บทที่ 1

บทนำ

ที่มาและความสำคัญของปัญหา

การหาความสัมพันธ์ของตัวแปรหรือการวิเคราะห์ความสัมพันธ์เป็นกระบวนการหนึ่งของการวิเคราะห์ทางสถิติที่หาความสัมพันธ์ระหว่างตัวแปรสองตัวขึ้นไป โดยค่าที่ได้จะหมายถึงตัวแปรมีความสัมพันธ์กันในทิศทางใด มาตราวัดความสัมพันธ์ระหว่างตัวแปรได้แก่สัมประสิทธิ์สหสัมพันธ์ (correlation coefficient) ซึ่งเป็นค่าความสัมพันธ์ระหว่างตัวแปรสุ่มสองตัว โดยทั่วไปค่าสัมประสิทธิ์สหสัมพันธ์นิยมใช้ค่าสัมประสิทธิ์สหสัมพันธ์เชิงเส้นแบบเพียร์สัน (pearson productmoment correlation) ซึ่งจะประมาณได้ด้วยค่า r โดยค่า r จะมีค่าตั้งแต่ -1 จนถึง +1 ถ้า r = -1 หมายถึงตัวแปรสองตัวมีความสัมพันธ์กันในทิศทางตรงกันข้าม ถ้า r = 1 หมายถึง ตัวแปรสองตัวมีความสัมพันธ์กันในทิศทางเดียวกัน ถ้า r = 0 ตัวแปรสองตัวไม่มีความสัมพันธ์กัน

ในบางครั้งค่าสัมประสิทธิ์สหสัมพันธ์ (ρ) ที่มีค่าน้อย ($\rho \in [0.3,0.5]$)ทำให้ไม่อาจตัดสินใจจากค่า ρ ได้ว่าตัวแปรสองตัวมีความสัมพันธ์กันหรือไม่ จึงได้มีการทดสอบสมมุติฐานเพื่อทดสอบตัวแปรสองตัวมีความสัมพันธ์กันหรือไม่และได้ทำการประมาณค่า ρ โดยใช้ค่า ρ แบบเพียร์สัน ในบางครั้งข้อมูลที่เก็บมาเป็นข้อมูลที่ถูกตัดทิ้งซึ่งส่วนใหญ่ข้อมูลชนิดนี้จะเกิดขึ้นในทางชีววิทยาและทางการแพทย์ เช่น การศึกษาความสัมพันธ์ระหว่างโรคเส้นเลือดตีบ (arteriosclerosis) กับระยะที่มีชีวิตอยู่ (lenght of life) ของลิงชนิดหนึ่ง ผู้ทดลองกำหนดจำนวนค่าสังเกตที่ต้องการไว้ เมื่อค่าสังเกตครบตามจำนวนที่กำหนดไว้จะทำการหยุดการทดลองแล้วเก็บค่าสังเกตที่ได้จากการทดลอง ในการกำหนดจำนวนค่าสังเกตไว้ล่วงหน้านั้นจะเกิดขึ้นในกรณีที่ผู้ทดลองไม่สามารถรอค่าสังเกตจนครบได้ ดังนั้นการหาค่า ρ แบบเพียร์สันจึงนำมาใช้ในการประมาณค่า ρ ในกรณีนี้ไม่ได้ เนื่องจากข้อมูลที่เก็บได้ในการทดลองนี้เป็นข้อมูลที่ถูกตัดทิ้ง ดังนั้นในปี ค.ศ.1989 Tiku และ Gill ได้ใช้ตัวประมาณภาวะน่าจะเป็นสูงสุดแก้ไข (modified maximum likelihood : MML) ประมาณค่า ρ ขึ้นมาใหม่

การแจกแจงของ ρ จะสมมาตรโดยปกติเมื่อ $\rho = 0$ แต่จะเบ้เมื่อ $\rho \neq 0$ จากคุณ-

