บทที่ ๔

สหสัมพันธ์

(Correlation)

หอสมุดกลาง จุฬาลงกรณมหาวิทยาลัย

๔.๑ ความหมายของสหสัมพันธ์

สหสัมพันธ์ เป็นเรื่องราวของความสัมพันธ์ระหว่างข้อมูลตั้งแต่ ๒ ชุดขึ้นไป เช่น ต้องการทราบว่าคนที่เรียนวิชาคณิตศาสตร์เก่ง จะเรียนวิชาสถิติและวิชาวิทยาศาสตร์เก่งหรือไม่ เป็นต้น จะต้องนำข้อมูลที่รวบรวมได้แต่ละเรื่องมาศึกษาดูความสัมพันธ์ และจะต้องสร้างความสัมพันธ์ขึ้นในรูปสมการก่อน ซึ่งทำให้สามารถพยากรณ์ได้ว่า คนที่เรียนวิชาหนึ่ง ๆ เก่ง จะเรียนวิชาอื่น ๆ เก่งหรือไม่ เช่น

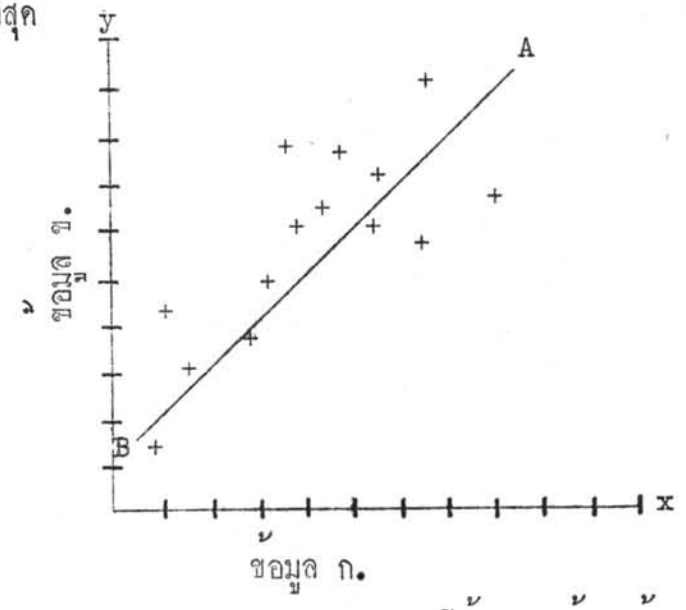
$y = x - z + 80$ เมื่อ y เป็นคะแนนวิชาคณิตศาสตร์, x เป็นคะแนนวิชาสถิติ และ z เป็นคะแนนวิชาวิทยาศาสตร์ ถ้าทราบค่า x และ z ก็คำนวณหาค่า y ออกมาได้ แสดงว่าสามารถพยากรณ์คะแนนคณิตศาสตร์ได้ แต่ความแน่นอนของการพยากรณ์จะมีเพียงใดนั้น สุดแต่ความละเอียดของการสร้างสมการ

ในการหาสมการแสดงความสัมพันธ์ระหว่างข้อมูล ๒ ชุด อาจหาได้หลายวิธี สมการจะแสดงออกในรูปสมการกำลังหนึ่ง ซึ่งแสดงถึงความสัมพันธ์เป็นเส้นตรง หรือสมการกำลังสอง ซึ่งแสดงถึงความสัมพันธ์เป็นพาราโบลา หรือสมการกำลังสามกำลังสี่ ต่อ ๆ ไปก็ได้ จากการพิจารณาข้อมูลในขั้นต้นประกอบกับประสพการณ์ของผู้ทำงานในด้านนำระเหยมาก่อน เห็นว่าสมการกำลังหนึ่งสามารถนำมาใช้ในงานดังกล่าวนี้ได้โดยมีความถ่วนถี่ในการใช้งานได้

๔.๒ การปรับเส้นตรงให้เหมาะกับแผนภาพกระจาย

แผนภาพกระจาย (Scatter diagram) คือ จุดกระจัดกระจายทั่ว ๆ ไป

บนกราฟซึ่งเมื่อมองดูอาจจะเห็นทางโน้มในการลากเส้นตรงเส้นหนึ่งแสดงทางโน้มประมาณได้ เส้นตรงนี้อาจลากขึ้นได้หลายทาง แต่เส้นไหนจะเป็นเส้นที่ดีที่สุดนั้นขึ้นอยู่กับการปรับให้ได้ความคลาดเคลื่อนหรือผลต่างของค่าที่อ่านได้จากจุดบนเส้นตรงกับจุดจริงน้อยที่สุด



