การตรวจสอบค่านอกเกณฑ์ในตัวอย่างสุ่มจากประชากรที่มีการแจกแจงปรกติ โดยใช้สัมประสิทธิ์ความเบ้

A Detection of Outliers in Random Sample from Normally Distributed Population Using Coefficient of Skewness

วรพรรณ เจริญขำ*

Woraphan Jareankam*

สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครสวรรค์

Mathematics and Statistics Department, Faculty of Science and Technology,

NakhonSawanRajabhat University

Received : 6 July 2019

Revised : 2 October 2019

Accepted : 9 October 2019

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อนำเสนอตัวสถิติทดสอบสำหรับตรวจสอบค่านอกเกณฑ์ โดยเป็นตัวสถิติที่พัฒนามาจากแนวคิดของตัวสถิติ GESD ของรอสเนอร์ โดยมีข้อตกลงเบื้องต้นว่า ตัวอย่างสุ่มมาจากประชากรที่มีการแจกแจงปรกติ ทำการตรวจสอบคุณสมบัติตัวสถิติทดสอบที่นำเสนอโดยพิจารณาค่าความน่าจะเป็นของความผิดพลาดแบบที่ 1 เมื่อสมมุติฐานว่างเป็นจริงเปรียบเทียบกับเกณฑ์ที่กำหนด และร้อยละของการตัดสินใจถูกต้องเมื่อสมมุติฐานทางเลือกเป็นจริง โดยทดสอบค่านอกเกณฑ์ด้านขวาโดยใช้สถิติทดสอบที่นำเสนอ ทำการจำลองข้อมูลจากประชากรที่มีการแจกแจงปรกติ กำหนดขนาดตัวอย่างคือ 10 , 20 , 30 , 50 , 100 และ 200 กำหนดค่านอกเกณฑ์ตามสถานการณ์ต่าง ๆ กัน ในแต่ละสถานการณ์จำลองข้อมูลมีการทำซ้ำจำนวน 1,000 รอบ ระดับนัยสำคัญที่ใช้คือ 0.05 ผลการวิจัยพบว่า กรณีตัวอย่างสุ่มไม่มีค่านอกเกณฑ์ เมื่อสมมุติฐานว่างเป็นจริง ทุกขนาดตัวอย่างสถิติทดสอบที่นำเสนอสามารถควบคุมความน่าจะเป็นของความผิดพลาดแบบที่ 1 ได้ และกรณีตัวอย่างสุ่มมีค่านอกเกณฑ์ด้านขวา k ค่า โดยค่านอกเกณฑ์ลักษณะเป็นค่าผิดปกติมาก เมื่อสมมุติฐานทางเลือกเป็นจริงพบว่า ทุกขนาดตัวอย่างมีค่าร้อยละของการตัดสินใจถูกต้องมากกว่าร้อยละ 95

คำสำคัญ : ค่านอกเกณฑ์, การแจกแจงปรกติ สัมประสิทธิ์ความเบ้

Abstract

The objective of this research is to the propose a detection of outliers in a test statistic. Such a test statistic is developed from the concept of GESD by Rosner, where the assumption of this test statistic is a normally distributed population. In this paper, we shall show a comparison of an ability of control type I error and power of the test under a simulation of normal distributions induced by the sample sizes 10, 20, 30, 50, 100 and 200. The outliers in this simulation are defined in many different situations, which the situations are replicated 1,000 times. The significance level is 0.05. The results show that the proposed test statistic can control the probability of type I error. The percentage of correct decision are greater than 95 when the alternative hypothesis is true.

Keywords : outliers, normal distribution, coefficient of skewness

***Coresponding author. E-mail** : aj_woraphan@hotmail.com

บทนำ

ปัจจุบันข้อมูลสารสนเทศมีความสำคัญมากในทุกศาสตร์วิชา วิชาทางด้านสถิติจึงเข้ามามีบทบาทสำคัญในการวิเคราะห์ข้อมูล ข้อมูลที่นำมาวิเคราะห์ควรมีคุณภาพและมีความถูกต้องเชื่อถือได้ ปัญหาอย่างหนึ่งในการวิเคราะห์ข้อมูลทางสถิติคือ ข้อมูลที่เก็บรวบรวมมาได้ มีค่าบางค่าสูงหรือต่ำมาก หรือไม่ได้มาจากประชากรเดียวกันกับข้อมูลส่วนใหญ่ เราเรียกค่าเหล่านี้ว่า ค่านอกเกณฑ์ (Outliers) ซึ่งสาเหตุของค่านอกเกณฑ์เหล่านี้อาจมาจากความคลาดเคลื่อนต่าง ๆ เช่น ความคลาดเคลื่อนจากความแปรผันที่มีอยู่ในประชากรที่ศึกษา ความคลาดเคลื่อนที่เกิดจากการวัดหรือจากการปฏิบัติการ เป็นต้น หากนำข้อมูลที่มีค่านอกเกณฑ์ไปทำการวิเคราะห์ อาจส่งผลต่อการกระจายของข้อมูล หรืออาจทำให้ค่าเฉลี่ยของข้อมูลเปลี่ยนไป ซึ่งจะส่งผลต่อการเลือกใช้สถิติที่ไม่ถูกต้องตามลักษณะของข้อมูล หรือไม่ถูกต้องตามข้อตกลงเบื้องต้น (Assumption) ของระเบียบวิธีสถิติที่นำมาวิเคราะห์ ผลการวิเคราะห์ขาดคุณภาพและความน่าเชื่อถือ ทำให้ผลของการวิจัยไม่สามารถนำไปใช้ประโยชน์ได้อย่างเหมาะสม หรือไม่สามารถตอบคำถามวิจัยได้ ดังนั้น การตรวจสอบค่านอกเกณฑ์จึงถือเป็นเรื่องสำคัญที่ควรทำก่อนลงมือวิเคราะห์ข้อมูล

