บทที่ 1

บทนำ

1.1 ความเป็นมาของปัญหา

การวิเคราะห์ความสัมพันธ์ (Correlation Analysis) เป็นการศึกษาเกี่ยวกับความสัมพันธ์ระหว่างตัวแปรว่ามีความสัมพันธ์กันขนาดไหน หรือเป็นการศึกษาถึงระดับของความสัมพันธ์ของตัวแปรนั่นเอง ซึ่งมาตราวัดความสัมพันธ์ระหว่างตัวแปรซึ่งเรียกว่า สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) มีอยู่หลายแบบ ซึ่งขึ้นอยู่กับลักษณะของตัวแปรว่าจะเป็นแบบใด โดยทั่ว ๆ ไปแบ่งประเภทของตัวแปรเป็น 4 แบบคือ แบบอัตราส่วน (ratio scale) แบบช่วง (interval scale) แบบจัดอันดับ (ordinal scale) และแบบแบ่งกลุ่ม (nominal scale) ในกรณีที่ตัวแปรเป็นแบบอัตราส่วน หรือแบบช่วง สัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรสองตัวที่รู้จักกันดีที่สุดคือ สัมประสิทธิ์สหสัมพันธ์เชิงเส้นแบบ Pearson (Pearson Product Moment Correlation Coefficient) ที่ใช้สัญลักษณ์ ρ ซึ่งสามารถคำนวณหาได้โดย

$$\rho = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right) \left(\frac{Y-\mu_Y}{\sigma_Y}\right) \right]$$

$$= \frac{N\Sigma XY-(\Sigma X)(\Sigma Y)}{\sqrt{\{N\Sigma X^2-(\Sigma X)^2\}\{N\Sigma Y^2-(\Sigma Y)^2\}}}$$

โดยมีตัวประมาณค่าของ ρ คือ r ซึ่งคำนวณหาได้จาก

$$r = \frac{1}{n-1} E\left[\left(\frac{X-\bar{X}}{S_X}\right) \left(\frac{Y-\bar{Y}}{S_Y}\right) \right]$$

$$= \frac{n\Sigma XY-(\Sigma X)(\Sigma Y)}{\sqrt{\{n\Sigma X^2-(\Sigma X)^2\}\{n\Sigma Y^2-(\Sigma Y)^2\}}}$$

