บทที่ 2 สถิติเชิงพรรณนา

2.1 การวัดแนวโน้มเข้าสู่ส่วนกลาง

การวัดแนวโน้มเข้าสู่ส่วนกลางเป็นวิธีการทางสถิติเชิงพรรณนาที่ใช้ในการหาค่ากลางหรือค่าเฉลี่ยเพื่อใช้เป็นตัวแทนแสดงขนาดและลักษณะของข้อมูลชุดนั้น ประโยชน์ของการวัดแนวโน้มเข้าสู่ส่วนกลางคือทำให้ได้ตัวแทนของข้อมูลที่เป็นตัวเลขจำนวนเดียวที่แทนค่าทั้งหมดของข้อมูลชุดนั้นมาเสนอรายงาน โดยไม่จำเป็นต้องนำข้อมูลทั้งชุดมาพิจารณา ค่าวัดแนวโน้มเข้าสู่ส่วนกลางมีอยู่หลายค่าด้วยกัน เช่น ค่าเฉลี่ย ค่ามัธยฐาน และค่าฐานนิยม เป็นต้น

2.1.1 ค่าเฉลี่ย (Mean)

เป็นค่ากลางแบบหนึ่งที่นิยมใช้เป็นตัวแทนของข้อมูล เนื่องจากใช้ข้อมูลทุกตัวในการคำนวณ ซึ่งคำนวณได้จากนำข้อมูลทุกตัวมารวมกันแล้วหารด้วยจำนวนข้อมูลทั้งหมด ดังสูตรต่อไปนี้

ค่าเฉลี่ยของตัวอย่าง ($\bar{x}$)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

ค่าเฉลี่ยของประชากร (μ)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

เมื่อ n คือ จำนวนข้อมูลทั้งหมดในตัวอย่าง
N คือ จำนวนข้อมูลทั้งหมดในประชากร

ตัวอย่างที่ 2.1 จงหาค่าเฉลี่ยของข้อมูลตัวอย่างต่อไปนี้ 16, 8, 15, 12, 14, 7

$$\bar{x} = \frac{16 + 8 + 15 + 12 + 14 + 7}{6} = 12$$

คำสั่ง R

```
> age=c(14,16,14,17,16,14,18,17)
> mean(age)
[1] 15.75
```

หมายเหตุ ฟังก์ชัน mean() ใช้หาค่าเฉลี่ยเลขคณิตสำหรับข้อมูลเชิงปริมาณ

ตัวอย่างที่ 2.2 ข้อมูลเวลาในการออกกำลังกาย/วัน (นาที) ของนาย A ในสัปดาห์ที่ 1 ที่ถูกสุ่มมาเป็นตัวอย่าง ดังนี้

30, 36, 25, 18, 45, 55, 22

จงหาเวลาเฉลี่ยที่นาย A ใช้ออกกำลังกายใน 1 สัปดาห์

2.1.2 ค่าเฉลี่ยแบบถ่วงน้ำหนัก (weighted mean)

ถ้าข้อมูลแต่ละค่ามีความสำคัญไม่เท่ากันหรือมีค่าถ่วงน้ำหนักไม่เท่ากัน จะต้องนำน้ำหนักมาคิดในการหาค่าเฉลี่ยด้วยเสมอ โดยการนำน้ำหนักของข้อมูลแต่ละค่ามาคูณกับข้อมูลตัวนั้น จากนั้นนำมารวมกันทุกตัวแล้วหารด้วยผลรวมของน้ำหนักทั้งหมด สามารถคำนวณได้จากสูตรต่อไปนี้

ค่าเฉลี่ยแบบถ่วงน้ำหนัก

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

เมื่อ w_i คือ น้ำหนักของ x_i

ตัวอย่างที่ 2.3 ให้หาเกรดเฉลี่ยของนิสิตคนหนึ่ง ซึ่งมีผลการเรียนดังนี้

วิชา	หน่วยกิต	เกรด
ภาษาไทย	2	3
คณิตศาสตร์	3	4
ภาษาอังกฤษ	4	2
วิทยาศาสตร์	1	4

2.1.3 ค่าเฉลี่ยเลขคณิตรวม

ในกรณีที่มีข้อมูลอยู่หลาย ๆ กลุ่มและเราทราบค่าเฉลี่ยของข้อมูลแต่ละกลุ่มจะสามารถหาค่าเฉลี่ยเลขคณิตรวมทุกกลุ่มได้ดังนี้

ค่าเฉลี่ยเลขคณิตรวม

$$\bar{x} = \frac{\sum_{i=1}^{a} n_i \bar{x}_i}{\sum_{i=1}^{a} n_i}$$

เมื่อ $\bar{x}_i$ คือ ค่าเฉลี่ยของกลุ่มที่ i
n_i คือ จำนวนข้อมูลของกลุ่มที่ i
a คือ จำนวนกลุ่มทั้งหมด

ตัวอย่างที่ 2.4 นิสิตสาขาวิชาสถิติของมหาวิทยาลัยแห่งหนึ่งมีจำนวน 90 คน โดยมีนิสิตชายทั้งหมด 40 คน ซึ่งนิสิตชายสอบได้คะแนนเฉลี่ย 65 คะแนน ส่วนนิสิตหญิงสอบได้คะแนนเฉลี่ย 72 คะแนน จงหาคะแนนเฉลี่ยของนิสิตทั้งสาขา

2.1.4 ค่ามัธยฐาน (Median)

มัธยฐานเป็นค่าที่อยู่กึ่งกลางของข้อมูล เมื่อเรียงลำดับข้อมูลจากน้อยไปหามาก และแบ่งครึ่งข้อมูลออกเป็นสองส่วนเท่ากัน ค่ามัธยฐานอาจจะเป็นค่าใดค่าหนึ่งของชุดข้อมูล หรือมีค่าอยู่ระหว่างสองค่าใด ๆ ก็ได้ สัญลักษณ์ที่ใช้แทนค่ามัธยฐานคือ MD สามารถคำนวณหาตำแหน่งของมัธยฐานได้ดังนี้

ขั้นที่ 1 เรียงลำดับข้อมูล

ขั้นที่ 2 หาค่ามัธยฐาน

- เมื่อจำนวนข้อมูลเป็นเลขคี่

$$MD = \text{ค่าของข้อมูลลำดับที่ } \frac{n+1}{2}$$

- เมื่อจำนวนข้อมูลเป็นเลขคู่

$$MD = \frac{\text{ค่าของข้อมูลลำดับที่ } \frac{n}{2} + \text{ค่าของข้อมูลลำดับที่ } \frac{n+1}{2}}{2}$$

ตัวอย่างที่ 2.5 จำนวนนิสิตที่ลงทะเบียนเรียนวิชาสถิติสำหรับวิทยาศาสตร์ในเทอมนี้มี 7 กลุ่ม คือ 25, 35, 55, 74, 28, 54 และ 50 จงหาค่ามัธยฐานของข้อมูลชุดนี้

