บทที่ 2

ความรู้และทฤษฎีพื้นฐานสำหรับการวิจัย

ในการวิจัยเรื่องนี้ สถิติที่ใช้ศึกษาคือ สัมประสิทธิ์สหสัมพันธ์ และการทดสอบแบบไคสแควร์ สำหรับสัมประสิทธิ์สหสัมพันธ์ ตัวแปรที่นำมาศึกษาจะต้องมีการแจกแจงแบบปกติสองตัวแปร สำหรับการทดสอบแบบไคสแควร์ในกรณีศึกษาตัวแปรแบบไม่ต่อเนื่อง ตัวแปรทั้งสองจะมีการแจกแจงร่วมแบบพหุนามที่มาจากประชากรเดียวกัน ในการจำลองข้อมูลทั้งสองแบบทำได้โดยวิธีการจำลองแบบ ดังนั้น ในบทนี้จะกล่าวถึงรายละเอียดของสถิติแต่ละวิธี พร้อมทั้งความรู้ที่เกี่ยวข้องอย่างละเอียด ดังนี้

2.1 <u>การแจกแจงแบบปกติสองตัวแปร</u> (The bivariate normal distribution)

ถ้า (x, y) เป็นตัวแปรที่มีฟังก์ชันการแจกแจงความน่าจะเป็นร่วม

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[- \frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right)+\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\}\right]$$

ภายใต้เงื่อนไข

1. $-\infty < x < \infty$, $-\infty < y < \infty$

2. σ_1 , σ_2 , μ_1 , μ_2 และ ρ เป็นค่าพารามิเตอร์คงที่

โดยที่ ρ คือค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร x, y $(-1 < \rho < 1)$

μ_1 คือค่าเฉลี่ยของตัวแปร x $(-\infty < \mu_1 < \infty)$

μ_2 คือค่าเฉลี่ยของตัวแปร y $(-\infty < \mu_2 < \infty)$

σ_1^2 คือความแปรปรวนของตัวแปร x $(\sigma_1^2 > 0)$

σ_2^2 คือความแปรปรวนของตัวแปร y $(\sigma_2^2 > 0)$

เราจะเรียก (x, y) ว่าเป็นตัวแปรที่มีการแจกแจงแบบปกติสองตัวแปร

ที่มีค่าเฉลี่ย $\underset{\sim}{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$

และเมทริกซ์ความแปรปรวนร่วม $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

2.2 การแจกแจงแบบพหุนามร่วม (Joint multinomial distribution)

ในการศึกษาคุณลักษณะทางประชากร 2 ลักษณะที่สนใจคือ A และ B และจากผลการศึกษาคุณลักษณะ A สามารถแบ่งออกได้ r ระดับ คือ $A_1, A_2, \ldots, A_r$ และคุณลักษณะ B สามารถแบ่งได้ c ระดับ คือ $B_1, B_2, \ldots, B_c$ สมมติทำการทดลอง n ครั้งอย่างเป็นอิสระต่อกัน ปรากฏผลการทดลองดังนี้คือ A_1และ B_1 จะเกิดขึ้น x_{11} ครั้ง A_1 และ B_2 เกิดขึ้น x_{12} ครั้ง ... A_i และ B_j จะเกิดขึ้น x_{ij} ครั้ง ด้วยความน่าจะเป็น $p_{11}, p_{12}, \ldots, p_{ij}$ $(i = 1, 2, \ldots, r$ และ $j = 1, 2, \ldots, c)$ ตามลำดับ ดังนั้นฟังก์ชั่นการแจกแจงความน่าจะเป็นร่วมคือ $f(x_{11}, x_{12}, \ldots, x_{rc} ; p_{11}, p_{12}, \ldots, p_{rc}) = \binom{n}{x_{11}, x_{12}, \ldots, x_{rc}} \prod_i \prod_j p_{ij}^{x_{ij}}$

เมื่อ x_{ij} = จำนวนหน่วยตัวอย่าง (ความถี่) ที่สอดคล้องกับระดับที่ i ของคุณลักษณะ A และสอดคล้องกับระดับที่ j ของคุณลักษณะ B (จำนวนความถี่ที่ตกในเซล (i, j))

p_{ij} = ความน่าจะเป็นที่หน่วยตัวอย่างใด ๆ จะตกในเซล (i, j) หรือนัยหนึ่งคือ ความน่าจะเป็นที่หน่วยตัวอย่างใด ๆ จะมีคุณลักษณะ A ระดับที่ i และคุณลักษณะ B ระดับที่ j

และ $\sum_i \sum_j x_{ij} = n$ (จำนวนหน่วยตัวอย่างทั้งหมด)

2.3 สัมประสิทธิ์สหสัมพันธ์[1] (Correlation coefficient)

