# บทที่ 10
# การวิเคราะห์ข้อมูล

<u>วัตถุประสงค์</u> เพื่ออธิบายถึงปัจจัยสำคัญในการวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูล คือ การวิเคราะห์ค่าของตัวแปรได้มาจากการรวบรวมข้อมูลเพื่อที่จะทราบลักษณะของตัวแปรพื้น ๆ แล้วนำผลจากการวิเคราะห์นั้นมาตีความ เพื่อหาคำตอบให้กับปัญหาการวิจัยนั่นเอง จุดสำคัญของการวิเคราะห์ข้อมูล คือ การหาความสัมพันธ์ระหว่างข้อมูลเรื่องต่าง ๆ ที่เก็บมา ข้อมูลจะเป็นหลักฐานยืนยันหรือพิสูจน์ว่า x กับ y สัมพันธ์กันหรือไม่เพียงใด

ขั้นตอนของการวิเคราะห์เริ่มแต่การนำข้อมูลหรือคำตอบที่ได้มาจัดระเบียบเป็นหมวดหมู่ ซึ่งเรามักจะเรียกว่าการจัดรูปข้อมูล หรือการกระทำกับข้อมูล (Data processing) ข้อมูลที่ได้มานี้ในตอนแรกเรียกว่าข้อมูลดิบนั้นจะเป็นข้อมูลที่มีจำนวนมาก และยุ่งยากแก่ความเข้าใจ จำต้องทำให้อยู่ในแบบฟอร์มง่าย ๆ แบบใดแบบหนึ่งเพื่อให้ง่ายแก่การอธิบายหรือทำความเข้าใจแล้วจึงนำข้อมูลที่จัดรูปแล้วมาจัดเรียงลำดับตามความต้องการเพื่อให้ง่ายต่อการวิเคราะห์ การเปลี่ยนรูปข้อมูลนี้ เราใช้ความรู้ในการจัดกลุ่มทางสถิติ

ในการจัดรูปข้อมูลนั้น เราจะต้องวางแบบไว้ก่อนเก็บข้อมูล ซึ่งจะอยู่ในแบบของการวิจัยว่าจะใช้แบบใด เมื่อเก็บข้อมูลแล้วจึงนำมาจัดส่งตามแผนที่วางไว้ ซึ่งมักจะทำเป็นตารางเปล่า ( Dummy table) ไว้ก่อน

การจัดรูปข้อมูลนั้นขึ้นอยู่กับธรรมชาติของข้อมูลด้วยว่าเป็นข้อมูลชนิดใด ซึ่งโดยทั่วไปเราแบ่งข้อมูลเป็น 2 ชนิด

1. Categorical data หรือ discrete data หมายถึง ข้อมูลที่

มีลักษณะขาดตอน เช่น มา-ไม่มา ชาย-หญิง ชอบ-ไม่ชอบ เป็นต้น ซึ่งถ้าเทียบกับค่าของตัวแปรแล้วจะอยู่ในนามมาตรา และองศามาตรา (nominal and ordinal scales) การจัดรูปข้อมูลมักจัดในรูปอัตราส่วน (proportion) เช่น อัตราร้อยละ

2. Continuous data หมายถึงข้อมูลที่เราสามารถจัดได้ โดยไม่มีการขาดตอน (มีทศนิยมได้) เช่น ความสูง น้ำหนัก อายุ จำนวนเงิน ฯลฯ เป็นข้อมูลที่อาจวัดอยู่ในระดับอันตรมาตรา และสูญมาตรา ( Interval and Ratio scales) การจัดรูปข้อมูลมักจัดในรูปของการวัดแนวโน้มเข้าสู่ศูนย์กลาง เช่น ตัว-กลางเลขคณิต มัธยฐาน ฐานนิยม และพิสัย ฯลฯ

การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรสองตัวหรือมากกว่าสองตัวอาจทำได้โดยการเปรียบเทียบจำนวนร้อยละ เปรียบเทียบค่าเฉลี่ย จนถึงการคำนวณค่าสหสัมพันธ์แบบต่าง ๆ

การเลือกใช้สถิติแบบใดในการวิเคราะห์ต้องพิจารณาว่าข้อมูลนั้นจัดอยู่ในระดับการวัด (level of Measurement) ระดับใด ระดับของการวัดตัวแปรแบ่งออกเป็น 4 ระดับ คือ nominal ordinal, interval และ ratio แต่ละระดับเหมาะกับสถิติต่างกัน ในการใช้สถิตินักวิจัยไม่จำเป็นต้องรู้สถิติอย่างนักสถิติ นักวิจัยเพียงแต่รู้ว่า การวิเคราะห์ข้อมูลของตนควรใช้สถิติอะไรเท่านั้นก็นับว่าเพียงพอแล้ว การคำนวณทางสถิติที่ยุ่งยากซับซ้อนอาจให้นักสถิติหรือใช้เครื่องจักรช่วยได้

จุดมุ่งหมายสำคัญในการวิเคราะห์ข้อมูลก็คือ การหาความสัมพันธ์ที่เป็นเหตุเป็นผลต่อกัน ระหว่างตัวแปร ( Causal relationship) สำหรับการวิจัยที่ไม่ใช่เป็นการทดลอง วิธีพิจารณาว่าตัวแปรสองตัวหรือมากกว่าสองตัวมีความสัมพันธ์กันหรือไม่เพียงใด เป็นเรื่องที่จะต้องใช้เหตุผลประกอบกับหลักสถิติมาช่วย ในขั้นต้นพิจารณาจากตารางแจกแจงความถี่ ซึ่งประกอบด้วยสองตัวแปรคือ ตัวแปรอิสระ และตัวแปรตาม

( X และ Y ) ตารางแสดงความสัมพันธ์ระหว่างสองตัวแปรสมมติมีลักษณะดังนี้ (ตามตัวอย่างความสัมพันธ์ระหว่างฐานะทางเศรษฐกิจกับการมีส่วนร่วมในกิจกรรมพัฒนาสตรี)

