การวิเคราะห์ข้อมูลด้วยวิธี Exploratory

*ปรีชา วิจิตรธรรมรส**

1. คำนำ

การวิเคราะห์ข้อมูลในงานวิจัยคงจะไม่มีใครปฏิเสธว่า สถิติพรรณนาหรือเทคนิคการวิเคราะห์ข้อมูลเบื้องต้นด้วยค่าวัดแนวโน้มสู่ส่วนกลาง ค่าวัดการกระจาย แผนภาพฮิสโตแกรมและตารางแจกแจงความถี่เป็นวิธีการทางสถิติที่จำเป็นและนิยมใช้กันในการวิเคราะห์ข้อมูลเบื้องต้นก่อนที่จะใช้เทคนิคการวิเคราะห์อื่น ๆ ทางสถิติ ถึงแม้ว่าเทคนิคการวิเคราะห์เบื้องต้นจะเป็นที่รู้จักกันดีแต่ก็ยังมีปัญหาอยู่บ้างในการวิเคราะห์ข้อมูลด้วยเทคนิคดังกล่าวเนื่องจากเทคนิคบางเทคนิคมีความยุ่งยากในการคำนวณและการพรรณนาลักษณะของข้อมูลก็ยังมีข้อบกพร่อง ตัวอย่างเช่น แผนภาพฮีสโตแกรมในการสร้างมีความยุ่งยากพอสมควรเพราะจะต้องมีการกำหนดจำนวนและความกว้างของช่วงข้อมูล กำหนดค่าขอบเขตของช่วงข้อมูลแต่ละช่วงแล้วจึงทำการแจงนับและสร้างแผนภาพ การคำนวณค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานซึ่งค่าที่คำนวณได้อาจจะไม่ใช่ค่าที่เหมาะสมในการสรุปลักษณะของข้อมูลหากข้อมูลมีค่าผิดปกติ (Outliers) เป็นต้น เพื่อลดปัญหาที่อาจเกิดขึ้นจากการวิเคราะห์ด้วยสถิติพรรณนาดังกล่าวผู้วิเคราะห์อาจจะใช้วิธีการพรรณนาข้อมูลด้วยวิธี Exploratory (Exploratory Data Analysis--EDA) EDA เป็นเทคนิคหนึ่งทางสถิติที่สามารถใช้ในการวิเคราะห์ข้อมูลซึ่งเทคนิคที่มีการใช้กันมากคือ การพรรณนาข้อมูลโดยอาศัยแผนภาพเข้าช่วยซึ่งนับเป็นข้อดีประการหนึ่งของ EDA เนื่องจากการใช้แผนภาพในการพรรณนาข้อมูลจะช่วยให้ผู้รับข่าวสารจากการวิเคราะห์สามารถเข้าใจลักษณะของ ข้อมูลได้โดยง่าย นอกจากนี้การคำนวณและการสร้างแผนภาพก็มิได้ยุ่งยาก

ความจริงแล้ว EDA เป็นเทคนิคที่ถูกคิดค้นมานานแล้วโดย John W. Tukey ในปี 1977 (Tukey, 1977) ซึ่งในยุคนั้นยังไม่มีคอมพิวเตอร์ใช้กันเหมือนปัจจุบัน เทคนิค EDA ที่ Tukey คิดขึ้นจึงเป็นการพรรณนาข้อมูลแบบง่าย ๆ ไม่ต้องใช้การคำนวณที่ยุ่งยากแต่มีประสิทธิภาพในการนำเสนอข้อมูลได้ดี จะเห็นได้จากในปัจจุบันถึงแม้มีการนำคอมพิวเตอร์เข้ามาใช้ช่วยในการวิเคราะห์ข้อมูลทางสถิติที่มีสูตรการคำนวณยุ่งยาก เทคนิค EDA กลับได้รับความนิยมมีการนำมาใช้กันแพร่หลายมากขึ้น โปรแกรมสำเร็จ

---

* ผู้ช่วยศาสตราจารย์ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์

รูปทางสถิติต่าง ๆ ที่มีการนำมาใช้กันแพร่หลายเช่น SAS SPSS MINITAB เป็นต้น ได้มีการพัฒนาให้สามารถวิเคราะห์ข้อมูลด้วยเทคนิค EDA