ในการศึกษาระดับความสัมพันธ์ระหว่างตัวแปรสองตัว หรือลักษณะที่สนใจศึกษาสองลักษณะ โดยที่ไม่มีตัวแปรใดหรือลักษณะใดถูกกำหนดค่าไว้ล่วงหน้า จะพิจารณาได้จากค่าสัมประสิทธิ์สหสัมพันธ์(Correlation coefficient) สัมประสิทธิ์สหสัมพันธ์ประชากรสัญญลักษณ์ที่ใช้คือ ρ (rho) ค่าประมาณของสัมประสิทธิ์สหสัมพันธ์ ρ คือ สัมประสิทธิ์สหสัมพันธ์ตัวอย่าง (sample correlation) ใช้สัญญลักษณ์ r หรือ r_{xy}

---

[1] ในที่นี้จะหมายถึงสัมประสิทธิ์สหสัมพันธ์เชิงเส้นเท่านั้น

สัมประสิทธิ์สหสัมพันธ์ระหว่างสองตัวแปร ที่รู้จักกันดีที่สุดคือ สัมประสิทธิ์สหสัมพันธ์เชิงเส้นแบบเพียร์สัน ซึ่ง คาร์ล เพียร์สัน (Karl Pearson) เป็นผู้เริ่มใช้มาตั้งแต่ปี ค.ศ. 1895 บางทีจึงเรียกสัมประสิทธิ์สหสัมพันธ์นี้ว่า สัมประสิทธิ์สหสัมพันธ์ผลคูณโมเมนต์ของ เพียร์สัน (Pearson Product-Moment Correlation Coefficient) สูตรที่ใช้คำนวณ คือ

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

โดยที่ r คือ สัมประสิทธิ์สหสัมพันธ์ตัวอย่างระหว่างตัวแปร X กับตัวแปร Y

ΣX คือ ผลรวมของค่าข้อมูลของตัวแปร X

ΣY คือ ผลรวมของค่าข้อมูลของตัวแปร Y

ΣXY คือ ผลรวมของผลคูณระหว่างค่าข้อมูลของตัวแปร X และตัวแปร Y

ΣX^2 คือ ผลรวมของกำลังสองของค่าข้อมูลของตัวแปร X

ΣY^2 คือ ผลรวมของกำลังสองของค่าข้อมูลของตัวแปร Y

n คือ จำนวนคู่ของข้อมูล

และค่าของ X และ Y จะต้องเป็นค่าที่วัดในลักษณะควบคู่กัน (bivariate (x, y) data)

<u>ข้อตกลงเบื้องต้น</u> (Assumption)

1. ตัวแปรทั้งสองต้องเป็นค่าต่อเนื่อง และมีการแจกแจงปกติสองตัวแปร (bivariate normal distribution)

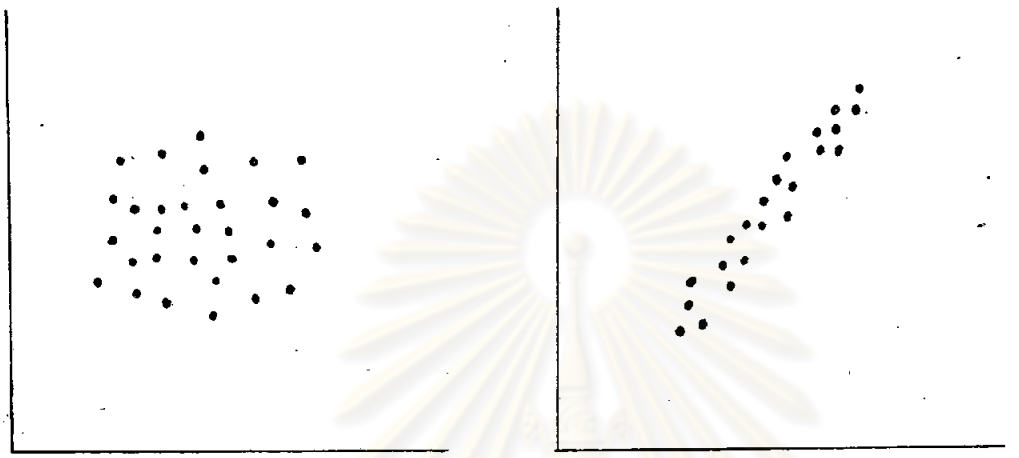
2. ความสัมพันธ์ระหว่างตัวแปรทั้งสองเป็นแบบเส้นตรง (linear relationship)

3. ข้อมูล (x_i, y_i) $(i = 1, 2, \ldots, n)$ เป็นตัวอย่างเชิงสุ่ม

สัมประสิทธิ์สหสัมพันธ์มีค่าอยู่ระหว่าง -1 ถึง +1 ถ้าสัมประสิทธิ์สหสัมพันธ์มีค่าสัมบูรณ์เข้าใกล้ 1 แสดงว่าตัวแปรคู่ที่กำลังศึกษามีความสัมพันธ์เชิงเส้นในระดับสูง ถ้าสัมประสิทธิ์สหสัมพันธ์มีค่าสัมบูรณ์ไม่มากนัก (ประมาณ 0.3 - 0.7) แสดงว่ามีความสัมพันธ์เชิงเส้นในระดับปานกลาง แต่ถ้าสัมประสิทธิ์สหสัมพันธ์มีค่าเข้าใกล้ 0 แสดงว่ามีความสัมพันธ์กันน้อย และถ้าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0 แสดงว่าตัวแปรทั้งสองไม่มีความสัมพันธ์เชิงเส้น

