บทที่ 6

ทฤษฎีทางสถิติที่ใช้ในการตรวจสอบเครื่องมือและวิเคราะห์ข้อมูล

ในการตรวจสอบเครื่องมือและวิเคราะห์ข้อมูล ใช้ทฤษฎีทางสถิติที่สำคัญดังนี้

1. ค่าเฉลี่ยหรือตัวกลางเลขคณิต (Arithmetic mean)

เมื่อมีข้อมูลอยู่ชุดหนึ่ง เช่นค่าฟลักซ์จากเครื่องกำเนิดสัญญาณเทียมตามลำดับวันต่างๆ ถ้าเราต้องการหาตัวเลขตัวหนึ่งซึ่งใช้แทนข้อมูลชุดนี้ โดยปกติจะใช้ค่าเฉลี่ยของข้อมูลชุดนี้ ซึ่งหาได้ดังนี้

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (6.1)$$

เมื่อ $\bar{x}$ = ค่าเฉลี่ย

x_i = ค่าของข้อมูลแต่ละตัว

n = จำนวนข้อมูล

(Topping ,1971)

2. ความเบี่ยงเบนมาตรฐาน (Standard deviation)

เมื่อเราใช้ค่าเฉลี่ยของข้อมูลชุดหนึ่งเป็นตัวแทนของข้อมูลชุดนั้น ข้อมูลแต่ละตัวอาจมีค่าเท่ากับหรือแตกต่างไปจากค่าเฉลี่ย ปริมาณที่ต่างออกไปจากค่าเฉลี่ยคือค่าเบี่ยงเบนเฉลี่ย (d_i) เขียนเป็นสูตรได้ว่า

$$d_i = x_i - \bar{x} \qquad (6.2)$$

ขนาดของการกระจาย (dispersion)ของค่าเบี่ยงเบนนี้ จะบอกโดยค่าเบี่ยงเบนมาตรฐาน (σ) ซึ่งเขียนเป็นสูตรได้ว่า

$$\sigma = \left(\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)\right)^{\frac{1}{2}} \qquad (6.3)$$

ในกรณีที่มีข้อมูล 2 ชุด คือ $x_1, x_2, \ldots, x_n$ และ $y_1, y_2, \ldots, y_n$ ซึ่งเป็นอิสระต่อกัน มีความเบี่ยงเบนมาตรฐานเป็น σ_x และ σ_y ตามลำดับ ถ้า $Z=x\pm y$ และ σ_z เป็นความเบี่ยงเบนมาตรฐานของ Z จะได้ว่า

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 \qquad \text{(Topping, 1971)} \; (6.4)$$

3. <u>การเปรียบเทียบการกระจายระหว่างหมู่</u>

ในการเปรียบเทียบการกระจายของข้อมูลตั้งแต่ 2 ชุดขึ้นไป เราจะพิจารณาเฉพาะค่าความเบี่ยงเบนมาตรฐานเพียงอย่างเดียวไม่ได้ เพราะว่าข้อมูลแต่ละชุดอาจมีขนาดหรือสเกลแตกต่างกัน ดังนั้นจึงบอกขนาดของการกระจายในหน่วยของค่าเฉลี่ยของข้อมูลชุดนั้น ซึ่งเรียกว่าสัมประสิทธิ์การกระจาย (D) (coefficient of dispersion) เขียนเป็นสูตรได้ว่า

$$D = \frac{\sigma}{\bar{x}} \qquad (6.5)$$

ถ้าสัมประสิทธิ์การกระจายของข้อมูลชุดหนึ่งมีค่ามากกว่าอีกชุดหนึ่ง แสดงว่าข้อมูลชุดแรกมีการกระจายมากกว่าชุดหลัง (โดยปกติข้อมูล 2 ชุดนี้จะเป็นข้อมูลประเภทเดียวกัน) (Topping ,1971)

4. <u>การกระจายของความผิดพลาด</u> (Distribution of error)

ในการวัดปริมาณทางฟิสิกส์ใด ๆ ค่าที่สังเกตการณ์ได้ (observed values) หรือค่าที่วัดได้ อาจแตกต่างจากค่าที่แท้จริง (actual values) ได้มากน้อยต่างกัน เรียกความแตกต่างนี้ว่าความผิดพลาดของการสังเกตการณ์ (error of observation) ซึ่งแบ่งได้เป็น 2 แบบ คือ ความผิดพลาดที่เป็นระบบ (systematic error) เป็น