ตารางที่ 1 การมีส่วนร่วมในกิจกรรมพัฒนาสตรีจำแนกตามฐานะทางเศรษฐกิจ

| การมีส่วนร่วมในกิจกรรมพัฒนาสตรี | ฐานะทางเศรษฐกิจ | |
|---|---|---|
| | สูง (N = 200) | ต่ำ (N = 300) |
| ร่วม | 60% | 40% |
| ไม่ร่วม | 40% | 60% |
| รวม | 100% | 100% |

การวิเคราะห์โดยพิจารณาความถี่หรือร้อยละจะบอกเราได้เพียงว่าตัวแปรชุดนี้มีความสัมพันธ์กันหรือไม่เท่ากัน เราไม่อาจพูดได้ว่าความสัมพันธ์นั้นมากน้อยเพียงใด หรือมีความมั่นใจเพียงใด เรามีทางที่จะแสดงถึงความมั่นใจในความสัมพันธ์ระหว่างตัวแปรได้โดยการคำนวณค่าทางสถิติ เช่น โดยการทดสอบที่เรียกว่า Chi-Square Test และสามารถบอกถึงขนาดความมากน้อยของความสัมพันธ์ได้โดยการคำนวณค่าสหสัมพันธ์ ( Correlation)

ความสัมพันธ์ระหว่าง X และ Y ที่พบในตอนแรกนั้น อาจเป็นความสัมพันธ์ลวง ( squrious relationship) หรือความสัมพันธ์นั้นอาจมีเงื่อนไขบางอย่างแฝงอยู่ก็ได้ เช่น อาจมีตัวแปรหรือปัจจัยอื่นเกี่ยวข้อง เช่น มีตัวแปรอื่นเป็นสาเหตุ

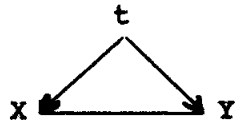
ร่วมทำให้เกิด Y หรือตัวแปรอื่นเข้ามาแทรกระหว่าง X กับ Y หรือตัวแปรอื่น ๆ เป็นเงื่อนไขทำให้ความสัมพันธ์ระหว่าง X กับ Y เพิ่มขึ้นหรือลดลงก็ได้ การวิเคราะห์โดยการนำตัวแปรเข้ามาพิจารณาร่วมด้วยเรียกว่า การวิเคราะห์ตัวแปรหลายตัว ( multivariate analysis) ตัวแปรอื่นที่นำเข้ามาวิเคราะห์เรียกว่า ตัวแปรทดสอบ ( Test variable)

การนำตัวแปรทดสอบเข้ามาวิเคราะห์ หรือเรียกว่า การควบคุมตัวแปรที่สามทำได้โดย การแยกตารางเดิมที่เป็นการวิเคราะห์ความสัมพันธ์ระหว่างสองตัวแปรออกเป็นตารางย่อยตามประเภทของตัวแปรที่สาม เช่น จากตารางเดิม สมมติให้ระดับการศึกษาเป็นตัวแปรที่สาม ซึ่งนำมาทดสอบ เราจะได้ตารางใหม่ลักษณะดังนี้

ตารางที่ 2 การมีส่วนร่วมในกิจกรรมพัฒนาสตรีจำแนกตามฐานะเศรษฐกิจและระดับการศึกษา

| การมีส่วนร่วมในกิจกรรมพัฒนาสตรี | การศึกษาสูง ฐานะทางเศรษฐกิจ | | การศึกษาต่ำ ฐานะทางเศรษฐกิจ | |
|---|---|---|---|---|
| | สูง | ต่ำ | สูง | ต่ำ |
| | N = 120 | N = 50 | N = 80 | N = 250 |
| ร่วม | 83% | 60% | 25% | 36% |
| ไม่ร่วม | 17% | 40% | 75% | 64% |
| รวม | 100% | 100% | 100% | 100% |

จากตารางข้างบนนี้ จะเห็นว่าเมื่อนำระดับการศึกษามาทดสอบปรากฏว่าความสัมพันธ์ระหว่าง X กับ Y เปลี่ยนแปลงไปจากเดิม คือในกลุ่มผู้ที่มีการศึกษาสูงมีจำนวนร้อยละของผู้มีส่วนร่วมในกิจกรรมฯ มากกว่าผู้ไม่ร่วมในกิจกรรมฯ ไม่ว่าจะเป็นผู้ที่มีฐานะเศรษฐกิจสูงหรือต่ำก็ตาม และในกลุ่มที่มีการศึกษาต่ำไม่ว่ามีฐานะทางเศรษฐกิจสูงหรือต่ำ ส่วนใหญ่เป็นผู้ที่ไม่ร่วมในกิจกรรมฯ แสดงว่า ฐานะทางเศรษฐกิจไม่ใช่ตัวการทำให้คนร่วมหรือไม่ร่วมในกิจกรรมฯ แต่ตัวกำหนดการมีส่วนร่วมในกิจกรรมฯ คือ ระดับการศึกษา กรณีที่เป็นเช่นนี้เราสรุปได้ว่า ความสัมพันธ์ที่พบในตอนแรกระหว่างฐานะทางเศรษฐกิจกับการมีส่วนร่วมในกิจกรรมฯ นั้น เป็นความสัมพันธ์ลวง ( spurious relationship) ที่เราพบว่าตัวแปรทั้งสองสัมพันธ์กันในตอนแรกความจริงเป็นเพราะผู้ที่มีฐานะทางเศรษฐกิจสูงส่วนใหญ่เป็นผู้ที่มีระดับการศึกษาสูง และพวกที่มีฐานะทางเศรษฐกิจต่ำส่วนใหญ่เป็นผู้ที่มีระดับการศึกษาต่ำนั่นเอง ระดับการศึกษาเป็นตัวกำหนดของทั้งฐานะทางเศรษฐกิจ และการมีส่วนร่วมในกิจกรรมพัฒนาสตรี อาจโยงความสัมพันธ์ระหว่างตัวแปรทั้งสาม ดังนี้