ในที่นี้จะกล่าวถึงเทคนิค EDA ซึ่งมีการใช้กันมาก 2 เทคนิค คือ แผนภาพลำต้นและใบ (Stem and Leaf Display) และแผนภาพกล่อง (Box Plots or Box-and-Whisker Plot) เทคนิคทั้งสองเหมาะสำหรับการพรรณนาข้อมูลเชิงปริมาณ (Quantitative Data) หรือข้อมูลมาตรวัดแบบช่วงและอัตราส่วน (Interval and Ratio Scales)

2. แผนภาพลำต้นและใบ (Stem and Leaf Display)

แผนภาพลำต้นและใบเป็นแผนภาพที่ใช้พรรณนาลักษณะข้อมูลโดยมีลักษณะคล้ายกับการพรรณนาข้อมูลโดยใช้ฮิสโตแกรม แต่การพรรณนาข้อมูลโดยใช้แผนภาพลำต้นและใบจะสูญเสียข่าวสาร (Information) ที่มีอยู่ในข้อมูลเดิม (Raw Data) น้อยกว่าการใช้ฮิสโตแกรม (Histogram) เหตุที่เรียกแผนภาพนี้ว่าลำต้นและใบเนื่องจากในการสร้างแผนภาพลำต้นและใบจะมีการแบ่งข้อมูลแต่ละค่าออกเป็น 2 ส่วน เรียกว่าส่วนลำต้นและส่วนของใบ ตัวอย่างเช่น สมมุติว่ามีข้อมูลอายุพนักงานอยู่ 16 คน ดังนี้

18, 22, 32, 21, 19, 40, 64, 57, 41, 29, 28, 29, 39, 38, 56, 33

การสร้างแผนภาพลำต้นและใบมีขั้นตอนดังนี้

1. เลือกหน่วยของลำต้นและใบ ในที่นี้ข้อมูลเป็นเลข 2 หลัก ดังนั้นจะเลือกเลขหลักสิบเป็นลำต้นและเลขหลักหน่วยเป็นใบ
2. ในส่วนของลำต้นจะเห็นว่าค่าต่ำสุดและสูงสุดของเลขหลักสิบในข้อมูลคือ 1 และ 6 ตามลำดับ การสร้างลำต้นทำได้โดยการเขียนเลขลำต้นเรียงลำดับจากค่าน้อยไปหาค่ามาก (หรืออาจจะเรียงลำดับจากค่ามากไปหาค่าน้อยก็ได้) จะได้ลำต้นดังนี้

ลำต้น

1
2
3
4
5
6

3. อ่านค่าข้อมูลครั้งละ 1 ค่า แล้วเติมใบให้กับลำต้น เช่น ข้อมูลตัวแรกคือ 18 ที่ค่าลำต้นเท่ากับ 1 ให้เติมใบให้ลำต้น 1 มีค่าเท่ากับ 8 ข้อมูลตัวถัดไปคือ 22 ที่ค่าลำต้นเท่ากับ 2 ให้เติม

ใบให้ลำต้น 2 มีค่าเท่ากับ 2 ทำอย่างนี้จนกระทั่งครบจำนวนข้อมูลทั้ง 16 ค่าจะได้แผนภาพดังนี้

ลำต้น	ใบ
1	89
2	21989
3	2983
4	01
5	76
6	4

4. เพื่อให้แผนภาพมีระเบียบยิ่งขึ้นให้เรียงค่าตัวเลขที่เป็นส่วนของใบในแต่ละค่าของลำต้นใหม่ โดยเรียงจากน้อยไปหามาก จะได้แผนภาพลำต้นและใบ ดังนี้

ความถี่	ลำต้น	ใบ
2	1	89
5	2	12899
4	3	2389
2	4	01
2	5	67
1	6	4

5. ในแต่ละค่าของลำต้นให้นับความถี่ของข้อมูล (หรือจำนวนใบ) แล้วใส่ค่าความถี่ไว้ทางซ้ายมือของลำต้น ก็จะได้แผนภาพที่สมบูรณ์

จากแผนภาพลำต้นและใบที่ได้จะเห็นว่ามีลักษณะของการใช้สถิติพรรณนาที่เป็นที่รู้จักกันทั่วไป คือ ฮีสโตแกรมและตารางแจกแจงความถี่ ในตัวอย่างดังกล่าวแผนภาพลำต้นและใบเหมือนการนำเสนอข้อมูลด้วยฮีสโตแกรมที่มีการกำหนดช่วงของข้อมูลออกเป็นช่วง 6 ช่วง แต่ละช่วงมีความกว้างเท่ากับ 10 โดยช่วงที่ 1 มีค่าของข้อมูลระหว่าง 10 ถึง 19 ช่วงที่ 2 มีค่าของข้อมูลระหว่าง 20 ถึง 29 เป็นต้น