การตรวจสอบค่านอกเกณฑ์ที่มีและใช้กันอยู่ในปัจจุบันมีหลายวิธี เช่น การตรวจสอบค่านอกเกณฑ์ด้วยวิธีกราฟหรือการทดสอบสมมุติฐาน การตรวจสอบค่านอกเกณฑ์ด้วยวิธีกราฟ เช่น ฮีสโทแกรม (Histogram) แผนภาพกล่อง (Box Plot) แผนภาพการกระจาย (Scatter Plot) การตรวจสอบค่านอกเกณฑ์ด้วยวิธีกราฟเป็นวิธีการตรวจสอบที่ง่ายและสะดวกแต่บางครั้งการใช้วิธีกราฟอาจให้ผลสรุปที่ไม่ชัดเจน ทั้งนี้ขึ้นอยู่กับองค์ประกอบหลายประการรวมทั้งความเชี่ยวชาญและประสบการณ์ของผู้วิเคราะห์ การตรวจสอบค่านอกเกณฑ์โดยการทดสอบสมมุติฐานว่า ค่าสังเกตลำดับที่ k ($x_{(k)}$) เป็นค่านอกเกณฑ์หรือไม่ ซึ่งมีผู้ศึกษาและพัฒนาตัวสถิติทดสอบอย่างต่อเนื่องในหลากหลายสถานการณ์ บาร์เนทและเลวิส (Barnett & Lewis, 1984) ได้ทำการรวบรวมสถิติที่ใช้ทดสอบค่านอกเกณฑ์กรณีตัวอย่างสุ่มมาจากการแจกแจงปรกติตามสถานการณ์ต่าง ๆ ตัวอย่างเช่น ตัวสถิติทดสอบ $T_{N1}=\frac{x_{(n)}-\bar{x}}{s}$ ใช้ทดสอบว่าค่าสังเกตที่มีค่ามากที่สุด ($x_{(n)}$) เป็นค่านอกเกณฑ์หรือไม่ โดยที่ไม่ทราบค่าเฉลี่ยและความแปรปรวนประชากร หรือตัวสถิติทดสอบ $T_{N8}=\max\left(\frac{x_{(n)}-x_{(n-1)}}{x_{(n)}-x_{(1)}},\frac{x_{(2)}-x_{(1)}}{x_{(n)}-x_{(1)}}\right)$ ใช้ทดสอบว่าข้อมูลมีค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติมาก (Extreme outlier) หรือไม่ เมื่อตัวอย่างสุ่มมาจากการแจกแจงปรกติโดยที่ไม่ทราบค่าความแปรปรวนประชากร ตัวสถิติทดสอบ $T_{N\sigma^3}=\frac{x_{n-k+1}+\ldots+x_{n-1}+x_{n}-k\bar{x}}{\sigma}$ ใช้ทดสอบว่าค่าสังเกตตัวที่ n-k+1,..., ตัวที่ n-1 จนถึงตัวที่ n ($x_{n-k+1}+\ldots+x_{n-1}+x_{n}$) เป็นค่านอกเกณฑ์หรือไม่ โดยที่ k มีค่ามากกว่า 1 เมื่อไม่ทราบค่าเฉลี่ยประชากร แต่ทราบค่าความแปรปรวนประชากร เป็นต้น

ปัญหาหนึ่งที่พบในการตรวจสอบค่านอกเกณฑ์คือ การตรวจสอบค่านอกเกณฑ์ด้วยวิธีตรวจสอบทีละ 1 ค่า หรือ 2 ค่าแล้วไม่สามารถอธิบายได้ว่าแท้จริงแล้วในตัวอย่างสุ่มมีค่านอกเกณฑ์ทั้งหมดกี่ค่า ซึ่งอาจเป็นไปได้ว่ามีค่านอกเกณฑ์มากกว่า 2 ค่า เมื่อนำตัวสถิติทดสอบที่พัฒนาขึ้นไปใช้จริงนักสถิติส่วนใหญ่ที่อิงแนวคิดแบบดั้งเดิมต่างตระหนักถึงปัญหานี้ มีความพยายามพัฒนาวิธีการและตัวสถิติทดสอบที่สามารถตรวจสอบค่านอกเกณฑ์ได้ทีละหลายค่า สมมติ k ค่า โดยที่ k มากกว่า 2 รอสเนอร์ (Rosner, 1975) ได้เปรียบเทียบวิธีการตรวจสอบค่านอกเกณฑ์ทีละ 1 ค่า และ 2 ค่า กับวิธีการตรวจสอบค่านอกเกณฑ์ทีละหลายค่า โดยใช้ตัวสถิติ ESD, STR , Kurtosis และ RST ผลการศึกษาพบว่ากรณีที่มีค่านอกเกณฑ์ 1 ค่าอยู่ในข้อมูล วิธีการตรวจสอบหาค่านอกเกณฑ์ทีละค่ามีกำลังการทดสอบสูงกว่าวิธีการตรวจสอบหาค่า

