

CHAPTER 6

การลดมิติและการฝังตริงข้อมูล

6.1 การคัดเลือกคุณลักษณะ

6.1.1 วิธีการตัวกรอง

6.1.1.1 การทดสอบไคกำลังสอง

การทดสอบไคกำลังสอง (Chi-square Test) เป็นวิธีการทางสถิติวิธีหนึ่งที่สามารถนำมาใช้ทดสอบความสัมพันธ์ระหว่างตัวแปรจำแนกประเภท (Categorical Variable) 2 ตัวแปร โดยทำการทดสอบความเป็นอิสระระหว่างตัวแปร 2 ตัวแปร ภายใต้การตั้งสมมติฐาน ดังนี้

H_0 : ตัวแปรสุ่มทั้ง 2 ตัวเป็นอิสระต่อกัน

H_1 : ตัวแปรสุ่มทั้ง 2 ตัวมีความสัมพันธ์กัน

ค่าไคกำลังสองระหว่างตัวแปร 2 ตัวแปร สามารถคำนวณได้โดยเริ่มจากสร้างตารางการจร (Contingency table) ที่แสดงความถี่จำแนกตามตัวแปรทั้งสอง สมมติตัวแปร X_a และ X_b เป็นตัวแปรจำแนกประเภท โดยค่าที่เป็นไปได้ของตัวแปร X_a มีทั้งหมด c ค่า ในเซต $\{a_1, a_2, \dots, a_c\}$ และค่าที่เป็นไปได้ของตัวแปร X_b มีทั้งหมด r ค่า ในเซต $\{b_1, b_2, \dots, b_r\}$ จะสามารถสร้างตารางการจร ได้ดังนี้

		X_a				Total
		a_1	a_2	$\dots$	a_c	
X_b	b_1	$O_{1,1}$	$O_{1,2}$	$\dots$	$O_{1,c}$	$R_1 = \sum_{j=1}^c O_{1,j}$
	b_2	$O_{2,1}$	$O_{2,2}$	$\dots$	$O_{2,c}$	$R_2 = \sum_{j=1}^c O_{2,j}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	b_r	$O_{r,1}$	$O_{r,2}$	$\dots$	$O_{r,c}$	$R_r = \sum_{j=1}^c O_{r,j}$
Total		$C_1 = \sum_{i=1}^r O_{i,1}$	$C_2 = \sum_{i=1}^r O_{i,2}$	$\dots$	$C_c = \sum_{i=1}^r O_{i,c}$	$T = \sum_{i=1}^r \sum_{j=1}^c O_{i,j}$

ค่าไคกำลังสอง ของตัวแปร A และ B คำนวณได้ ดังนี้

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

เมื่อ $O_{i,j}$ คือ ความถี่ที่ได้จากการสังเกตข้อมูลที่ค่าตัวแปร $X_a = a_i$ และ $X_b = b_j$ $E_{i,j}$ คือ ความถี่คาดหวังที่ค่าตัวแปร $A = a_i$ และ $B = b_j$ โดย $E_{i,j} = \frac{R_i C_j}{T}$ และองศาอิสระ (Degree of Freedom: df) มีค่าเท่ากับ $(r - 1)(c - 1)$

จะสรุปว่า ตัวแปร X_a และ X_b มีความเป็นอิสระต่อกัน (ยอมรับสมมติฐาน H_0) ก็ต่อเมื่อค่าไคกำลังสองมีค่าน้อยกว่าหรือเท่ากับค่าวิกฤต $\chi^2_{\alpha,df}$ และจะสรุปว่า ตัวแปร A และ B มีความสัมพันธ์กัน (ปฏิเสธสมมติฐาน H_0) ก็ต่อเมื่อค่าไคกำลังสองมีค่ามากกว่าค่าวิกฤต $\chi^2_{\alpha,df}$ ซึ่งค่าวิกฤต $\chi^2_{\alpha,df}$ สามารถหาได้จากตารางการแจกแจงไคกำลังสองที่ระดับนัยสำคัญ α และองศาอิสระ df อีกทางหนึ่งคือการพิจารณาค่า p -value ในการทดสอบไคกำลังสอง ซึ่งสามารถหาได้จากตารางการแจกแจงไคกำลังสอง โดยถ้าค่า p -value น้อยกว่าระดับนัยสำคัญ α แล้วจะสรุปว่าแปรสุ่มทั้ง 2 ตัวมีความสัมพันธ์กัน (ปฏิเสธสมมติฐาน H_0) หากค่า p -value มากกว่าหรือเท่ากับระดับนัยสำคัญ α แล้วจะสรุปว่าแปรสุ่มทั้ง 2 ตัวมีความเป็นอิสระต่อกัน (ยอมรับสมมติฐาน H_0) โดยทั่วไปมักกำหนดค่า α เท่ากับ 0.05