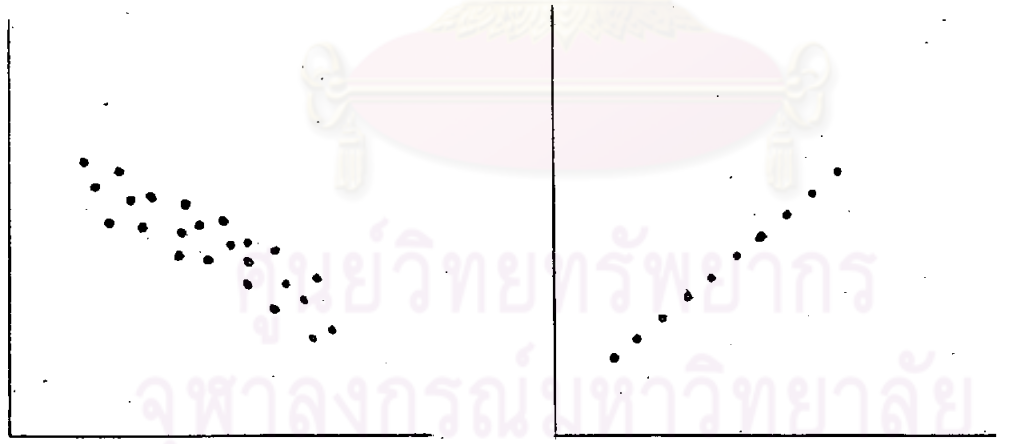
สำหรับทิศทางของความสัมพันธ์ จะพิจารณาได้จากเครื่องหมายของสัมประสิทธิ์สหสัมพันธ์ ถ้าสัมประสิทธิ์สหสัมพันธ์มีค่าเป็นบวก แสดงว่าเมื่อตัวแปรตัวหนึ่งมีค่าเพิ่มขึ้น ตัวแปรอีกตัวหนึ่งจะมีค่าเพิ่มขึ้นตาม แต่ถ้าสัมประสิทธิ์สหสัมพันธ์มีค่าเป็นลบ แสดงว่าเมื่อตัวแปรตัวหนึ่งมีค่าเพิ่มขึ้น ตัวแปรอีกตัวหนึ่งจะมีค่าลดลง นั่นคือเป็นไปในทางตรงกันข้าม อย่างไรก็ตามผลที่ได้จากการคำนวณค่า χ^2 แล้วตีความในรูปของค่าสัมประสิทธิ์สหสัมพันธ์ ยังไม่สามารถบอกทิศทางของความสัมพันธ์ได้ ต้องอาศัยวิธีการอื่น ๆ ประกอบ จึงจะสามารถบ่งชี้ได้

รูปที่ 2.1 แสดงค่าสัมประสิทธิ์สหสัมพันธ์ในระดับความสัมพันธ์ต่าง ๆ



r = 0

r = .8



r = -.7

r = 1

สัมประสิทธิ์สหสัมพันธ์เป็นตัวเลขไม่มีหน่วย ไม่ขึ้นอยู่กับขนาดมากน้อยของค่าตัวแปร ไม่จำเป็นที่ตัวแปรทั้งสองจะต้องมีหน่วยเดียวกัน และความสัมพันธ์ที่ได้เป็นการแสดงแต่เพียงว่าการเกิดของสองปรากฏการณ์นั้นมีความสัมพันธ์เชิงบวกหรือเชิงลบเท่านั้น มิได้หมายความว่าอะไรเป็นเหตุหรือเป็นผลซึ่งกันและกันแต่อย่างใด

2.4 <u>การจำลองแบบ</u> (Simulation)

ในการศึกษาปัญหาซึ่งมีข้อจำกัดว่าข้อมูลที่นำมาวิเคราะห์นั้น จะต้องมีคุณสมบัติตามข้อสมมติเบื้องต้น เช่น สามารถระบุค่าเฉลี่ย ความแปรปรวน หรือลักษณะการแจกแจงของประชากรได้นั้น วิธีการสร้างข้อมูลที่มีคุณสมบัติตามต้องการ กระทำได้โดยวิธีการที่เรียกว่าเทคนิคการจำลองแบบ (Simulation technique) เป็นเทคนิคการทดลองพื้นฐานที่มีความรวดเร็ว และเสียค่าใช้จ่ายถูกกว่าในการทดลองจริง ซึ่งอาจทำได้ยาก หรือไม่สามารถทำได้เพราะมีปัจจัยหรือข้อจำกัดมากมาย