ขั้นที่ 1 เรียงลำดับข้อมูลจากน้อยไปหามาก 25 28 35 50 54 55 74

ขั้นที่ 2 หาค่ามัธยฐาน

$$\text{มัธยฐาน} = \text{ค่าของข้อมูลลำดับที่ } \frac{7+1}{2} = 4 = 50$$

คำสั่ง R

```
> num=c(25,28,35,50,54,55,74)
> median(num)
[1] 50
```

หมายเหตุ ฟังก์ชัน median() ใชหาค่ามัธยฐานสำหรับข้อมูลเชิงปริมาณ

2.1.5 ฐานนิยม (Mode)

ฐานนิยม คือ ค่าของข้อมูลที่เกิดขึ้นบ่อยครั้งที่สุดหรือมีความถี่มากที่สุด ชุดข้อมูลที่มีค่าของข้อมูลที่เกิดขึ้นบ่อยที่สุดเพียงค่าเดียว จะเรียกว่า **unimodal** ส่วนชุดข้อมูลที่มีค่าที่เกิดขึ้นบ่อยที่สุดสองค่า จะเรียกว่า **bimodal** และเมื่อชุดข้อมูลใดที่ไม่มีค่าสังเกตใดเกิดขึ้นมากกว่าหนึ่งครั้งเราจะเรียกว่าไม่มีฐานนิยม

ตัวอย่างที่ 2.6 จงหาฐานนิยมของข้อมูลต่อไปนี้

A: 5, 8, 6, 5, 9, 5

B: 10, 15, 2, 10, 9, 2, 8

C: 6, 9, 7, 2, 3, 11

คำสั่ง R

```
> a=c(11, 11, 12, 12, 12, 13, 13, 13, 13, 13, 14, 14, 14, 15, 15, 16, 16, 17, 17, 18)
> which.max(table(a))
13
3
```

หมายเหตุ ฟังก์ชัน table() ใชนับจำนวนความถี่ของข้อมูล

ฟังก์ชัน which.max() เป็นการหาค่าที่มีจำนวนสูงที่สุด โดยที่เราต้องใส่ตัวแปรที่หาความถี่เรียบร้อยแล้วให้กับฟังก์ชันนี้ฟังก์ชัน which.max() จึงต้องทำงานคู่กับ table()

2.1.6 คุณสมบัติและการใช้ค่าวัดแนวโน้มเข้าสู่ส่วนกลาง

ค่าเฉลี่ย (Mean)

1. ค่าเฉลี่ยนั้นคำนวณมาจากค่าสังเกตทุกค่าในชุดข้อมูล
2. ค่าเฉลี่ยจะมีค่าเปลี่ยนแปลงน้อยมากเมื่อเทียบกับค่ามัธยฐานหรือฐานนิยม เมื่อชุดข้อมูลตัวอย่างถูกสุ่มมาจากประชากรเดียวกัน
3. ค่าเฉลี่ยจะถูกนำไปใช้ในการคำนวณค่าสถิติอื่น ๆ เช่น ความแปรปรวน เป็นต้น
4. สำหรับชุดข้อมูลชุดหนึ่งจะมีค่าเฉลี่ยเพียงค่าเดียวเท่านั้น และไม่จำเป็นต้องเป็นค่าสังเกตค่าใดค่าหนึ่งในชุดข้อมูล
5. ค่าเฉลี่ยจะถูกกระทบด้วยค่าสังเกตที่มีค่าสูงหรือต่ำผิดปกติ ซึ่งเราเรียกค่าเหล่านี้ว่า outliers ซึ่งใน กรณีนี้ค่าเฉลี่ยอาจจะไม่ใช่ค่าวัดแนวโน้มเข้าสู่ส่วนกลางที่เหมาะสม

มัธยฐาน (Median)

1. การหาค่ามัธยฐานนั้นเราต้องหาค่ากึ่งกลางของข้อมูล
2. ค่ามัธยฐานจะถูกนำมาใช้เมื่อเราต้องการทราบว่าค่าสังเกตนั้นตกอยู่ในครึ่งล่างหรือครึ่งบนของการแจกแจง
3. ค่ามัธยฐานจะถูกกระทบด้วยค่าสังเกตที่สูงหรือต่ำผิดปกติน้อยกว่าค่าเฉลี่ย

ฐานนิยม (Mode)

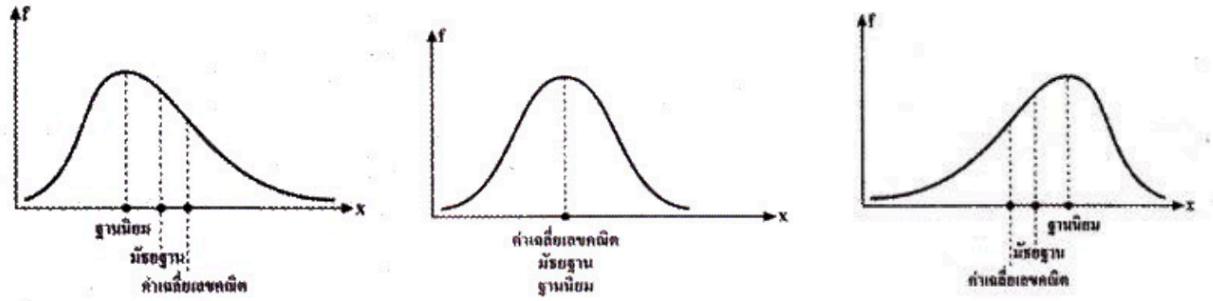
1. การหาค่าฐานนิยมจะใช้เพียงค่าสังเกตที่เกิดขึ้นบ่อยครั้งมากที่สุดเท่านั้น
2. ฐานนิยมเป็นค่าวัดแนวโน้มเข้าสู่ส่วนกลางที่คำนวณได้ง่ายที่สุด
3. ฐานนิยมนั้นสามารถใช้ได้กับข้อมูลที่อยู่ในระดับนามบัญญัติ เช่น เพศ หรือ ศาสนา เป็นต้น
4. ข้อมูลชุดหนึ่งอาจมีฐานนิยมได้มากกว่าหนึ่งค่า หรืออาจจะไม่มีฐานนิยมเลยก็ได้

2.2 ลักษณะการกระจายของข้อมูล

สิ่งที่สำคัญที่จะช่วยให้ทราบลักษณะของการแจกแจงมีอยู่ 4 ประการคือ ตำแหน่งกึ่งกลาง, ความแปรปรวน, ความเบ้และความโด่ง เมื่อเราทราบค่าทั้ง 4 แล้ว เราจะสามารถทราบลักษณะของการแจกแจงของข้อมูลได้