สมบัตินี้เอง ในปี ค.ศ. 1921 Fisher และ ค.ศ. 1951 Gayen ได้เสนอตัวสถิติ Z_f[1] เพื่อขจัดความเบ้ของการแจกแจงของ ρ ในกรณีที่เราทดสอบเมื่อ $\rho \neq 0$ แต่ก็ยังมีปัญหาในเรื่องการประมาณค่าเฉลี่ยและค่าความแปรปรวนที่ยังไม่ถูกต้อง ดังนั้นในปี ค.ศ. 1978 Konishi จึงได้เสนอตัวสถิติ Z_k[2] โดยได้ปรับปรุงส่วนค่าความแปรปรวนเสียใหม่ซึ่งทำให้ Z_k มีประสิทธิภาพมากขึ้น ส่วนในกรณีที่จำนวนข้อมูล (n) มีค่าสู่เข้าสู่อนันต์และได้กำหนดจำนวนข้อมูลขาดหายทางซ้าย (left censored data : r_1) จำนวนข้อมูลขาดหายทางขวา (right censored data : r_2) เราพบว่าตัวประมาณภาวะน่าจะเป็นสูงสุดแก้ไข (modified maximum likelihood estimator : MML) จะคล้ายคลึงตัวประมาณภาวะน่าจะเป็นสูงสุด (maximum likelihood estimator : ML) ดังนั้นการแจกแจงของค่า ρ การแจกแจงปกติทำให้ได้ตัวสถิติ Z_v[3] ในปี ค.ศ. 1990 D.C. Vanghan นำตัวสถิติทดสอบ Z_f , Z_k และ Z_v มาเปรียบเทียบค่าอำนาจการทดสอบสำหรับการทดสอบ ค่า ρ เมื่อกำหนดให้ข้อมูลมีการแจกแจงปกติทวิซึ่งพบว่า ตัวสถิติทดสอบ Z_k มีอำนาจการทดสอบสูงกว่าตัวสถิติทดสอบตัวอื่นในกรณีที่ทดสอบค่าสัมประสิทธิ์สหสัมพันธ์ไม่เท่ากับ 0 เมื่อวิเคราะห์ข้อมูลสมบูรณ์ (complete data) และข้อมูลที่ถูกตัดทิ้งทางขวา (right censored data)

ผู้วิจัยจึงได้สนใจทำการศึกษาเปรียบเทียบอำนาจการทดสอบของตัวสถิติทดสอบ 3 ตัวดังนี้

1. Z_f **(Fisher statistics)**
2. Z_k **(Konishi statistics)**
3. Z_v **(Vaughan statistics)**

โดยทำการทดสอบภายใต้การวิเคราะห์ข้อมูลสมบูรณ์ (complete data) และข้อมูลที่ถูกตัดทิ้งทางขวา (right censored data)

---

[1] Z_f เป็นตัวสถิติทดสอบที่อาศัยการแปลงข้อมูลเดิม r โดยใช้ ln กับค่ารากที่สองของค่าความแปรปรวน

[2] Z_k เป็นตัวสถิติทดสอบที่อาศัยการแปลงข้อมูลเดิม r โดยใช้ ln กับค่ารากที่สองของค่าความแปรปรวนที่ขึ้นอยู่กับค่า ρ

[3] Z_v เป็นตัวสถิติทดสอบที่พิจารณาจากค่าอัตราส่วนของข้อมูล r กับค่ารากที่สองของค่าความแปรปรวน

วัตถุประสงค์ของการวิจัย

เพื่อศึกษาเปรียบเทียบอำนาจการทดสอบของตัวสถิติทดสอบทั้ง 3 ตัวซึ่งใช้ทดสอบค่าสัมประสิทธิ์สหสัมพันธ์ เมื่อข้อมูลมีการแจกแจงปกติทวิและแกมมาทวิ

