ใบความรู้
หน่วยการเรียนรู้ที่ 2 เรื่อง การวิเคราะห์ข้อมูลเบื้องต้น

การวัดการกระจายของข้อมูล

การวัดการกระจายของข้อมูล แบ่งได้เป็น 2 วิธี คือ

1. การวัดการกระจายสัมบูรณ์
2. การวัดการกระจายสัมพัทธ์

การวัดการกระจายสัมบูรณ์ หมายถึง การวัดการกระจายของข้อมูลเพียงชุดเดียว เพื่อดูว่าค่าจากการสังเกตแต่ละค่าของข้อมูล มีความแตกต่างกันมากน้อยเพียงไร

การวัดการกระจายสัมบูรณ์ที่นิยมใช้มี 4 แบบ ได้แก่ พิสัย ส่วนเบี่ยงเบนควอร์ไทล์ ส่วนเบี่ยงเบนเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน

1) พิสัย (Range) พิสัยของข้อมูลคือ ค่าที่ใช้วัดการกระจายของข้อมูลที่ได้จากผลต่างระหว่างค่าจากการสังเกตที่มีค่าสูงสุดและต่ำสุด

ดังนั้น

$$\text{พิสัย} = x_{max} - x_{min}$$

เมื่อ x_{max} เป็นค่าจากการสังเกตที่มีค่าสูงสุด

x_{min} เป็นค่าจากการสังเกตที่มีค่าต่ำสุด

2) ส่วนเบี่ยงเบนควอร์ไทล์ (Quratile Deviation) เขียนแทนด้วยสัญลักษณ์ Q.D. คือค่าที่ใช้วัดการกระจายของข้อมูล ซึ่งเท่ากับครึ่งหนึ่งของผลต่างระหว่าง ควอร์ไทล์ที่ 3 และควอร์ไทล์ที่ 1

ดังนั้น

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

3) ส่วนเบี่ยงเบนเฉลี่ย (Mean Deviation) เขียนแทนด้วยสัญลักษณ์ M.D. คือค่าที่ใช้วัดการกระจายของข้อมูลที่ได้จากการเฉลี่ยค่าสัมบูรณ์ของผลต่างของค่าจากการสังเกตแต่ละค่ากับค่ากลางของข้อมูล (ค่ากลางที่ใช้ อาจเป็นค่าเฉลี่ยเลขคณิตหรือมัธยฐาน แต่ที่นิยมใช้คือ ค่าเฉลี่ยเลขคณิต)

ดังนั้น

$$\text{M.D.} = \frac{\sum |x - \bar{x}|}{N}$$

เมื่อ x แทน ค่าในข้อมูล

N แทน จำนวนค่าในข้อมูล

$\bar{x}$ แทน ค่าเฉลี่ยเลขคณิตของข้อมูล

4) ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) เขียนแทนด้วยสัญลักษณ์ s หรือ S.D. คือ รากที่สองที่ไม่เป็นจำนวนลบของค่าเฉลี่ยของกำลังสองของผลต่างระหว่างค่าในข้อมูลกับค่าเฉลี่ยเลขคณิตของข้อมูลนั้น

ดังนั้น

$$S.D. = \sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{N}}$$

เราสามารถปรับสูตร S.D. เป็นสูตรใหม่ ซึ่งสามารถหา S.D. ได้เร็วกว่า คือ

$$S.D. = \sqrt{\frac{\sum x^2}{N}-\left(\bar{x}\right)^2}$$

เมื่อ x แทนแต่ละค่าในข้อมูล N เป็นจำนวนข้อมูล และ $\bar{x}$ แทนค่าเฉลี่ยเลขคณิตของข้อมูล

หมายเหตุ

1) ในกรณีที่ $\bar{x}$ เป็นจำนวนเต็ม ควรใช้สูตรที่ 1 แต่ถ้า $\bar{x}$ เป็นทศนิยม หรือเศษส่วน ควรใช้สูตรที่ 2
2) ส่วนเบี่ยงเบนมาตรฐาน ถือว่าเป็นวิธีวัดการกระจายได้ดีที่สุด เนื่องจากต้องใช้ข้อมูลทุกค่า หรือ มีตัวแทนของข้อมูลทุกค่ามาคำนวณ และขจัดปัญหาในการที่ต้องใช้ค่าสัมบูรณ์ให้หมดไป