ตัวอย่าง 6.1.1

จากชุดข้อมูลผู้ป่วยโรคเบาหวาน สามารถวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพศ (Gender) และผลลัพธ์ (Outcome) โดยกำหนดสมมติฐาน

H_0 : ตัวแปรเพศและผลลัพธ์เป็นอิสระต่อกัน

H_1 : ตัวแปรเพศและผลลัพธ์มีความสัมพันธ์กัน

จากข้อมูล สามารถสร้างตารางการจรได้ ดังนี้

Gender	Outcome = 0	Outcome = 1	Total
Female	45	54	99
Male	60	41	101
Total	105	95	200

คำนวณค่าความถี่คาดหวัง $E_{i,j}$ ได้จากสูตร $E_{i,j} = \frac{R_i C_j}{T}$ ได้ดังนี้

$$E_{1,1} = \frac{99 \times 105}{200} = 51.9750$$

$$E_{1,2} = \frac{99 \times 95}{200} = 47.0250$$

$$E_{2,1} = \frac{101 \times 105}{200} = 53.0250$$

$$E_{2,2} = \frac{101 \times 95}{200} = 47.9750$$

จากสูตรการหาไคกำลังสอง

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

จะได้

$$\begin{aligned}\chi^2 &= \frac{(45 - 51.9750)^2}{51.9750} + \frac{(54 - 47.0250)^2}{47.0250} + \frac{(60 - 53.0250)^2}{53.0250} + \frac{(41 - 47.9750)^2}{47.9750} \\ &= \frac{(-6.9750)^2}{51.9750} + \frac{(6.9750)^2}{47.0250} + \frac{(6.9750)^2}{53.0250} + \frac{(-6.9750)^2}{47.9750} \\ &= \frac{48.6506}{51.9750} + \frac{48.6506}{47.0250} + \frac{48.6506}{53.0250} + \frac{48.6506}{47.9750} \\ &= 0.9360 + 1.0346 + 0.9175 + 1.0141 \\ &= 3.9022\end{aligned}$$

โดยค่าองศาอิสระ $df = (2 - 1) \times (2 - 1) = 1$ กำหนดระดับนัยสำคัญ α เท่ากับ 0.05 สามารถหาค่าวิกฤต $\chi^2_{0.05,1}$ จากตารางการแจกแจงไคกำลังสองได้ เท่ากับ 3.841

เนื่องจากค่าไคกำลัง เท่ากับ 3.9022 มีค่ามากกว่าค่าวิกฤต ($3.9022 > 3.841$) จึงปฏิเสธสมมติฐาน H_0 สามารถสรุปได้ว่าตัวแปรเพศและผลลัพธ์มีความสัมพันธ์กัน ดังนั้น ตัวแปรเพศจึงถูกเลือกใช้เป็นตัวทำนาย (Predictor) ในการสร้างแบบจำลองเพื่อทำนายค่าตัวแปรผลลัพธ์

6.1.1.2 การวิเคราะห์ความแปรปรวน

การวิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA) เป็นวิธีการวิเคราะห์ทางสถิติที่ใช้สำหรับตรวจสอบค่าเฉลี่ยของข้อมูลตั้งแต่ 3 กลุ่มขึ้นไปว่ามีความแตกต่างกันอย่างมีนัยยะสำคัญหรือไม่ กำหนดให้ ข้อมูล k กลุ่มมีความเป็นอิสระต่อกัน และข้อมูลแต่ละกลุ่มมีการแจกแจงปกติ ที่มีค่าเฉลี่ย $\mu_1, \mu_2, \dots, \mu_k$ จะสามารถตั้งสมมติฐานทางสถิติได้ดังนี้

H_0 คือ ค่าเฉลี่ยของข้อมูลทุกกลุ่มมีค่าเท่ากัน นั่นคือ $\mu_1 = \mu_2 = \dots = \mu_k$

H_1 คือ ค่าเฉลี่ยของข้อมูลอย่างน้อย 2 กลุ่มข้อมูลมีค่าไม่เท่ากัน

การวิเคราะห์ความแตกต่างของค่าเฉลี่ยของแต่ละกลุ่มข้อมูล จะใช้การเปรียบเทียบความแปรปรวนระหว่างกลุ่มข้อมูลกับความแปรปรวนของข้อมูลภายในกลุ่ม ผ่านการคำนวณค่าสถิติ F (F -Statistic) ซึ่งแสดงอัตราส่วนของค่าเฉลี่ยกำลังสองของความแปรปรวนระหว่างกลุ่ม MS_{Between} ต่อค่าเฉลี่ยกำลังสองของความแปรปรวนภายในกลุ่ม MS_{Within} กำหนดให้ $\{x_i^j\}_{i=1}^{n_j}$ เป็นเซตของข้อมูลในกลุ่ม j โดย $j = 1, 2, \dots, k$ ค่าสถิติ F สามารถคำนวณได้ดังนี้