ความผิดพลาดที่เกิดขึ้นอย่างเป็นระบบ มีกฎเกณฑ์แน่นอน ถ้าทราบกฎเกณฑ์นี้ก็จะสามารถคำนวณแก้ไขได้ ความผิดพลาดอีกแบบหนึ่ง คือ ความผิดพลาดแบบไม่แน่นอน (random error) เป็นความผิดพลาดที่ไม่เป็นระบบ ไม่มีระเบียบ และไม่มีกฎเกณฑ์ที่แน่นอนสำหรับการวัดแต่ละครั้ง ไม่สามารถคำนวณแก้ไขได้ สำหรับการวัดปริมาณทางฟิสิกส์ใด ๆ ถ้าเกิดความผิดพลาดแบบไม่แน่นอนเพียงอย่างเดียว ค่าที่วัดได้จะมีการกระจายในรูปของการกระจายแบบปกติ (normal distribution) หรือการกระจายตามแบบของเกาส์ (Gaussian distribution) ซึ่งเขียนเป็นสมการได้ดังนี้

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2} \qquad (6.6)$$

เมื่อ $\bar{x}$ = ค่าเฉลี่ย

σ = ความเบี่ยงเบนมาตรฐาน

ในการบอกปริมาณมากน้อยของความผิดพลาดที่เกิดขึ้นเป็นร้อยละ เราสามารถบอกได้ด้วยอัตราส่วนระหว่างความเบี่ยงเบนมาตรฐานกับค่าเฉลี่ยของข้อมูลที่วัดได้ ซึ่งเขียนเป็นสูตรได้ว่า

$$E = 100\,\frac{\sigma}{\bar{x}} \qquad (6.7)$$

เมื่อ E = ความผิดพลาดสัมพัทธ์เป็นร้อยละ (Topping, 1971)

5. <u>การปรับเส้นตรงด้วยวิธีกำลังสองน้อยที่สุด</u> (Linear fitting)

หลักการกำลังสองน้อยที่สุด (least square) กล่าวโดยสรุปได้ว่า ในการวัดปริมาณ ๆ หนึ่งหลาย ๆ ครั้งจะได้ข้อมูลกลุ่มหนึ่ง เราอาจหาค่าที่จะมาแทนข้อมูลกลุ่มนี้คือค่าที่น่าจะเป็นมากที่สุด (most probable value) ซึ่งถ้าเราวัดปริมาณดังกล่าวหลาย ๆ ครั้ง ค่าที่วัดได้จะมีแนวโน้มเข้าใกล้ค่า ๆ นี้มากกว่าค่าอื่น ๆ มีหลักการเป็นคณิตศาสตร์ว่า ผลบวกกำลังที่สองของความเบี่ยงเบนจากค่าที่น่าจะเป็นนี้ จะมีค่าน้อยที่สุด โดยอาศัยหลักการดังกล่าวหาแนวโน้มของความสัมพันธ์ระหว่างปริมาณ 2 ชุดได้ดังนี้

ถ้ามีผลการทดลองชุดหนึ่งวัดปริมาณ y ได้ $y_1, y_2, \ldots, y_n$ ซึ่งสอดคล้องกับปริมาณ $x_1, x_2, \ldots, x_n$ เมื่อนำค่าต่าง ๆ ของ y มาเขียนกราฟกับ x ถ้ากราฟที่ได้มีแนวโน้มเป็นเส้นตรง จากหลักการกำลังสองน้อยที่สุด เราสามารถหาสมการเส้นตรงที่แทนแนวโน้มของความสัมพันธ์ระหว่าง x กับ y ได้ว่า

$$y = a + bx \tag{6.8}$$

เมื่อ

$$a = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} ; b = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

6. สหสัมพันธ์ (COrrelation)

6.1 การหาสหสัมพันธ์ระหว่างข้อมูล 2 ชุด ในข้อ 5 นอกจากเราจะสามารถหาเส้นตรงซึ่งใช้แทนแนวโน้มดังกล่าวแล้ว ยังสามารถบอกว่าข้อมูล 2 ชุดนี้มีความสัมพันธ์กันแบบเส้นตรงมากน้อยเพียงใด โดยหาค่าสัมประสิทธิ์สหสัมพันธ์ (coefficient of correlation) ซึ่งเขียนเป็นสูตรได้ว่า

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{6.9}$$

เมื่อ r = สัมประสิทธิ์สหสัมพันธ์ ค่าของ r จะอยู่ระหว่าง 1 และ −1 ถ้า r มีค่าใกล้ 1 มาก แสดงว่า x_i และ y_i มีสหสัมพันธ์กันมาก (Hoel,1967)

6.2 การทดสอบนัยสำคัญของสัมประสิทธิ์สหสัมพันธ์ เมื่อมีข้อมูลจับคู่อยู่ชุดหนึ่ง

$\{x_1, y_1; x_2, y_2; x_3, y_3; \ldots\ldots; x_i, y_i; \ldots\ldots; x_n, y_n\}$

ซึ่งประชากร (population) มีจำนวนข้อมูลมาก แต่ละส่วนคือ $\{x_i\}$ กับ $\{y_i\}$ มีการกระจายแบบปกติ ถ้าข้อมูล 2 ส่วนนี้มีสหสัมพันธ์กัน โดยมีค่าสัมประสิทธิ์สหสัมพันธ์