- ตำแหน่งกึ่งกลาง เป็นการวัดค่าที่อยู่ตรงกลางของการแจกแจง ในที่นี้คือค่าเฉลี่ย
- ความแปรปรวน บอกขนาดของกลุ่ม ถ้าข้อมูลทั้งหมดมีค่าเข้าใกล้ค่าเฉลี่ย ความแปรปรวนจะมีค่าน้อยที่สุด
- ความเบ้บอกความสมมาตรหรือไม่สมมาตรของการแจกแจงความถี่ ถ้าการแจกแจงไม่สมมาตรแล้ว ความถี่ส่วนใหญ่มีค่าต่ำและความถี่ส่วนน้อยมีค่าสูง การแจกแจงจะเป็นเบ้ทางบวก (positively skewed) หรือเบ้ขวา (right skewed) ในทางตรงกันข้าม ถ้าความถี่ส่วนใหญ่มีค่าสูงและความถี่ส่วนน้อยมีค่าต่ำ การแจกแจงความถี่จะเป็นเบ้ทางลบ (negatively skewed) หรือเบ้ซ้าย (left skewed)
- ความโด่ง บอกลักษณะของการแจกแจงว่าโด่งมากหรือโด่งน้อย ซึ่งอธิบายการกระจายของข้อมูล หากข้อมูลมีความโด่งมาก นั่นคือข้อมูลมีการกระจายน้อย แต่หากข้อมูลมีความโด่งน้อย นั่นคือข้อมูลมีการกระจายมาก

รูปร่างการการจายของข้อมูลมีได้หลายลักษณะขึ้นอยู่กับข้อมูล เมื่อข้อมูลมีการแจกแจงแบบสมมาตร (symmetry) ค่าเฉลี่ย มัธยฐาน และฐานนิยม จะอยู่ที่ตำแหน่งเดียวกัน หรือมีค่าเท่ากันทั้งหมด ถ้าข้อมูลมีการแจกแจงแบบเบ้ซ้าย ค่าเฉลี่ยจะมีค่าน้อยกว่ามัธยฐาน และโดยปกติค่ามัธยฐานมักจะมีค่าน้อยกว่าฐานนิยม ถ้าข้อมูลมีการแจกแจงเบ้ขวา ค่าเฉลี่ยจะมีค่ามากกว่ามัธยฐาน และ มัธยฐานจะมีค่ามากกว่าฐานนิยม ดังรูปที่ 2.1



รูปที่ 2.1 การกระจายของข้อมูลในแบบต่าง ๆ

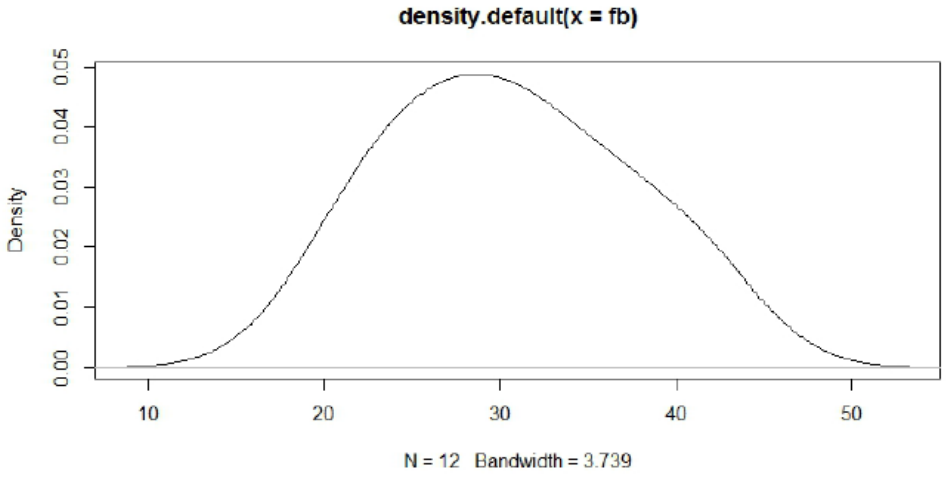
ตัวอย่างที่ 2.7 ข้อมูลต่อไปนี้เป็นจำนวนครั้งที่นิสิตคนหนึ่งเข้าเว็บไซต์ Facebook ใน 1 วัน จำนวน 12 วัน

40 35 24 28 26 29 36 31 42 20 23 32

จงนำเสนอค่ากลางของข้อมูลที่เหมาะสม

คำสั่ง R

```
> fb=c(40,35,24,28,26,29,36,31,42,20,23,32)
> plot(density(fb))
> mean(fb)
[1] 30.5
> median(fb)
[1] 30
```



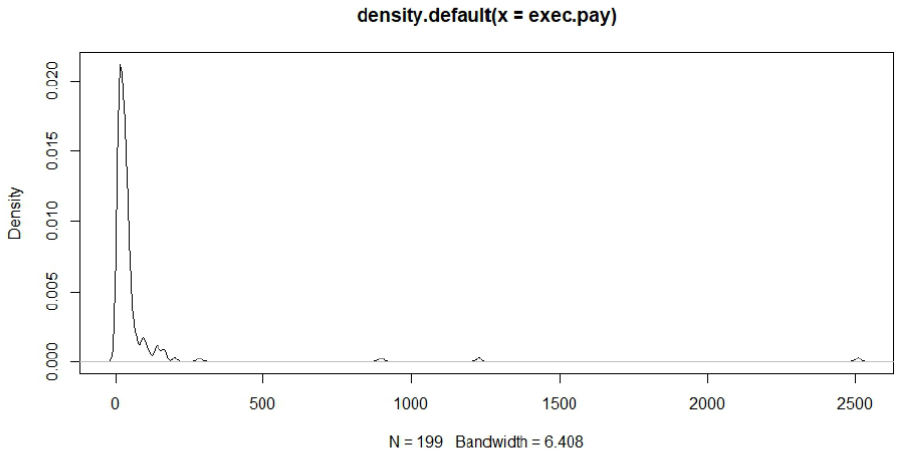
รูปที่ 2.2 กราฟหนาแน่นของข้อมูลจำนวนครั้งในการเข้าเว็บไซต์ Facebook

จากการวิเคราะห์จะได้ว่า จำนวนครั้งที่นิสิตเข้าเว็บไซต์ Facebook ใน 1 วัน มีค่าเฉลี่ยเป็น 30.5 ครั้ง และมีค่ามัธยฐานเป็น 30 ครั้ง ค่าทั้งสองมีค่าใกล้เคียงกัน ซึ่งสอดคล้องกับการกระจายของข้อมูลซึ่งมีลักษณะสมมาตรดังรูปที่ 2.2 ดังนั้นเราจะเลือกใช้ค่าเฉลี่ยในการนำเสนอข้อมูล นั่นคือ โดยเฉลี่ยแล้วนิสิตเข้าเว็บไซต์ Facebook 30.5 ครั้งต่อวัน

ตัวอย่างที่ 2.8 จากข้อมูล exec.pay (UsingR) แสดงค่าตอบแทนของผู้บริหาร ในประเทศสหรัฐอเมริกา จำนวน 199 คน ในปี 2000 (มีหน่วยเป็น 10000 ดอลลาร์)

คำสั่ง R

```
> install.packages("UsingR")
> library(UsingR)
> exec.pay
> plot(density(exec.pay))
> mean(exec.pay)
[1] 59.88945
> median(exec.pay)
[1] 27
```