นอกเกณฑ์ ทีละหลายค่า แต่ในกรณีที่มีค่านอกเกณฑ์มากกว่า 1 ค่าอยู่ในข้อมูล วิธีการตรวจสอบหาค่านอกเกณฑ์ทีละหลายค่ามีกำลังการทดสอบสูงกว่าวิธีการตรวจสอบหาค่านอกเกณฑ์ทีละค่า ในกรณีที่ค่านอกเกณฑ์มีมากกว่า 1 ค่า ตัวสถิติ ESD เป็นสถิติที่มีกำลังการทดสอบสูงที่สุดและสามารถควบคุมความผิดพลาดแบบที่ 1 ได้ดีกว่าการทดสอบด้วยตัวสถิติอื่นอีก 3 ตัว รอสเนอร์ชี้ให้เห็นถึงปัญหาของวิธีการตรวจสอบค่านอกเกณฑ์ทีละหลายค่าโดยใช้ตัวสถิติ ESD ในการตรวจสอบค่าผิดปกติกล่าวคือ เริ่มแรกของการพัฒนาวิธีการตรวจสอบค่านอกเกณฑ์ทีละหลายค่าโดยใช้ตัวสถิติ ESD ได้กำหนดให้ค่าวิกฤตของตัวสถิติทั้งหมดอยู่ในระดับเดียวกันเพื่อความสะดวก ปัญหาจากการทำเช่นนี้ ฮอว์คินส์ (Hawkins, 1978) ได้อภิปรายไว้ถึงความไม่เหมาะสม นั่นคือ วิธีการนี้จะมีความผิดพลาดแบบที่ 1 ที่เหมาะสมในกรณีที่ไม่มีค่านอกเกณฑ์ในข้อมูล แต่ถ้ามีค่านอกเกณฑ์ในข้อมูลวิธีการนี้อาจให้ความผิดพลาดแบบที่ 1 ที่ไม่เหมาะสม รอสเนอร์ได้พัฒนาวิธีการตรวจสอบค่านอกเกณฑ์ใช้ตัวสถิติ ESD โดยการพัฒนานี้ทำให้มีแนวโน้มที่จะตรวจสอบค่านอกเกณฑ์ที่มีอยู่ในข้อมูลได้ถูกต้องแม่นยำกว่าเดิม เพราะสามารถควบคุมความผิดพลาดแบบที่ 1 ได้เหมาะสมกว่า ตัวสถิติที่เสนอคือ ตัวสถิติ GESD (Generalized Extreme Studentized Deviate) ซึ่งมีพื้นฐานจากตัวสถิติ $R_1 , R_2 , \dots , R_k$ ที่คำนวณจากตัวอย่างที่ลดขนาดต่อเนื่องกันไปคือ n , n-1 , … , n-k+1 โดยที่ k เป็นจำนวนค่าผิดปกติสูงสุดที่คาดว่าจะมีในชุดข้อมูล 1 ชุด แสดงได้ดังนี้

ข้อมูลเริ่มต้นคือ $x_1 , x_2 , \dots , x_n$ กำหนดให้ $I_0 = \{ x_1 , x_2 , \dots , x_n \}$ นั่นคือ I_0 มีขนาดตัวอย่างเท่ากับ n จะได้

$$R_1 = \frac{\underset{x_i \in I_0}{\text{Max}} \, | x_i - \bar{x}_1 |}{s_1} \qquad \text{โดยที่} \qquad \bar{x}_1 = \frac{\sum_{x_i \in I_0} x_i}{n} \quad \text{และ} \quad s_1^2 = \frac{\sum_{x_i \in I_0} x_i - \bar{x}_1^{\,2}}{n - 1}$$

จากนั้นตัดค่า x_i ที่ให้ค่า $\underset{x_i \in I_0}{\text{Max}} \, | x_i - \bar{x}_1 |$

กำหนดให้ $x^{(0)}$ เป็น x_i ที่ $\underset{x_i \in I_0}{\text{Max}} \, | x_i - \bar{x}_1 |$ และ $I_1 = I_0 - x^{(0)}$ นั่นคือ จากข้อมูลเริ่มต้นจะได้ $x^{(0)}, x^{(1)} , \dots , x^{(n-1)}$ ซึ่งคือ ค่าสังเกตที่มีระยะห่างจากค่าเฉลี่ยมากที่สุดในแต่ละ $I_0 , I_1 , \dots , I_{n-1}$ และจะได้

$$R_2 = \frac{\underset{x_i \in I_1}{\text{Max}} \, | x_i - \bar{x}_2 |}{s_2} \qquad \text{โดยที่} \qquad \bar{x}_2 = \frac{\sum_{x_i \in I_1} x_i}{n - 1} \quad \text{และ} \quad s_2^2 = \frac{\sum_{x_i \in I_1} x_i - \bar{x}_2^{\,2}}{n - 2}$$

คำนวณ $R_3 , R_4 , \dots , R_k$ ได้ในทำนองเดียวกัน

ค่าวิกฤตของการทดสอบหาได้โดยกำหนด α แล้วหา λ_i เมื่อ i = 1, 2, … , k ที่ทำให้

$$\Pr\left\{ \bigcup_{i=L+1}^{k} R_i > \lambda_i \mid H_L \right\} = \alpha \qquad ; \qquad L = 0 , 1 , \dots , k-1$$

ถ้าทั้งหมดของ $R_i \leq \lambda_i$ แล้วแสดงว่า ไม่มีค่านอกเกณฑ์ในข้อมูล

ถ้าบางตัว $R_i > \lambda_i$ แล้วกำหนด $C = \text{Max} \quad i ; R_i > \lambda_i$ จะถือว่า $x^{(0)}, x^{(1)} , \dots , x^{(C-1)}$ เป็นค่านอกเกณฑ์ เมื่อ $x^{(0)}, x^{(1)} , \dots , x^{(C-1)}$ เป็นค่าสังเกตที่ทำให้ค่า $\text{Max} \, | x_i - \bar{x}_1 |$ ในข้อมูลที่ค่อย ๆ ลดขนาดลงตามลำดับ

วิธีการนี้จะตรวจสอบค่าผิดปกติได้ตั้งแต่ 1 ถึง k ค่าและสามารถควบคุมความผิดพลาดแบบที่ 1 ได้อย่างดีทั้งภายใต้สมมุติฐานว่าง คือ ไม่มีค่านอกเกณฑ์ในข้อมูล และภายใต้สมมุติฐานทางเลือกของค่านอกเกณฑ์ 1 , 2 , … , k-1 ค่า ตามลำดับ ทั้งนี้ตัวสถิติทดสอบ GESD (Generalized Extreme Studentized Deviate) สามารถตรวจสอบค่านอก