สมมุติฐานของการวิจัย

ตัวสถิติทดสอบ Z_k จะให้อำนาจการทดสอบสูงสุดเมื่อทดสอบค่าสัมประสิทธิ์สหสัมพันธ์ไม่เท่ากับ 0 ทั้งในกรณีวิเคราะห์ข้อมูลสมบูรณ์และข้อมูลที่ถูกตัดทิ้งทางขวา ในสถานการณ์ต่างๆ ตามหลักการของตัวสถิติทดสอบซึ่งอยู่ในบทที่ 2

ข้อตกลงเบื้องต้น

1. การแจกแจงสองตัวแปรซึ่งมีความสัมพันธ์กันเชิงเส้นและมีการแจกแจงปกติทวิ

ถ้า $(X,Y)'$ เป็นเวกเตอร์สุ่มที่มีการแจกแจงปกติทวิ ฟังก์ชันความน่าจะเป็นร่วมของตัวแปรทั้งสองจะอยู่ในรูปของ

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}}\exp\left[\frac{1}{-2(1-\rho^2)}\left\{(x-\mu_X)^2 - 2\rho(x-\mu_X)(y-\mu_Y)+(y-\mu_Y)^2\right\}\right]$$

เราเรียก $(X,Y)'$ ว่าเป็นเวกเตอร์สุ่มที่มีการแจกแจงปกติทวิที่มีเวกเตอร์ค่าเฉลี่ย $\underset{\sim}{\mu} = \begin{bmatrix}\mu_X \\ \mu_Y\end{bmatrix}$

และมีเมตริกซ์ความแปรปรวนร่วม $\Sigma = \begin{bmatrix}\sigma_X{}^2 & \rho\,\sigma_X\,\sigma_Y \\ \rho\,\sigma_X\,\sigma_Y & \sigma_Y{}^2\end{bmatrix}$

เมื่อ ρ คือ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรสุ่ม X และ Y

μ_X คือ ค่าเฉลี่ยของตัวแปรสุ่ม X

μ_Y คือ ค่าเฉลี่ยของตัวแปรสุ่ม Y

$\sigma_X{}^2$ คือ ค่าความแปรปรวนของตัวแปรสุ่ม X

$\sigma_Y{}^2$ คือ ค่าความแปรปรวนของตัวแปรสุ่ม Y

2. การแจกแจงสองตัวแปรซึ่งมีความสัมพันธ์กันเชิงเส้นและมีการแจกแจงแกมมาทวิ

ถ้า $(X_1, X_2)'$ เป็นเวกเตอร์สุ่มที่มีการแจกแจงแกมมาทวิ ฟังก์ชันความน่าจะเป็นร่วมของตัวแปรทั้งสองจะแบ่งออกเป็น 2 กรณี

<u>กรณีที่ 1</u> $\alpha_1 = \alpha_2$ และ $\beta_1 = \beta_2 = 1$

$$\overset{*}{f}(x_1, x_2) = \prod_{j=1}^{2}\left[\left(\frac{1}{\Gamma(\alpha_j)}\right)x_j^{\alpha-1}e^{-x_j}\right]\left[1 + \sum_{j=1}^{\infty}\rho^{j} L_j^{\alpha-1}(x_1)\, L_j^{\alpha-1}(x_2)\right]$$

$$;\ \alpha > 0, 0 \le \rho < 1, x_1 \text{ และ } x_2 > 0$$

<u>กรณีที่ 2</u> $\alpha_1 > \alpha_2$ และ $\beta_1 = \beta_2 = 1$

$$\overset{**}{f}(x_1, x_2) = \prod_{j=1}^{2}\left[\left(\frac{1}{\Gamma(\alpha_j)}\right)x_j^{\alpha_j-1}e^{-x_j}\right]\left[1 + \sum_{j=1}^{2} a_j L_j^{\alpha_1-1}(x_1)\, L_j^{\alpha_2-1}(x_2)\right]$$

$$;\ x_1 \text{ และ } x_2 > 0$$

เมื่อ α คือ พารามิเตอร์แสดงสเกล (scale parameter)

β คือ พารามิเตอร์แสดงรูปร่าง (shape parameter)