$$F = \frac{MS_{\text{Between}}}{MS_{\text{Within}}}$$

ค่าเฉลี่ยกำลังสองของความแปรปรวนระหว่างกลุ่มและค่าเฉลี่ยกำลังสองของความแปรปรวนภายในกลุ่มคำนวณได้โดย

$$MS_{\text{Between}} = \frac{SS_{\text{Between}}}{df_{\text{Between}}}$$

$$MS_{\text{Within}} = \frac{SS_{\text{Within}}}{df_{\text{Within}}}$$

ค่าผลรวมกำลังสองของความแปรปรวนระหว่างกลุ่ม $SS_{\text{Between}} = \sum_{j=1}^k n_j (\bar{x}^j - \bar{x})^2$ โดยค่าองศาอิสระระหว่างกลุ่ม $df_{\text{Between}} = k - 1$ ค่าผลรวมกำลังสองของความแปรปรวนภายในกลุ่ม $SS_{\text{Within}} = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_i^j - \bar{x}^j)^2$ โดยค่าองศาอิสระภายในกลุ่ม $df_{\text{Within}} = n - k$ เมื่อ $\bar{x}$ คือ ค่าเฉลี่ยข้อมูลรวม $\bar{x}^j$ คือ ค่าเฉลี่ยข้อมูลในกลุ่ม j และ n คือ จำนวนข้อมูลทั้งหมด

จะสรุปว่า ค่าเฉลี่ยของข้อมูลทุกกลุ่มมีค่าเท่ากัน (ยอมรับสมมติฐาน H_0) ก็ต่อเมื่อค่าสถิติ F มีค่าน้อยกว่าหรือเท่ากับค่าวิกฤต F_{α, df_1, df_2} และจะสรุปว่า ค่าเฉลี่ยของข้อมูลอย่างน้อย 2 กลุ่มมีค่าไม่เท่ากันอย่างมีนัยยะสำคัญ (ปฏิเสธสมมติฐาน H_0) ก็ต่อเมื่อค่าสถิติ F มีค่ามากกว่าค่าวิกฤต F_{α, df_1, df_2} ซึ่งค่าวิกฤต F_{α, df_1, df_2} สามารถหาได้จากตารางการแจกแจงค่าสถิติ F ที่ระดับนัยสำคัญ α องศาอิสระระหว่างกลุ่ม df_1 และองศาอิสระภายในกลุ่ม df_2 เมื่อ $df_1 = k - 1$ และ $df_2 = n - k$ อีกทางหนึ่งคือการพิจารณาค่า p -value ในการทดสอบ ซึ่งสามารถหาได้จากตารางค่าสถิติ F โดยถ้าค่า p -value น้อยกว่าระดับนัยสำคัญ α แล้วจะสรุปว่าค่าเฉลี่ยของข้อมูลอย่างน้อย 2 กลุ่มมีค่าไม่เท่ากันอย่างมีนัยยะสำคัญ (ปฏิเสธสมมติฐาน H_0) หากค่า p -value มากกว่าหรือเท่ากับระดับนัยสำคัญ α แล้วจะสรุปว่าค่าเฉลี่ยของข้อมูลทุกกลุ่มมีค่าเท่ากัน (ยอมรับสมมติฐาน H_0) โดยทั่วไปมักกำหนดค่า α เท่ากับ 0.05

หากผลการวิเคราะห์ความแปรปรวนบ่งชี้ว่าค่าเฉลี่ยข้อมูลอย่างน้อย 2 กลุ่มมีค่าไม่เท่ากันอย่างมีนัยยะสำคัญแล้ว แสดงว่า ข้อมูลนั้นมีความสามารถในการแยกแยะกลุ่มได้ หรือกลุ่มข้อมูลมีความสามารถระบุค่าตัวแปรได้ เนื่องจากมีค่าข้อมูลที่แตกต่างกันอย่างชัดเจนอย่างน้อย 2 กลุ่ม ค่าตัวแปรที่ระบุกลุ่มข้อมูลและค่าตัวแปรเชิงตัวเลขที่ศึกษาสามารถใช้เป็นตัวทำนายในการพยากรณ์ของอีกตัวแปรหนึ่งได้

ตัวอย่าง 6.1.2

การเก็บข้อมูลผลกระทบของยาลดความดันโลหิต 3 ชนิดต่อระดับความดันโลหิตหลังการรักษา 8 สัปดาห์ ในผู้ป่วยทั้งหมด 30 คน มีข้อมูล ดังนี้

กลุ่ม A ได้รับยาลดความดันโลหิตชนิด ACE inhibitor จำนวน 15 คน มีระดับความดันโลหิต ดังนี้

142, 145, 140, 143, 141, 144, 139, 147, 138, 146, 143, 140, 142, 141, 144

กลุ่ม B ได้รับยาลดความดันโลหิตชนิด Beta-blocker จำนวน 15 คน มีระดับความดันโลหิต ดังนี้