ข้อดีของแผนภาพลำต้นและใบคือ แผนภาพยังคงแสดงค่าของข้อมูลดั้งเดิมอยู่ ขณะที่ถ้าใช้ฮีสโตแกรมในการนำเสนอจะมีการรวมข้อมูลเข้าเป็นช่วง ๆ เหมือนกันแต่ค่าของข้อมูลเดิมจะไม่ได้แสดงในแผนภาพฮีสโตแกรม ทำให้สูญเสียสารสนเทศบางส่วนของข้อมูลไป ตัวอย่างเช่น ถ้าใช้ฮีสโตแกรมจะบอกได้แต่เพียงว่า ค่าต่ำสุดของข้อมูลคือ 10 (ค่าเริ่มต้นของช่วงที่ 1) แต่หากใช้แผนภาพลำต้นและใบดังตัวอย่างข้างต้น จะเห็นว่าค่าต่ำสุดของข้อมูลคือ 18 นอกจากนี้แผนภาพลำต้นและใบยังสามารถช่วยในการ

คำนวณค่าวัดแนวโน้มสู่ส่วนกลางโดยเฉพาะค่าวัดตำแหน่งที่ได้ง่ายขึ้นเนื่องจากมีการเรียงลำดับของข้อมูลแล้ว เช่น จากแผนภาพจะเห็นว่า ค่าฐานนิยมเท่ากับ 29 และค่ามัธยฐานเท่ากับ 32.5 โดยคำนวณได้จากค่าเฉลี่ยระหว่าง 32 และ 33 ซึ่งเป็นค่าของข้อมูลที่อยู่ในตำแหน่งที่ 8 และ 9 เมื่อมีการเรียงลำดับข้อมูลแล้ว

นอกจากการใช้ความกว้างของช่วงเท่ากับ 10 แล้ว การสร้างแผนภาพลำต้นและใบอาจใช้ลำต้นซึ่งมีความกว้างของแต่ละช่วงเท่ากับ 5 โดยช่วงที่ 1 มีค่าของข้อมูลระหว่าง 10 ถึง 14 ช่วงที่ 2 มีค่าข้อมูลระหว่าง 15 ถึง 19 ในกรณีที่มีปริมาณข้อมูลมาก การสร้างแผนภาพก็อาจจะใช้ความกว้างของลำต้นเท่ากับ 2 โดยในการระบุค่าของข้อมูลในแต่ละช่วงจะใช้สัญลักษณ์

- แทนลำต้นที่มีใบเท่ากับ 0 และ 1
- t แทนลำต้นที่มีใบเท่ากับ 2 และ 3 (two and three)
- f แทนลำต้นที่มีใบเท่ากับ 4 และ 5 (four and five)
- s แทนลำต้นที่มีใบเท่ากับ 6 และ 7 (six and seven) และ
- . แทนลำต้นที่มีใบเท่ากับ 8 และ 9

จากข้อมูลข้างต้นหากสร้างแผนภาพโดยใช้ความกว้างของช่วงเท่ากับ 5 จะได้แผนภาพลำต้นและใบดังนี้

ลำต้น	ใบ
1*	
1·	89
2*	12
2·	899
3*	23
3·	89
4*	01
4·	
5*	
5·	67
6	4

ถ้าใช้ความกว้างของช่วงข้อมูลเท่ากับ 2 ลักษณะของแผนภาพในช่วงของข้อมูลระหว่าง 10 ถึง 19 จะเป็นดังนี้

ลำต้น	ใบ
1	*
1	t
1	f
1	s
1	.