รูปที่ 2.3 กราฟความหนาแน่นของข้อมูลค่าตอบแทน

จากรูปที่ 2.3 จะเห็นได้ว่าข้อมูลมีการกระจายแบบเบ้ขวา และเมื่อคำนวณหาค่าเฉลี่ยได้ค่าเป็น 59.88945 และค่ามัธยฐานมีค่าเป็น 27 ซึ่งมีค่าแตกต่างกันอย่างมาก ดังนั้นหากพิจารณาเลือกใช้สถิติในการอธิบายค่ากลางของข้อมูลชุดนี้จะเลือกใช้ค่ามัธยฐานเป็นค่ากลาง ดังนั้นค่าตอบแทนของผู้บริหารในประเทศสหรัฐอเมริกามีค่ากลางเป็น 27,000 ดอลลาร์

2.3 การวัดตำแหน่งของข้อมูล (Measure of position)

การจำแนกข้อมูลหมายถึง การแบ่งข้อมูลออกเป็นส่วน ส่วนละเท่า ๆ กัน โดยการจำแนกข้อมูลนี้ ต้องนำค่าของข้อมูลมาเรียงลำดับจากค่าน้อยไปหาค่ามาก หรืออาจจะเรียงจากค่ามากไปหาค่าน้อย การจำแนกข้อมูลแบ่งเป็น 3 ประเภท คือ

1. **ควอไทล์ (Quartile)** เป็นการแบ่งข้อมูลออกเป็น 4 ส่วน เท่า ๆ กัน โดยแต่ละส่วนจะมีจำนวนข้อมูล 25% ของจำนวนข้อมูลทั้งหมด ซึ่งควอไทล์ในข้อมูลชุดหนึ่งๆ จะมี 3 ค่า คือ ควอไทล์ที่ 1 (Q_1) ควอไทล์ที่ 2 (Q_2) และ ควอไทล์ที่ 3 (Q_3)
   การหาค่าควอไทล์
   1.1 เรียงลำดับข้อมูลทั้งหมดจากน้อยไปหามาก
   1.2 คำนวณควอไทล์ที่ r (Q_r) ซึ่งก็คือข้อมูลที่อยู่ในตำแหน่ง $\frac{r(n+1)}{4}$
2. **เดไซล์ (Decile)** เป็นการแบ่งข้อมูลออกเป็น 10 ส่วน เท่า ๆ กัน โดยแต่ละส่วนจะมีจำนวนข้อมูลส่วนละ 10% ของจำนวนข้อมูลทั้งหมด และจะมี 9 ค่า คือ เดไซล์ที่ 1 (D_1) เดไซล์ที่ 2 (D_2) ....เดไซล์ที่ 9 (D_9)
   การหาค่าเดไซล์
   2.1 เรียงลำดับข้อมูลทั้งหมดจากน้อยไปหามาก
   2.2 คำนวณเดไซล์ที่ r (D_r) ซึ่งก็คือข้อมูลที่อยู่ในตำแหน่ง $\frac{r(n+1)}{10}$
3. **เปอร์เซ็นไทล์ (Percentile)** เป็นการแบ่งข้อมูลออกเป็น 100 ส่วน เท่า ๆ กัน โดยแต่ละส่วนจะมีจำนวนข้อมูล ส่วนละ 1% ของจำนวนข้อมูลทั้งหมด และจะมี 99 ค่า คือ เปอร์เซ็นไทล์ที่ 1 (P_1) เปอร์เซ็นไทล์ที่ 2 (P_2) . . . และ ปอร์เซ็นไทล์ที่ 99 (P_{99}) ซึ่ง P_{25} จะเทียบได้กับ Q_1, P_{50} จะเทียบได้กับ Q_2 และ P_{75} เทียบได้กับ Q_3
   การหาค่าเปอร์เซ็นไทล์
   3.1 เรียงลำดับข้อมูลทั้งหมดจากน้อยไปหามาก
   3.2 คำนวณเปอร์เซ็นไทล์ที่ r (P_r) ซึ่งก็คือข้อมูลที่อยู่ในตำแหน่ง $\frac{r(n+1)}{100}$

ตัวอย่างที่ 2.9 ผลการสอบวิชาสถิติเบื้องต้นของนิสิตคณะวิทยาการสารสนเทศ 20 คน (คะแนนเต็ม 100 คะแนน) ดังนี้

40	78	80	50	60	45	70	63	55	66
71	45	74	72	75	61	65	53	42	52

จงหาคะแนนที่อยู่ในตำแหน่งเปอร์เซ็นไทล์ที่ 40, เปอร์เซ็นไทล์ที่ 75, เดไซล์ที่ 6 และ ควอไทล์ที่ 3

วิธีทำ

1. เรียงลำดับจากน้อยไปมาก

40	42	45	45	50	52	53	55	60	61
63	65	66	70	71	72	74	75	78	80

2. หาคะแนนในตำแหน่งเปอร์เซ็นไทล์ที่ 40

$$\frac{r(n+1)}{100} = \frac{40(20+1)}{100} = 8.4$$

$$x_{8.4} = x_8 + 0.4(x_9 - x_8)$$

ดังนั้น คะแนน $P_{40} = 55 + 2 = 57$

3. ตำแหน่งเปอร์เซ็นไทล์ที่ 75

4. เดไซล์ที่ 6

5. ควอไทล์ที่ 3

2.4 การวัดการกระจายของข้อมูล (Measures of Variation)

การพิจารณาหรือสรุปลักษณะของข้อมูลโดยใช้ค่ากลางหรือค่าเฉลี่ยเพียงอย่างเดียวอาจทำให้ไม่ทราบถึงลักษณะของข้อมูลได้อย่างชัดเจน เนื่องจากอาจมีข้อมูลที่มีค่ากลางเท่ากัน แต่ลักษณะของข้อมูลแตกต่างกัน เช่น การตัดสินใจเลือกซื้อหุ้นจากบริษัท A และ B โดยพิจารณาจากเปอร์เซ็นต์ของ ผลกำไรต่อปี ในช่วง 5 ปีที่ผ่านมา ได้ข้อมูลดังนี้

บริษัท A	10	12	15	18	20
บริษัท B	2	8	15	22	28

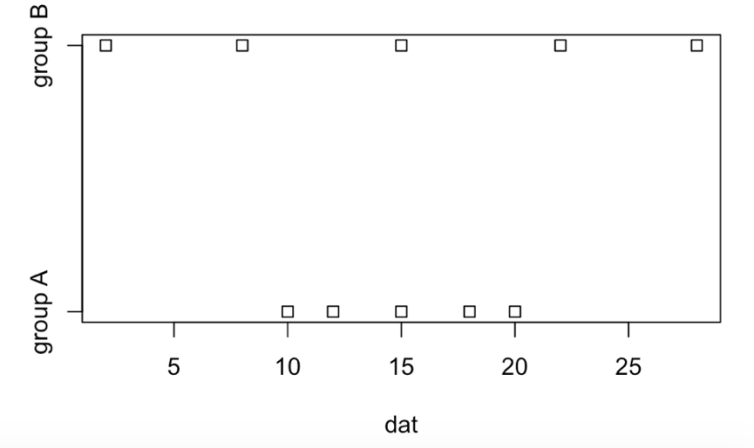
คำสั่ง R

```
> set=c(rep("group A",5),rep("group B",5))
> dat=c(10,12,15,18,20,2,8,15,22,28)
> info=data.frame(set,dat)
> tapply(info$dat,set,mean) #คำนวณค่าเฉลี่ยของข้อมูล dat ในแต่ละกลุ่มของตัวแปร set
group A group B
      15      15
> stripchart(dat~set,data=info)
```