เกณฑ์ได้อย่างมีประสิทธิภาพในตัวอย่างที่มีขนาดมากกว่า 25 แต่ถ้าตัวอย่างที่นำมาตรวจสอบมีขนาดน้อยกว่า 25 ความน่าเชื่อถือของตัวสถิติทดสอบ GESD อาจไม่เหมาะสม

ในงานวิจัยครั้งนี้ผู้วิจัยมีความสนใจนำเสนอตัวสถิติทดสอบสำหรับตรวจสอบค่านอกเกณฑ์ทีละหลายค่า เพื่อให้คำนวณข้อมูลได้ง่ายและสะดวกขึ้น โดยประยุกต์จากแนวคิดของรอสเนอร์ที่ได้เสนอตัวสถิติทดสอบ GESD (Generaized Extreme Studentized Deviate) และจากสมบัติของการแจกแจงปรกติในกรณีที่ตัวอย่างสุ่มมาจากประชากรปรกติ

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อนำเสนอตัวสถิติทดสอบสำหรับตรวจสอบค่านอกเกณฑ์ โดยเป็นตัวสถิติที่พัฒนามาจากแนวคิดของตัวสถิติ GESD ของรอสเนอร์ โดยมีข้อตกลงเบื้องต้นว่า ตัวอย่างสุ่มมาจากประชากรที่มีการแจกแจงปรกติ เมื่อได้ตัวสถิติทดสอบแล้ว จะทำการตรวจสอบคุณสมบัติตัวสถิติที่นำเสนอโดยพิจารณาจากความสามารถในการควบคุมความน่าจะเป็นของความผิดพลาดแบบที่ 1 และร้อยละของการตัดสินใจถูกต้อง (กำลังการทดสอบ)

สำหรับตัวอย่างสุ่มจากการแจกแจงปรกติ สัมประสิทธิ์ความเบ้มีค่าเท่ากับ 0 การแจกแจงของข้อมูลมีลักษณะสมมาตร แต่ในกรณีที่มีค่านอกเกณฑ์ในตัวอย่างสุ่ม เช่น มีค่านอกเกณฑ์ทางขวา จะทำให้การแจกแจงของข้อมูลมีลักษณะเบ้ขวา สัมประสิทธิ์ความเบ้จะมีค่ามากกว่า 0 หรือถ้ามีค่านอกเกณฑ์ทางซ้าย จะทำให้การแจกแจงของข้อมูลมีลักษณะเบ้ซ้าย สัมประสิทธิ์ความเบ้จะมีค่าน้อยกว่า 0 แม้ว่าความเบ้จะสามารถสังเกตได้ด้วยสายตา แต่โดยทั่วไปนิยมวัดความเบ้ในรูปของสัมประสิทธิ์ความเบ้ ซึ่งมีอยู่หลายวิธีได้แก่ วิธีโมเมนต์ ค่าสัมประสิทธิ์ความเบ้ของเพียร์สัน (Pearson's Coefficient of Skewness) ค่าความเบ้สัมพัทธ์ของบาวลี (Bowley's Measure of Skewness) ค่าความเบ้สัมพัทธ์ของเคลลี (Kelly's Measure of Skewness) เป็นต้น การคำนวณสัมประสิทธิ์ความเบ้ของโปรแกรมทางสถิติบางโปรแกรม หรือของนักสถิติบางคนอาจใช้สูตรการคำนวณที่แตกต่างกันไปตามวิธีต่าง ๆ เช่น โจเนสและกิล (Joanes and Gill 1998) ใช้สูตรดังนี้

สัมประสิทธิ์ของความเบ้คือ

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3} \qquad (1)$$

หรือ

$$b_1 = \left(\frac{n-1}{n}\right)^{\frac{3}{2}} \frac{m_3}{m_2^{\frac{3}{2}}} \qquad (2)$$

โดยที่ $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ และ $m_r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^r}{n}$ เมื่อ r = 2 , 3

การศึกษาพัฒนาสถิติทดสอบสำหรับการตรวจสอบค่านอกเกณฑ์ของรอสเนอร์ การคำนวณค่าวิกฤตของการทดสอบค่อนข้างยุ่งยาก นอกจากนี้หากตัวอย่างที่นำมาตรวจสอบมีขนาดน้อยกว่า 25 ความน่าเชื่อถือของตัวสถิติทดสอบ GESD อาจไม่เหมาะสม ข้อจำกัดด้านนี้จึงจำเป็นต้องมีการศึกษาเพื่อให้ใช้งานได้ครอบคลุมมากขึ้น

จากทฤษฎีบทขีดจำกัดส่วนกลาง (Central Limit Theorem) เมื่อ $x_1, x_2, ..., x_n$ เป็นตัวแปรสุ่มที่เป็นอิสระกันและมีการแจกแจงความน่าจะเป็นเหมือนกัน $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ มีค่าจำกัด ให้ $T(x)$ ซึ่งเป็นสถิติตัวหนึ่ง แล้ว $\frac{T(x) - \mu}{\sigma}$ ลู่เข้าในการแจกแจงไปหาตัวแปรสุ่มปรกติมาตรฐาน Z (Suwattee, 2010)

ในการศึกษาครั้งนี้ พิจารณาการคำนวณสัมประสิทธิ์ความเบ้ตามสูตรในสมการที่ 1 เป็นสถิติทดสอบ ซึ่งมีสมบัติที่สำคัญตามทฤษฎีบทที่ 1 ดังนี้

