บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในการศึกษาเชิงวิจัยโดยทั่วไป เช่น ทางด้านสังคมศาสตร์ หรือทางด้านธุรกิจ ผู้ที่ทำการศึกษามีจุดมุ่งหมายที่ต้องการหาข้อสรุปเกี่ยวกับลักษณะของประชากร เช่น ค่าเฉลี่ยของประชากร ลักษณะดังกล่าวนี้ เราสามารถพิจารณาได้จากพารามิเตอร์แสดงตำแหน่ง (Location Parameter) และถ้าต้องการทราบลักษณะการกระจายของประชากร เราสามารถพิจารณาได้จากพารามิเตอร์แสดงสเกล (Scale Parameter) โดยการประมาณค่าพารามิเตอร์ทั้งสองนี้ อยู่บนพื้นฐานข้อมูลตัวอย่าง

ในทางปฏิบัติลักษณะของข้อมูลหรือค่าสังเกตที่ถูกเก็บรวบรวมมาเป็นหน่วยตัวอย่างนั้น อาจจะมีค่าสูงกว่าหรือต่ำกว่าค่าสังเกตส่วนใหญ่มาก หรืออาจจะเป็นค่าสังเกตที่ไม่ได้มาจากประชากรเดียวกันกับค่าสังเกตโดยส่วนใหญ่ ซึ่งลักษณะข้อมูลดังกล่าวจะเรียกว่า **ค่าผิดปกติ** (Outliers) โดยสาเหตุการเกิดค่าผิดปกติอาจจะเกิดจากความคลาดเคลื่อนของการวัด เช่น การบันทึกข้อมูล การใช้เครื่องมือที่มีคุณภาพต่ำ เป็นต้น หรืออาจจะเกิดจากเหตุการณ์ผิดปกติที่ไม่สามารถคาดการณ์ได้ล่วงหน้า เช่น ไฟไหม้ คนงานนัดหยุดงาน และภาวะเศรษฐกิจตกต่ำ เป็นต้น หรืออาจจะเกิดจากการที่หน่วยตัวอย่างที่ให้ข้อมูลมาจากประชากรอื่นที่การแจกแจงแตกต่างไปจากการแจกแจงของประชากรที่สนใจศึกษา โดยลักษณะของค่าผิดปกติอาจจะส่งผลกระทบต่อการประมาณค่าพารามิเตอร์แสดงตำแหน่งและพารามิเตอร์แสดงสเกล เนื่องจากค่าประมาณที่ได้จะถูกปรับทิศทางไปตามลักษณะของค่าผิดปกติ

ดังนั้นถ้านำข้อมูลที่ใช้ในการประมาณค่ามาตรวจสอบค่าผิดปกติ และพบว่ามีข้อมูลที่มีค่าผิดปกติ และมีอิทธิพลต่อการประมาณค่าพารามิเตอร์แสดงตำแหน่งหรือพารามิเตอร์แสดงสเกล ผู้วิเคราะห์ไม่ควรที่จะทำการตัดข้อมูลที่มีค่าผิดปกติออกทันที แต่ควรพิจารณาถึงสาเหตุของการเกิดค่าผิดปกติ เช่น ถ้าทราบว่าข้อมูลที่มีค่าผิดปกติเกิดจากการบันทึกข้อมูลผิดพลาด วิธีแก้ไขอาจใช้การตัดข้อมูลที่มีค่าผิดปกติออกไป(มักจะใช้กับกรณีที่มีข้อมูลมาก) หรืออาจทำการปรับแก้ค่าให้ถูกต้อง จากนั้นนำข้อมูลที่เหลือมาประมาณค่าพารามิเตอร์แสดงตำแหน่ง และพารามิเตอร์แสดงสเกลต่อไป แต่ถ้าข้อมูลที่มีค่าผิดปกติเหล่านี้ถูกบันทึกมาอย่างถูกต้องแล้ว ผู้วิเคราะห์ควรพิจารณาหาตัวประมาณค่าที่เหมาะสม เพื่อให้ค่าที่ประมาณใกล้เคียงกับค่าพารามิเตอร์ที่แท้จริง