สำหรับการศึกษาในเรื่องนี้ ได้ใช้เทคนิคการจำลองแบบสร้างข้อมูลที่มีการแจกแจงแบบปกติสองตัวแปร และพหุนามแบบสองตัวแปร เครื่องมือที่สำคัญในเทคนิคการจำลองแบบ คือเทคนิคการผลิตเลขสุ่ม (Technique of random number generation) ซึ่งเป็นพื้นฐานในการสร้างลักษณะการแจกแจงแบบต่าง ๆ เช่น การแจกแจงแบบทวินาม การแจกแจงแบบปัวซอง การแจกแจงแบบแกมมา เป็นต้น การจำลองแบบจะใกล้เคียงความเป็นจริงเพียงใดนั้น ขึ้นอยู่กับคุณภาพของตัวเลขสุ่มเป็นสำคัญ ตัวเลขสุ่มที่ได้จะถูกสร้างให้มีลักษณะสุ่ม (random) และให้แต่ละหมายเลขมีโอกาสเกิดขึ้นเท่าเทียมกัน

การสร้างข้อมูลที่มีการแจกแจงแบบปกติสองตัวแปร จะสร้างตัวเลขสุ่มที่มีการแจกแจงแบบสม่ำเสมอ (Uniform distribution) ขึ้นมาก่อน โดยใช้สับรูทีนแรนดู (Subroutine Randu) จากตัวเลขสุ่มที่มีการแจกแจงแบบสม่ำเสมอในช่วง (0, 1) นี้ จะนำมาสร้างตัวแปรปกติมาตรฐานโดยใช้สับรูทีนเกาส์ (Subroutine Gauss) เรียกค่าที่ได้ครั้งแรกนี้ว่า x ต่อจากนั้นจะสร้างตัวเลขสุ่มที่มีการแจกแจงแบบสม่ำเสมอขึ้นมาอีกชุดหนึ่ง ซึ่งไม่ขึ้นกับตัวเลขสุ่มที่มีการแจกแจงแบบสม่ำเสมอในชุดแรก โดยใช้สับรูทีนแรนดูเช่นกัน จากนั้นจะนำตัวเลขสุ่มนี้มาสร้างตัวแปรปกติมาตรฐาน และเรียกค่าที่ได้นี้ว่า y โดยค่า y จะมีลักษณะการแจกแจงที่ขึ้นอยู่กับค่า x ที่ได้จากครั้งแรก ผลลัพธ์ที่ได้คือค่า (x, y) จะเป็นตัวแปรสุ่มปกติสองตัวแปร

การสร้างตัวแปรปกติโดยใช้สับรูทีนเกาส์ อาศัยทฤษฎีสถิติ คือ ทฤษฎีการโน้มสู่ค่ากลาง (Central limit theorem)

ทฤษฎี 2.4.1[1] กำหนดให้ $y_1, y_2 \ldots, y_n$ ซึ่งเป็นอิสระต่อกัน เป็นตัวอย่างสุ่มขนาด n จากประชากรที่มีค่าเฉลี่ย μ และค่าความแปรปรวน σ^2 การแจกแจงของค่าสถิติ

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

จะประมาณได้ด้วยการแจกแจงแบบปกติที่มีค่าเฉลี่ย μ และค่าความแปรปรวน $\frac{\sigma^2}{n}$

เนื่องจากเกาส์ ใช้ตัวแปรที่มีการแจกแจงแบบสม่ำเสมอในช่วง (0, 1) 12 ตัว ($n = 12$) $\bar{y}$ จะมีค่าเฉลี่ย $\frac{1}{2}$ และค่าความแปรปรวน $\frac{1}{144}$

จะได้ $\bar{y} = \frac{1}{12} \sum_{i=1}^{12} Y_i \sim N(\frac{1}{2}, \frac{1}{144})$

$$\therefore x = \frac{\bar{y} - \frac{1}{2}}{\sqrt{1/144}}$$

$$= \sum_{i=1}^{12} y_i - 6.0$$

จะมีการแจกแจงโดยประมาณแบบปกติมาตรฐาน

2.5 ตารางการณ์จร (Contingency table)

การทดสอบความเป็นอิสระ หรือทดสอบความสัมพันธ์ระหว่างตัวแปร 2 ตัว โดยใช้การทดสอบแบบไคสแควร์ จะต้องนำเสนอข้อมูลให้อยู่ในรูปตารางแจกแจงความถี่ของตัวแปรทั้งสอง (Bivariate frequency table) หรือที่เรียกกันว่า ตารางการณ์จร (Contingency table)

ตารางการณ์จร[2] เป็นตารางแสดงถึงการแจกแจงความถี่ของข้อมูล ซึ่งจำแนกตามตัวแปร 2 ตัวพร้อม ๆ กัน สำหรับตัวแปรแต่ละตัวอาจจำแนกออกเป็น 2, 3, 4... ลักษณะ

---

[1] ดูพิสูจน์จาก Robert V. Hogg and Allen T. Craig "Introduction to Mathematical Statistics หน้า 182-183

[2] วีนัส พีชวณิชย์ และ สมจิต วัฒนาชยากูล "สถิติสำหรับนักสังคมศาสตร์" พิมพ์ครั้งที่ 5 กรุงเทพมหานคร สำนักพิมพ์ประกายพรึก, 2527