เมื่อ ρ คือสัมประสิทธิ์สหสัมพันธ์ของประชากรระหว่างตัวแปร X และ Y

r คือสัมประสิทธิ์สหสัมพันธ์ของตัวอย่างระหว่างตัวแปร X และ Y

ΣX คือผลรวมของค่าตัวแปร X

ΣY คือผลรวมของค่าตัวแปร Y

ΣXY คือผลรวมของผลคูณระหว่างค่าตัวแปร X และตัวแปร Y

ΣX^2 คือผลรวมของกำลังสองของค่าตัวแปร X

ΣY^2 คือผลรวมของกำลังสองของค่าตัวแปร Y

N คือจำนวนคู่ของค่าตัวแปร X และ Y

n คือจำนวนคู่ของค่าตัวแปร X และ Y จากตัวอย่าง

σ_X คือส่วนเบี่ยงเบนมาตรฐานของประชากรของตัวแปร X

σ_Y คือส่วนเบี่ยงเบนมาตรฐานของประชากรของตัวแปร Y

S_X คือส่วนเบี่ยงเบนมาตรฐานจากตัวอย่างของตัวแปร X

S_Y คือส่วนเบี่ยงเบนมาตรฐานจากตัวอย่างของตัวแปร Y

โดยค่าของ ρ และ r เป็นไปได้ตั้งแต่ -1 ถึง 1 ซึ่งถ้าค่าคำนวณเป็นลบ แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันในทิศทางตรงกันข้าม นั่นคือถ้าตัวแปรหนึ่งมีค่ามาก อีกตัวหนึ่งจะมีค่าน้อย แต่ถ้าค่าคำนวณเป็นบวก แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์เชิงเส้นในทิศทางเดียวกัน นั่นคือถ้าตัวแปรตัวหนึ่งมีค่ามาก (หรือน้อย) อีกตัวหนึ่งจะมีค่ามาก (หรือน้อย) เช่นเดียวกัน ในกรณีที่ค่าคำนวณเท่ากับ ± 1 แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์เชิงเส้นกันมาก หรือมีความสัมพันธ์กันอย่างสมบูรณ์ (Perfect Correlation) แต่ถ้าค่าคำนวณเท่ากับศูนย์ แสดงว่าตัวแปรทั้งสองไม่มีความสัมพันธ์ (Uncorrelation) เชิงเส้น

ส่วนในการวิจัยทางสังคมศาสตร์ส่วนมากแล้ว ข้อมูลที่เก็บรวบรวมมาวิเคราะห์มักจะอยู่ในรูปของข้อมูลแบบแบ่งกลุ่ม หรือข้อมูลเชิงคุณภาพ (Qualitative data) ซึ่งสามารถนำมาจำแนกให้อยู่ในรูปตารางการณ์จร (Contingency table) ได้ ค่าที่ปรากฏในตารางการณ์จรจะเป็นความถี่ของค่าสังเกตที่เก็บรวบรวมมาได้ ซึ่งเรียกว่าข้อมูลจำนวนนับ (Counted data) หรือข้อมูลจำแนกประเภท (Categorical data) ข้อมูลประเภทนี้ส่วนมากการวิเคราะห์จะใช้การทดสอบแบบไคสแควร์ (Chi-square test) โดยมีตัวสถิติทดสอบไคสแควร์ (χ^2) ซึ่งมีสูตรคำนวณดังนี้

$$\chi^2 = \sum_{i}^{r}\sum_{j}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

เมื่อ O_{ij} คือความถี่ของค่าสังเกตในแถวที่ i และสดมภ์ที่ j

E_{ij} คือความถี่คาดหวังของค่าสังเกตในแถวที่ i และสดมภ์ที่ j โดย

$$E_{ij} = \frac{R.C}{T}$$

R คือผลรวมของความถี่ของตัวแปรในแถวที่ i

C คือผลรวมของความถี่ของตัวแปรในสดมภ์ที่ j

T คือผลรวมของความถี่ทั้งหมด

r คือจำนวนแถวของตารางแจกแจง

c คือจำนวนสดมภ์ของตารางแจกแจง

ตัวสถิติไคสแควร์นี้ จะใช้สำหรับทดสอบว่าตัวแปรทางด้านแถวและตัวแปรทางด้านสดมภ์เป็นอิสระต่อกันหรือไม่ ถ้าผลการทดสอบปรากฏว่าตัวแปรทางด้านแถว และตัวแปรทางด้านสดมภ์เป็นอิสระต่อกัน แสดงว่าตัวแปรทั้งสองไม่มีความสัมพันธ์กัน แต่ถ้าผลการทดสอบปรากฏว่าตัวแปรทางด้านแถว และตัวแปรทางด้านสดมภ์ไม่เป็นอิสระต่อกัน แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กัน หรือตัวแปรด้านหนึ่งสามารถอธิบายตัวแปรอีกด้านหนึ่งได้ แต่ในการทดสอบความเป็นอิสระกันของข้อมูลนั้น ค่าไคสแควร์ที่คำนวณได้เพียงแต่เป็นตัวทดสอบว่าตัวแปรด้านแถวและด้านสดมภ์ มีความสัมพันธ์กันหรือไม่เท่านั้น เพราะค่าของไคสแควร์ที่ใช้ในการทดสอบนั้นไม่ได้เป็นตัวชี้ว่า ถ้าค่าไคสแควร์มากตัวแปรทั้งสองจะต้องมีความสัมพันธ์กันมาก หรือถ้าค่าไคสแควร์น้อย ก็สรุปไม่ได้เช่นเดียวกันว่าตัวแปรทั้งสองจะต้องมีความสัมพันธ์กันน้อยตามไปด้วย เพราะได้มีการพิสูจน์แล้วว่าถ้าความถี่ของข้อมูลเพิ่มขึ้นเป็นสัดส่วนเท่า ๆ กันแล้ว ค่าของไคสแควร์ที่คำนวณได้ก็จะเพิ่มขึ้นเท่ากับจำนวนเท่าของสัดส่วนที่เพิ่มขึ้นนั้น[1] ดังนั้นการที่จะนำค่าไคสแควร์มาเป็นตัวบอกความสัมพันธ์ของตัวแปรทั้งสองจึงยังไม่ถูกต้องนัก จึงได้มีนักสถิติหลายท่านได้คิดค้นตัวสถิติที่จะใช้วัดความสัมพันธ์ของข้อมูลที่อยู่ในรูปแบบแบ่งกลุ่ม ดังนี้