สมบัติบางประการของส่วนเบี่ยงเบนเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน

1. ส่วนเบี่ยงเบนเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานของข้อมูลใด ๆ จะต้องเป็นจำนวนจริงบวกหรือศูนย์เสมอ และมีหน่วยเดียวกับค่าของข้อมูล
2. ส่วนเบี่ยงเบนเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานของข้อมูลจะเท่ากับศูนย์ เมื่อ ค่าทุกค่าในข้อมูลเท่ากันหมดและเท่ากับค่าเฉลี่ยเลขคณิตของข้อมูลชุดนั้น
3. ถ้านำจำนวนจริง b ไปบวกกับแต่ละค่าของข้อมูลเดิม ส่วนเบี่ยงเบนเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานของข้อมูลใหม่ จะเท่ากับส่วนเบี่ยงเบนเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานของข้อมูลเดิม
4. ถ้านำจำนวนจริง a ไปคูณแต่ละค่าในข้อมูลเดิมแล้ว ส่วนเบี่ยงเบนเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานของข้อมูลใหม่ จะเท่ากับ $|a|$ เท่า ของส่วนเบี่ยงเบนเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน

\

5. ถ้า x แทนค่าในข้อมูลชุดที่หนึ่ง และ y แทนค่าในข้อมูลชุดที่สอง โดยที่

$$y = ax + b$$

$$M.D._y = |a| M.D._x$$

$$S.D._y = |a| S.D._x$$

6. ถ้าคำนวณหาส่วนเบี่ยงเบนมาตรฐานโดยใช้ค่ากลางของข้อมูลอย่างอื่นที่ไม่ใช่ค่าเฉลี่ยเลขคณิต ส่วนเบี่ยงเบนมาตรฐานที่หาได้ จะมีค่ามากกว่าส่วนเบี่ยงเบนมาตรฐานที่หาได้จากค่าเฉลี่ยเลขคณิตเสมอ

ความแปรปรวน (Variance)

ทบทวนความรู้พื้นฐาน

ให้ $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$ เป็นค่าเฉลี่ยเลขคณิตของข้อมูลชุดที่ 1 , 2 , 3 , ...

$N_1, N_2, N_3, \ldots$ เป็นจำนวนข้อมูลของข้อมูลแต่ละชุดที่ 1 , 2, 3 , ...

$$\bar{x}_{รวม} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + N_3\bar{x}_3 + \ldots}{N_1 + N_2 + N_3 + \ldots}$$

ถ้า S.D. คือ ส่วนเบี่ยงเบนมาตรฐานของข้อมูล แล้ว $S.D.^2$ คือ ความแปรปรวนของข้อมูล

1) ถ้า $S.D. = \sqrt{\frac{\sum x^2}{N} - \bar{x}^2}$ แล้ว $S.D.^2 = \frac{\sum x^2}{N} - \bar{x}^2$

2) $S.D._{รวม}^2 = \frac{\sum x_{รวม}^2}{N_{รวม}} - \bar{x}_{รวม}^2$ (ในกรณีที่ข้อมูลสองชุดมี $\bar{x}$ ไม่เท่ากัน)

3) $S.D._{รวม}^2 = \frac{N_1 S.D._1 + N_2 S.D._2}{N_1 + N_2}$ (ในกรณีที่ข้อมูลสองชุดมี $\bar{x}$ เท่ากัน)

หมายเหตุ ความแปรปรวนไม่มีหน่วย

การวัดการกระจายสัมพัทธ์ หมายถึง การวัดการกระจายของข้อมูลมากกว่าหนึ่งชุด และนำข้อมูลแต่ละชุดมาเปรียบเทียบกันว่าข้อมูลชุดใดมีการกระจายมากกว่ากัน

การวัดการกระจายสัมพัทธ์ที่นิยมใช้มี 4 แบบ ได้แก่

1. สัมประสิทธิ์ของพิสัย $= \dfrac{x_{max} - x_{min}}{x_{max} + x_{min}}$
2. สัมประสิทธิ์ของส่วนเบี่ยงเบนควอร์ไทล์ $= \dfrac{Q_3 - Q_1}{Q_3 + Q_1}$
3. สัมประสิทธิ์ของส่วนเบี่ยงเบนเฉลี่ย $= \dfrac{M.D.}{\bar{x}}$
4. สัมประสิทธิ์ของการแปรผัน $= \dfrac{S.D.}{\bar{x}}$