รูปที่ ๔.๑ แผนภาพกระจายและทางโน้มแสดงด้วยเส้นตรง

วิธีปรับเส้นตรงที่ใช้กันมาก คือ

๔.๒.๑ การกะด้วยตา

ซึ่งอาจให้ผลคลาดเคลื่อนน้อยถ้าจุดที่เขียนลงไปไม่กระจัดกระจายกันมาก แต่ถ้าจุดเหล่านี้กระจัดกระจายมาก การกะด้วยตาก็จะทำให้ผลคลาดเคลื่อนมาก เพราะไม่ทราบว่าเส้นตรงเส้นใดเป็นเส้นตรงที่ดีที่สุด

๔.๒.๒ โดยวิธีเฉลี่ยครึ่งต่อครึ่ง

วิธีนี้แบ่งข้อมูลออกเป็น ๒ กลุ่มเท่า ๆ กัน และหาค่าเฉลี่ยในแต่ละกลุ่ม ดังนั้นจะได้จุด ๒ จุดบนกราฟ แล้วลากเส้นตรงผ่านจุดทั้งสอง

๔.๒.๓ โดยวิธีเฉลี่ยเคลื่อนที่

วิธีนี้จะต้องกำหนดอันดับ (order) และจะช่วยลดจำนวนของตัวแปรผันที่ปรากฏในชุดของข้อมูล

๔.๒.๔ โดยวิธีกำลังน้อยที่สุด

ซึ่งเป็นวิธีที่ดีที่สุดและใช้กันมากที่สุด โดยวิธีนี้จะได้สมการทางโน้มในรูป

$$y = \frac{\sum xy}{\sum x^2} x \qquad (๔.๑)$$

เส้นตรงที่ปรับแล้วนี้โดยวิธีกำลังน้อยที่สุด เรียกว่าเส้นถดถอย y เทียบกับ x (Regression Line of y on x) หมายถึงเส้นถดถอยที่มี x เป็นหลัก ความคลาดเคลื่อนที่เกิดขึ้นเป็นความคลาดเคลื่อนของ y

ด้วยวิธีเดียวกันนี้ อาจหาสมการเส้นถดถอยของ X เทียบกับ Y (Regression Line of X on Y) ได้ ค่าของ b หรือความชัน (Slope) ของเส้นถดถอยทั้งสองจะต่างกัน ดังจะได้ใช้ค่า b_1 และ b_2 แทน ดังนี้

เส้นถดถอย X เทียบกับ Y

$$X-\overline{X} = b_1(Y-\overline{Y}) \qquad (๔.๒)$$

$$b_1 = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (Y-\overline{Y})^2} \qquad (๔.๓)$$

เส้นถดถอยของ Y เทียบกับ X

$$Y-\overline{Y} = b_2(X-\overline{X}) \qquad (๔.๔)$$

$$b_2 = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (X-\overline{X})^2} \qquad (๔.๕)$$

b_1 และ b_2 นี้มีชื่อเฉพาะว่า สัมประสิทธิ์แห่งการถดถอย (Coefficient of Regression)

จะเห็นได้ว่าสมการ (๔.๒) และ (๔.๔) นี้ให้เส้นถดถอย ๒ เส้น ซึ่งไม่ใช่เส้นตรงเดียวกัน จะเป็นเส้นตรงเดียวกันเมื่อจุดทุกจุดบนกราฟบังเอิญอยู่บนเส้นตรงเดียวกันเท่านั้น การพยากรณ์ค่าของ X และ Y จึงขึ้นอยู่กับค่าที่กำหนดให้ ถ้ากำหนด X ให้พยากรณ์ Y ก็ต้องใช้สมการเส้นถดถอยของ Y เทียบกับ X แต่ถ้ากำหนด Y ให้พยากรณ์ X ก็ต้องใช้สมการเส้นถดถอย X เทียบกับ Y

๔.๓ สัมประสิทธิ์แห่งสหสัมพันธ์ (Correlation Coefficient)

แม้จะได้เส้นถดถอยทั้งคู่แสดงความสัมพันธ์ของข้อมูล ๒ ชุดแล้วก็ตาม ก็ยังหาได้ทราบไม่ว่าข้อมูล ๒ ชุดนั้นมีความสัมพันธ์กันมากน้อยเพียงใด ดังนั้นจึงจำเป็นต้องสร้างดัชนีขึ้นตัวหนึ่ง เพื่อใช้วัดความมากน้อยของสหสัมพันธ์ ดัชนีที่สร้างขึ้นนี้เรียกว่า สัมประสิทธิ์แห่งสหสัมพันธ์ โดยมีข้อมูล x และ y มีส่วนร่วมในดัชนีเท่า ๆ กัน จะได้สัมประสิทธิ์แห่งสหสัมพันธ์ (r) ดังนี้