$L_j{}^{\alpha-1}(x)$ คือ ลาแกรพหุนาม (laguerre polynomial)

a_j คือ สหสัมพันธ์ระหว่าง X_1 กับ X_2

---

* มีชื่อว่า a symmetric gamma distribution สร้างโดย Sarmanov

** มีชื่อว่า asymmetric bivariate gamma distribution สร้างโดย Sarmanov

3. ความสัมพันธ์ของสองตัวแปรเป็นเชิงเส้น (linear relationship)

4. ในการวิจัยครั้งนี้เราจะใช้อำนาจการทดสอบ (power of test) และความสามารถในการควบคุมความคลาดเคลื่อนประเภทที่ 1 (type I error) เป็นเกณฑ์ในการเลือกตัวสถิติทดสอบ

ขอบเขตการวิจัย

ในการวิจัยครั้งนี้จะทำภายใต้ขอบเขตดังนี้

1. ข้อมูลที่นำมาวิจัยครั้งนี้เป็นการแจกแจงปกติทวิ ซึ่งเรากำหนดให้ μ_1 และ μ_2 เป็นค่าเฉลี่ยของ X และ Y ตามลำดับ σ_1^2และ σ_2^2 เป็นค่าความแปรปรวนของ X และ Y ตามลำดับ และ ρ เป็นค่าสัมประสิทธิ์สหสัมพันธ์ โดยที่ผู้วิจัยสนใจศึกษาเมื่อ $\mu_1 = \mu_2 = 0$ และ $\sigma_1^2 = \sigma_2^2 = 1$

2. ข้อมูลที่นำมาวิจัยครั้งนี้เป็นการแจกแจงแกมมาทวิ ซึ่งเรากำหนดให้ α_1 และ α_2 เป็นสเกลพารามิเตอร์ของ X และ Y ตามลำดับ β_1 และ β_2 เป็นเชพพารามิเตอร์ของ X และ Y ตามลำดับและ ρ เป็นค่าสัมประสิทธิ์สหสัมพันธ์ โดยที่ผู้วิจัยสนใจศึกษาเมื่อ$\alpha_1 = \alpha_2 = 5$,7 และ $\beta_1 = \beta_2 = 1$ ตามลำดับ*

3. ขนาดตัวอย่างที่ใช้ทำการศึกษามี 3 ระดับ 10 , 15 และ 20 ตามลำดับ**

4. กำหนดระดับความเชื่อมั่น 2 ระดับคือ α = 0.05 และ 0.1

---

* ผู้วิจัยทดลองศึกษาเมื่อ $\alpha = 5$ และ 7 จากความสัมพันธ์ระหว่างค่าสัมประสิทธิ์สหสัมพันธ์กับค่าสัมประสิทธิ์การแปรผัน (coefficient of variation : CV) โดยที่ค่าสัมประสิทธิ์การแปรผันจะแปรผกผันกับค่ารากที่สองของ α พบว่าค่า $\alpha = 5$ ซึ่งจะได้ค่า CV = 44.72% แสดงถึงประชากรที่มีการกระจายมากกว่าเมื่อกำหนดค่า $\alpha = 7$ ซึ่งจะได้ค่า CV = 37.79% ดังนั้นเมื่อเรากำหนดค่า $\alpha = 5$ จะแสดงถึงประชากรที่มีความสัมพันธ์กันน้อยในขณะที่เรากำหนดค่า $\alpha = 7$ จะแสดงถึงประชากรที่มีความสัมพันธ์กันชัดเจนขึ้น ส่วนค่า $\beta = 1$ เนื่องจากเมื่อค่า β เปลี่ยนแปลงไปโดยที่ค่า α คงที่พบว่าลักษณะของเส้นโค้งไม่เปลี่ยนแปลง

** ในทางปฏิบัติทางการทดลองเกี่ยวกับข้อมูลอายุของสัตว์ทดลอง ผู้ทดลองมักจะกำหนดขนาดตัวอย่างไม่ใหญ่มากเพราะถ้ากำหนดขนาดตัวอย่างใหญ่มากในกรณีที่มีข้อมูลที่ถูกตัดทิ้ง ผู้ทดลองจะต้องเสียค่าใช้จ่าย และใช้เวลามากในการคำนวณหาค่าความแปรปรวนของตัวสถิติอันดับ