ทฤษฎีบทที่ 1 ให้ $x_1,x_2,...,x_n$ เป็นตัวแปรสุ่มที่มาจากการแจกแจงปรกติ พบว่า

$$T = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{ns^3}}{\sqrt{\frac{n-1}{n}^3 \frac{6(n-2)}{(n+1)(n+3)}}} \xrightarrow{\mathscr{L}} Z \sim N(0,1) \qquad (3)$$

พิสูจน์ ให้ $b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{ns^3} = \left(\frac{n-1}{n}\right)^{\frac{3}{2}} \frac{m_3}{m_2^{\frac{3}{2}}}$ โดยที่ $m_r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^r}{n}$

Fisher (1930) และ Cramér (1946) แสดงให้เห็นว่า

$$E\left(\frac{m_3}{m_2^{\frac{3}{2}}}\right) = 0 \quad \text{และ} \quad Var\left(\frac{m_3}{m_2^{\frac{3}{2}}}\right) = \frac{6(n-2)}{(n+1)(n+3)}$$

$$\text{ดังนั้น} \quad E\left(\frac{n-1}{n}^{\frac{3}{2}} \frac{m_3}{m_2^{\frac{3}{2}}}\right) = E(b_1) = 0 \quad \text{และ} \quad Var(b_1) = \left(\frac{n-1}{n}\right)^3 \frac{6(n-2)}{(n+1)(n+3)}$$

เนื่องจากโมเมนต์ตัวอย่างเป็นตัวประมาณภาวะน่าจะเป็นสูงสุด จึงเห็นได้ชัดเจนว่า เมื่อ $n \longrightarrow \infty$ แล้ว สมการ (3) เป็นจริง

จากทฤษฎีบทที่ 1 และแนวคิดจากตัวสถิติ GESD ของรอสเนอร์ ผู้วิจัยขอนำเสนอตัวสถิติทดสอบสำหรับการตรวจสอบค่านอกเกณฑ์ โดยศึกษาเฉพาะการทดสอบค่านอกเกณฑ์ด้านขวาดังนี้

การทดสอบค่านอกเกณฑ์ด้านขวา

ข้อตกลงเบื้องต้นคือ ตัวอย่างสุ่มมาจากประชากรที่มีการแจกแจงปรกติ

สมมุติฐานในการทดสอบคือ

H_0: ตัวอย่างสุ่มไม่มีค่านอกเกณฑ์

H_1: ตัวอย่างสุ่มมีค่านอกเกณฑ์ด้านขวา k ค่า

ตัวสถิติทดสอบ คือ

ให้ $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ เป็นตัวอย่างสุ่มขนาด n คำนวณตัวสถิติ $T_1, T_2, \ldots, T_k$ จากตัวอย่างสุ่มที่ลดขนาดต่อเนื่องกันไปคือ n , n-1 , … , n-k+1 โดยที่ k เป็นจำนวนค่านอกเกณฑ์สูงสุดในตัวอย่างสุ่มโดยที่

$$T_1 = \frac{\sum_{i=1}^{n}(x_{(i)} - \bar{x})^3}{\sqrt{\frac{n-1}{n}^3 \frac{6(n-2)}{(n+1)(n+3)}}} \text{ เมื่อ } \bar{x} = \frac{\sum_{i=1}^{n} x_{(i)}}{n} \text{ , } s^2 = \frac{\sum_{i=1}^{n}(x_{(i)} - \bar{x})^2}{n-1}$$

คำนวณ T_1 จากตัวอย่างสุ่มขนาด n และคำนวณ $T_2, T_3, \ldots, T_k$ ได้ในลักษณะเดียวกัน แต่ลดขนาดตัวอย่างเป็น n-1, n-2, … ,n-k+1

เกณฑ์การทดสอบสมมุติฐานคือ

ถ้า $T_i \leq Z_\alpha$ แสดงว่า เซตของตัวอย่างสุ่ม $O_{Normal} = \{ x_{(1)}, x_{(2)}, ..., x_{(n-i+1)} \}$ เป็นข้อมูลที่ไม่มีค่านอกเกณฑ์

ถ้า $T_i > Z_\alpha$ แสดงว่า เซตของตัวอย่างสุ่ม $O_{Outliers} = \{ x_{(n-i+1)}, x_{(n-i+2)}, ... ,x_{(n)} \}$ เป็นข้อมูลที่มีค่านอกเกณฑ์ i ค่า

เมื่อ Z_α คือ ค่าควอนไทล์ของการแจกแจงปรกติมาตรฐาน

ในการวิจัยครั้งนี้ใช้วิธีการจำลองข้อมูลเพื่อพิจารณาค่าความน่าจะเป็นของความผิดพลาดแบบที่ 1 เมื่อสมมุติฐานว่างเป็นจริงเปรียบเทียบกับเกณฑ์ที่กำหนด และร้อยละของการตัดสินใจถูกต้องเมื่อสมมุติฐานทางเลือกเป็นจริง โดยใช้โปรแกรมสำเร็จรูปทางสถิติคือ Minitab ในการจำลองข้อมูลและทดสอบสมมุติฐาน มีขั้นตอนการดำเนินงาน ดังนี้

1) สุ่มตัวอย่างจากประชากรที่มีการแจกแจงปรกติทั้งหมด 9 การแจกแจงคือ N(1,0.25) , N(1,1) , N(1,4) , N(3,0.25) , N(3,1) , N(3,4) , N(5,0.25) , N(5,1) และ N(5,4) โดยขนาดตัวอย่างที่ทำการศึกษามี 3 ขนาดคือ ขนาดเล็ก (n = 10 , 20) ขนาดกลาง (n = 30 , 50) และขนาดใหญ่ (n = 100 , 200)