สำหรับการประมาณค่าพารามิเตอร์แสดงตำแหน่งด้วยค่าเฉลี่ยของตัวอย่าง และการประมาณค่าพารามิเตอร์แสดงสเกลด้วยค่าเบี่ยงเบนมาตรฐานของตัวอย่างนั้นจะมีความเที่ยงตรง(Accuracy)เมื่อข้อมูลไม่มีค่าผิดปกติ แต่ความเที่ยงตรงจะลดลงเมื่อข้อมูลมีค่าผิดปกติ ซึ่งได้มีผู้คิดค้นหาวิธีการแก้ปัญหาที่เกิดขึ้นไว้หลายวิธี แต่ที่รู้จักกันดีคือ วิธีการทางสถิติที่ไม่ใช้พารามิเตอร์(Nonparamatric Method) เช่น วิธีบูตสแตรป (Bootstrap Method) เป็นวิธีหนึ่งทางสถิติที่ไม่ใช้พารามิเตอร์ โดยวิธีการทางสถิติที่ไม่ใช้พารามิเตอร์นั้นจะไม่มีข้อจำกัดทางด้านลักษณะการแจกแจงของประชากร

ในปี ค.ศ. 1993 Seki, Takashina , and Yoikoyama ได้เสนอวิธีการประมาณค่าพารามิเตอร์แสดงตำแหน่งและพารามิเตอร์แสดงสเกลของการแจกแจงแบบปกติ โดยทำการเปรียบเทียบความเที่ยงตรงระหว่างตัวประมาณค่าพารามิเตอร์แสดงตำแหน่ง และตัวประมาณค่าพารามิเตอร์แสดงสเกลที่ได้จากวิธีบูตสแตรปกับตัวประมาณค่าที่ไม่เอนเอียงพบว่า ตัวประมาณค่าที่ได้จากวิธีบูตสแตรปมีความเที่ยงตรงสูงกว่าตัวประมาณค่าที่ไม่เอนเอียง

ในการวิจัยครั้งนี้ ผู้วิจัยมีความสนใจตัวประมาณค่าพารามิเตอร์แสดงตำแหน่ง(ค่าเฉลี่ย)และพารามิเตอร์แสดงสเกล(ค่าเบี่ยงเบนมาตรฐาน) ของการแจกแจงแบบปกติที่มีค่าผิดปกติปลอมปน โดยตัวประมาณที่ศึกษาคือ ตัวประมาณค่าที่มีความแกร่ง ตัวประมาณค่าที่ได้จากวิธีบูตสแตรป และตัวประมาณค่าที่ไม่เอนเอียง แล้วทำการเปรียบเทียบความเที่ยงตรงของตัวประมาณต่างๆ ด้วยความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) ดังนั้นถ้าตัวประมาณค่าด้วยวิธีใดที่มีความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดจะถือว่าตัวประมาณค่านั้นมีความเที่ยงตรงสูงสุดในกลุ่มตัวประมาณที่เปรียบเทียบ

1.2 วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบความเที่ยงตรงของตัวประมาณค่าพารามิเตอร์แสดงตำแหน่งและพารามิเตอร์แสดงสเกล เมื่อข้อมูลมีการแจกแจงแบบปกติที่มีค่าผิดปกติ โดยจำแนกดังนี้

1. เพื่อศึกษาตัวประมาณค่าพารามิเตอร์แสดงตำแหน่ง โดยพิจารณาตัวประมาณ 3 วิธีคือ

1.1 ตัวประมาณค่าที่มีความแกร่งสำหรับพารามิเตอร์แสดงตำแหน่ง (Robust estimator of location)

1.2 ตัวประมาณค่าที่ได้จากวิธีบูตสแตรปสำหรับพารามิเตอร์แสดงตำแหน่ง (Bootstrap estimator of location)