138, 135, 136, 137, 134, 139, 133, 140, 132, 137, 136, 135, 138, 134, 136

กลุ่ม C ได้รับยาลดความดันโลหิตชนิด CCB จำนวน 15 คน มีระดับความดันโลหิต ดังนี้

130, 128, 129, 131, 127, 132, 126, 133, 125, 130, 129, 128, 131, 127, 129

สามารถวิเคราะห์ความแปรปรวน เพื่อเปรียบเทียบค่าเฉลี่ยระดับความดันโลหิตในแต่ละกลุ่มผู้ป่วยได้ โดยตั้งสมมติฐาน ดังนี้

H_0 : ค่าเฉลี่ยความดันโลหิตหลังการรักษาของทั้งสามกลุ่ม ไม่แตกต่างกัน

H_1 : มีอย่างน้อยหนึ่งกลุ่มที่ค่าเฉลี่ย แตกต่างกัน

จากข้อมูลข้างต้น มีจำนวนกลุ่ม $k = 3$ กลุ่ม มีขนาดตัวอย่างทั้งหมด $n = 45$ คน แต่ละกลุ่มมีตัวอย่าง $n_j = 15$ คน คำนวณค่าระดับความดันโลหิตเฉลี่ยของแต่ละกลุ่ม ได้ ดังนี้

$$\bar{x}^A = 142.333$$

$$\bar{x}^B = 136.000$$

$$\bar{x}^C = 129.000$$

และค่าเฉลี่ยข้อมูลรวม $\bar{x}$ เท่ากับ 135.778

ต่อมา คำนวณค่าผลรวมกำลังสองของความแปรปรวนระหว่างกลุ่ม

$$\begin{aligned} SS_{\text{Between}} &= \sum_{j=1}^k n_j (\bar{x}^j - \bar{x})^2 \\ &= 15 (142.333 - 135.778)^2 + 15 (136.000 - 135.778)^2 \\ &\quad + 15 (129.000 - 135.778)^2 \\ &= 15 \times 6.555^2 + 15 \times 0.222^2 + 15 \times (-6.778)^2 \\ &= 15 \times 42.968 + 15 \times 0.049 + 15 \times 45.941 \\ &= 644.520 + 0.735 + 689.115 \\ &= 1334.370 \end{aligned}$$

โดยค่าองศาอิสระระหว่างกลุ่ม $df_{\text{Between}} = k - 1 = 3 - 1 = 2$ และคำนวณค่าผลรวมกำลังสองของความแปรปรวนภายในกลุ่ม

$$\begin{aligned} SS_{\text{Within}} &= \sum_{j=1}^k \sum_{i=1}^{n_j} (x_i^j - \bar{x}^j)^2 \\ &= 92.333 + 70.000 + 71.000 \\ &= 233.333 \end{aligned}$$

โดยค่าองศาอิสระภายในกลุ่ม $df_{\text{Within}} = n - k = 45 - 3 = 42$

จากนั้น คำนวณค่าเฉลี่ยกำลังสองของความแปรปรวนระหว่างกลุ่มและค่าเฉลี่ยกำลังสองของความแปรปรวนภายในกลุ่มคำนวณ ได้ดังนี้

$$\begin{aligned} MS_{\text{Between}} &= \frac{SS_{\text{Between}}}{df_{\text{Between}}} = \frac{1334.370}{2} = 667.185 \\ MS_{\text{Within}} &= \frac{SS_{\text{Within}}}{df_{\text{Within}}} = \frac{233.333}{42} = 5.556 \end{aligned}$$

คำนวณค่าสถิติ F จะได้

$$\begin{aligned} F &= \frac{MS_{\text{Between}}}{MS_{\text{Within}}} \\ &= \frac{667.185}{5.556} \\ &= 120.084 \end{aligned}$$

จากตารางค่าสถิติ F ค่าวิกฤต $F_{0.05,2,42}$ เท่ากับ 3.233 พบว่าค่าสถิติ F ที่คำนวณได้มีค่ามากกว่าค่าวิกฤตมาก จึงสรุปได้ว่ามีอย่างน้อยหนึ่งกลุ่มที่ค่าเฉลี่ยระดับความดันโลหิตแตกต่างกัน ดังนั้น ตัวแปรกลุ่มยาลดความดันโลหิตและระดับความดันโลหิตหลักการรักษา 8 สัปดาห์ สามารถใช้เป็นตัวแทนของกันและกันได้