2) กำหนดให้แต่ละขนาดตัวอย่างมีค่านอกเกณฑ์ทางด้านขวาจำนวน 10% ของขนาดตัวอย่าง

3) กำหนดค่านอกเกณฑ์โดยการสุ่มค่านอกเกณฑ์ให้อยู่ในช่วงที่กำหนดตามแต่ละการแจกแจงของประชากร สำหรับการแจกแจงปรกติ $Q_1 = -0.675$, $Q_2 = 0$ และ $Q_3 = 0.657$ การทดสอบค่านอกเกณฑ์ด้านขวาแบ่งค่านอกเกณฑ์ออกเป็น 2 ช่วง คือ

(1) ค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติเล็กน้อย ($Q_3+1.5IQR < x_{outliers} < Q_3+3IQR$)

(2) ค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติมาก ($x_{outliers} \geq Q_3+3IQR$)

4) ในแต่ละสถานการณ์มีการทำซ้ำจำนวน 1,000 รอบ

5) กำหนดระดับนัยสำคัญที่ใช้ในการทดสอบคือ 0.05

6) คำนวณสถิติทดสอบโดย การทดสอบค่านอกเกณฑ์ด้านขวา สถิติที่ใช้ทดสอบคือ สถิติทดสอบ T

7) นำค่าสถิติที่คำนวณได้ เทียบกับค่าวิกฤต เพื่อสรุปผลว่าจะปฏิเสธหรือยอมรับสมมุติฐานว่าง

8) บันทึกจำนวนครั้งที่ปฏิเสธสมมุติฐานว่างที่ระดับนัยสำคัญที่กำหนด จากนั้นทำการประมาณความน่าจะเป็นของความผิดพลาดแบบที่ 1 และร้อยละของการตัดสินใจในแต่ละสถานการณ์ คำนวณค่าความน่าจะเป็นของความผิดพลาดแบบที่ 1 และร้อยละของการตัดสินใจถูกต้องของสถิติทดสอบที่นำเสนอ

เกณฑ์ในการตัดสินใจเมื่อสมมุติฐานว่างเป็นจริงคือ

ความน่าจะเป็นของความผิดพลาดแบบที่ 1 = $\frac{(\text{จำนวนครั้งที่ปฏิเสธสมมุติฐานว่างเมื่อสมมุติฐานว่างเป็นจริง})}{\text{จำนวนครั้งที่ทำการทดสอบ}}$

เกณฑ์ในการตัดสินใจเมื่อสมมุติฐานทางเลือกเป็นจริงคือ

ร้อยละของการตัดสินใจถูกต้อง (กำลังการทดสอบ) = $\frac{(\text{จำนวนครั้งที่ปฏิเสธสมมุติฐานว่างเมื่อสมมุติฐานทางเลือกเป็นจริง})}{\text{จำนวนครั้งที่ทำการทดสอบ}}$ x 100

ทำการตรวจสอบสมบัติของสถิติทดสอบที่นำเสนอ โดยเปรียบเทียบความน่าจะเป็นของความผิดพลาดแบบที่ 1 กับเกณฑ์ของ Cochran (Cochran,1954) หากอยู่ในเกณฑ์ คือ เมื่อกำหนดระดับนัยสำคัญ α = 0.05 และผลการคำนวณค่าความน่าจะเป็นของความผิดพลาดแบบที่ 1 อยู่ระหว่าง 0.04 – 0.06 จะสรุปได้ว่าการทดสอบนั้นสามารถควบคุมความน่าจะเป็นของความผิดพลาดแบบที่ 1 ได้ และในแต่ละสถานการณ์ พิจารณาร้อยละของการตัดสินใจถูกต้อง (กำลังการทดสอบ) ของสถิติทดสอบที่นำเสนอ

ผลการวิจัย

จากตารางที่ 1 กรณีตัวอย่างสุ่มไม่มีค่านอกเกณฑ์ เมื่อสมมุติฐานว่างเป็นจริงพบว่า ที่การแจกแจงปรกติค่าเฉลี่ยเท่ากับ 1 , 3 และ 5 ($\mu = 1,3,5$) ทั้งกรณีความแปรปรวนน้อย ($\sigma^2 = 0.25$) และความแปรปรวนมาก ($\sigma^2 = 4$) ทุกขนาดตัวอย่างมีค่าความน่าจะเป็นของความผิดพลาดแบบที่ 1 อยู่ระหว่าง 0.04 – 0.06 แสดงให้เห็นว่าการทดสอบโดยใช้สถิติทดสอบ T ที่นำเสนอสามารถควบคุมความน่าจะเป็นของความผิดพลาดแบบที่ 1 ได้ตามเกณฑ์ของ Cochran

ตารางที่ 1 ความน่าจะเป็นของความผิดพลาดแบบที่ 1 ของสถิติทดสอบ T ที่ ระดับนัยสำคัญ 0.05 จำแนกตามการแจกแจงและขนาดตัวอย่าง

n	การแจกแจง								
	N(1,0.25)	N(1,1)	N(1,4)	N(3,0.25)	N(3,1)	N(3,4)	N(5,0.25)	N(5,1)	N(5,4)
10	0.057	0.053	0.053	0.051	0.056	0.043	0.045	0.062	0.045
20	0.041	0.045	0.055	0.044	0.047	0.046	0.048	0.059	0.055
30	0.044	0.060	0.044	0.046	0.042	0.051	0.047	0.053	0.051
50	0.050	0.052	0.058	0.048	0.057	0.061	0.049	0.043	0.049
100	0.045	0.052	0.055	0.053	0.047	0.046	0.048	0.047	0.049
200	0.059	0.045	0.054	0.042	0.060	0.038	0.058	0.050	0.049