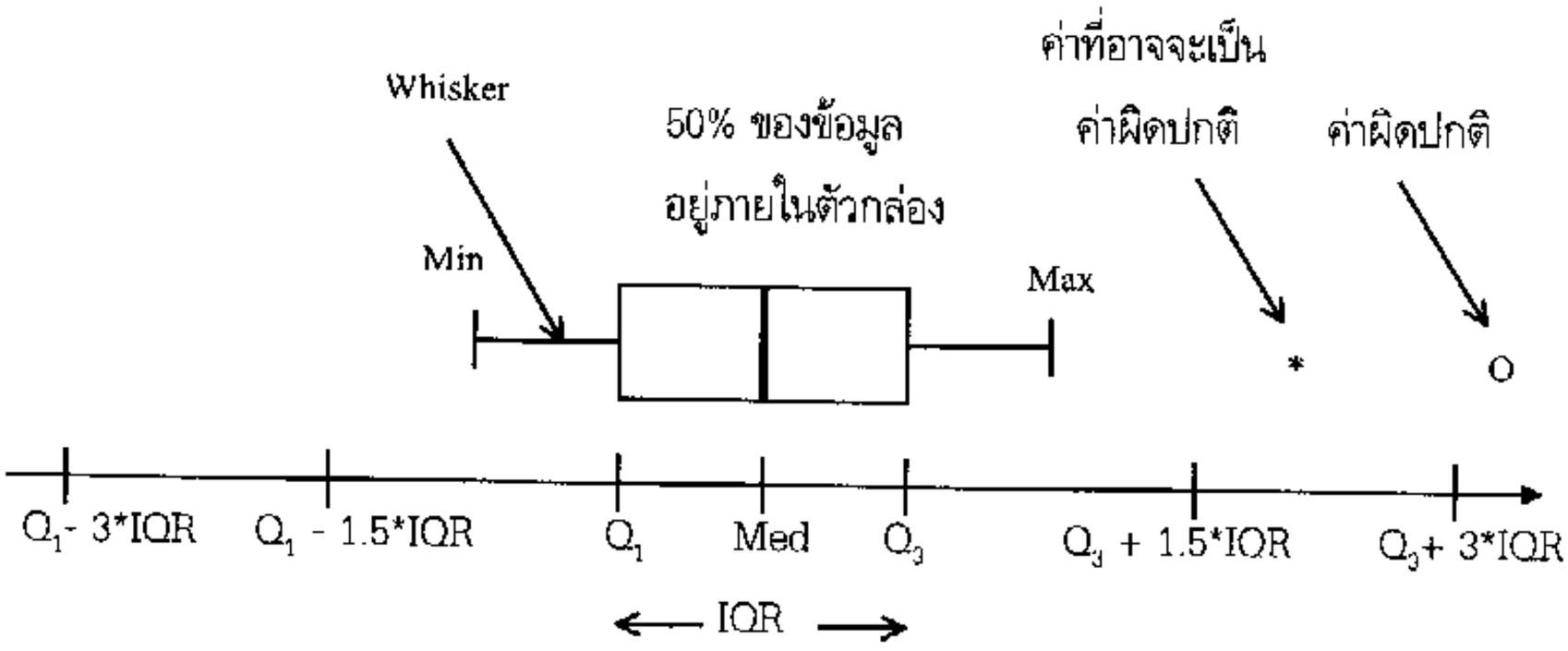
89

หากข้อมูลมีหลายหลักการแบ่งค่าตัวเลขเพื่อสร้างลำต้นอาจใช้เลขมากกว่า 1 หลักโดยใช้เลขหลักสุดท้ายเป็นใบ เช่น ข้อมูลประกอบด้วยค่า 105 106 107 107 109 การสร้างแผนภาพโดยใช้เลขหลักร้อยและหลักสิบเป็นลำต้น และเลขหลักหน่วยเป็นใบ จะได้แผนภาพดังนี้

ลำต้น	ใบ
10	56779

3. แผนภาพกล่อง (Box Plot or Box-and-Whisker Plot)

แผนภาพกล่องเป็นการพรรณนาข้อมูลโดยใช้ค่า 5 ค่าในการสรุปลักษณะข้อมูล (Five Summary Measures) ได้แก่ ค่าควอไทล์ที่ 1 (Q_1) ค่าควอไทล์ที่ 2 หรือค่ามัธยฐาน (Q_2 or Med) ค่าควอไทล์ที่ 3 (Q_3) ค่าต่ำสุด (Min) และค่าสูงสุด (Max) ของข้อมูล การนำเสนอลักษณะของข้อมูลจะนำเสนอเป็นแผนภาพซึ่งประกอบด้วยตัวกล่อง (Box) และทางหรือหนวดของกล่อง (Whiskers) ลักษณะของแผนภาพเป็นดังนี้