เป็น ρ และถ้าสุ่มเอาตัวอย่างจากประชากรชุดนี้มา 1 กลุ่ม เป็น $\{x_\alpha, y_\alpha, x_\beta, y_\beta, \ldots\}$ มีจำนวนข้อมูล N คู่นำมาหาสัมประสิทธิ์สหสัมพันธ์ได้ r ซึ่ง r ไม่จำเป็นต้องเท่ากับ ρ เมื่อสุ่มตัวอย่างหลาย ๆ ครั้ง (สมมติว่าจำนวนข้อมูลในกลุ่มเท่ากันทุกครั้ง) ก็จะได้ค่า r แตกต่างกันไป โดยทั่วไปแล้วการกระจายของค่า r ไม่เป็นแบบปกติ (normal distribution) ในทางปฏิบัติ การหาสหสัมพันธ์ของประชากรของข้อมูลจับคู่ มักจะหาจากตัวอย่างที่สุ่มออกมาจากประชากร ค่าสัมประสิทธิ์สหสัมพันธ์ที่คำนวณได้ถึงแม้มีค่าไม่เท่า ρ แต่เพื่อความสะดวกเรามักจะใช้ r แทน ρ เนื่องจากค่า r จะกระจายรอบ ๆ ค่า $r = \rho$ เราสามาถบอกขอบเขตการกระจายของค่า r ได้โดยการแปลงค่า r ให้อยู่ในรูปของ z ด้วยสมการ 6.10 และ 6.11 ซึ่งค่า z นี้จะมีการกระจายใกล้เคียงกับการกระจายแบบปกติ (Hoel, 1967) และเพื่อความสะดวกในที่นี้จะยอมรับว่า การกระจายของ z เป็นแบบปกติจริง

ในการแปลง r เป็น z ดังกล่าวคือ การแปลงของฟีชเชอร์ (Fisher transformation) (Hoel, 1967) เขียนเป็นสมการได้ดังนี้

$$z = 0.5 \ln \frac{1+r}{1-r} \quad \text{(Hoel, 1967)} \quad (6.10)$$

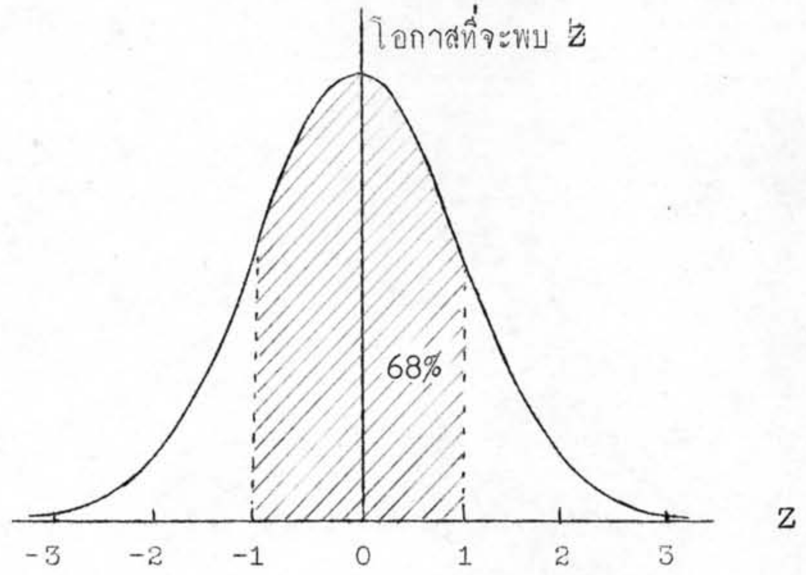
ได้มัชฌิมเลขคณิตของ z เป็น $\mu_z = 0.5 \ln \frac{1+\rho}{1-\rho}$ (Hoel, 1967) (6.11)

และได้ความเบี่ยงเบนมาตรฐานของ z เป็น $\sigma_z = \frac{1}{\sqrt{N-3}}$ (Hoel, 1967) (6.12)

เมื่อ N เป็นจำนวนค่าของ r หรือจำนวนค่าของ z เราอาจแปลง z ต่อไปให้อยู่ในรูปของ Z ซึ่งมีการกระจายแบบปกติเช่นกัน เพื่อนำไปใช้ในการทดสอบ ได้ว่า

$$Z = \frac{z - \mu_z}{\sigma_z} \qquad (6.13)$$

เนื่องจากค่าของ Z มีการกระจายแบบปกติดังในรูป 6.1 จึงสามารถคำนวณหาโอกาสหรือความน่าจะเป็น (probability) ที่จะพบ Z ในช่วงต่าง ๆ ได้ ความน่าจะเป็นที่จะพบ Z ในช่วงต่าง ๆ นี้ได้มีการคำนวณไว้เป็นตารางสำเร็จแล้ว