นอกจากตัวอย่างข้างต้น ตัวแปรที่สามอาจเข้ามามีส่วนทำให้ความสัมพันธ์เดิมเปลี่ยนไปเนื่องจากตัวแปรที่สามเป็นตัวแปรแทรก ( intervening variable) ระหว่าง X กับ Y ลักษณะของตัวเลขในตารางจะเป็นทำนองเดียวกันกับตัวอย่างก่อน แต่แทนที่ t จะเป็นตัวกำหนด X และ Y มันกลับเข้ามาแทรกอยู่ระหว่างกลางดังนี้

$$X \longrightarrow t \longrightarrow Y$$

การตัดสินว่า t เป็นตัวแปรที่กำหนด X และ Y หรือเป็นตัวแปร

แทรกระหว่าง x กับ y นั้น ก็โดยพิจารณาจากเหตุผลที่ว่า t เป็นปรากฏการณ์ที่เกิดขึ้นก่อนหรือหลัง x บางกรณีตัดสินยาก เช่น ระหว่างระดับการศึกษากับฐานะทางเศรษฐกิจอาจมีผู้เห็นว่า ผู้ที่มีระดับการศึกษาสูงได้ต้องมีฐานะทางเศรษฐกิจดีเสียก่อน บางคนอาจเห็นว่าเมื่อมีการศึกษาสูงแล้ว จึงทำให้ฐานะทางเศรษฐกิจดีตามมา ตัวแปรบางตัวสามารถระบุได้โดยง่าย เช่น เพศ อายุ เชื้อชาติ ภูมิหลังทางครอบครัว เป็นต้น เหล่านี้ย่อมมาก่อนระดับการศึกษา ฐานะทางเศรษฐกิจหรือการมีส่วนร่วมในกิจกรรมทางสังคม การวิเคราะห์เกี่ยวกับเรื่องเหล่านี้ สถิติไม่สามารถช่วยอะไรได้ ขึ้นอยู่กับเหตุผลของนักวิจัยเป็นสำคัญ

นอกจากนี้ตัวแปรทดสอบยังอาจเข้ามาเกี่ยวข้องในอีกลักษณะหนึ่งคือเป็นตัวแปรที่ทำให้ความสัมพันธ์ระหว่าง x กับ y เพิ่มขึ้นหรือลดลงจากเดิม ตัวแปร t ไม่ได้ทำให้ความสัมพันธ์ระหว่าง x กับ y หายไปเหมือนสองกรณีแรก แต่ t เป็นตัวที่ช่วยชี้ให้เห็นว่า x กับ y จะสัมพันธ์กันมากขึ้นหรือน้อยลงภายใต้เงื่อนไขอะไร เป็นการขยายรายละเอียดของความสัมพันธ์เดิม โดยการระบุเงื่อนไข (specification) เช่น อาจพบว่าฐานะทางเศรษฐกิจกับการมีส่วนร่วมในกิจกรรมพัฒนาสตรีมีความสัมพันธ์เด่นชัดยิ่งขึ้น เมื่อใช้อายุมาทดสอบหรือควบคุม

ตารางที่ 3 การมีส่วนร่วมในกิจกรรมพัฒนาสตรี จำแนกตามฐานะทางเศรษฐกิจและอายุ

| การมีส่วนร่วมในกิจกรรมพัฒนาสตรี | อายุน้อย | | อายุมาก | |
|---|---|---|---|---|
| | ฐานะทางเศรษฐกิจ | | ฐานะทางเศรษฐกิจ | |
| | สูง | ต่ำ | สูง | ต่ำ |
| | N = 100 | N = 120 | N = 100 | N = 180 |
| ร่วม | 80% | 17% | 40% | 56% |
| ไม่ร่วม | 20% | 83% | 60% | 44% |
| รวม | 100% | 100% | 100% | 100% |

จากตารางตัวอย่างนี้ จะเห็นว่า ความสัมพันธ์ระหว่างฐานะทางเศรษฐกิจกับการมีส่วนร่วมในกิจกรรมพัฒนาสตรีเด่นชัดขึ้นกว่าเดิมในกลุ่มผู้ที่มีอายุน้อย (เปรียบเทียบกับตารางที่ 1) แสดงว่า อายุเป็นตัวแปรตัวหนึ่งที่เป็นเงื่อนไขของความสัมพันธ์ระหว่างฐานะทางเศรษฐกิจกับการมีส่วนร่วมในกิจกรรมพัฒนาสตรี ถ้าหากนำตัวแปรอื่น ๆ มาทดสอบอาจพบว่า ตัวแปรบางตัวทำให้ความสัมพันธ์ระหว่างฐานะทางเศรษฐกิจกับการมีส่วนร่วมในกิจกรรมพัฒนาสตรีลดน้อยลงก็ได้

การเลือกตัวแปรทดสอบว่าควรนำตัวแปรอะไรบ้างมาวิเคราะห์ขึ้นอยู่กับเหตุของผู้วิจัยเอง ผู้วิจัยต้องคิดวางแผนไว้ก่อนว่า มีปัจจัยอะไรบ้างที่จะเกี่ยวข้องกับ x และ y อาจกำหนดไว้ในตอนต้นสมมติฐาน หรืออาจไม่จำเป็นต้องเป็นสมมติฐาน แต่จะต้องคิดเผื่อไว้ และต้องเก็บข้อมูลเหล่านั้นไว้ด้วย

การวิเคราะห์ข้อมูลโดยใช้สถิติแบบพรรณา (Descriptive Statistics)

นักวิจัยนิยมใช้สถิติพรรณา ( descriptive statistics) เช่น เปอร์เซนต์ ( percentages) มัชฌิม ( mean) ฐานนิยม ( mode) มัธยฐาน (median) พิสัย (range) และค่าเบี่ยงเบนมาตรฐาน (standard deviation) ช่วยในการวิเคราะห์ข้อมูล เพื่อสรุปลักษณะสำคัญต่าง ๆ ของกลุ่มตัวอย่าง และอธิบายหรือคาดลักษณะของประชากรที่ศึกษา