6.1.1.3 สัมประสิทธิ์สหสัมพันธ์เพียร์สัน

สัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson Correlation Coefficient) คือ ค่าทางสถิติที่ใช้วัดความแข็งแกร่งและทิศทางของความสัมพันธ์เชิงเส้นระหว่างตัวแปรเชิงตัวเลข (Numerical Variable) 2 ตัวแปร กำหนดให้ชุดข้อมูล D ประกอบด้วยข้อมูลคู่ตัวแปร X_a และ X_b จำนวน n ข้อมูล โดย $D = \{(x_{ia}, x_{ib})\}_{i=1}^n$ ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน ระหว่างตัวแปร A และ B แทนด้วยสัญลักษณ์ r_{X_a, X_b} สามารถคำนวณได้จาก

$$r_{X_a, X_b} = \frac{s_{X_a, X_b}}{s_{X_a} s_{X_b}} = \frac{\sum_{i=1}^n (x_{ia} - \bar{x}_a)(x_{ib} - \bar{x}_b)}{\sqrt{\sum_{i=1}^n (x_{ia} - \bar{x}_a)^2} \sqrt{\sum_{i=1}^n (x_{ib} - \bar{x}_b)^2}}$$

เมื่อ s_{X_a, X_b} คือ ค่าความแปรปรวนร่วมระหว่างตัวแปร X_a และ X_b s_{X_a} คือ ค่าส่วนเบี่ยงเบนมาตรฐานของตัวแปร X_a s_{X_b} คือ ค่าส่วนเบี่ยงเบนมาตรฐานของตัวแปร X_b $\bar{x}_a$ คือ ค่าเฉลี่ยของตัวแปร X_a และ $\bar{x}_b$ คือ ค่าเฉลี่ยของตัวแปร X_b

ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน มีค่าที่เป็นไปได้อยู่ระหว่าง -1 ถึง 1 ค่าสัมประสิทธิ์สหสัมพันธ์ ที่มีค่าเท่ากับ 1 บ่งบอกถึงความสัมพันธ์เชิงเส้นตรงอย่างสมบูรณ์แบบระหว่างค่าตัวแปรทั้งสอง เมื่อค่าของตัวแปรหนึ่งเพิ่มขึ้นค่าของอีกตัวแปรหนึ่งจะเพิ่มขึ้นตามไปด้วย ส่วนค่าสัมประสิทธิ์สหสัมพันธ์ที่มีค่าเท่ากับ -1 แสดงถึงความสัมพันธ์เชิงเส้นตรงในลักษณะตรงข้ามอย่างสมบูรณ์ของตัวแปรทั้งสอง โดยเมื่อค่าของตัวแปรหนึ่งเพิ่มขึ้น ค่าของอีกตัวแปรหนึ่งจะลดลง ในกรณีที่ค่าสัมประสิทธิ์สหสัมพันธ์มีค่าเท่ากับ 0 แสดงถึงค่าของตัวแปรทั้งสองตัวแปรไม่มีความสัมพันธ์เชิงเส้นตรงต่อกัน ค่าสัมบูรณ์ (Absolute Value) ของ ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันยังชี้วัดระดับความสัมพันธ์ของค่าตัวแปรทั้งสองด้วย กล่าวคือ ค่าสัมบูรณ์ของค่าสัมประสิทธิ์สหสัมพันธ์ที่มีค่ามาก บ่งบอกถึงค่าตัวแปรทั้งสองมีความสัมพันธ์กันอย่างมาก ในขณะที่ค่าสัมบูรณ์ของค่าสัมประสิทธิ์สหสัมพันธ์ที่มีค่าน้อย แสดงถึงความสัมพันธ์ในระดับต่ำระหว่างตัวแปรทั้งสองด้วย แสดงแผนภาพการกระจายของค่าข้อมูลระหว่าง 2 ตัวแปร และสัมประสิทธิ์สหสัมพันธ์เพียร์สันที่สัมพันธ์กัน ดังภาพ 6.1

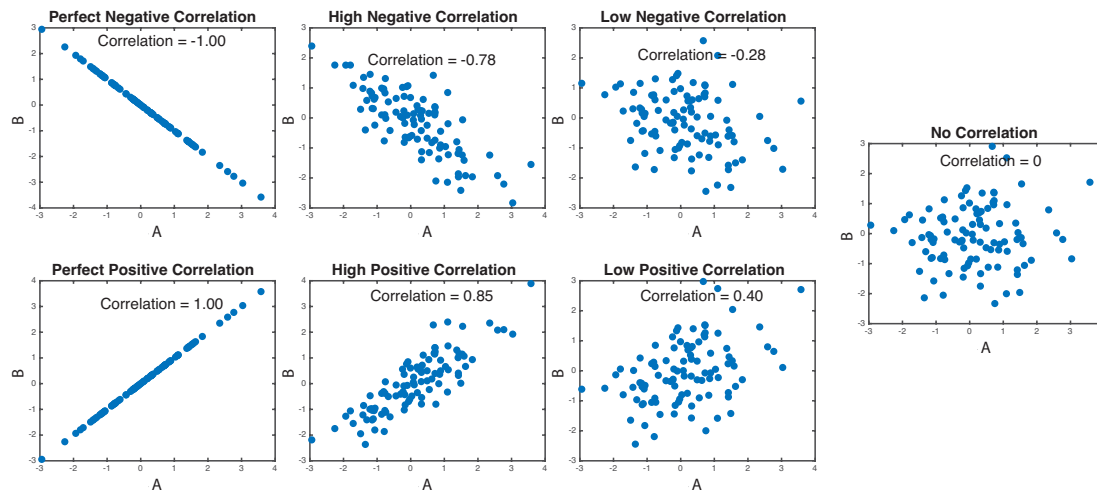


Figure 6.1. แผนภาพ การกระจาย ของ ค่า ข้อมูล ที่ สัมพันธ์ กับ ค่า สัมประสิทธิ์ สหสัมพันธ์เพียร์สัน 7 ลักษณะ