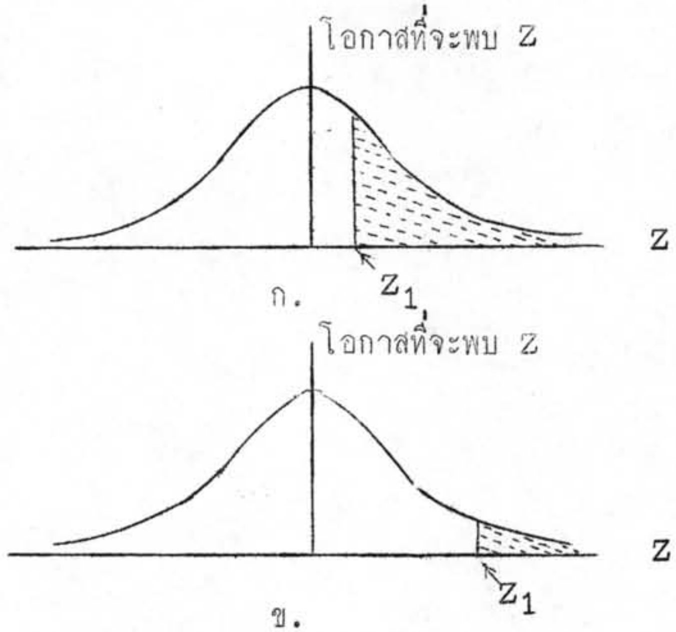


รูป 6.1 แสดงลักษณะของการกระจายแบบปกติ

จากการคำนวณหาโอกาสที่จะได้ค่า Z ในช่วงต่าง ๆ พบว่า โอกาสที่ Z จะมีค่าอยู่ระหว่าง Z_1 และ Z_2 ก็คือพื้นที่ใต้โค้งปกติที่มีขอบเขตอยู่ระหว่าง Z_1 และ Z_2 สำหรับช่วงจาก –1 ถึง 1 จะมีโอกาสที่จะพบค่า Z อยู่ในช่วงนี้เท่ากับ 68% อาศัยสมการ 6.13 หาได้ว่า ช่วงจาก –1 ถึง 1 จะสอดคล้องกับค่า z ในช่วงจาก $\mu_z - \sigma_z$ ถึง $\mu_z + \sigma_z$ ดังนั้นโอกาสที่ z จะอยู่ระหว่าง $\mu_z - \sigma_z$ ถึง $\mu_z + \sigma_z$ จึงมีค่าเท่ากับ 68% ซึ่งโดยทั่วไปถือว่า z ที่มีค่าอยู่ในช่วงนี้ไม่แตกต่างกันอย่างมีนัยสำคัญ

เนื่องจากค่า r ที่หาได้จากตัวอย่างที่สุ่มจากประชากร อาจมาจากประชากรที่ 2 ส่วนไม่มีสหสัมพันธ์กันเลยก็ได้ (คือ $\rho = 0$) ถ้าเราได้ค่า r อยู่ค่าหนึ่ง เช่นค่าเป็น r_1 เราต้องการทดสอบว่า ประชากรที่มี $\rho = 0$ จะมีโอกาสได้ค่า $r = r_1$ มากน้อยเพียงใด การทดสอบทำโดยใช้สมการ 6.10 แปลงค่า r_1 เป็น z_1 และถือว่า z มีการกระจายแบบปกติ ซึ่งมี $\mu_z = 0$ (เพราะ $\rho = 0$) และมี $\sigma_z = \frac{1}{\sqrt{N-3}}$ หลังจากนั้นอาศัยสมการ 6.13 แปลง z_1 ให้เป็น Z_1 แล้วพิจารณาว่า โอกาสที่ $Z \geq Z_1$ มีค่ามากน้อยเพียงใด

(สมมติกรณีนี้ $Z_1 > 0$) ถ้าโอกาสที่ $Z \geq Z_1$ มีค่าน้อย (โปรดดูรูป 6.2 ก และ ข ประกอบ) ก็แสดงว่า r_1 ไม่น่าจะมาจากประชากรซึ่ง 2 ส่วนไม่มีสหสัมพันธ์กันเลย และสรุปได้ว่า ประชากรที่สุ่มตัวอย่างมานั้น มีสหสัมพันธ์กัน



รูป 6.2 กราฟเปรียบเทียบโอกาสที่ z_1 หรือ r_1 จะมาจากประชากรที่ $\rho = 0$

ก. แสดงว่า z_1 หรือ r_1 น่าจะมาจากประชากรที่ $\rho = 0$

ข. แสดงว่า z_1 หรือ r_1 ไม่น่าจะมาจากประชากรที่ $\rho = 0$