1.3 ตัวประมาณค่าที่ไม่เอนเอียง (Unbiased estimator)

2 เพื่อศึกษาตัวประมาณค่าพารามิเตอร์แสดงสเกลโดยพิจารณาตัวประมาณ 3 วิธี คือ

2.1 ตัวประมาณค่าที่มีความแกร่งสำหรับพารามิเตอร์แสดงสเกล (Robust estimator of scale)

2.2 ตัวประมาณค่าที่ได้จากวิธีบูตสแตรปสำหรับพารามิเตอร์แสดงสเกล(Bootstrap estimator of scale)

2.3 ตัวประมาณค่าที่ไม่เอนเอียง (Unbiased estimator)

1.3 ข้อตกลงเบื้องต้น

สำหรับข้อมูลที่มีค่าผิดปกตินั้น ผู้วิจัยได้ทำการศึกษาในกรณีที่ข้อมูลมีการแจกแจงแบบปกติที่มีค่าผิดปกติ โดยศึกษากรณีที่ข้อมูลมีค่าผิดปกติ 2 ระดับ คือ ระดับไม่รุนแรง (Mild Outliers) และระดับรุนแรง (Extreme Outliers) ตามเงื่อนไขการตรวจสอบค่าผิดปกติ โดยใช้เกณฑ์ดังนี้[1]

1. ข้อมูลที่มีค่าผิดปกติระดับไม่รุนแรง คือ ข้อมูลที่มีค่าตกอยู่ในช่วง

   $(q_1 - 3(Iqr), q_1 - 1.5(Iqr))$ หรือ ตกอยู่ในช่วง $(q_3 + 1.5(Iqr), q_3 + 3(Iqr))$

2. ข้อมูลที่มีค่าผิดปกติระดับรุนแรง คือ ข้อมูลที่มีค่าตกอยู่ในช่วง

   $(-\infty, q_1 - 3(Iqr))$ หรือ ตกอยู่ในช่วง $(q_3 + 3(Iqr), \infty)$

เมื่อ q_1 คือ ค่าควอไทล์ที่ 1 (The First Quartile) ของข้อมูล

q_3 คือ ค่าควอไทล์ที่ 3 (The Third Quartile) ของข้อมูล

Iqr คือ ค่าพิสัยระหว่างควอไทล์ (The Interquartile Range) ซึ่งเท่ากับ $q_3 - q_1$

---

[1] เกณฑ์พิจารณาข้อมูลมีค่าผิดปกติปรากฏในหนังสือสถิติเบื้องต้นทั่วไป เช่น ในหนังสือ Introductory Statistics ของ Prem S. Mann. เป็นต้น

1.4 ขอบเขตของการวิจัย

1. กำหนดลักษณะข้อมูลที่มีค่าผิดปกติออกตามการแจกแจงของข้อมูลดังนี้

1.1 การแจกแจงแบบปกติ (Normal Distribution)

ฟังก์ชันความหนาแน่นอยู่ในรูป

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-1}{2}.\frac{(x-\mu)^2}{\sigma^2}\right) \quad ; \quad -\infty < x < \infty, \sigma > 0, -\infty < \mu < \infty$$

μ คือ พารามิเตอร์แสดงตำแหน่ง

σ คือ พารามิเตอร์แสดงสเกล

ในกรณีนี้จะศึกษาเมื่อค่าเฉลี่ยเท่ากับ 2 และกำหนดค่าเบี่ยงเบนมาตรฐานเท่ากับ 2, 4, 6, 8, 10*

1.2 การแจกแจงแบบปกติปลอมปน (Contaminated Normal Distribution) แบ่งเป็น

1.2.1 การแจกแจงแบบปกติปลอมปนในตำแหน่งมีรูปแบบดังนี้

- ฟังก์ชันความหนาแน่นอยู่ในรูป

$$f(x) = (1-p).N(1,2^2) + p.N(C,2^2) \quad ; \quad -\infty < x < \infty$$

- ฟังก์ชันความหนาแน่นอยู่ในรูป

$$f(x) = (1-p).N(1,2^2) + p.L(\theta,\beta) \quad ; \quad -\infty < x < \infty, \beta > 0$$