ร้อยละ (Percentage) เป็นระเบียบวิธีการทางสถิติที่ใช้ในการเปรียบเทียบอย่างหนึ่ง ซึ่งบางครั้งไม่สามารถที่จะวัดพฤติกรรม ทัศนคติ หรือคุณลักษณะบางอย่างได้โดยตรง แต่อาจจะคิดเป็นร้อยละได้โดยการกะประมาณจากตัวอย่างหรือข้อมูลที่มีอยู่ การใช้ร้อยละเพื่อดูการกระจายของตัวแปร และเพื่อนำเสนอข้อมูลทั่วไปของกลุ่มตัวอย่างในแต่ละหมวดหมู่ของตัวแปร อย่างไรก็ดีการใช้ร้อยละสรุปเรื่องราวบางอย่างจากข้อมูล ควรต้องระมัดระวังพอสมควร เพราะถ้าใช้อย่างสุ่มเพื่อยไม่ระมัดระวัง อาจทำให้เกิดการตีความผิดจากข้อมูลสถิติได้มากเช่นกัน ในมหาวิทยาลัยแห่งหนึ่ง มีนักศึกษาตายอดอยู่เพียง 3 คน และปรากฏว่าภายหลังนักศึกษาตายอด 1 คน สอบได้เกียรตินิยมอันดับ 1 ก็ไม่ควรจะเอาไปวิเคราะห์เพื่อสรุปว่า นักศึกษาตายอดมหาวิทยาลัยนั้น $33\frac{1}{3}$ % สอบได้เกียรตินิยมอันดับ 1 เพราะความหมายของร้อยละเรามักจะใช้สรุปสถานการณ์จากข้อมูลจำนวนมาก ๆ เช่น การวิเคราะห์อัตราการว่างงานของประชากรในเมืองใหญ่ การวิเคราะห์อัตราภาวะเจริญพันธุ์ของสตรีในชนบท การวิเคราะห์ภาวะการถือครองที่ดินของเกษตรกรในจังหวัด การวิเคราะห์ผลการหยั่งเสียงประชามติของคนจำนวนมาก การวิเคราะห์ข้อมูลเกี่ยวกับสาเหตุการตายของประชากร การวิเคราะห์ข้อมูลเกี่ยวกับสาเหตุการย้ายถิ่นของประชากร เป็นต้น

ค่าเฉลี่ย ( Average or mean) ในทางสถิติถือเป็นค่าที่วัดสภาพปานกลาง หรือแนวโน้มเข้าสู่ส่วนกลาง ( Measure of central Tendency) ของข้อมูลเป็นตัวแทนของข้อมูลต่าง ๆ ซึ่งอาจจะเป็นจำนวนเลข จำนวนคะแนน จำนวนคน จำนวนอายุ ที่แสดงให้เห็นถึงค่ากลาง ๆ ของจำนวนเลขทั้งหมด หรือจำนวนคะแนนทั้งหมด เพราะในวงการสถิตินั้น บางครั้งเราอาจจะสรุปหรือย่อข้อมูลจำนวนมาก ๆ โดยการหาตัวแทนของข้อมูลที่เหมาะสมเพียงบางตัวขึ้นมาก็พอจะทำให้ผู้สนใจในตัวเลข หรือข้อมูลนั้น ๆ ได้ข้อสรุปโดยไม่ต้องเอาข้อมูลทั้งหมดไปดู และการรู้ค่าเฉลี่ยนี้จะทำให้ผู้ทำการศึกษาสามารถเทียบค่าต่ำของจำนวนตัวเลข หรือคะแนนในแต่ละหน่วยได้ โดยถือค่ากลางคือ ค่าเฉลี่ยเป็นมาตรการในการเปรียบเทียบ

ในทางสถิติเรามีวิธีการหาค่าเฉลี่ยเพื่อใช้สรุปย่อข้อมูลได้หลายวิธีที่สำคัญและใช้กันมากกับการวิเคราะห์ค่าเฉลี่ยข้อมูลทั่วไป ได้แก่

1. มัชฌิมเลขคณิต (Arithmetic Mean)
2. มัธยฐาน (Median)

ส่วนค่าเฉลี่ยอื่น ๆ ที่ใช้กันอยู่บ้างในบางกรณี คือ

3. มัชฌิมเรขาคณิต (Geometric Mean)
4. มัชฌิมฮาร์โมนิค (Harmonic Mean)
5. ฐานนิยม (Mode)

1. <u>มัชฌิมเลขคณิต</u> ( Arithmetic Mean)

ในการศึกษาลักษณะใด ๆ ในทางสังคมศาสตร์ หรือแขนงวิชาอื่น ๆ ข้อมูลที่ได้ประกอบด้วยสังเกตต่าง ๆ กัน จำนวนมากหรือน้อยขึ้นอยู่กับการทดลองหาม

กลางความแตกต่างของค่าเหล่านี้ จะมีค่าหนึ่งที่แสดงให้เห็นว่าส่วนมากค่าทั้งหมดจะตกอยู่ใกล้เคียงกับค่านี้ ดังนั้นจึงเป็นเครื่องชี้ให้เห็นศูนย์กลางของลักษณะนั้น

มัชฌิมเลขคณิต (Arithmetic Mean) ซึ่งคำนวณได้จากผลรวมของค่าทั้งหมด หารด้วยจำนวนข้อมูล ถ้ามีข้อมูลอยู่ N ค่าซึ่งอาจเขียนแทนด้วย $X_1, X_2, X_3, \ldots\ldots\ldots X_n$ ค่าเฉลี่ยโดยมัชฌิมเลขคณิต (Arithmetic Mean) ของข้อมูลคือ

$$\bar{X} = \frac{X_1 + X_2 + \ldots\ldots\ldots\ldots + X_n}{N}$$

ซึ่งมักจะเขียนสั้น ๆ ด้วย

$$\bar{X} = \sum_{i=1}^{N} X_i/N$$

ในกรณีที่มีข้อมูลจำนวนมาก ๆ และได้มีการแจกแจงความถี่ของข้อมูลออกเป็นชั้นต่าง ๆ แล้ว สูตรที่ใช้ในการคำนวณค่าเฉลี่ยของข้อมูลโดยมัชฌิมเลขคณิต คือ

$$\bar{X} = \frac{f_1x_1 + f_2x_2 + \ldots\ldots\ldots\ldots + f_kx_k}{f_1 + f_2 \ldots\ldots + f_k}$$