การสร้างแผนภาพกล่องมีขั้นตอนดังนี้

1. คำนวณค่าวัดตำแหน่งที่ 3 ค่าคือ ค่าควอไทล์ที่ 1, 2 และ 3 เพื่อสร้างตัวกล่อง
2. คำนวณค่าพิสัยระหว่างควอไทล์ (Interquartile Range: IQR)
3. คำนวณค่าขอบเขตภายในของแผนภาพ (Inner Fences) โดย

   ค่าขอบเขตภายในด้านล่าง (Lower Inner Fence) เท่ากับ Q_1 - 1.5*IQR

   ค่าขอบเขตภายในด้านบน (Upper Inner Fence) เท่ากับ Q_3 + 1.5*IQR
4. คำนวณค่าขอบเขตภายนอกของแผนภาพ (Inner Fences) โดย

   ค่าขอบเขตภายนอกด้านล่าง (Lower Outer Fence) เท่ากับ Q_1 - 3*IQR

   ค่าขอบเขตภายนอกด้านบน (Upper Outer Fence) เท่ากับ Q_3 + 3*IQR
5. วาดหางของกล่องโดยลากเส้นหางของกล่องทั้งสองข้างยาวไปจนถึงข้อมูลที่มีค่าต่ำสุดและสูงสุดซึ่งอยู่ภายในช่วงของขอบเขตภายใน
6. สำหรับค่าของข้อมูลที่มีค่าสูงหรือต่ำกว่าค่าขอบเขตภายในแต่ยังคงอยู่ในช่วงของขอบเขตภายนอกให้ถือว่า ค่าของข้อมูลดังกล่าวอาจจะเป็นค่าสูงหรือต่ำผิดปกติ (Suspected Outlier) ให้แสดงค่าดังกล่าวด้วยสัญลักษณ์ * หากค่าของข้อมูลมีค่าสูงหรือต่ำผิดปกติมาก กล่าวคือ มีค่าอยู่นอกช่วงขอบเขตภายนอกให้ถือว่า ค่าดังกล่าวเป็นค่าผิดปกติ (Outlier) ซึ่งจะแสดงบนแผนภาพด้วยสัญลักษณ์ O