หมายเหตุ

ฟังก์ชัน tapply() เป็นฟังก์ชันที่ใช้ในการคำนวณค่าสถิติของข้อมูลในแต่ละกลุ่ม

ฟังก์ชัน stripchart ()เป็นฟังก์ชันพล็อตค่าของข้อมูลแต่ละค่าเพื่อให้เห็นการกระจายของข้อมูล



รูปที่ 2.4 กราฟการกระจายของข้อมูล 2 ชุด

จากข้อมูลจะพบว่ากำไรของทั้ง 2 บริษัทมีค่าเท่ากันคือ 15% ถ้าพิจารณาเพียงแค่นี้ก็อาจจะตัดสินใจซื้อหุ้นของบริษัทใดก็ได้ แต่ถ้าวิเคราะห์ให้ละเอียดมากขึ้นจะพบว่ากำไรของบริษัท A แต่ละปีจะแตกต่างกันน้อยมาก เมื่อพิจารณาทั้งกำไรเฉลี่ยและการกระจายของกำไรจะทำให้ตัดสินใจได้ว่าควรซื้อหุ้นจากบริษัท A ดังนั้นถ้าข้อมูลมีค่าเฉลี่ยเท่ากันแล้วให้พิจารณาการกระจายควบคู่กันไปด้วย

ในการจะทราบความแตกต่างของข้อมูลในแต่ละกลุ่มเราเรียกว่า การวัดการกระจาย โดยข้อมูลที่ดีจะต้องมีการกระจายต่ำสุด การวัดการกระจายของข้อมูลสามารถทำได้หลายวิธีดังนี้

การวัดการกระจายสัมบูรณ์ (Absolute Variation) คือการวัดการกระจายของข้อมูลเพียงชุดเดียวเพื่อดูว่าข้อมูลชุดนี้ แต่ละค่ามีความแตกต่างกันมากหรือน้อยเพียงไร นิยมใช้กันอย่างน้อย 4 ชนิด ได้แก่ ค่าพิสัย ค่าความแปรปรวน ส่วนเบี่ยงเบนมารตฐาน และพิสัยควอไทล์

การวัดการกระจายสัมพัทธ์ (Relative Variation) การหาค่าเพื่อเปรียบเทียบการกระจายระหว่างข้อมูลมากกว่าหนึ่งชุด โดยใช้อัตราส่วนการเปรียบเทียบการกระจายของข้อมูลระหว่างชุด ที่นิยมใช้มี 2 ชนิด ได้แก่ ค่าสัมประสิทธิ์การแปรผัน และสัมประสิทธิ์ส่วนเบี่ยงเบนควอไทล์

2.4.1 พิสัย (Range)

พิสัยเป็นการวัดการกระจายที่ง่ายที่สุด เป็นการหาความแตกต่างของข้อมูลสูงสุดและต่ำสุดของกลุ่ม

พิสัย = ค่าสูงสุดของข้อมูล – ค่าต่ำสุดของข้อมูล

ตัวอย่างที่ 2.10

พิสัยของข้อมูลบริษัท A ซึ่งมีข้อมูลคือ 10, 12, 15, 18 และ 20 คำนวณหาพิสัยคือ 20-10=10

พิสัยของข้อมูลบริษัท B ซึ่งมีข้อมูลคือ 2, 8, 15, 22 และ 28 คำนวณหาพิสัยคือ 28-2=26

จะเห็นว่าข้อมูลบริษัท B จะมีค่าการกระจายมากกว่าข้อมูลบริษัท A

จากข้อมูล info สามารถคำนวณค่าพิสัย ได้ดังนี้

คำสั่ง R

```
> tapply(info$dat,set,range)
$`group A`
[1] 10 20
$`group B`
[1] 2 28
```

หากสร้างตัวแปร 2 ตัวเพื่อเก็บค่าข้อมูล ดังนี้

คำสั่ง R

```
> groupA=c(10,12,15,18,20)
> groupB=c(2,8,15,22,28)
> range(groupA)
[1] 10 20
> range(groupB)
[1] 2 28
```

หมายเหตุ ฟังก์ชัน range() ให้ผลลัพธ์เป็นค่าต่ำสุดและค่าสูงสุด ดังนั้นจะหาค่าพิสัยต้องนำค่าทั้งสองมาลบกัน

ในกรณีใช้พิสัยกับข้อมูลที่มีจำนวนมาก การวัดจะไม่แน่นอน และค่าของพิสัยจะขึ้นอยู่กับขนาดของข้อมูล ถ้าข้อมูลมีจำนวนมากพิสัยจะมาก ถ้าข้อมูลมีจำนวนน้อยพิสัยจะน้อย

จากข้อมูล exec.pay(UsingR) คำนวณหาค่าพิสัย

```
> install.packages("UsingR")
> library(UsingR)
> exec.pay
> diff(range(exec.pay))
[1]2510
```

2.4.2 ค่าความแปรปรวน และส่วนเบี่ยงเบนมาตรฐาน (Variance and Standard deviation)

เนื่องจากการใช้พิสัยเป็นการวัดการกระจายอย่างหยาบเท่านั้นเพราะใช้ข้อมูลเพียง 2 ตัว คือค่าข้อมูลที่มีค่าสูงสุดและต่ำสุด แต่การวัดความแปรปรวนจะใช้ข้อมูลทุกตัวจึงเป็นค่าที่นิยมใช้วัดการกระจายมากที่สุด โดยจะพิจารณาจากผลรวมของค่าแตกต่างระหว่างค่าของข้อมูลกับค่าเฉลี่ย ถ้าค่าแตกต่างนั้นมากแสดงว่าข้อมูลกระจายมาก ค่าแตกต่างระหว่างข้อมูลกับค่าเฉลี่ยอาจมีค่าเป็นบวกหรือลบก็ได้ ซึ่งอาจทำให้ผลรวมเป็นศูนย์ จึงต้องกำหนดให้นำค่าแตกต่างระหว่างข้อมูลกับค่าเฉลี่ยมายกกำลังสองด้วย

ความแปรปรวนของตัวอย่าง (Sample Variance)

$$s^2 = \frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n - 1}$$

ความแปรปรวนของประชากร (population variance)