ซึ่งมักจะเขียนสั้น ๆ ด้วย

$$\bar{X} = \frac{\sum_{i=1}^{k} f_ix_i}{\sum_{i=1}^{k} f_i}$$

ในเมื่อ $f_1, f_2, \ldots\ldots, f_k$ คือ ความถี่ของค่า $x_1, x_2, \ldots\ldots, x_k$ ตามลำดับ

ตัวอย่าง ข้อมูลต่อไปนี้เป็นข้อมูลเกี่ยวกับรายได้ต่อบุคคลของไทยในช่วง 5 ปี

(เหรียญสหรัฐ)

| รายได้ต่อบุคคล | | | | |
|---|---|---|---|---|
| พ.ศ. 2514 | พ.ศ. 2515 | พ.ศ. 2516 | พ.ศ. 2517 | พ.ศ. 2518 |
| 167 | 188 | 245 | 302 | 368 |

ที่มา : Monthly. Bulletin of Statistic ; United Nations; June 1977

ในที่นี้

$$x_1 = 167,\ x_2 = 188,\ \ldots\ldots\ldots\ldots x_5 = 368$$

และค่าเฉลี่ยโดยมัชฌิมเลขคณิตของรายได้ต่อบุคคลของไทยดังกล่าวในช่วง 5 ปี คือ

$$\bar{x} = \frac{167 + 188 + 245 + 302 + 368}{5}$$

$$= 254$$

ตัวอย่าง ข้อมูลเกี่ยวกับจำนวนบุตรที่มีภายหลังการแต่งงานของสตรี 206 คน (ตัวเลขสมมติ)

| จำนวนบุตรที่มี | จำนวนสตรีที่รายงาน |
|---|---|
| 0 | 10 |
| 1 | 32 |
| 2 | 41 |
| 3 | 26 |
| 4 | 23 |
| 5 | 10 |
| 6 | 19 |
| 7 | 10 |
| 8 | 15 |
| 9 | 12 |
| 10 | 8 |
| รวม | 206 |

ถ้า $x_i$ = จำนวนบุตรที่มี ซึ่งมีค่าได้ตั้งแต่ 0, 1, 2,....,10

$f_i$ = จำนวนสตรีที่รายงาน

ค่าเฉลี่ยของ x's ซึ่งในที่นี้คือจำนวนบุตรที่มีโดยเฉลี่ยภายหลังการแต่งงานของสตรีดังกล่าว อาจหาได้จากสูตร

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

$$= 10(0) + 32(1) + 41(2) + \cdots\cdots\cdots + 12(9) + 8(10)$$

$$= 10 + 32 + 41 + \cdots\cdots\cdots + 12 + 8$$

$$= \frac{836}{264}$$

$$= 405 \quad \text{คน}$$

ตัวอย่าง ข้อมูลเกี่ยวกับปริมาณการขายโดยเฉลี่ยของร้านอาหาร 8 ร้าน ใน มหาวิทยาลัยรามคำแหง (ตัวเลขสมมติ)

| ร้านที่ | บริเวณการขายต่อวัน (บาท) |
|---|---|
| 1 | 200 |
| 2 | 300 |
| 3 | 500 |
| 4 | 400 |
| 5 | 800 |
| 6 | 1300 |
| 7 | 900 |
| 8 | 600 |

ถ้าเฉลี่ยโดยมัชฌิมเลขคณิตได้

$$\bar{X} = \frac{200 + 300 + 500 + 4.0 + 800 + 1300 + 900 + 6.0}{8}$$

$$= 625 \text{ บาท/วัน}$$

ถ้าเฉลี่ยโดยมัธยฐานจะได้ค่าเฉลี่ย คือ

$$= \frac{500 + 600}{2}$$

$$= 550 \text{ บาท/วัน}$$

ซึ่งค่าที่อยู่ระหว่าง 2 ค่ากลางของข้อมูลชุดนี้คือระหว่าง 500 และ 600 บาท โดยพิจารณาจากการเรียงค่าจากน้อยไปหามาก คือ (200, 300, 400, 500, 600, 800, 900, 1300)

2. มัธยฐาน ( Median)

มัธยฐาน (Median) หมายถึงค่าที่เมื่อเรียงค่าต่าง ๆ ตามลำดับจากต่ำไปหาสูงแล้ว ค่านั้นจะตกอยู่ตรงกลาง คือ จะมีค่าต่ำกว่าและสูงกว่าอย่างละครึ่ง หรือกล่าวอีกนัยหนึ่งก็คือ ค่าซึ่งแบ่งข้อมูลออกเป็น 2 พวกเท่า ๆ กัน โดยพวกหนึ่งมีค่าต่ำกว่านี้ และอีกพวกหนึ่งมีค่าสูงกว่านี้ ถ้าหากมีจำนวนเลขเป็นจำนวนคู่ Median เท่ากับ ค่าเฉลี่ยของค่า ซึ่งอยู่ตรงกลางสองค่า

ตัวอย่าง ข้อมูลเกี่ยวกับอายุการทำงานของอาจารย์มหาวิทยาลัย 14 คน เป็นดังนี้

1, 2, 3, 5, 4, 6, 7, 10, 8, 3, 6, 14, 4, 5

ถ้าวิเคราะห์อายุการทำงานโดยเฉลี่ยโดยมัธยฐานอาจทำได้โดยเรียงค่าของข้อมูลจากน้อยไปหามาก คือ

1, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 8, 10, 14

และค่าที่แบ่งข้อมูลทั้ง 14 ตัวเป็น 2 ส่วนเท่า ๆ กัน ก็คือ "5" ซึ่งเป็นค่าที่ "7" ของข้อมูลเมื่อเรียงจากน้อยไปหามาก