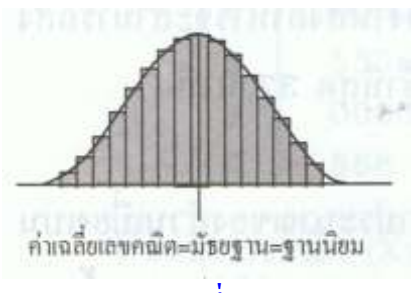
หมายเหตุ

1) ผลลัพธ์ของการกระจายสัมพัทธ์ จะเป็นลบ หรือ บวก หรือ ศูนย์ ก็ได้

2) สัมประสิทธิ์ของการแปรผัน เรียกอีกอย่างหนึ่งว่า สัมประสิทธิ์ของการกระจาย

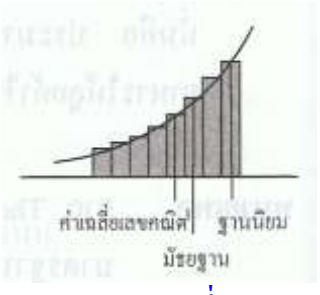
ความสัมพันธ์ระหว่างการแจกแจงความถี่ ค่ากลาง และการกระจายของข้อมูล

ลักษณะของการกระจายของข้อมูล อาจแบ่งได้เป็น 3 แบบ ให้พิจารณาจากฮิสโทแกรมต่อไปนี้



รูปที่ 1

รูปที่ 2



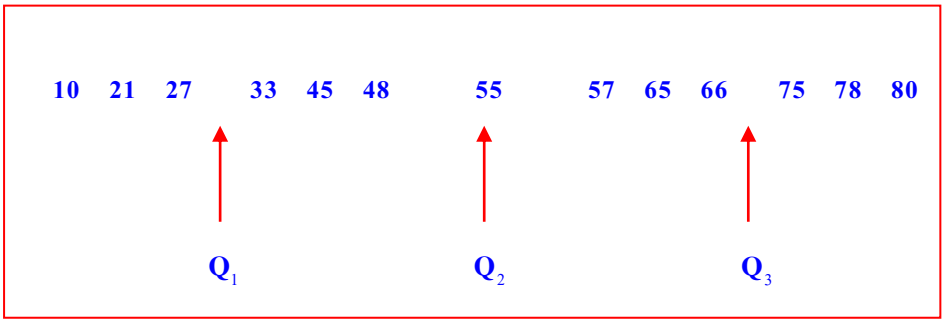
รูปที่ 3

รูปที่ 1 ลักษณะการกระจายของข้อมูลในแบบที่ 1 เป็นการกระจายแบบสมมาตร (Symmetric distribution) คือ ค่าเฉลี่ยเลขคณิต มัธยฐาน และฐานนิยมมีค่าเท่ากัน หรืออยู่ที่จุดเดียวกัน คือจุดที่มีความถี่สูงสุด