จากตัวอย่างข้อมูล ข้อมูลอายุพนักงาน 16 คน สามารถสร้างแผนภาพกล่องได้ดังนี้

1. คำนวณหาค่า ควอไทล์ที่ 1, 2 และ 3

$$Q_1 = (22 + 28)/2 = 25 \qquad Q_2 = (32 + 33)/2 = 32.5$$

$$Q_3 = (40 + 41)/2 = 40.5$$

2. คำนวณค่าพิสัยระหว่างควอไทล์

$$IQR = Q_3 - Q_1 = 15.5$$

3. คำนวณค่าขอบเขตภายในและภายนอก

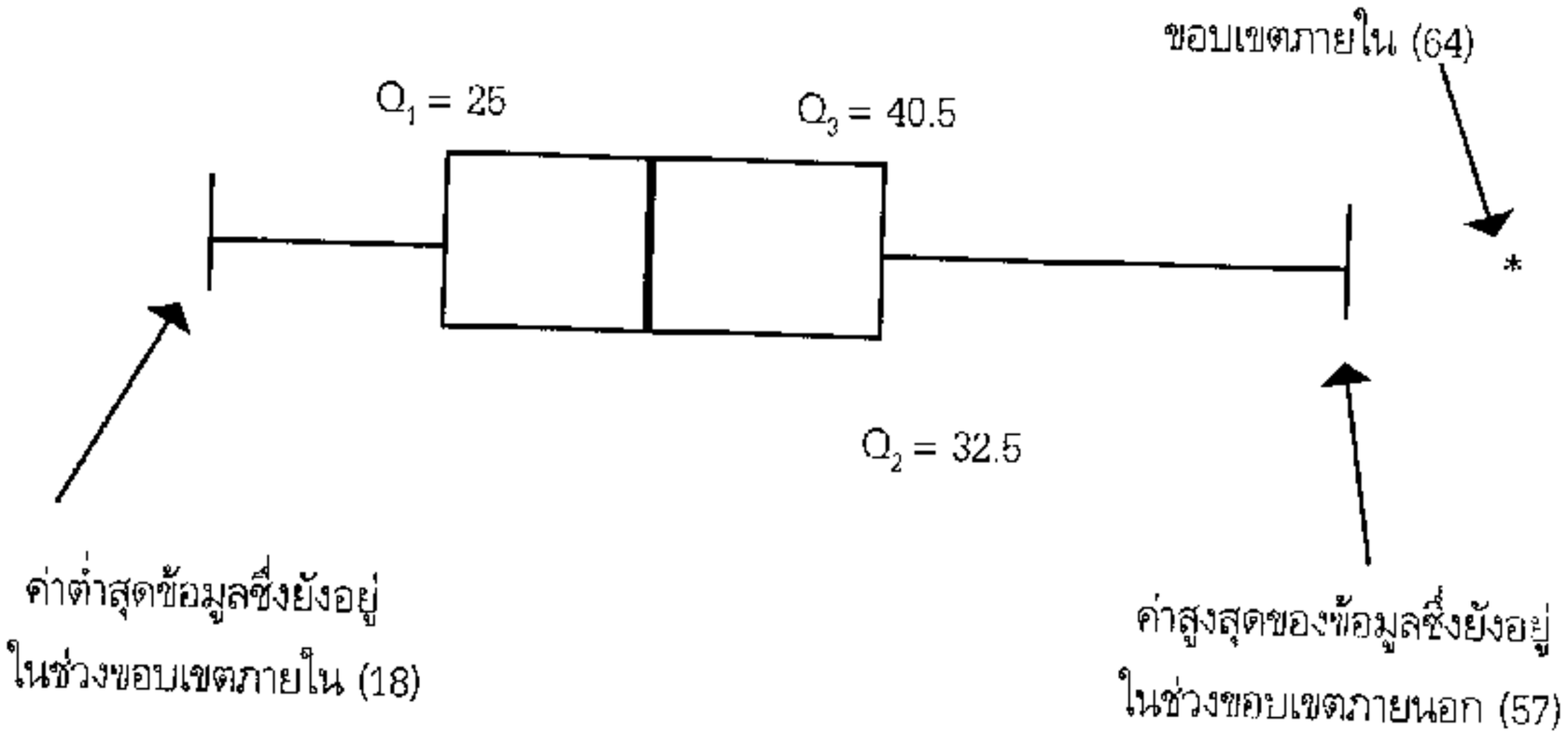
ค่าขอบเขตภายในด้านล่าง = Q_1 - 1.5*IQR = 1.75