ในกรณีนี้จะศึกษาเมื่อ $\theta = C$, $\beta = 2$

เมื่อ $L(\theta,\beta)$ คือ การแจกแจงแบบลาปลาซ ซึ่งมีฟังก์ชันความหนาแน่นอยู่ในรูป

$$f(x) = \frac{1}{2\beta}\exp\left(-\left|\frac{x-\theta}{\beta}\right|\right) \quad ; \quad -\infty < x < \infty, \ -\infty < \theta < \infty, \beta > 0$$

$E(X) = \theta$

$V(X) = 2\beta^2$

---

* การกำหนดค่าเบี่ยงเบนมาตรฐานให้มีค่าต่างๆ เนื่องจากถ้ายิ่งค่าเบี่ยงเบนมาตรฐานมากขึ้นโอกาสที่ข้อมูลจะมีค่าผิดปกติก็จะมากขึ้นด้วย

พอสมุดกลาง สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

1.2.2 การแจกแจงแบบปกติปลอมปนในสเกลมีรูปแบบดังนี้

- ฟังก์ชันความหนาแน่นอยู่ในรูป

$$f(x) = (1-p).N(2,2^2) + p.N(2,C^2 2^2) \quad ; \quad -\infty < x < \infty$$

- ฟังก์ชันความหนาแน่นอยู่ในรูป

$$f(x) = (1-p).N(2,2^2) + p.L(\theta, C.\beta) \quad ; \quad -\infty < x < \infty, -\infty < \theta < \infty, \beta > 0$$

ในกรณีนี้จะศึกษาเมื่อ $\theta = 2$, $\beta = 2$

- ฟังก์ชันความหนาแน่นอยู่ในรูป

$$f(x) = (1-p).N(2,2^2) + p.Expo(\lambda) \quad ; \quad -\infty < x < \infty, \lambda > 0$$

ในกรณีนี้จะศึกษาเมื่อ

$$\lambda = \frac{1}{2C}$$

เมื่อ EXPO(λ) คือ การแจกแจงแบบแบบเลขชี้กำลัง ซึ่งมีฟังก์ชันความหนาแน่นอยู่ในรูป

$$f(x) = \lambda \exp(-\lambda x) \quad ; \quad x > 0, \ \lambda > 0$$

$$E(X) = \frac{1}{\lambda}$$

$$V(X) = \frac{1}{\lambda^2}$$

เมื่อ p คือ อัตราส่วนการปลอมปน ในกรณีนี้จะศึกษาเมื่ออัตราส่วนการปลอมปนเท่ากับ 0.05, 0.1, 0.15 และ 0.20

C คือ สเกลแฟกเตอร์(Scale Factor) ซึ่งจากการทดลองซ้ำพบว่าค่าสเกลแฟกเตอร์เท่ากับ 3 สอดคล้องกับข้อมูลที่มีค่าผิดปกติอยู่ในระดับไม่รุนแรง และค่าสเกลแฟกเตอร์เท่ากับ 10 สอดคล้องกับข้อมูลที่มีค่าผิดปกติอยู่ในระดับรุนแรง ดังนั้นกำหนด C = 3, 10

2. กำหนดขนาดตัวอย่างที่ใช้ในการศึกษาเท่ากับ 20, 30, 50 และ 70

3. ในการศึกษาครั้งนี้จะทำการประมาณพารามิเตอร์แสดงตำแหน่ง (μ) และพารามิเตอร์แสดงสเกล (σ) ของประชากรที่มีการแจกแจงแบบปกติที่มีค่าผิดปกติ