ฉะนั้น ตารางการณ์จรอาจเป็นชนิด 2 x 2 2 x 3 3 x 2 3 x 3 ... ฯลฯ ตารางการณ์จรที่ได้จากการจำแนกประเภทของหน่วยตัวอย่าง โดยอาศัยค่าของตัวแปร 2 ตัวนี้ จะเรียกได้อีกชื่อหนึ่งว่า ตารางการณ์จรสองทาง (Two-way contingency table)

ถ้ากำหนดสัญญลักษณ์สำหรับตารางการณ์จรว่า ให้ตารางมีแถวนอน r แถว และแถวตั้ง c แถว นั่นคือจะพิจารณาตารางการณ์จรชนิด r x c และให้ O_{ij} เป็นค่าของข้อมูลที่เก็บรวบรวมได้ในแถวนอนที่ i และแถวตั้งที่ j ซึ่งตารางการณ์จรชนิด r x c แสดงได้ดังนี้

ตารางที่ 2.1 แสดงตารางการณ์จรชนิด r x c

	จำแนกแถวตั้ง 1	2	...	j	...	c	รวมแถวนอน
จำแนกแถวนอน 1	O_{11}	O_{12}	...	O_{1j}	...	O_{1c}	$O_{1.}$
2	O_{21}	O_{22}	...	O_{2j}	...	O_{2c}	$O_{2.}$
.							
.							
.							
i	O_{i1}	O_{i2}	...	O_{ij}	...	O_{ic}	$O_{i.}$
.							
.							
.							
r	O_{r1}	O_{r2}	...	O_{rj}	...	O_{rc}	$O_{r.}$
รวมแถวตั้ง	$O_{.1}$	$O_{.2}$	...	$O_{.j}$	...	$O_{.c}$	n

ผลรวมของแต่ละแถวนอนคือ $O_{1.}, O_{2.}, \ldots, O_{r.}$

ผลรวมของแต่ละแถวตั้งคือ $O_{.1}, O_{.2}, \ldots, O_{.c}$

นั่นคือ $O_{i.} = \sum_{j=1}^{c} O_{ij}$, $O_{.j} = \sum_{i=1}^{r} O_{ij}$

n คือจำนวนความถี่ทั้งหมดที่เก็บรวบรวมมา หรือคือขนาดตัวอย่าง

โดยที่ $\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} = \sum_{i=1}^{r} O_{i.} = \sum_{j=1}^{c} O_{.j} = n$

ในการจำแนกข้อมูลตามแถวนอนและตามแถวตั้งนี้ ข้อมูลในแต่ละช่องจะไม่ซ้ำซ้อนกัน (mutually exclusive and exhaustive class) ดังนั้นข้อมูลแต่ละค่าจะต้องตกในแถวนอนแถวใดแถวหนึ่ง และแถวตั้งแถวใดแถวหนึ่ง

2.6 การทดสอบแบบไคสแควร์ (Chi-square test)

ในการทดสอบความเป็นอิสระระหว่างตัวแปรสองตัว ในกรณีข้อมูลเชิงคุณภาพหรือแม้แต่ข้อมูลเชิงปริมาณ นักวิจัยมักจะนิยมใช้การทดสอบแบบไคสแควร์ทดสอบ ซึ่งตัวสถิติของการทดสอบนี้คือ

$$W = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad , W \sim \chi^2$$

โดยที่ $E_{ij} = \frac{O_{i.}\, O_{.j}}{n}$

ตัวสถิตินี้จะประมาณได้ด้วยการแจกแจงแบบไคสแควร์ ดังนั้น ในส่วนนี้จะกล่าวถึงการแจกแจงแบบไคสแควร์ก่อน แล้วจึงจะกล่าวถึงรายละเอียดของการทดสอบแบบไคสแควร์ดังนี้

2.6.1 การแจกแจงแบบไคสแควร์ (Chi-square distribution)

ตัวแปรสุ่มไคสแควร์ คือตัวแปรสุ่มที่พัฒนาขึ้นมาจากฟังก์ชันของตัวแปรสุ่มปกติ กล่าวคือ ถ้า $x_1, x_2, \ldots, x_n$ เป็นตัวแปรสุ่มที่มีการแจกแจงปกติโดยมี $E(x_i) = \mu$ และ $V(x_i) = \sigma^2$, $i = 1, 2, \ldots, n$ ถ้าสุ่มตัวอย่างขึ้นมา 1 ตัวแล้วแปลงเป็นคะแนนปกติมาตรฐาน Z ดังนี้

$$Z = \frac{x - \mu}{\sigma} \sim N(0, 1)$$

ถ้ายกกำลังสอง $Z^2 = (\frac{x - \mu}{\sigma})^2 = W \quad , \quad W \sim \chi^2_{(1)}$