ตัวอย่าง 6.1.3

ชุดข้อมูลจากการสำรวจระดับเชาว์ปัญญา (IQ) และประสิทธิภาพของงาน (Job Performance: JP) จากตัวอย่างพนักงานบริษัท จำนวน 10 คน ดังนี้

IQ	Job Performance
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

ค่าส่วนเบี่ยงเบนมาตรฐานของค่าตัวแปรระดับเชาว์ปัญญา (s_{IQ}) เท่ากับ 9.70 ค่าส่วนเบี่ยงเบนมาตรฐานของค่าตัวแปรประสิทธิภาพของงาน (s_{JP}) เท่ากับ 2.71 และค่าความแปรปรวนร่วมระหว่างตัวแปรระดับเชาว์ปัญญาและประสิทธิภาพของงาน ($s_{IQ,JP}$) เท่ากับ 15.64 จะสามารถหาค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปรระดับเชาว์ปัญญาและประสิทธิภาพของงาน ได้ดังนี้

จากสูตรการหาค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันของตัวอย่าง

$$r_{IQ,JP} = \frac{s_{IQ,JP}}{s_{IQ}s_{JP}}$$

แทนค่าในสูตร จะได้

$$\begin{aligned} r_{IQ,JP} &= \frac{15.64}{9.70 \times 2.71} \\ &= \frac{15.64}{26.287} \\ &= 0.59 \end{aligned}$$

ดังนั้น ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปรระดับเซวาร์ปัญญาและประสิทธิภาพของงาน เท่ากับ 0.59 บ่งชี้ว่าค่าตัวแปรระดับเซวาร์ปัญญา และประสิทธิภาพของงานมีความสัมพันธ์เชิงเส้นตรงต่อกัน โดยเมื่อค่าตัวแปรระดับเซวาร์ปัญญามีค่าเพิ่มขึ้น ค่าตัวแปรประสิทธิภาพของงานจะมีค่าเพิ่มขึ้นตามด้วย ในขณะที่เดียวกันเมื่อค่าตัวแปรระดับเซวาร์ปัญญามีค่าลดลง ค่าตัวแปรประสิทธิภาพของงานจะมีค่าลดลงด้วย

6.1.1.4 สารสนเทศข้อมูลร่วม

สารสนเทศข้อมูลร่วม (Mutual Information) เป็นค่าวัดปริมาณข้อมูลที่ตัวแปรสุ่มตัวหนึ่งมีเกี่ยวข้องกับตัวแปรสุ่มอีกตัวหนึ่ง โดยนำความรู้เกี่ยวกับทฤษฎีสารสนเทศ (Information Theory) มาประยุกต์ใช้ นั่นคือ เอนโทรปี (Entropy) ซึ่งนำเสนอโดย Claude Shannon ในปี ค.ศ. 1948 สำหรับใช้วัดระดับความไม่แน่นอนของตัวแปรสุ่มตัวหนึ่ง ถ้าความน่าจะเป็นของผลลัพธ์มีการกระจายตัวอย่างสม่ำเสมอ ย่อมหมายถึงการคาดเดาผลลัพธ์ทำได้ยาก ค่าเอนโทรปีจะมีค่าสูง ในทางกลับกันถ้าผลลัพธ์ใดผลลัพธ์หนึ่งมีความน่าจะเป็นสูง ย่อมสามารถคาดเดาผลลัพธ์ได้ง่ายขึ้น ค่าเอนโทรปีจะมีค่าต่ำลง ค่าเอนโทรปีของตัวแปรสุ่ม X แทนด้วยสัญลักษณ์ $H(X)$ คำนวณโดย

$$H(X) = - \sum_{x \in X} p(x) \log p(x)$$

เมื่อ $p(x)$ คือ ความน่าจะเป็นที่ตัวแปรสุ่ม X มีค่าเท่ากับ x

กำหนดให้ X และ Y เป็นตัวแปรสุ่ม ค่าเอนโทรปีร่วม (Joint Entropy) ของตัวแปรสุ่ม X_a และ X_b แทนด้วยสัญลักษณ์ $H(X, Y)$ นิยามโดย

$$H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$

เมื่อ $p(x, y)$ คือ ความน่าจะเป็นร่วม (Joint Propability) ที่ตัวแปรสุ่ม X และ Y มีค่าเท่ากับ x และ y ตามลำดับ