จากตารางที่ 2 กรณีตัวอย่างสุ่มมีค่านอกเกณฑ์ด้านขวา k ค่า โดยค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติเล็กน้อย ($Q_3+1.5IQR < x_{outliers} < Q_3+3IQR$) เมื่อสมมุติฐานทางเลือกเป็นจริงพบว่า ที่การแจกแจงปรกติค่าเฉลี่ยเท่ากับ 1 , 3 และ 5 ($\mu = 1,3,5$) ทั้งกรณีความแปรปรวนน้อย ($\sigma^2 = 0.25$) และความแปรปรวนมาก ($\sigma^2 = 4$) ที่ขนาดตัวอย่างเล็ก (n = 10 , 20) และขนาดกลาง (n = 30 , 50) ร้อยละของการตัดสินใจถูกต้องโดยส่วนใหญ่มีค่าอยู่ในช่วงร้อยละ 50 ถึง ร้อยละ 60 ส่วนที่ตัวอย่างขนาดใหญ่ (n = 100 , 200) ร้อยละของการตัดสินใจถูกต้องโดยส่วนใหญ่มีค่าอยู่ในช่วงร้อยละ 30 ถึง ร้อยละ 50 แสดงให้เห็นว่าที่ตัวอย่างขนาดใหญ่ร้อยละของการตัดสินใจถูกต้องจะน้อยกว่าที่ตัวอย่างขนาดเล็กและขนาดกลาง

ตารางที่ 2 ร้อยละของการตัดสินใจถูกต้องของสถิติทดสอบ T ที่ระดับนัยสำคัญ 0.05 จำแนกตามการแจกแจงและขนาดตัวอย่าง กรณีค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติเล็กน้อย ($Q_3+1.5IQR < x_{outliers} < Q_3+3IQR$)

n	การแจกแจง								
	N(1,0.25)	N(1,1)	N(1,4)	N(3,0.25)	N(3,1)	N(3,4)	N(5,0.25)	N(5,1)	N(5,4)
10	57.10	56.10	56.20	57.30	60.10	54.70	57.20	56.10	55.30
20	60.90	57.30	58.20	59.30	61.30	59.80	59.60	59.60	58.60
30	58.70	57.70	59.70	62.90	60.70	60.60	58.50	56.70	59.60
50	57.30	58.50	54.50	54.00	54.80	53.50	55.70	56.00	54.30
100	48.40	46.20	46.40	46.30	46.50	44.80	48.20	46.90	46.20
200	33.90	34.90	34.10	33.80	36.50	37.00	36.70	31.00	35.00

จากตารางที่ 3 กรณีตัวอย่างสุ่มมีค่านอกเกณฑ์ด้านขวา k ค่า โดยค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติมาก ($x_{outliers} \geq Q_3+3IQR$) เมื่อสมมุติฐานทางเลือกเป็นจริงพบว่า ที่การแจกแจงปรกติค่าเฉลี่ยเท่ากับ 1 , 3 และ 5 ($\mu = 1,3,5$) ทั้งกรณีความแปรปรวนน้อย ($\sigma^2 = 0.25$) และความแปรปรวนมาก ($\sigma^2 = 4$) ทุกขนาดตัวอย่างร้อยละของการตัดสินใจถูกต้องมีค่ามากกว่าร้อยละ 95 แสดงให้เห็นว่ากรณีตัวอย่างสุ่มมีค่านอกเกณฑ์ด้านขวา k ค่า โดยค่านอกเกณฑ์เป็นค่าผิดปกติมากร้อยละของการตัดสินใจถูกต้องมีค่าสูงเกือบถึงร้อยละ 100 นั่นคือ การทดสอบโดยใช้สถิติทดสอบ T เหมาะสมกับกรณีตัวอย่างสุ่มมีค่านอกเกณฑ์ด้านขวาโดยค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติมาก

ตารางที่ 3 ร้อยละของการตัดสินใจถูกต้องของสถิติทดสอบ T ที่ระดับนัยสำคัญ 0.05 จำแนกตามการแจกแจงและขนาด ตัวอย่าง กรณีค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติมาก ($x_{outliers} \geq Q_3+3IQR$)

n	การแจกแจง								
	N(1,0.25)	N(1,1)	N(1,4)	N(3,0.25)	N(3,1)	N(3,4)	N(5,0.25)	N(5,1)	N(5,4)
10	95.80	96.80	96.10	95.90	97.10	95.80	96.30	96.50	96.40
20	98.80	99.40	98.80	99.30	99.10	99.00	98.70	99.10	99.00
30	99.40	99.40	99.80	99.80	99.40	98.90	99.40	98.80	99.50
50	99.40	99.60	99.30	99.60	99.80	99.20	99.60	99.30	99.60
100	99.10	99.00	99.00	99.00	98.50	98.80	99.50	98.50	98.10
200	95.40	96.70	96.90	97.10	97.00	96.50	97.40	96.30	96.90

วิจารณ์ผลการวิจัย

กรณีตัวอย่างสุ่มไม่มีค่านอกเกณฑ์ สถิติทดสอบ T ที่นำเสนอสามารถควบคุมความน่าจะเป็นของความผิดพลาดแบบที่ 1 ได้ตามเกณฑ์ของ Cochran ในทุกสถานการณ์ โดยมีค่าความน่าจะเป็นของความผิดพลาดแบบที่ 1 อยู่ระหว่าง 0.04 – 0.06 สำหรับสถิติทดสอบ GESD ของรอสเนอร์สามารถควบคุมความน่าจะเป็นของความผิดพลาดแบบที่ 1 ได้ เมื่อขนาดตัวอย่างมากกว่า 25 ซึ่งแสดงให้เห็นว่า กรณีตัวอย่างสุ่มไม่มีค่านอกเกณฑ์ สถิติทดสอบ T สามารถควบคุมความน่าจะเป็นของความผิดพลาดแบบที่ 1 ได้ดีกว่าสถิติทดสอบ GESD ของรอสเนอร์