จะได้ว่า Z^2 จะมีการแจกแจงแบบไคสแควร์ที่มีขั้นแห่งความเป็นอิสระเท่ากับ 1 ในทำนองเดียวกัน ถ้าสุ่มกลุ่มตัวอย่างขนาด n ซึ่งเป็นอิสระต่อกันจะได้ว่าผลรวมของ Z^2 จะมีการแจกแจงแบบไคสแควร์ที่มีขั้นแห่งความเป็นอิสระเท่ากับ n ดังนี้

$$W = Z_1^2 + Z_2^2 + \ldots + Z_n^2 \quad , \quad W \sim \chi^2_{(n)}$$

สำหรับฟังก์ชั่นความน่าจะเป็นของการแจกแจงไคสแควร์แสดงได้ดังนี้

$$f(x) = \frac{1}{2^{n/2}\left(\frac{n}{2}-1\right)!}\, x^{\frac{n}{2}-1}\, e^{-x/2} \quad \text{เมื่อ } x > 0$$

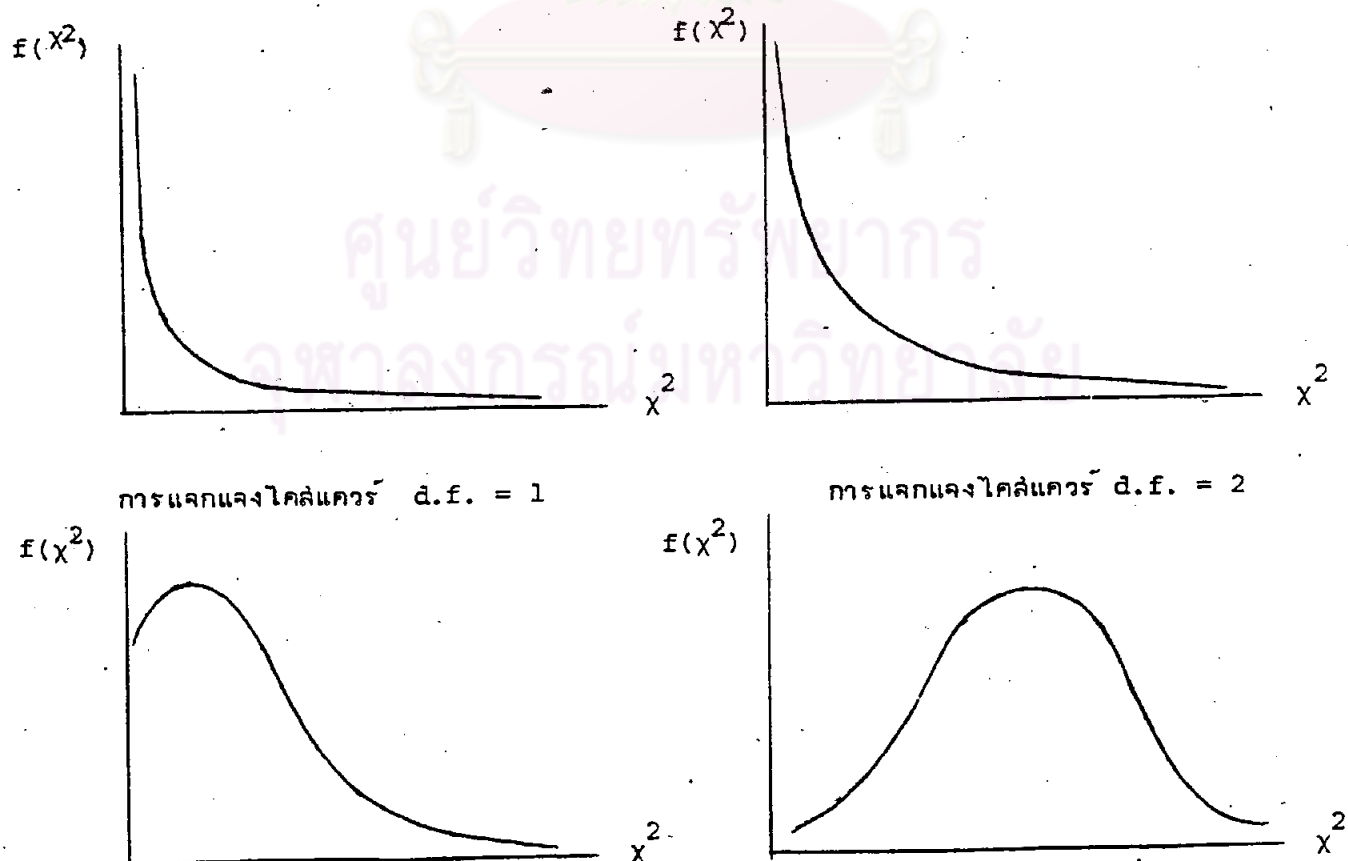
n เป็นจำนวนขั้นแห่งความเป็นอิสระ

e เป็นจำนวนอตรรยะ(irrational) มีค่าเท่ากับ 2.71828 (โดยประมาณ)

จากฟังก์ชั่นความน่าจะเป็นของการแจกแจงไคสแควร์ จะเห็นว่าการแจกแจงไคสแควร์ขึ้นอยู่กับจำนวนขั้นแห่งความเป็นอิสระ ส่วนโค้งการแจกแจงจะมีค่าไคสแควร์เป็นแอบซิสซ่า (Abcissa) และมีค่าฟังก์ชั่นความน่าจะเป็น (f(x)) เป็นออร์ดิเนต (Ordinate)