สารสนเทศข้อมูลร่วม วัดว่าตัวแปรสุ่มสองตัวแบ่งปันสารสนเทศร่วมกันมากน้อยเพียงใด โดยหากทราบค่าของตัวแปรหนึ่งแล้วจะช่วยลดความไม่แน่นอนของอีกตัวแปรได้มากน้อยเพียง สารสนเทศข้อมูลร่วมของ

ตัวแปรสุ่ม X และ Y แทนด้วยสัญลักษณ์ $I(X; Y)$ สามารถคำนวณจากค่าเอนโทรปีได้โดย

$$I(X; Y) = H(X) + H(Y) - H(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

ค่า $I(X; Y)$ จะเท่ากับ 0 เมื่อการทราบค่าตัวแปรสุ่มตัวใดตัวหนึ่งไม่สามารถช่วยลดความไม่แน่นอนของอีกตัวแปรหนึ่งได้ นั่นหมายถึงตัวแปรสุ่มทั้งสองมีความเป็นอิสระต่อกัน เมื่อ $I(X; Y)$ มีค่าสูงขึ้นหมายถึงตัวแปรสุ่มทั้งสองมีระดับความสัมพันธ์กันสูงขึ้นด้วย คุณสมบัตินี้ทำให้ MI เป็นตัวชี้วัดความสัมพันธ์ที่มีความยืดหยุ่นสูง สามารถจับความสัมพันธ์ได้ทั้งเชิงเส้นและไม่เชิงเส้น แตกต่างจากสัมประสิทธิ์สหสัมพันธ์เพียร์สันที่ตรวจจับได้เฉพาะความสัมพันธ์เชิงเส้น

ตัวอย่าง 6.1.4

จากการรวบรวมข้อมูลพฤติกรรมกรรมการสูบบุหรี่ (Smoking) และการเกิดโรคปอด (Lung Disease) สามารถสร้างตารางการจร ได้ดังนี้

		Lung Disease		Total
		Yes	No	
Smoking	Yes	30	10	40
	No	10	50	60
Total		40	60	100

จากข้อมูล ค่าความน่าจะเป็นของการเกิดโรคปอด คำนวณโดย

$$p(\text{Lung Disease} = \text{Yes}) = \frac{40}{100} = 0.4$$

$$p(\text{Lung Disease} = \text{No}) = \frac{60}{100} = 0.6$$

และค่าความน่าจะเป็นของพฤติกรรมกรรมการสูบบุหรี่ คำนวณโดย

$$p(\text{Smoking} = \text{Yes}) = \frac{40}{100} = 0.4$$

$$p(\text{Smoking} = \text{No}) = \frac{60}{100} = 0.6$$

ค่าความน่าจะเป็นร่วมระหว่างพฤติกรรมการสูบบุหรี่และการเกิดโรคปอด คำนวณได้ดังนี้

$$\begin{aligned} p(\text{Smoking} = \text{Yes}, \text{Lung Disease} = \text{Yes}) &= \frac{30}{100} = 0.3 \\ p(\text{Smoking} = \text{Yes}, \text{Lung Disease} = \text{No}) &= \frac{10}{100} = 0.1 \\ p(\text{Smoking} = \text{No}, \text{Lung Disease} = \text{Yes}) &= \frac{10}{100} = 0.1 \\ p(\text{Smoking} = \text{No}, \text{Lung Disease} = \text{No}) &= \frac{50}{100} = 0.5 \end{aligned}$$

จะหาค่าความสารสนเทศข้อมูลร่วมระหว่างพฤติกรรมการสูบบุหรี่และการเกิดโรคปอด ได้จาก

$$\begin{aligned} H(\text{Smoking}, \text{Lung Disease}) &= - \sum_{x \in \{\text{Yes}, \text{No}\}} \sum_{y \in \{\text{Yes}, \text{No}\}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= 0.30 \cdot \log \left(\frac{0.30}{0.40 \cdot 0.40} \right) + 0.10 \cdot \log \left(\frac{0.10}{0.40 \cdot 0.60} \right) \\ &\quad + 0.10 \cdot \log \left(\frac{0.10}{0.60 \cdot 0.40} \right) + 0.50 \cdot \log \left(\frac{0.50}{0.60 \cdot 0.60} \right) \\ &= 0.30 \cdot \log(1.875) + 0.10 \cdot \log(0.417) \\ &\quad + 0.10 \cdot \log(0.417) + 0.50 \cdot \log(1.389) \\ &= 0.270 + (-0.126) + (-0.126) + 0.235 \\ &= 0.253 \end{aligned}$$