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x - \mu)^2}{N}$$

โดยปกติเรามักนิยมใช้ส่วนเบี่ยงเบนมาตรฐานเป็นค่าวัดความผันแปรของข้อมูลมากกว่าความแปรปรวน สาเหตุนั้นเกี่ยวข้องกับหน่วยของค่าสถิติทั้งสองนี้ จะเห็นได้ว่าหน่วยของความแปรปรวนนั้นจะเป็นหน่วยของ x ยกกำลังสอง เช่น ถ้า x มีหน่วยเป็นคะแนน ความแปรปรวนจะมีหน่วยเป็นคะแนนกำลังสอง ในขณะที่ส่วนเบี่ยงเบนมาตรฐานนั้นเป็นรากที่สองของความแปรปรวน จะมีหน่วยเดียวกับ x นั่นคือมีหน่วยเป็นคะแนน ดังนั้นการอธิบายถึงการกระจายของข้อมูลด้วยส่วนเบี่ยงเบนมาตรฐานจึงเข้าใจได้ง่ายกว่าการใช้ความแปรปรวน

ส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง (Sample Standard Deviation)

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n - 1}}$$

ส่วนเบี่ยงเบนมาตรฐานของประชากร (population standard deviation)

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(x - \mu)^2}{N}}$$

จากสูตรข้างต้นเห็นได้ว่าต้องใช้เวลาในการคำนวณมากเพราะต้องใช้ค่าเฉลี่ยเลขคณิตดังนั้นจึงมีการปรับปรุงสูตรนี้เพื่อให้ใช้คำนวณได้รวดเร็วขึ้นดังนี้

ความแปรปรวนของตัวอย่าง

$$s^2 = \frac{\sum_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)^2}{n}}{n-1}$$

ส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง

$$s = \sqrt{\frac{\sum_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)^2}{n}}{n-1}}$$

x คือค่าของข้อมูล

$\bar{x}$ คือค่าเฉลี่ยของตัวอย่าง

μ คือค่าเฉลี่ยของประชากร

n คือขนาดของข้อมูลตัวอย่าง

N คือขนาดของข้อมูลประชากร

ตัวอย่างที่ 2.11 ข้อมูลต่อไปนี้เป็นน้ำหนักของนักกีฬายกน้ำหนัก 10 คนที่ถูกสุ่มมาเป็นตัวอย่าง

58 62 50 65 68 48 52 60 49 75

จงหาความแปรปรวนและส่วนเบี่ยงเบนมาตรฐานของข้อมูลชุดนี้

ตัวอย่างที่ 2.12 จงหาค่าความแปรปรวนและส่วนเบี่ยงเบนมาตรฐานของข้อมูลจำนวนนิสิต (คน) ที่ลงทะเบียนเรียนวิชาสถิติ 7 กลุ่ม

25 35 55 74 28 54 50

คำสั่ง R

```
> num=c(25,28,35,50,54,55,74)
> var(num)
[1] 305.1429
> sd(num)
[1] 17.46834
```

หมายเหตุ

ฟังก์ชัน var() คำนวณค่าความแปรปรวนของตัวอย่าง

ฟังก์ชัน sd() คำนวณค่าเบี่ยงเบนมาตรฐานของตัวอย่าง

จากการคำนวณได้ค่าความแปรปรวนของข้อมูลจำนวนนิสิตที่เรียนวิชาสถิติเป็น 305.14 คน^2 และมี ค่าเบี่ยงเบนมาตรฐานเป็น 17.47 คน

2.4.3 พิสัยควอไทล์ (Inter quartile range :IQR)

ค่าพิสัยควอไทล์เป็นค่าที่บอกความผันแปรของข้อมูลได้อย่างหยาบๆ โดยค่าพิสัยควอไทล์หาได้จากผลต่างระหว่าง Q_1 และ Q_3 ซึ่งก็คือพิสัยของข้อมูลจำนวน 50 เปอร์เซ็นต์ที่อยู่กึ่งกลางของชุดข้อมูลนั่นเอง พิสัยควอไทล์นี้เป็นการวัดการกระจายที่เหมาะกับข้อมูลที่การแจกแจงแบบเบ้ ซึ่งสังเกตได้จากการคำนวณจากค่า Q_1 และ Q_3 ซึ่งไม่ได้นำข้อมูลที่มีค่าสูงมาก ๆ หรือต่ำมาก ๆ มาคำนวณ

$$IQR = Q_3 - Q_1$$

ตัวอย่างที่ 2.13 จงหาค่าพิสัยควอไทล์ของข้อมูลจำนวนนิสิต (คน) ที่ลงทะเบียนเรียนวิชาสถิติ 7 กลุ่ม

25 35 55 74 28 54 50

จากข้อมูล exec.pay

คำสั่ง R

```
> IQR(exec.pay)
[1] 27.5
> sd(exec.pay)
[1] 207.0435
> summary(exec.pay)
Min. 1st Qu. Median Mean 3st Qu. Max
0.00 14.00   27.00    59.89  41.50  2510.00
```

จะได้ว่า ค่าพิสัยควอไทล์ของข้อมูลมีค่าเป็น 27.5 แต่เมื่อเปรียบเทียบกับค่าส่วนเบี่ยงเบนมาตรฐานของข้อมูลจะได้ว่ามีค่าเป็น 207.04 ซึ่งมีค่าแตกต่างกันมาก ดังนั้นหากพิจารณาลักษณะของข้อมูลโดยรวมพบว่าข้อมูลมีการกระจายเบ้ขวามาก และด้วยคำสั่ง summary แล้วจะพบว่าข้อมูลมีความแตกต่างกันมากจากค่าสูงสุดและต่ำสุด แต่ข้อมูลส่วนใหญ่มีการกระจายไม่มากอยู่ใกล้ ๆ 0 มีเพียงบางค่าที่มีค่าสูงมาก ๆ ดังนั้นเราจะเห็นได้ว่าการใช้ค่าพิสัยควอไทล์เป็นการวัดการกระจายของข้อมูลชุดนี้

2.4.4 สัมประสิทธิ์การแปรผัน (Coefficient of variation)

หากเราต้องการเปรียบเทียบการกระจายของข้อมูลมากกว่าหนึ่งชุด และนำข้อมูลแต่ละชุดมาเปรียบเทียบกันว่าข้อมูลชุดใดมีการกระจายมากกว่ากัน โดยที่ข้อมูลทั้งสองชุดนั้นจะเป็นข้อมูลในเรื่องเดียวกันหรือต่างเรื่องกันก็ได้ เช่น ต้องการเปรียบเทียบข้อมูลน้ำหนัก กับส่วนสูง ซึ่งจะเห็นได้ว่า ข้อมูลน้ำหนักมีหน่วยเป็นกิโลกรัม และส่วนสูงมีหน่วยเป็นเซนติเมตร ซึ่งโดยปกติเราไม่สามารถเปรียบเทียบการกระจายของข้อมูลทั้งสองชุดนี้โดยพิจารณาจากค่าส่วนเบี่ยงเบนมาตรฐานโดยตรงได้ เนื่องจากหน่วยของข้อมูลนั้นแตกต่างกันอย่างไรก็ตามค่าที่จะช่วยวัดการกระจายของข้อมูลเหล่านี้ เรียกว่า สัมประสิทธ์ความแปรผัน ซึ่งเป็นค่าที่ไม่มีหน่วย ต่างจากค่าสถิติตัวอื่น ๆ ที่ใช้วัดการกระจาย ซึ่งมีหน่วยเป็นหน่วยเดียวกันกับข้อมูล ค่าสัมประสิทธิ์การแปรผันคือ ค่าเบี่ยงเบนมาตรฐานหารด้วยค่าเฉลี่ย เนื่องจากหน่วยของค่าเบี่ยงเบนมาตรฐานและค่าเฉลี่ยของข้อมูลจะเป็นหน่วยเดียวกัน ทำให้ค่าสัมประสิทธิ์การแปรผันไม่มีหน่วย