ค่าขอบเขตภายในด้านบน = Q_3 + 1.5*IQR = 63.75

ค่าขอบเขตภายนอกด้านล่าง = Q_1 - 3*IQR = -21.5

ค่าขอบเขตภายนอกด้านบน = Q_3 + 3*IQR = 87

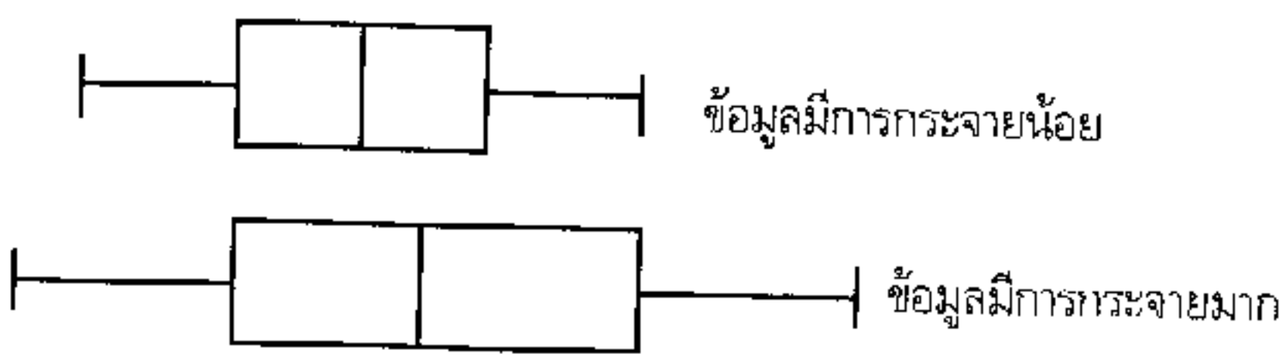
4. สร้างแผนภาพกล่องได้ดังนี้



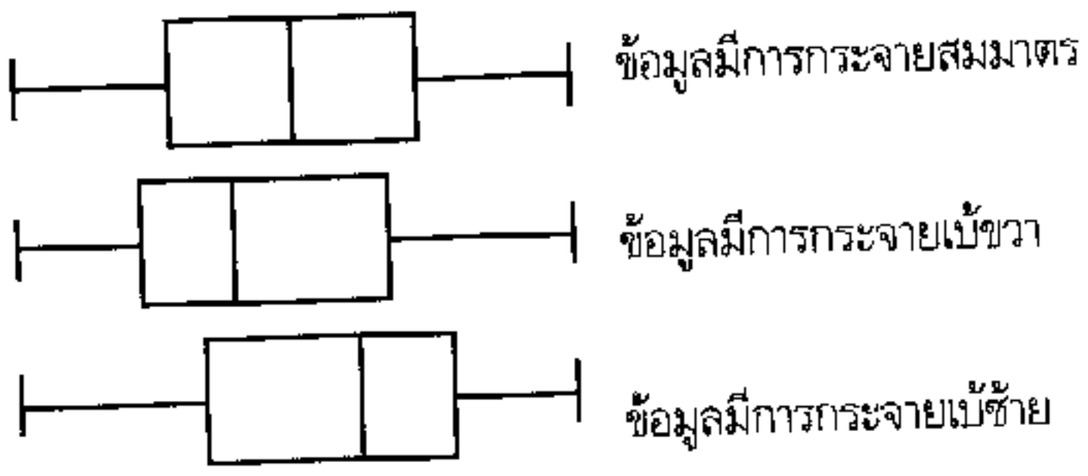
จะเห็นว่าแผนภาพกล่องมีประโยชน์ในการพรรณนาข้อมูลดังนี้

1. แสดงค่าวัดตำแหน่งที่ ควอไทล์ โดยจากตัวกล่องจะทราบว่า 50 เปอร์เซนต์หรือครึ่งหนึ่งของข้อมูลอยู่ภายในตัวกล่อง จากตัวอย่างจะสรุปได้ว่า 50 เปอร์เซนต์ของพนักงานมีอายุอยู่ในช่วง 20 ถึง 40.5 ปี

2. แสดงลักษณะการกระจายของข้อมูล โดยพิจารณาได้จากความยาวของกล่องและหาง ถ้าหากมีความยาวมากแสดงว่า ข้อมูลมีการกระจายมาก ดังภาพตัวอย่าง



3. แสดงลักษณะการกระจายของข้อมูล (Distribution) โดยพิจารณาจากตัวกล่อง หากค่ามัธยฐานหรือเส้นกลางของกล่องอยู่ตรงกึ่งกลางระหว่างค่าควอไทล์ที่ 1 และ 3 หรือขอบกล่องทั้งสองด้านแสดงว่า ข้อมูลมีลักษณะการกระจายที่มีสมมาตร (Symmetrically Distributed) ถ้าเส้นกลางของกล่องอยู่ค่อนไปทางด้านซ้ายของขอบกล่องหรือค่าควอไทล์ที่ 1 แสดงว่า ข้อมูลมีการกระจายเบ้ขวา (Right-Skewed) ในทำนองกลับกันหากเส้นกลางของกล่องอยู่ค่อนไปทางด้านขวาของขอบกล่องหรือค่าควอไทล์ที่ 3 แสดงว่า ข้อมูลมีการกระจายเบ้ซ้าย (Left-Skewed) ดังภาพตัวอย่าง



4. แสดงค่า สูง - ต่ำ ผิดปกติของข้อมูลโดยพิจารณาจากขอบเขตภายในและขอบเขตภายนอก

จะเห็นว่าแผนภาพกล่องเป็นแผนภาพที่สามารถใช้แสดงลักษณะของข้อมูลซึ่งมีประโยชน์มากโดยเฉพาะอย่างยิ่งเมื่อต้องการเปรียบเทียบลักษณะของข้อมูลก่อนที่จะทำการทดสอบสมมติฐานเปรียบค่าเฉลี่ยของ 2 ประชากรหรือการวิเคราะห์ความแปรปรวน แผนภาพดังกล่าวจะแสดงให้เห็นว่าลักษณะของข้อมูลจำแนกตามกลุ่มประชากรมีลักษณะการกระจายอย่างไร มีค่าผิดปกติหรือไม่

บรรณานุกรม

Tukey, J.W. 1977. **Exploratory Data Analysis**. New York: Addison-Wesley.