4. ในการวิจัยนี้เป็นการวิจัยเชิงทดลอง โดยใช้เทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) บนเครื่องคอมพิวเตอร์ เขียนโปรแกรมด้วยภาษาฟอร์แทน (Fortran) ทำการจำลองซ้ำ 500 รอบในแต่ละสถานการณ์ของการทดลอง

1.5 สมมติฐานของการวิจัย

ภายใต้ข้อมูลที่มีการแจกแจงแบบปกติที่มีค่าผิดปกติ ตัวประมาณค่าพารามิเตอร์แสดงตำแหน่งและพารามิเตอร์แสดงสเกล ด้วยวิธีตัวประมาณค่าที่มีความแกร่งจะให้ความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำกว่าตัวประมาณค่าที่ได้จากวิธีบูตสแตรป และตัวประมาณค่าที่ไม่เอนเอียง

1.6 คำจำกัดความ

1. ความแกร่ง (Robustness) ของตัวประมาณ หมายถึง คุณสมบัติของตัวประมาณที่ไม่แสดงความไวต่อการเปลี่ยนแปลงของปัจจัยอื่นที่ไม่ใช่ปัจจัยที่สนใจ เช่น ค่าข้อมูลบางค่าเบี่ยงเบนไปจากค่าข้อมูลโดยส่วนใหญ่

2. ความเที่ยงตรง (Accuracy) หมายถึง เกณฑ์ที่พิจารณาว่าค่าที่เป็นไปได้ของตัวประมาณแตกต่างจากค่าประชากรหรือค่าจริงเพียงใด

1.7 ขั้นตอนการดำเนินงานวิจัย

ขั้นตอนการดำเนินการวิจัยมีดังนี้ คือ

1.สร้างข้อมูลให้มีการแจกแจงแบบปกติที่มีค่าผิดปกติในลักษณะต่างๆ และขนาดตัวอย่างตามที่ต้องการศึกษา

2. คำนวณค่าตัวประมาณด้วยวิธีการประมาณค่าทั้ง 3 วิธี

3. คำนวณความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณค่า

4. เปรียบเทียบความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณค่าในแต่ละสถานการณ์

5.สรุปผลการวิจัยในแต่ละสถานการณ์

1.8 เกณฑ์การตัดสินใจ

เพื่อศึกษาความเที่ยงตรงของตัวประมาณค่าพารามิเตอร์แสดงตำแหน่ง และพารามิเตอร์แสดงสเกล จะทำการพิจารณาจากความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) โดยเกณฑ์การพิจารณาความ

เที่ยงตรงกำหนดว่า ถ้าตัวประมาณค่าที่ได้จากวิธีใดให้ความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำกว่า จะถือว่าตัวประมาณค่านั้นเป็นตัวประมาณค่าที่มีความเที่ยงตรงสูงกว่าตัวประมาณค่าอื่นๆ โดยสูตรการคำนวณความคลาดเคลื่อนกำลังสองเฉลี่ยเป็นดังนี้

$$MSE = \frac{\sum_{i=1}^{500}(\hat{\theta}_i - \theta)^2}{500}$$

เมื่อ θ แทน ค่าพารามิเตอร์ที่ต้องการประมาณ

$\hat{\theta}_i$ แทนค่าประมาณของพารามิเตอร์ θ ในรอบที่ i

1.9 ประโยชน์ที่คาดว่าจะได้รับ

ในงานวิจัยครั้งนี้ คาดว่าประโยชน์ที่จะได้รับคือ

1.จากผลการวิจัยครั้งนี้จะเป็นแนวทางในการเลือกตัวประมาณค่าพารามิเตอร์แสดงตำแหน่ง (μ) และพารามิเตอร์แสดงสเกล (σ)ของการแจกแจงแบบปกติเมื่อมีค่าผิดปกติได้เหมาะสมในแต่ละสถานการณ์

2. เพื่อใช้เป็นแนวทางในการศึกษาเปรียบเทียบกับตัวประมาณอื่นๆ ต่อไป