การแจกแจงไคสแควร์จะเบ้ขวา เมื่อจำนวนขั้นแห่งความเป็นอิสระมีค่าน้อย และความเบ้จะลดลงไปเมื่อจำนวนขั้นแห่งความเป็นอิสระมีค่าเพิ่มขึ้น และถ้าจำนวนขั้นแห่งความเป็นอิสระมีค่ามาก (ประมาณ 20) การแจกแจงค่อนข้างจะสมมาตร จึงอาจประมาณการแจกแจงไคสแควร์ด้วยการแจกแจงปกติได้

รูปที่ 2.2 แสดงการแจกแจงไคสแควร์เมื่อจำนวนขั้นแห่งความเป็นอิสระแตกต่างกัน



คุณสมบัติของการแจกแจงไคสแควร์

1. การแจกแจงไคสแควร์เป็นการแจกแจงแบบต่อเนื่อง

2. ค่าไคสแควร์เป็นได้ตั้งแต่ 0 ไปจนถึง ∞ นั่นคือค่าไคสแควร์เป็นค่าบวกเสมอ

3. มี moment generating function (m.g.f.) = $(\frac{1}{1-2t})^{n/2}$ มีค่าเฉลี่ยเท่ากับ n ค่าความแปรปรวนเท่ากับ 2n และมี mode ที่ n-1 (n = d.f.)

4. ลักษณะของส่วนโค้งจะเบ้ไปทางขวามือ

5. เป็นส่วนโค้งแบบเพียร์สันชนิดที่สาม (Pearson Type III Curve)

6. การกระจายเป็นไปตามกฎของการแจกแจงแบบมัลติโนเมียล (Multinomial Distribution)

ค่าวิกฤติของการแจกแจงไคสแควร์

เนื่องจากการแจกแจงไคสแควร์แตกต่างกันไปตามค่าของจำนวนชั้นแห่งความเป็นอิสระ ดังนั้นในการสร้างตารางไคสแควร์จึงกำหนดค่าวิกฤติ (Critical Value) ในรูปของจำนวนชั้นแห่งความเป็นอิสระ และระดับความมีนัยสำคัญต่าง ๆ กันตั้งแต่ระดับ 0.99 จนถึงระดับ 0.001

จำนวนชั้นแห่งความเป็นอิสระ (degrees of freedom)

ในการทดสอบความเป็นอิสระระหว่างสองตัวแปร จะต้องนำเสนอข้อมูลในรูปของตารางการณ์จรชนิด r x c ดังนั้นจำนวนชั้นแห่งความเป็นอิสระจึงขึ้นอยู่กับขนาดของตาราง ดังนี้

จำนวนพารามิเตอร์ที่จำเป็นต้องประมาณจากตาราง $= r-1+c-1 = r+c-2$

สำหรับตารางการณ์จรชนิด r x c ข้อมูลที่เก็บรวบรวมมาคือ rc ตัว

$\therefore$ จำนวนชั้นแห่งความเป็นอิสระ $= rc-1-(r+c-2)$

$= rc-r-c+1$

$= (r-1)(c-1)$

r คือ จำนวนแถวนอน (row)

c คือ จำนวนแถวตั้ง (column)

ความสำคัญของการแจกแจงไคสแควร์

การแจกแจงไคสแควร์เป็นการแจกแจงชนิดต่อเนื่อง แต่มักใช้กับข้อมูลชนิดไม่ต่อเนื่อง โดยเฉพาะเกี่ยวกับความถี่ของข้อมูลหรือช่วงของข้อมูล ใช้ได้กับการทดสอบที่ใช้พารามิเตอร์ (parameter) และที่ไม่ใช้พารามิเตอร์ (non-parameter) เช่น ใช้ในการหาช่วงความเชื่อมั่น (confidence interval) ของความแปรปรวนประชากร การทดสอบภาวะสารูปสนิทดี (test of goodness of fit) การทดสอบความเป็นอิสระ (test of independent) เป็นต้น

2.6.2 การทดสอบความเป็นอิสระ (Test of independent)

การทดสอบความเป็นอิสระ เป็นวิธีการทางสถิติที่ใช้ทดสอบว่าตัวแปรสองตัวมีความเป็นอิสระต่อกันหรือไม่ โดยมีสมมติฐาน คือ

Ho : ตัวแปรทั้งสองเป็นอิสระต่อกัน

Ha : ตัวแปรทั้งสองไม่เป็นอิสระต่อกัน

สถิติทดสอบ

การทดสอบแบบไคสแควร์ตัวสถิติที่ใช้ในการทดสอบ คือ

$$W = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

W จะมีการแจกแจงแบบไคสแควร์

เมื่อ O_{ij} = ความถี่ของข้อมูลที่ได้จากการสังเกตหรือทดลองระดับที่ i ของตัวแปรตัวที่ 1 และระดับที่ j ของตัวแปรตัวที่ 2

E_{ij} = ความถี่ที่คาดหวังของระดับที่ i ของตัวแปรตัวที่ 1 และระดับที่ j ของตัวแปรตัวที่ 2 ภายใต้สมมติฐานว่าง (Ho)