ฉะนั้นอายุการทำงานโดยเฉลี่ยของอาจารย์มหาวิทยาลัย 14 คน โดยมัธยฐานจะได้ 5 ปี

แต่ถ้าใช้มัชฌิมเลขคณิตเป็นค่าเฉลี่ยจะได้

= 1 + 2 + 3 + 3 + .....+ 10 + 14 ปี

= 5.56 ปี

ซึ่งพออนุโลมได้ว่าไม่แตกต่างกันมาก

แต่ถ้าเป็นข้อมูลซึ่งมีค่าบางค่าที่สูงหรือต่ำผิดปกติ ค่าเฉลี่ยโดยมัธยฐานจะมีความเป็นตัวแทนได้ดีกว่าค่าเฉลี่ยโดยมัชฌิมเลขคณิต

3. ฐานนิยม (Mode)

ฐานนิยม (Mode) เป็นค่าที่แสดงความโน้มเข้าสู่ส่วนกลางอีกอย่างหนึ่ง อันหมายถึงค่าที่เกิดขึ้นบ่อยที่สุด

ตัวอย่างเช่น มีข้อมูลอยู่ 1 ชุด ซึ่งมีจำนวนข้อมูล 10 ค่า

1, 2, 3, 3, 4, 4, 4, 5, 5, 6

ในข้อมูลชุดนี้ ฐานนิยมเท่ากับ 4

แนวทางในการเลือกใช้ค่าเฉลี่ย (Average)

ในการเลือกใช้ค่าเฉลี่ยเพื่อสรุปย่อข้อมูลนั้น จำเป็นจะต้องทราบข้อดีและข้อเสียของค่าเฉลี่ยเหล่านี้เสียก่อน ซึ่งมีอยู่หลายประการด้วยกัน

มัชฌิมเลขคณิต (Arithmetic Mean)

ข้อได้เปรียบ

1. คำนวณได้ง่ายสำหรับข้อมูลทั่ว ๆ ไป
2. เข้าใจง่าย
3. ค่าของมัชฌิมเลขคณิตเป็นค่าที่แน่นอน
4. เหมาะสมสำหรับการเข้ากับการคำนวณในขั้นต่อไป

ข้อเสียเปรียบ

1. ต้องทราบค่าของข้อมูลทุกค่าจึงจะคำนวณได้
2. ค่าของมัชฌิมเลขคณิตจะตรงกับค่าที่เป็นจริงของข้อมูลเพียงไม่กี่ราย หรือไม่ตรงกันเลย

ประโยชน์

1. เหมาะสำหรับวัดค่าความเป็นกลางสำหรับข้อมูลที่มีค่าไม่แตกต่างกันมากนัก

มัธยฐาน (Median)

ข้อได้เปรียบ

1. คำนวณง่าย
2. เข้าใจง่าย
3. ไม่กระทบกระเทือนมากนัก เมื่อมีค่าที่สูงหรือต่ำผิดปกติในข้อมูล
4. ไม่ต้องใช้ค่าของข้อมูลทุกค่า

ข้อเสียเปรียบ

1. ไม่สามารถหาค่าของมัธยฐานได้จากข้อมูลที่มิได้จัดเรียงลำดับ
2. ไม่สะดวกในการใช้คำนวณสถิติอื่น ๆ เช่น การรวมค่าเฉลี่ยโดยมัธยฐานต้องหาใหม่หมด

ฐานนิยม (Mode)

ข้อได้เปรียบ

1. เข้าใจง่าย
2. สามารถหาค่าเฉลี่ยของฐานนิยมได้จากข้อมูลที่ได้จัดเรียงลำดับแล้ว
3. เป็นส่วนเฉลี่ยที่อาศัยเสียงข้างมาก

ข้อเสียเปรียบ

1. คำนวณได้ยาก
2. ไม่เหมาะสำหรับเข้าสมการคำนวณในขั้นต่อไป
3. ค่าของฐานนิยมเปลี่ยนไปตามขนาดของอันตรภาคชั้น

การวิเคราะห์ข้อมูลเพื่อสรุปย่อข้อมูล โดยการใช้ค่าเฉลี่ย (Average) ในกรณีที่ข้อมูลมีเป็นจำนวนมากเพียงอย่างเดียวยังไม่เพียงพอ เพราะว่าข้อมูลที่มีจำนวนมาก ๆ จะมีความแปรปรวนในค่าตัวเลขอยู่เป็นอันมาก ถ้าหากเราใช้ค่าเฉลี่ยเพียงอย่างเดียวไม่เพียงพอที่จะอธิบายให้เห็นสภาพความแปรปรวนต่าง ๆ ภายในชุดของข้อมูล ดังนั้นในการวิเคราะห์ข้อมูลเพื่อสรุปย่อข้อมูล จึงจำเป็นต้องคำนวณค่าที่ใช้วัดความแปรปรวน (Variation) หรือกระจาย (Dispersion) ของข้อมูลนั้น ๆ กำกับไปด้วย ค่าเฉลี่ยด้วยค่าทางสถิติที่ใช้วัดความแปรปรวนของข้อมูลที่นิยมใช้กันมีอยู่ 2 ตัว คือ พิสัย (Range) และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation)

พิสัย (Range) เป็นค่าที่ใช้วัดการกระจายหรือความแปรปรวนในค่าตัวเลขของข้อมูลที่ง่ายที่สุด และคำจำกัดความของพิสัยของข้อมูลชุดหนึ่งชุดใดคือ