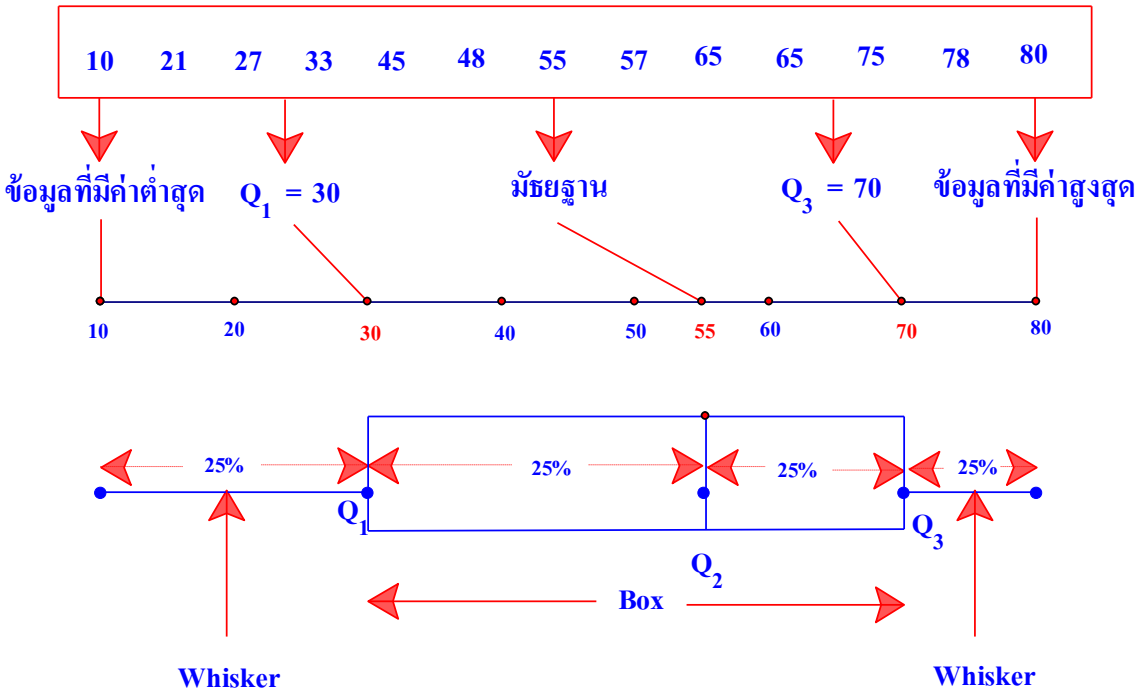
รูปที่ 2 เป็นการกระจายที่เบ้ทางขวา (Right distribution) แท่งสี่เหลี่ยมมุมฉากของฮิสโทแกรมที่มีความถี่น้อยและน้อยที่สุดอยู่ทางด้านขวา ค่าเฉลี่ยเลขคณิตจะมีค่ามากที่สุด รองลงมาเป็นมัธยฐาน และฐานนิยมตามลำดับ

รูปที่ 3 เป็นการกระจายแบบเบ้ทางซ้าย (Left distribution) แท่งสี่เหลี่ยมมุมฉากของฮิสโทแกรม ที่มีความถี่น้อยและน้อยที่สุดอยู่ทางด้านซ้าย ฐานนิยมจะมีค่ามากที่สุด รองลงมาคือมัธยฐาน และค่าเฉลี่ยเลขคณิตจะมีค่าน้อยที่สุด

นอกจากการวัดการกระจายของข้อมูลโดยใช้พิสัย และส่วนเบี่ยงมาตรฐานแล้ว ยังสามารถใช้มัธยฐานในการสร้างแผนภาพเพื่อวัดการกระจายของข้อมูลจากมัธยฐานซึ่งเป็นค่ากลางของข้อมูลได้ ดังตัวอย่างต่อไปนี้



จากข้อมูลข้างต้น เมื่อหาข้อมูลที่มีค่าต่ำสุด ข้อมูลที่มีค่าสูงสุด ควอร์ไทล์ที่ 1 ควอร์ไทล์ที่ 2 และควอร์ไทล์ที่ 3 แล้ว จะสร้างแผนภาพที่เรียกว่า box - and – whisker plot หรือ box plot ซึ่งจะเรียกว่า แผนภาพกล่อง ได้ดังนี้



ตัวอย่างที่ 1 จากการตรวจปริมาณน้ำตาล (กรัม) ต่อปริมาณอาหาร 100 กรัม ของอาหารชนิดหนึ่ง จำนวน 31 จาน ได้ข้อมูลดังนี้

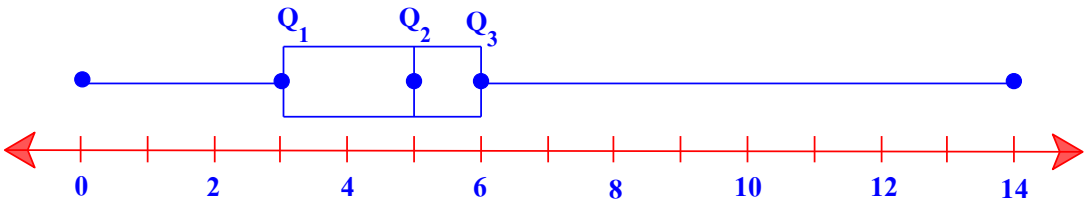
0 0 0 0 2 2 2 3 3 3 3 3 3 4 5 5 5 5 5 6 6 6 6 6 7 9 10 11 12 12 14