5. การวิเคราะห์จะพิจารณากรณีข้อมูลสมบูรณ์และกรณีข้อมูลที่ถูกตัดทิ้งทางขวา 10% และ 20% ตามลำดับ

6. ผู้วิจัยสนใจศึกษาเมื่อ $\rho = 0$, 0.05 , 0.10 , 0.15 , 0.3 , 0.5 และ 0.8 ตามลำดับ

เกณฑ์การตัดสินใจ

เกณฑ์ตัดสินใจว่าตัวสถิติใดให้อำนาจทดสอบสูงสุด เราจะพิจารณาภายใต้สมมติฐาน

1. $H_0 : \rho = 0$

เทียบกับ $H_1 : \rho \neq 0$

2. $H_0 : \rho = \rho_0 , \rho_0 \neq 0$

เทียบกับ $H_1 : \rho \neq \rho_0$

โดยมีเกณฑ์ดังนี้

1. พิจารณาจากความสามารถในการควบคุมความคลาดเคลื่อนประเภทที่ 1 (type I error) โดยใช้เกณฑ์ของ Bladley* โดยที่เกณฑ์ของ Bladley จะพิจารณาว่าถ้าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 จากการทดลองอยู่ในช่วง (0.025,0.075) และ (0.051,0.150) ณ ระดับนัยสำคัญ 0.05 และ 0.1 ตามลำดับ จะถือว่าการทดสอบนั้นสามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1ได้

2. พิจารณาอำนาจการทดสอบเฉพาะกรณีที่ตัวสถิติทดสอบสามารถควบคุมความน่า-จะเป็นของความคลาดเคลื่อนประเภทที่ 1 ได้เท่านั้น

---

* ในการเลือกเกณฑ์ของ Bradley ผู้วิจัยพิจารณาจากค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 จากการทดสอบเมื่อข้อมูลมีการแจกแจงปกติทวิพบว่าค่าโดยส่วนใหญ่ตกอยู่นอกขอบเขตช่วงของ Cochan แต่มีค่าใกล้เคียงกับขอบเขตช่วงของ Cochan มากจึงน่าที่จะยอมรับการทดสอบนั้นสามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ได้ ดังนั้นผู้วิจัยจึงได้พิจารณาเกณฑ์ของ Bradley ซึ่งมีช่วงที่กว้างกว่าแทน

คำจำกัดความ

1. ความคลาดเคลื่อนประเภทที่ 1 (type I error) คือ ความคลาดเคลื่อนที่เกิดจากการปฏิเสธสมมุติฐานว่าง (H_0) เมื่อสมมุติฐานว่างนั้นเป็นจริง

2. ความคลาดเคลื่อนประเภทที่ 2 (type II error) คือ ความคลาดเคลื่อนที่เกิดจากการยอมรับสมมุติฐานว่าง (H_0) เมื่อสมมุติฐานว่างนั้นไม่จริง

3. อำนาจการทดสอบ (power of test) คือ ค่าความน่าจะเป็นที่จะปฏิเสธสมมุติฐานว่าง (H_0) เมื่อสมมุติฐานว่างนั้นไม่จริง

ประโยชน์ที่คาดว่าจะได้รับ

ผลการศึกษาทำให้ทราบตัวสถิติที่เหมาะสม ซึ่งใช้ในการทดสอบสมมุติฐานค่าสัมประสิทธิ์สหสัมพันธ์ โดยที่ทดสอบค่า $\rho = 0$ หรือทดสอบค่า $\rho \neq 0$ เมื่อข้อมูลมีการแจกแจงปกติทวิและแกมมาทวิ ทั้งในกรณีข้อมูลสมบูรณ์และข้อมูลที่ถูกตัดทิ้งทางขวา