โดย $E_{ij} = \frac{O_{i.} \times O_{.j}}{n}$, i = 1, 2, ..., r และ j = 1, 2,..., c

$= \frac{(\text{ผลรวมของแถวนอนที่ i})(\text{ผลรวมของแถวตั้งที่ j})}{n}$

r = จำนวนแถวนอนของตัวแปรตัวที่ 1

c = จำนวนแถวตั้งของตัวแปรตัวที่ 2

N = จำนวนข้อมูลทั้งหมด

เกณฑ์การตัดสินใจ

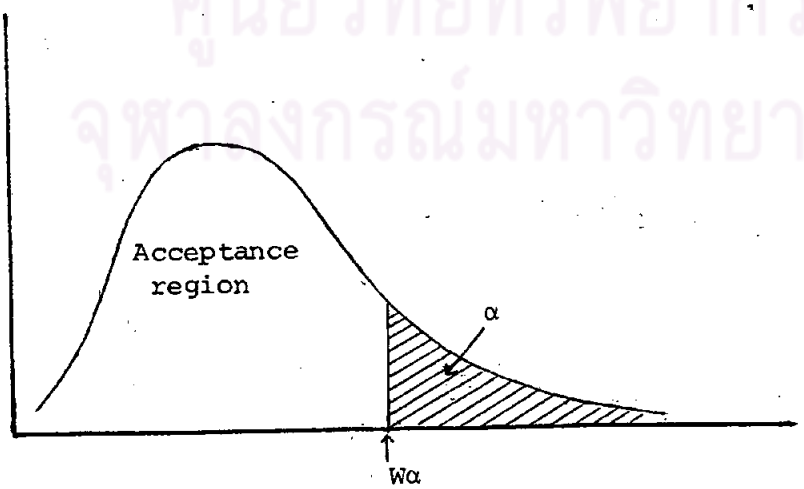
ในการตัดสินใจว่าจะยอมรับ Ho หรือไม่นั้น อาจทำได้ 2 แบบ คือ

1. ถ้าให้ W ที่คำนวณได้เป็น Wo เมื่อทราบค่า Wo และจำนวนขั้นแห่งความเป็นอิสระ (d.f.)แล้ว สามารถอ่านค่าของความน่าจะเป็นได้จากตาราง ถ้าค่าความน่าจะเป็นที่อ่านได้มีค่าน้อยกว่าระดับความมีนัยสำคัญ (α) ที่กำหนด แสดงว่าโอกาสที่สมมติฐานจะเป็นจริงมีน้อยมาก จึงปฏิเสธ Ho ถ้าความน่าจะเป็นที่อ่านได้จากตารางสูงกว่า α แสดงว่าความแตกต่างที่เกิดขึ้นไม่มีนัยสำคัญ จึงยอมรับ Ho ได้

2. อีกแบบคือ เมื่อกำหนดค่า α และทราบค่า d.f.แล้วก็อ่านค่า $W\alpha$ จากตาราง ถ้า $Wo \geq W\alpha$ ความแตกต่างมีนัยสำคัญ นั่นคือ จะปฏิเสธ Ho แต่ถ้า $Wo < W\alpha$ ความแตกต่างไม่มีนัยสำคัญ จะยอมรับ Ho

การทดสอบไคสแควร์เป็นการทดสอบแบบทางเดียวเสมอ ทางด้านมากกว่า ทั้งนี้เพราะว่า O_{ij} จะต่างกับ E_{ij} อย่างมีนัยสำคัญ ถ้า $(O_{ij} - E_{ij})^2$ มีค่ามาก ซึ่งค่าที่ได้จะมีเฉพาะค่าบวกอย่างเดียว

รูปที่ 2.3 แสดงบริเวณปฏิเสธและยอมรับสมมติฐาน



การแจกแจงไคสแควร์ เป็นการแจกแจงของตัวแปรสุ่มชนิดต่อเนื่อง แต่ความถี่ที่เก็บรวบรวมได้เป็นตัวแปรสุ่มชนิดไม่ต่อเนื่อง อย่างไรก็ดี การแจกแจงไคสแควร์ยังคงเป็นการแจกแจงโดยประมาณที่ดีของค่า χ^2 ที่คำนวณได้เมื่อจำนวนชั้นแห่งความเป็นอิสระมากกว่า 1 และค่าคาดหวังมีค่ามากพอ แต่ในกรณีที่จำนวนชั้นแห่งความเป็นอิสระเท่ากับ 1 นั่นคือ ตารางการณ์จรชนิด 2 x 2 จำเป็นต้องมีการแก้ไขโดยใช้ Yates' correction for continuity ตั้งชื่อเพื่อเป็นเกียรติแก่ Dr. Frank Yates ผู้คิด มีสูตรดังนี้

$$\chi^2_{(yates\text{-}correction)} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

โดยทั่ว ๆ ไป เมื่อขนาดตัวอย่างมากกว่าหรือเท่ากับ 50 ไม่จำเป็นต้องปรับค่า χ^2