Q_1 มัธยฐาน Q_3

ข้อมูลข้างต้นมี 0 เป็นค่าต่ำสุด และ 14 เป็นค่าสูงสุด

Q_1 = 3 มัธยฐาน = 5 Q_3 = 6

เขียนแผนภาพกล่องของข้อมูลข้างต้นได้ดังนี้



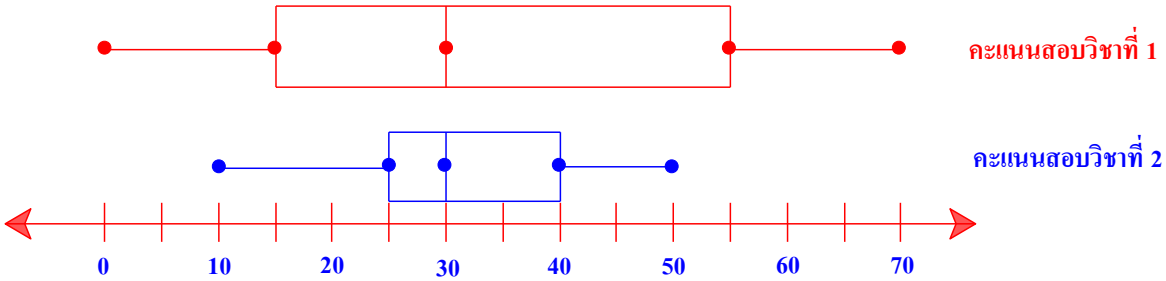
จากแผนภาพพบว่า

ข้อมูลที่มีค่าอยู่ระหว่าง 0 ถึง 3 และ 6 ถึง 14 มีเท่ากัน คือ ประมาณ 25% ของจำนวนข้อมูลทั้งหมด

ข้อมูลที่มีค่าอยู่ระหว่าง 3 ถึง 6 มีประมาณ 50% ของจำนวนข้อมูลทั้งหมด

ข้อมูลที่อยู่ระหว่าง Q_1 และ Q_2 มีการกระจายมากกว่าข้อมูลที่อยู่ระหว่าง Q_2 และ Q_3 แต่ข้อมูลที่อยู่ระหว่าง Q_3 ถึงค่าที่มากที่สุด มีการกระจายมากที่สุด

ตัวอย่างที่ 2 พิจารณาการกระจายของข้อมูลสองชุด ซึ่งเป็นคะแนนสอบของวิชาที่ 1 และวิชาที่ 2 ซึ่งมีคะแนนเต็ม 100 คะแนนเท่ากัน โดยใช้แผนภาพกล่องต่อไปนี้



จากแผนภาพข้างต้นจะพบว่า

ข้อมูลทั้งสองชุดมีมัธยฐานเท่ากันแต่มีการกระจายที่แตกต่างกัน

ข้อมูลชุดที่ 1 มีการกระจายของข้อมูลมากกว่าข้อมูลชุดที่ 2

ดังจะเห็นว่า คะแนนสอบวิชาที่ 1 มีค่าอยู่ระหว่าง 0 – 70 คะแนน

แต่ คะแนนสอบวิชาที่ 2 มีค่าอยู่ระหว่าง 10 – 50 คะแนน

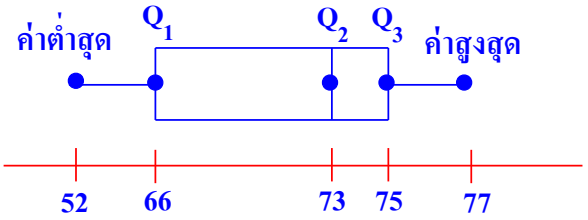
ตัวอย่างที่ 3 จากข้อมูลในแผนภาพต้น – ใบ สามารถนำข้อมูลดังกล่าวมาสร้างแผนภาพกล่องได้ดังนี้

5	2	4 5 6
6	6	6 7 9
7	1	2 2 3 3 4 4 4 4 5 5 6 6 7 7

จากแผนภาพต้น – ใบ จะได้ ค่าต่ำสุด คือ 52 ค่าสูงสุด คือ 77

และ $Q_1 = 66$, $Q_2 = 73$ และ $Q_3 = 75$

จากข้อมูลข้างต้นนำมาสร้างแผนภาพกล่องได้ดังนี้



จากแผนภาพจะเห็นว่า ข้อมูลที่อยู่ในช่วง Q_2 และ Q_3 มีค่าใกล้เคียงกัน

แต่ข้อมูลที่อยู่ในช่วง Q_1 และ Q_2 มีการกระจายมากกว่า

.....