---

[1] Hubert M. Blalock, JR "Social Statistics" p. 300-301.

1.1.1 สัมประสิทธิ์ฟาย (Phi-coefficient) เขียนแทนด้วยสัญลักษณ์ ϕ เมื่อ

$$\phi = \sqrt{\chi^2/n}$$

โดย ϕ คือสัมประสิทธิ์ที่ใช้วัดความสัมพันธ์ระหว่างตัวแปรทั้งสอง

χ^2 คือค่าสถิติทดสอบไคสแควร์ที่ใช้ทดสอบความเป็นอิสระระหว่างตัวแปร ตัวแปร 2 ตัว

n คือจำนวนข้อมูลทั้งหมด

1.1.2 สัมประสิทธิ์เงื่อนไขของเพียร์สัน (Pearson's Contingency Coefficient) เขียนแทนด้วยสัญลักษณ์ C เมื่อ

$$C = \sqrt{\frac{\chi^2}{\chi^2+n}}$$

โดย C คือสัมประสิทธิ์ที่ใช้วัดความสัมพันธ์ระหว่างตัวแปร

1.1.3 สัมประสิทธิ์เงื่อนไขของชูโพร (Tschuprow's Contingency Coefficient) เขียนแทนด้วยสัญลักษณ์ T เมื่อ

$$T = \sqrt{\frac{\chi^2/n}{\sqrt{(r-1)(c-1)}}}$$

โดย T คือสัมประสิทธิ์ที่ใช้วัดความสัมพันธ์ระหว่างตัวแปรทั้งสอง

r คือจำนวนแถวของตารางการณ์จร

c คือจำนวนสดมภ์ของตารางการณ์จร

1.1.4 สัมประสิทธิ์เงื่อนไขของคราเมร์ (Cramér's Contingency Coefficient) เขียนแทนด้วยสัญลักษณ์ V เมื่อ

$$V = \sqrt{\frac{\chi^2}{n\min\{(r-1),(c-1)\}}}$$

โดย V คือสัมประสิทธิ์ที่ใช้วัดความสัมพันธ์ระหว่างตัวแปรทั้งสอง

1.1.5 สัมประสิทธิ์การทำนายของกัทแมน (Guttman's coefficient of Optimal Predictability) เขียนแทนด้วยสัญลักษณ์ λ เมื่อ

$$\lambda = \frac{(\Sigma f_r + \Sigma f_c)-(F_r+F_c)}{2n-(F_r+F_c)}$$

โดย λ คือสัมประสิทธิ์ที่ใช้วัดความสัมพันธ์ระหว่างตัวแปรทั้งสอง

f_r คือความถี่สูงสุดที่พบในแต่ละแถวของตารางการณ์จร

f_c คือความถี่สูงสุดที่พบในแต่ละสดมภ์ของตารางการณ์จร

F_r คือความถี่สูงสุดที่พบในยอดรวมของแต่ละแถวในตารางการณ์จร

F_c คือความถี่สูงสุดที่พบในยอดรวมของแต่ละสดมภ์ในตารางการณ์จร

1.1.6 สัมประสิทธิ์แบบกูดแมนและครัสคัล (Goodman and Kruskal's Tau) เขียนแทนด้วยสัญญลักษณ์ τ เมื่อ