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 (y-\bar{y})^2}} \qquad (๔.๖)$$

๔.๔ การตีความหมายของสัมประสิทธิ์แห่งสหสัมพันธ์

ถ้าให้ E_1 = ค่าเฉลี่ยของความคลาดเคลื่อนที่วัดได้โดยวิธีกำลังสองน้อยที่สุด

E_0 = ค่าเฉลี่ยของกำลังสองของส่วนเบี่ยงเบนจากมัชฌิมเลขคณิต

จะได้ $r^2 = \frac{E_0 - E_1}{E_0}$ (๔.๗)

E_1 เป็นความคลาดเคลื่อนชนิดหนึ่งดังกล่าวแล้ว ส่วน E_0 ก็อาจนับเป็นความคลาดเคลื่อนได้เหมือนกัน ดังนั้น ค่าของ r^2 คือ อัตราส่วนของพวกความคลาดเคลื่อนนั่นเอง ซึ่งนับว่าเป็นความคลาดเคลื่อนสัมพันธ์ชนิดหนึ่ง r^2 จึงเหมาะที่จะเป็นดัชนีสำหรับการพิจารณาและเปรียบเทียบมาก แต่ใช้เพียงรากที่สองของค่า r^2 นี้เท่านั้น คือ r ซึ่งเรียกว่าสัมประสิทธิ์แห่งสหสัมพันธ์

จากสมการ (๔.๗) $r^2 = \frac{E_0 - E_1}{E_0}$ จะเห็นได้ว่าค่า $\left(\frac{E_0 - E_1}{E_0}\right)$ หรือ $\left(๑ - \frac{E_1}{E_0}\right)$ น้อยกว่า ๑ เสมอไป เนื่องจาก E_1 เป็นบวกมีค่าได้ตั้งแต่ศูนย์ขึ้นไป ดังนั้น r จะมีค่ามากที่สุดเมื่อ E_1 มีค่าน้อยที่สุด นั่นคือ เมื่อ $E_1 = 0$ แสดงว่าเส้นถดถอยผ่านทุก ๆ จุดบนกราฟ ไม่มีการคลาดเคลื่อนเลย ฉะนั้นค่าสูงสุดของ r = ๑ หรือ ๑๐๐ เปอร์เซ็นต์

และเนื่องจาก r มีค่าได้ทั้งบวกและลบ ดังนั้น r มีค่าน้อยที่สุดคือ -๑ หรือ -๑๐๐ เปอร์เซ็นต์ อาจสรุปได้ว่า ค่าของตัวเลขของ r อยู่ระหว่าง ๐ ถึง ๑ เครื่องหมายของ r จะเป็นเครื่องบอกแสดงทิศทางว่าข้อมูลแปรตามกันหรือกลับกัน

ดังนั้น จึงมีความหมายเป็นภาษาธรรมดาจากค่าของ r ว่า ถ้ามีข้อมูล ๑๒ ข้อมูลจากข้อมูลสองชุด ค่าตัวเลขของ r ใหญ่ใกล้ ๑ มาก หรือใกล้ ๑๐๐ เปอร์เซ็นต์ แสดงว่า ข้อมูลทั้งสองมีสหสัมพันธ์มากหรือกล่าวอีกนัยหนึ่งคือ ข้อมูลทั้งสองมีความสัมพันธ์กันมาก และแสดงว่า ถ้า x มีค่ามาก y จะมีค่ามากด้วย หรือ ถ้า x มีค่ามาก y อาจจะมีค่าน้อยลงทุกที สุดแล้วแต่เครื่องหมายของ r แต่ถ้าตัวเลขของ r ไม่มากนัก ประมาณ ๐.๖ หรือ ๖๐ เปอร์เซ็นต์ โดยมีข้อมูล ๑๒ ข้อมูล (n=12) และวัดโดยวิธีของสถิติ ที (statistic t) ซึ่งเท่ากับรากที่สองของเอฟ ($t=\sqrt{F}$) จะได้ข้อมูลทั้งสองมีสหสัมพันธ์กันพอสมควร และถ้าค่าตัวเลขของ r ต่ำกว่า ๐.๖ จากข้อมูล ๑๒ ข้อมูลเช่นเดียวกันจะได้ ข้อมูลทั้งสองมีสหสัมพันธ์น้อยจนน่าสงสัย และถ้าค่าตัวเลขของ r ต่ำมาก ประมาณต่ำกว่า ๐.๒ หรือ ๒๐ เปอร์เซ็นต์จากข้อมูลดังกล่าวข้างต้นก็นับว่าไม่มีสหสัมพันธ์กันเลย