พิสัย = ค่าสูงสุดของข้อมูล - ค่าต่ำสุดของข้อมูล

นั่นคือ

ถ้า $\{x_1, x_2, x_3, \ldots\ldots, x_n\}$ แทนค่าตัวเลขของข้อมูลชุดหนึ่งชุดใด

พิสัย = $X_{max.} - X_{min.}$

ตัวอย่าง ข้อมูล 3 ชุด ซึ่งมีจำนวนข้อมูลเท่ากัน คือ 7 ค่า

ชุดที่หนึ่ง : 19, 20, 22, 24, 26, 28, 32

ชุดที่สอง : 23, 24, 25, 25, 25, 26, 27

ชุดที่สาม : 25, 25, 25, 25, 25, 25, 25

พิสัยของข้อมูลชุดที่หนึ่ง = 32 - 19 = 13

พิสัยของข้อมูลชุดที่สอง = 27 - 23 = 4

พิสัยของข้อมูลชุดที่สาม = 25 - 25 = 0

ค่าพิสัยที่คำนวณออกมาจากข้อมูลทั้งสามชุด พอจะมีความหมายได้ว่า

ข้อมูลชุดที่หนึ่ง มีความแปรปรวนในค่าตัวเลขมาก เพราะฉะนั้นค่าของพิสัยจะใหญ่

ข้อมูลชุดที่สอง มีความแปรปรวนในค่าตัวเลขไม่มากนัก เพราะฉะนั้นค่าของพิสัยจะไม่ใหญ่นัก

ข้อมูลชุดที่สาม ไม่มีความแปรปรวนในค่าของตัวเลขเลย เพราะฉะนั้นค่าของพิสัยจะเป็นศูนย์

ฉะนั้น ในการวิเคราะห์เพื่อสรุปย่อข้อมูลจึงอาจจะใช้พิสัย (range) กำกับค่าเฉลี่ยไปด้วย ซึ่งจะทำให้ค่าสถิติทั้งสองค่าบอกความหมายจากข้อมูลได้ชัดเจนยิ่งขึ้น เช่น การวิเคราะห์ราคาข้าวเปลือก 2 ปี ถ้าสรุปว่า

ปี 2526 ราคาข้าวเปลือกที่ชาวนาขายได้โดยเฉลี่ยทั้งปี = 2,220 บาท

พิสัย = 400 บาท

ปี 2527 ราคาข้าวเปลือกที่ชาวนาขายได้โดยเฉลี่ยทั้งปี = 2,100 บาท

พิสัย = 600 บาท

ก็พอจะทำให้ผู้สนใจใช้สถิติข้อมูลพอเข้าใจความหมายแตกต่างในภาวะราคาข้าวเปลือกได้ดีพอสมควร โดยไม่ต้องกลับไปดูข้อมูลรายละเอียด

ส่วนเบี่ยงเบนมาตรฐาน ( Standard Deviation) เป็นเครื่องมือสำหรับใช้วัดการกระจายหรือความแปรปรวนของข้อมูลได้ดี และนิยมใช้กันมากในงานวิจัย วิธีการหาส่วนเบี่ยงเบนมาตรฐานทำได้ดังต่อไปนี้

ถ้า $X_1, X_2, X_3, \ldots\ldots, X_n$ เป็นข้อมูลชุดหนึ่ง

ส่วนเบี่ยงเบนมาตรฐานหรือ Standard Deviation ของ X's ก็คือ

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{N} (X_1 - \bar{X})^2}{N}}$$

ในเมื่อ $\sigma_X$ อ่านว่า(Sigma X) = standard deviation ของ X's

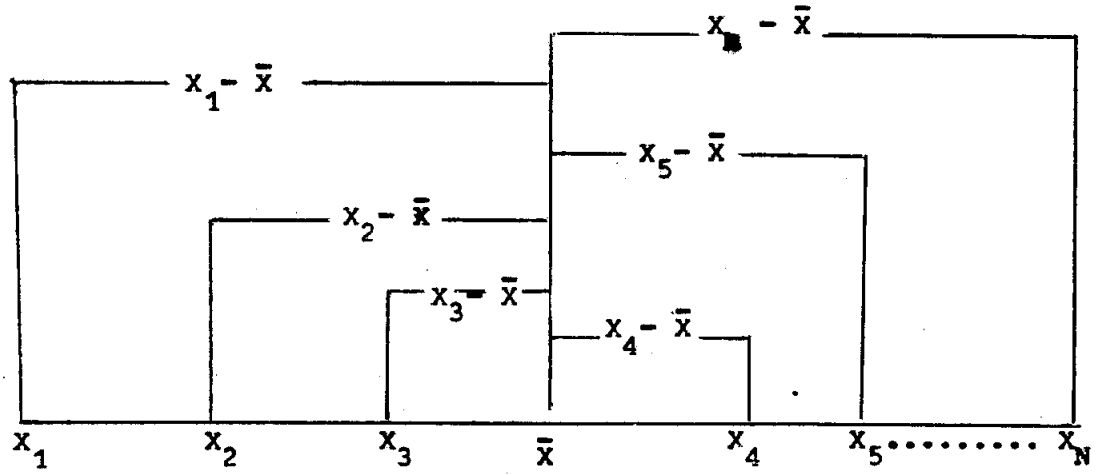
$$\bar{X} = \sum_{i=1}^{N} X_i/N = \text{Arithmetic mean ของ X's}$$

ข้อควรสังเกต ถ้าพิจารณาจากสูตรจะเห็นว่า standard deviation ของข้อมูลชุดหนึ่งชุดใด ก็คือ กรณีที่สองของค่าเฉลี่ยของกำลังสองของส่วนเบี่ยงเบนที่ X's อยู่ห่างจาก mean ของมันและถ้า X's มีความแปรปรวนหรือมีการกระจายมาก deviation "$X_i - \bar{X}$"

ซึ่งในแง่กราฟ ก็คือระยะที่ $X_i$ อยู่ห่างจาก $\bar{X}$ ก็จะมีมากฉะนั้น

$$\sigma_X = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$$

ก็จะมีค่ามากสำหรับข้อมูลที่มีความแปรปรวนมาก และมีค่าน้อยสำหรับข้อมูลที่มีความแปรปรวนน้อย (ดูรูป)




ในทางสถิติจึงใช้ "$\sigma x$" เป็นค่าที่วัดการกระจายของข้อมูลได้อย่างกว้างขวาง เพราะมีความหมายทางคณิตศาสตร์อยู่มากกว่า ค่าที่วัดการกระจายอื่น ๆ