สัมประสิทธิ์ความผันแปรของตัวอย่าง

$$cv = \frac{s}{\bar{x}} \times 100\%$$

สัมประสิทธิ์ความผันแปรของประชากร

$$cv = \frac{\sigma}{\mu} \times 100\%$$

ตัวอย่างที่ 2.14 จากข้อมูลของบริษัทจำหน่ายรถยนต์แห่งหนึ่ง ในรอบ 3 เดือน พบว่าจำนวนรถยนต์ ที่จำหน่ายได้เฉลี่ย 87 คัน มีค่าเบี่ยงเบนมาตรฐานเท่ากับ 5 คัน และค่าคอมมิชชั่น (commissions) เฉลี่ย $5225 มีค่าเบี่ยงเบนมาตรฐาน $773 จงเปรียบเทียบการกระจายของข้อมูลทั้งสอง

ตัวอย่างที่ 2.15 ข้อมูลต่อไปนี้เป็นส่วนสูง (cm) และน้ำหนัก (kg) ของนักกีฬา 10 คนที่ถูกสุ่มมาเป็นตัวอย่าง

น้ำหนัก	75	68	82	72	85	65	59	80	76	56
ส่วนสูง	172	169	185	170	180	173	165	182	175	166

จงเปรียบเทียบการกระจายของน้ำหนักและส่วนสูงของนักกีฬา

ตัวอย่างที่ 2.16 บริษัทแห่งหนึ่งแบ่งคนงานออกเป็น 2 กลุ่ม ๆ ละ 8 คน จำนวนชิ้นของสินค้าที่คนงานแต่ละคนในกลุ่มผลิตเป็นดังนี้

กลุ่มที่ 1 (X) : 13, 6, 8, 2, 15

กลุ่มที่ 2 (Y) : 8, 2, 7, 7, 8

จงหากลุ่มพนักงานใดมีการกระจายของความสามารถในการผลิตสินค้ามากกว่ากัน

คำสั่ง R

```
> g1=c(13,6,8,2,15)
> g2=c(8,2,7,7,8)
> CV1=sd(g1)/mean(g1)*100
> CV1
[1] 59.80772
> CV2=sd(g2)/mean(g2)*100
> CV2
[1] 39.21844
```

จากการคำนวณ กลุ่มที่ 1 มีค่าสัมประสิทธิ์ความแปรผันเป็น 59.80772 และกลุ่มที่ 2 มีค่าสัมประสิทธิ์ความ แปรผันเป็น 39.21844 จะได้ว่ากลุ่มที่ 1 มีการกระจายของความสามารถในการผลิตสินค้ามากกว่ากลุ่มที่ 2

2.4.5 สัมประสิทธิ์ส่วนเบี่ยงเบนควอไทล์ (Coefficient of quartile deviation)

เป็นการเปรียบเทียบการประจายของข้อมูลสองชุด ด้วยการกระจายสัมพัทธ์เมื่อข้อมูลไม่มีการแจกแจงสมมาตร โดยค่าสัมประสิทธิ์ส่วนเบี่ยงเบนควอไทล์คำนวณได้จาก

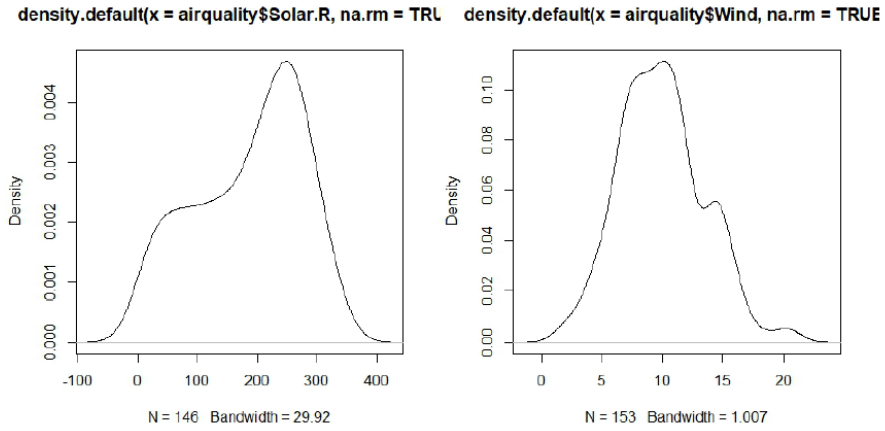
$$CD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

เมื่อ Q_3 และ Q_1 คือค่าควอไทล์ที่ 1 และ 3 ตามลำดับ

ตัวอย่างที่ 2.17 จากชุดข้อมูล airquality เปรียบเทียบการกระจายของข้อมูล Solar.R และ Wind

คำสั่ง R

```
> par(mfrow=c(1,2))
> plot(density(airquality$Solar.R,na.rm=TRUE))
> plot(density(airquality$Wind,na.rm=TRUE))
```



รูปที่ 2.4 กราฟการกระจายของ Solar.R และ Wind

ข้อมูล Solar.R และ Wind เป็นข้อมูลคนละประเภท มีหน่วยการวัดแตกต่างกัน อีกทั้งจากกราฟจะพบว่าข้อมูลไม่มีการกระจายสมมาตร ดังนั้นเราจะเปรียบเทียบการกระจายด้วยสัมประสิทธิ์ส่วนเบี่ยงเบนควอไทล์

คำสั่ง R

```
> Q3=quantile(airquality$Solar.R,0.75,na.rm = TRUE)
> Q1=quantile(airquality$Solar.R,0.25,na.rm = TRUE)
> CD1=(Q3-Q1)/(Q3+Q1)
> CD1
75% 0.3818425

> Q3_W=quantile(airquality$Wind,0.75,na.rm = TRUE)
> Q1_W=quantile(airquality$Wind,0.25,na.rm = TRUE)
> CD2=(Q3_W-Q1_W)/(Q3_W+Q1_W)
> CD2
75% 0.2169312
```

จากการคำนวณค่าสัมประสิทธิ์ส่วนเบี่ยงเบนควอไทล์ ได้ค่า CD1=0.38 และ CD2=0.22 แสดงว่าข้อมูล Wind มีการกระจายมากกว่าข้อมูล Solar.R

2.5 การสร้างแผนภาพกล่อง (Box plot)