กรณีตัวอย่างสุ่มมีค่านอกเกณฑ์ด้านขวา k ค่า โดยค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติเล็กน้อย ($Q_3+1.5IQR < x_{outliers} < Q_3+3IQR$) สถิติทดสอบ T ที่นำเสนอ เมื่อสมมุติฐานทางเลือกเป็นจริงพบว่า ร้อยละของการตัดสินใจถูกต้องโดยส่วนใหญ่มีค่าต่ำกว่าร้อยละ 60 อาจเนื่องมาจากระยะห่างของค่านอกเกณฑ์ห่างจากข้อมูลส่วนใหญ่เพียงเล็กน้อย ส่งผลให้ร้อยละของการตัดสินใจถูกต้องมีค่าน้อย ที่ขนาดตัวอย่างเล็ก (n = 10 , 20) และขนาดกลาง (n = 30 , 50) ร้อยละของการตัดสินใจถูกต้องโดยส่วนใหญ่มีค่าอยู่ในช่วงร้อยละ 50 ถึง ร้อยละ 60 ส่วนที่ตัวอย่างขนาดใหญ่ (n = 100 , 200) ร้อยละของการตัดสินใจถูกต้องโดยส่วนใหญ่มีค่าอยู่ในช่วงร้อยละ 30 ถึง ร้อยละ 50 แสดงให้เห็นว่าที่ตัวอย่างขนาดใหญ่ร้อยละของการตัดสินใจถูกต้องจะน้อยกว่าที่ตัวอย่างขนาดเล็กและขนาดกลาง กล่าวคือเมื่อขนาดตัวอย่างใหญ่ขึ้น ค่านอกเกณฑ์ไม่ส่งผลกระทบต่อข้อมูลส่วนใหญ่ทำให้ร้อยละของการตัดสินใจถูกต้องจะน้อยลง

กรณีตัวอย่างสุ่มมีค่านอกเกณฑ์ด้านขวา k ค่า โดยค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติมาก ($x_{outliers} \geq Q_3+3IQR$) สถิติทดสอบ T ที่นำเสนอ เมื่อสมมุติฐานทางเลือกเป็นจริงพบว่า สถิติทดสอบ T เหมาะสมกับกรณีค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติมากโดยให้ค่ามีค่าสูงเกือบถึงร้อยละ 100 ในทุกสถานการณ์ อาจเนื่องมาจากระยะห่างของค่านอกเกณฑ์ห่างจากข้อมูลส่วนใหญ่มาก ส่งผลให้ร้อยละของการตัดสินใจถูกต้องมีค่ามาก

กรณีตัวอย่างสุ่มมีค่านอกเกณฑ์ตั้งแต่ 1 ถึง k ค่า ตัวสถิติทดสอบ GESD ของรอสเนอร์ สามารถตรวจสอบค่านอกเกณฑ์ได้อย่างมีประสิทธิภาพในตัวอย่างที่มีขนาดมากกว่า 25 สำหรับสถิติทดสอบ T ที่นำเสนอ เมื่อสมมุติฐานทางเลือกเป็นจริงพบว่า สถิติทดสอบ T เหมาะสมกับกรณีค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติมากโดยให้ค่ามีค่าสูงเกือบถึงร้อยละ 100 ในทุกสถานการณ์ นอกจากนี้ก่อนทำการทดสอบสมมติฐาน ตัวสถิติทดสอบ GESD ของรอสเนอร์ต้องกำหนดจำนวนค่านอกเกณฑ์ (k) ไว้ล่วงหน้า แต่สถิติทดสอบ T ที่นำเสนอไม่ต้องกำหนด

สรุปผลการวิจัย

กรณีตัวอย่างสุ่มไม่มีค่านอกเกณฑ์ที่การแจกแจงปรกติค่าเฉลี่ยเท่ากับ 1 , 3 และ 5 ($\mu = 1,3,5$) และความแปรปรวนเท่ากับ 0.25 , 1 และ 4 ($\sigma^2 = 0.25,1,4$) สถิติทดสอบ T สามารถควบคุมความน่าจะเป็นของความผิดพลาดแบบที่ 1 ได้ เมื่อสมมุติฐานว่างเป็นจริง

กรณีตัวอย่างสุ่มมีค่านอกเกณฑ์ด้านขวา k ค่า สถิติทดสอบ T ให้ค่าร้อยละของการตัดสินใจถูกต้องกรณีค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติมากสูงกว่ากรณีค่านอกเกณฑ์ลักษณะที่เป็นค่าผิดปกติเล็กน้อย โดยให้ค่าร้อยละของการตัดสินใจถูกต้องสูงเกือบถึงร้อยละ 100

กิตติกรรมประกาศ

ขอขอบคุณคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครสวรรค์ ที่ให้การสนับสนุนทุนสำหรับการทำวิจัยครั้งนี้ และผู้วิจัยขอขอบคุณคณะอาจารย์สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครสวรรค์ที่ให้คำปรึกษา

เอกสารอ้างอิง

Barnett, V. & Lewis, T. (1984). *Outliers in Statistical Data.* (2nd Edition). Chichester: John Wiley & Sons.

Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101-129.

Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton: Princeton University Press.

Fisher, R. A. (1930). The moment of the distribution for normal samples of measures of departure from normality. In *Proceedings of the Royal Society A. 130*(812), 16-28.

Hawkin, D.M. (1980). *Identification of Outliers.* London: Chapman and Hall.

Joanes, D.N. & Gill C.A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society (Series D): The Statistician, 47*(1),183-189.

Rosner, B. (1975). On the Detection of Many Outliers. *Technometrics, 17*(2), 211-227.

Rosner, B. (1983). Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics, 25*(2), 165-172.

Suwattee, P. (2010). *Theory of statistical inference*. (3rd Edition). Bangkok: National Institute of Development Administration. (in Thai)