ดังนั้น ค่าสารสนเทศข้อมูลร่วมระหว่างพฤติกรรมการสูบบุหรี่และการเกิดโรคปอด มีค่าเท่ากับ 0.253 นั้นหมายถึง ค่าตัวแปรพฤติกรรมการสูบบุหรี่สามารถช่วยลดความไม่แน่นอนของค่าตัวแปรการเกิดโรคปอดได้ ซึ่งสะท้อนให้เห็นถึงความสัมพันธ์เชิงสารสนเทศระหว่างตัวแปรทั้งสอง

6.1.1.5 ค่าความไม่บริสุทธิ์แบบจีนิ

ค่าความไม่บริสุทธิ์แบบจีนิ (Gini Impurity) เป็นค่าวัดความไม่บริสุทธิ์ของข้อมูล โดยการวัดความผสมกันของข้อมูลที่มีความหลากหลายมากน้อยเพียงใด กำหนดให้ $C = \{c_1, c_2, \dots, c_k\}$ เป็นเซตของคลาสข้อมูลทั้งหมดที่เป็นไปได้ในชุดข้อมูล และ p_i คือ ความน่าจะเป็นของการสุ่มจุดข้อมูลหนึ่งจากชุดข้อมูลแล้วมีค่าคลาสเท่ากับ c_i ค่าความไม่บริสุทธิ์แบบจีนิ คำนวณได้โดย

$$G = 1 - \sum_{i=1}^C p_i^2$$

ค่าความไม่บริสุทธิ์แบบจินี จะเท่ากับ 0 ก็ต่อเมื่อในชุดข้อมูลนั้นมีข้อมูลเพียง 1 คลาส นั่นหมายถึงชุดข้อมูลนั้นมีความบริสุทธิ์สูง และจะมีค่าสูงสุด เมื่อชุดข้อมูลนั้นมีจำนวนข้อมูลในแต่ละคลาสเท่า ๆ กัน

สำหรับการประยุกต์ใช้ค่าความไม่บริสุทธิ์แบบจินี ในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรจำแนกประเภท 2 ตัวแปร สามารถทำได้โดยการแบ่งข้อมูลเป็นชุดย่อยตามค่าตัวแปรหนึ่ง แล้ววัดค่าความไม่บริสุทธิ์แบบจินีของตัวอีกแปรหนึ่ง กำหนดให้ X และ Y เป็นตัวแปรจำแนกประเภท ค่าความไม่บริสุทธิ์แบบจินีของค่าตัวแปร Y เมื่อแบ่งข้อมูลตามตัวแปร X แทนด้วยสัญลักษณ์ $G(Y|X)$ คำนวณได้โดย

$$G(Y|X) = \sum_x p(x) \left(1 - \sum_y p(y|x)^2 \right)$$

ตัวแปรทั้งสองจะมีความสัมพันธ์กันมากที่สุด เมื่อค่าความไม่บริสุทธิ์แบบจินีเท่ากับ 0 และระดับความสัมพันธ์นั้นจะน้อยลง เมื่อค่าความไม่บริสุทธิ์แบบจินีเพิ่มสูงขึ้น

ตัวอย่าง 6.1.5

จากข้อมูลข้อมูลพฤติกรรมกรรมการสูบบุหรี่ (Smoking) และการเกิดโรคปอด (Lung Disease) ในตัวอย่าง 6.1.4 สามารถหาค่าความไม่บริสุทธิ์แบบจินีของค่าตัวแปรการเกิดโรคปอด เมื่อแบ่งข้อมูลเป็นชุดย่อยตามค่าตัวแปรพฤติกรรมกรรมการสูบบุหรี่ ได้ดังนี้

$$\begin{aligned} G(\text{Lung Disease}|\text{Smoking}) &= \sum_{x \in \{\text{Yes}, \text{No}\}} p(x) \left(1 - \sum_{y \in \{\text{Yes}, \text{No}\}} p(y|x)^2 \right) \\ &= \left[\frac{40}{100} \left(1 - \left(\left(\frac{30}{40} \right)^2 + \left(\frac{10}{40} \right)^2 \right) \right) \right] \\ &\quad + \left[\frac{60}{100} \left(1 - \left(\left(\frac{10}{60} \right)^2 + \left(\frac{50}{60} \right)^2 \right) \right) \right] \\ &= (0.4 \cdot (1 - (0.75^2 + 0.25^2))) \\ &\quad + (0.6 \cdot (1 - (0.167^2 + 0.833^2))) \\ &= (0.4 \cdot (1 - 0.625)) + [0.6 \cdot (1 - 0.722)] \\ &= (0.4 \cdot 0.375) + (0.6 \cdot 0.278) \\ &= 0.150 + 0.167 \\ &= 0.316 \end{aligned}$$

ดังนั้น ค่าความไม่บริสุทธิ์แบบจินีของค่าตัวแปรการเกิดโรคปอด เมื่อแบ่งข้อมูลเป็นชุดย่อยตามค่าตัวแปรพฤติกรรมกรรมการสูบบุหรี่ มีค่าเท่ากับ 0.316