Box and whisker plot หรือ Boxplot เป็นการนำเสนอข้อมูลด้วยรูปแบบกราฟแบบหนึ่ง ซึ่งกราฟนี้จะแสดงค่ากลางของข้อมูล (มัธยฐาน) การกระจายของข้อมูล และบ่งบอกความเบ้ หรือสมมาตรของข้อมูล และ มากไปกว่านั้นยังสามารถตรวจสอบค่าผิดปกติของชุดข้อมูลด้วย เรามักนิยมใช้กราฟ boxplot ในการเปรียบเทียบข้อมูลตั้งแต่สองชุดขึ้นไป

การสร้าง boxplot

1. เรียงข้อมูลจากน้อยไปมาก
2. หาค่า Q_1, Q_2, Q_3
3. สร้างกล่อง
4. หาขอบเขตของค่าที่ยังไม่ผิดปกติ ได้แก่ $Q_3 + 1.5(IQR)$ และ $Q_1 - 1.5(IQR)$
5. สร้าง whisker ทั้ง 2 ด้าน โดยลากเส้นจากกึ่งกลางกล่องไปยังค่าสูงสุดของข้อมูลที่ยังไม่สูงผิดปกติ และ ค่าต่ำสุดของข้อมูลที่ยังไม่ต่ำผิดปกติ
6. ในกรณีที่มีค่าผิดปกติให้เขียนลงไปในแผนภาพโดยใช้สัญลักษณ์ ° หรือ *

- ความกว้างของ box เท่ากับ $Q_3 - Q_1(IQR)$ กล่าวได้ว่ามีข้อมูล 50% อยู่ใน box ถ้า box กว้างแสดงว่าข้อมูลมีการกระจายมาก ถ้า box แคบแสดงว่าข้อมูลมีการกระจายน้อย
- การดูลักษณะของข้อมูลว่า สมมาตร เบ้ซ้าย หรือ เบ้ขวา ให้ดูทั้งหมดของ box-plot ไปจนถึง whisker ถ้าด้านใดยาวแสดงว่าข้อมูลเบ้ไปทางด้านนั้น
- ค่าสูงสุดของข้อมูลที่ยังไม่สูงผิดปกติ คือ ค่าสูงสุดของข้อมูลที่มีค่าไม่เกิน $Q_3 + 1.5Z(IQR)$
- ค่าต่ำสุดของข้อมูลที่ยังไม่ต่ำผิดปกติ คือ ค่าต่ำสุดของข้อมูลที่มีค่าไม่เกิน $Q_1 - 1.5(IQR)$
- ถ้ามีข้อมูลใดมีค่าน้อยกว่า $Q_1 - 1.5(IQR)$ หรือมากกว่า $Q_3 + 1.5(IQR)$ จะเรียกข้อมูลนั้น ว่า **Outlier** แสดงด้วยเครื่องหมายวงกลม (°)
- ถ้ามีข้อมูลใดมีค่าน้อยกว่า $Q_1 - 3(IQR)$ หรือมากกว่า $Q_3 + 3(IQR)$ จะเรียกข้อมูลนั้นว่า **Extremes** แสดงด้วยเครื่องหมายดอกจัน (*)

ตัวอย่างที่ 2.18 ผลการสอบวิชาสถิติเบื้องต้นของนิสิตคณะวิทยาการสารสนเทศ 20 คน (คะแนนเต็ม 100 คะแนน) ดังนี้

40	78	80	50	60	45	70	63	55	66
71	45	74	72	75	61	65	53	42	52

วิธีทำ

1. เรียงลำดับจากน้อยไปมาก

40	42	45	45	50	52	53	55	60	61
63	65	66	70	71	72	74	75	78	80

2. หาค่า Q_1, Q_2, Q_3

   $Q_1 = 1(20+1)/4 = 5.25$ ข้อมูล $Q_1 = 50.5$

   $Q_2 = 2(20+1)/4 = 10.5$ ข้อมูล $Q_2 = 62$

   $Q_3 = 3(20+1)/4 = 15.75$ ข้อมูล $Q_3 = 71.75$

3. สร้างกล่อง

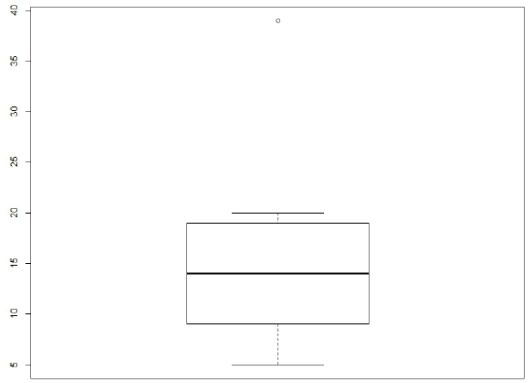
4.

ตัวอย่างที่ 2.19 ข้อมูลต่อไปนี้คือระดับความสูงของน้ำ (เซนติเมตร) ที่ท่วมบริเวณอำเภอต่างๆ 8 อำเภอในจังหวัดปราจีนบุรี

15 13 6 5 12 20 39 18

คำสั่ง R

```
> water=c(15,13,6,5,12,20,39,18)
> boxplot(water)
```



รูปที่ 2.6 Box and Whisker plot สำหรับข้อมูลความสูงของน้ำ

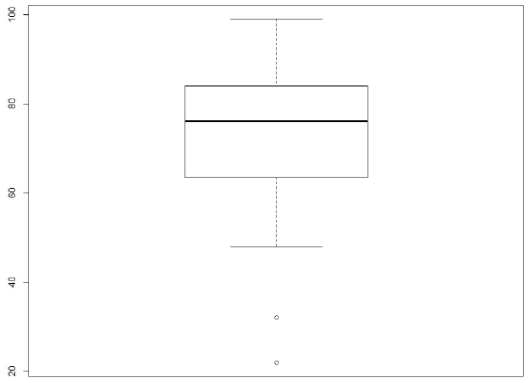
จากรูปที่ 2.6 จะเห็นได้ว่าข้อมูลชุดนี้มีค่าผิดปกติ 1 ค่า คือ 39 หากไม่พิจารณาค่าผิดปกติแล้วจะได้ว่าข้อมูลมีการกระจายค่อนข้างสมมาตร

ตัวอย่างที่ 2.20 ข้อมูลต่อไปนี้เป็นเวลาที่นิสิตใช้ในการเล่นอินเทอร์เน็ตต่อวัน (หน่วย:นาที) ของนิสิตจำนวน 50 คน

22 32 48 49 53 55 57 58 59 60 62 62 63
64 65 66 68 69 70 71 72 73 74 75 75 76
77 77 78 78 79 79 80 80 81 83 84 84 85
86 87 88 89 90 90 92 93 95 98 99

คำสั่ง R

```
> internet=c(22,32,48,49,53,55,57,58,59,60,62,62,63,64,65,66,68,69,70,71,72,73,74,75,75,76,
+ 77,77,78,78,79,79,80,80,80,81,83,84,84,85,86,87,88,89,90,90,92,93,95,98,99)
> boxplot(internet)
```



รูปที่ 2.7 Box and Whisker plot สำหรับข้อมูล internet