$$\tau = \frac{\{\sum_{ij}\sum f^2_{ij}/f_{.j} - \Sigma f^2_{i.}/n\}+\{\sum_{ij}\sum f^2_{ij}/f_{i.} - \Sigma f^2_{.j}/n\}}{2n-\{\sum_i f^2_{i.}/n + \sum_j f^2_{.j}/n\}}$$

โดย τ คือสัมประสิทธิ์ที่ใช้วัดความสัมพันธ์ระหว่างตัวแปรทั้งสอง

f_{ij} คือความถี่ของข้อมูลในแถวที่ i และสดมภ์ที่ j

$f_{.j}$ คือผลรวมของความถี่ของข้อมูลในสดมภ์ที่ j

$f_{i.}$ คือผลรวมของความถี่ของข้อมูลในแถวที่ i

n คือจำนวนข้อมูลทั้งหมด

ดังนั้นเมื่อมีตัวสถิติที่จะใช้วัดความสัมพันธ์ของข้อมูลในตารางการณ์จรหลายตัว จึงเป็นที่น่าสนใจว่า ตัวสถิติที่ใช้วัดความสัมพันธ์ดังกล่าวข้างต้น ตัวใดจะให้ผลการทดสอบที่มีความถูกต้องเชื่อถือได้มากกว่าตัวสถิติใดบ้าง และตัวสถิติแต่ละตัวนั้นเหมาะสมที่จะนำไปหาความสัมพันธ์เมื่อข้อมูลมีลักษณะใด

1.2 <u>วัตถุประสงค์ของการวิจัย</u>

เพื่อเปรียบเทียบความถูกต้องเชื่อถือได้ของตัวสถิติที่ใช้วัดความสัมพันธ์ระหว่างตัวแปรแบบกลุ่ม 2 ตัว 6 ชนิด คือ สัมประสิทธิ์ฟาย สัมประสิทธิ์เงื่อนไขของเพียร์สัน สัมประสิทธิ์เงื่อนไขของชูโพร สัมประสิทธิ์เงื่อนไขของคราแมร์ สัมประสิทธิ์การทำนายของกัทแมน และสัมประสิทธิ์แบบกูดแมนและครัสคัล

1.3 <u>สมมติฐานของการวิจัย</u>

ตัวสถิติที่ใช้วัดความสัมพันธ์ระหว่างตัวแปรแบบแบ่งกลุ่ม 2 ตัวทั้ง 6 ชนิด ให้ผลแตกต่างกัน โดยคาดว่าตัวสถิติคราแมร์น่าจะใช้วัดได้ดีที่สุด

1.4 <u>ขอบเขตของการวิจัย</u>

1.4.1 ศึกษาเฉพาะกรณีที่ข้อมูลมีการแจกแจงแบบปกติสองตัวแปร (Bivariate normal distribution)

1.4.2 ขนาดตัวอย่างที่ใช้ในการวิเคราะห์เท่ากับ 20 30 50 100 200 และ 500 และตารางการณ์จรที่ใช้ในการวิเคราะห์มีขนาด 2x2 2x3 2x4 3x3 3x4 3x5 4x4 4x5 และ 5x5

1.4.3 จำลองข้อมูลในแต่ละลักษณะจำนวน 500 ครั้ง

1.5 <u>ประโยชน์ที่คาดว่าจะได้รับ</u>

เพื่อช่วยให้นักวิจัยโดยเฉพาะทางด้านสังคมศาสตร์สามารถเลือกใช้ตัวสถิติที่ใช้วัดความสัมพันธ์ระหว่างตัวแปรแบบแบ่งกลุ่ม 2 ตัว ได้อย่างเหมาะสม