กันในปัจจุบัน นอกจากนี้ยังเป็นการเปรียบเทียบภายใต้สถานการณ์จำลองจากการสร้างข้อมูลที่แจกแจงปกติ ที่มีค่าเฉลี่ย (μ) เท่ากับ 30 และค่าเบี่ยงเบนมาตรฐานประชากร (σ) ต่าง ๆ ในตัวอย่างขนาดกลางและเล็กเท่านั้น ซึ่งอาจจะให้ผลแตกต่างหากอยู่ภายใต้เงื่อนไข หรือสถานการณ์ที่มีการแจกแจงปกติที่มีค่าเฉลี่ยอื่น หรือลักษณะการแจกแจงของข้อมูลแบบอื่น ๆ

2. ข้อเสนอแนะในการวิจัยครั้งต่อไป

2.1 ควรมีการเปรียบเทียบประสิทธิภาพตัวประมาณค่าเบี่ยงเบนมาตรฐานของประชากร สำหรับข้อมูลขนาดกลางและขนาดเล็กที่แจกแจงปกติ เมื่อกำหนดเงื่อนไขค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐานของประชากร ต่าง ๆ มากขึ้น หรือเพิ่มกรณีที่มีข้อมูลขนาดใหญ่ เพื่อหาวิธีการที่เหมาะสมที่สุดในประมาณค่าเบี่ยงเบนมาตรฐานของประชากร เมื่อข้อมูลแจกแจงปกติ

2.2 งานวิจัยนี้ได้เปรียบเทียบประสิทธิภาพตัวประมาณค่าเบี่ยงเบนมาตรฐานของประชากร จากกรณีข้อมูลแจกแจงปกติ ดังนั้นอาจพิจารณาเปรียบเทียบประสิทธิภาพจากข้อมูลที่มีการแจกแจงแบบอื่นที่มีลักษณะเบ้ เช่น การแจกแจงแกมมา (Gamma Distribution) หรือการแจกแจงไวบูล (Weibull Distribution) โดยอาจเพิ่มวิธีการประมาณค่าเบี่ยงเบนมาตรฐานของประชากร เช่น Median Absolute Deviation

เอกสารอ้างอิง

- Aggarwal, C. (2016). *An Introduction to Outlier Analysis*. New York: Springer, Cham.
- Anscombe, F. J. (1960). Rejection of Outliers. *American Statistical Association and American Society for Quality*, 2(2), 123-147.
- Barnett, V., and Lewis, T. (1995). *Outlier in Statistical Data*. 3rd ed. New York: John Wiley.
- Beckman, R. J., and Cook, R. D. (1983). Outliers. *American Statistical Association and American Society for Quality*, 2(25), 119-149.

- Boonyavaha, P., Pratchayalamlert, N., Boonpet, B., and Sinsomboonthong, J. (2018). Efficiency Comparison of the Standard Deviation Estimation Methods for a Normal Distribution When Data Set Containing Outliers. *Science and Technology*, 27(2), 213-225. (in Thai)
- Chelishchev, P., Popov, A., and Sorby, K. (2018). An investigation of Outlier Detection Procedures for CMM Measurement Data. *The Proceedings of the 2nd International Conference on Mechanical, System and Control Engineering (ICMSC 2018)*, 29 October 2018, 1-6.
<https://doi.org/10.1051/mateconf/201822004002>
- Christos, C. (2020). *Statistical techniques to identify and handle outliers in multivariate data* [Online]. Retrieved December 5, 2022, from: <https://idus.us.es/bitstream/handle/11441/115316/Grentzelos%2C%20Christos.pdf?sequence=1&isAllowed=y>
- Hubert, M., and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Comput Stat Data Anal*, 52(3), 5186-5201.
- Junsawang, P., Promwongsa, M., and Srisodaphol, M. (2021). Robust Outliers Detection Method for Skewed Distribution. *Thailand Statistician*, 19(3), 450-471.
- Leon, H. (1960). Tables of range and studentized range, *Ann. Math. Stat*, 31(4), 1122-1147.
- Ley, C., Klein, O., Bernard, P., and Licata., L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Experimental Social Psychology*, 49(4), 764-766.
- Montgomery, C. (2012). *Introduction to Statistical Quality Control*. 7th ed. New York: John Wiley.
- NIST/SEMATECH (2012). *Engineering Statistics*. [Online]. Retrieved December 5, 2022, from: <https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm>.
- Tukey, J. W. (1977). *Exploratory data analysis*. Massachusetts: Addison-Wesley.