ตัวอย่าง ส่วนเบี่ยงเบนมาตรฐาน

ชุดที่หนึ่ง

$$\sigma_1 = \sqrt{\frac{(19-25)^2 + (20-25)^2 + (22-25)^2 + (24-25)^2 + (26-25)^2 + (28-25)^2 + (32-25)^2}{7}}$$

$$= \sqrt{13.5714} = 3.6065$$

ชุดที่สอง

$$\sigma_2 = \sqrt{\frac{(23-25)^2 + (24-25)^2 + (25-25)^2 + (25-25)^2 + (25-25)^2 + (26-25)^2 + (27-25)^2}{7}}$$

$$= \sqrt{1.4285} = 1.1951$$

และชุดที่สาม

$$\sigma_3 = \sqrt{\frac{(25-25)^2 + (25-25)^2 + \cdots\cdots\cdots\cdots\cdots + (25-25)^2}{7}}$$

$$= \sqrt{0} \qquad = 0$$

ซึ่งจะเห็นได้ชัดว่า ข้อมูลที่มีความแปรปรวน หรือการกระจายมากกว่าจะมีค่าส่วนเบี่ยงเบนมาตรฐานที่ใหญ่กว่า และถ้าข้อมูลไม่มีความแปรปรวนเลย ส่วนเบี่ยงเบนมาตรฐานจะมีค่าเป็น 0

สำหรับการคำนวณค่า " x" จากข้อมูลที่มีการแจกจงความถี่แล้วอาจทำได้โดยหลักเกณฑ์เดียวกัน และสูตรจะอยู่ในรูป

$$= \frac{\sum_{i=1}^{k} f_i (X_i - \bar{X})^2}{\sum_{i=1}^{k} f_i}$$

การวิเคราะห์ข้อมูลด้วยวิธีการสถิติที่ไม่ใช้พารามิเตอร์ ( Non-Parametric Method)

สถิติที่ใช้ในการทดสอบความสัมพันธ์ระหว่างตัวแปร มีมากมายหลายชนิด ซึ่งแต่ละชนิดก็เหมาะสมที่จะใช้กับข้อมูลและลักษณะการกระจายของข้อมูลบางประเภท แต่ที่นิยมใช้สำหรับการวิจัยทางสังคมศาสตร์ ได้แก่

การทดสอบไคสแควร์ ( Chi-Square Test) เป็นระเบียบวิธีการทางสถิติที่ใช้ทดสอบความเป็นอิสระ ( test of Indepedence) คือ ดูว่าความสัมพันธ์ระหว่างตัวแปรสองตัวหรือดูความแตกต่างระหว่างตัวแปรสองตัวว่าต่างกันหรือเหมือนกันเพียงใด และใช้ทดสอบความมีนัยสำคัญทางสถิติด้วย ( Test of Statistical of Significance) ในการวิจัยที่มีระดับการวัดของตัวแปร ( level of measurement) เป็นนามสเกล ( noninal scale) และอันดับสเกล ( ordinal scale) และข้อมูลที่รวบรวมมาได้นั้นสามารถที่จะจำแนกคุณสมบัติของประชากรออกได้สองทางแล้ว การเลือกใช้ระเบียบวิธีการทางสถิติประเภทไคสแควร์ในการวิเคราะห์ข้อมูลจึงเป็นการเหมาะสม

สูตรทั่วไปของไคสแควร์ (Chi-Square)

$$X^2 = \frac{(O - E)^2}{E}$$

เมื่อ O = ค่าของข้อมูลที่รวบรวมได้ (observed value)

E = ค่าของข้อมูลที่คาดว่าจะเป็น (expected value)

$E = \frac{r.k}{N}$ = หมายความว่า คำนวณได้จากผลคูณของแต่ละ r ตามแถว ( row) และแต่ละค่า k สดมภ์ ( column) หารด้วยจำนวนทั้งหมด

**Degree of Freedom**

Degree of Freedom หมายถึง ระดับความเป็นอิสระของข้อมูล กล่าวคือ เป็นจำนวนค่าใน cell ที่เราสามารถหาค่าได้น้อยที่สุดแล้วสามารถหาค่าที่เหลืออยู่ได้

ยกตัวอย่าง เช่น d.f. = 2 แสดงว่าเราหาค่าใน cell เพียง 2 cell เราก็สามารถจะคำนวณหาค่าที่เหลือในแต่ละ cell ได้

สูตร d.f. = (r-1) (k-1)

r = จำนวน row

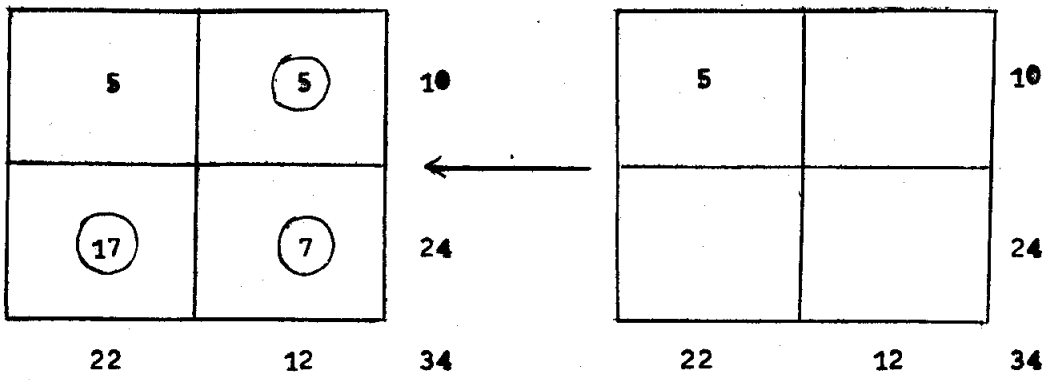
k = จำนวน column

ใช้ประโยชน์ในการเทียบค่าในตารางแสดงค่าความสัมพันธ์ต่าง ๆ หรือหมายถึง ความเป็นอิสระของข้อมูลในตาราง ใช้ในการหาค่า Chi-square เมื่อนำไปเปิดตารางค่า Chi-square ในระดับนัยสำคัญตามที่ต้องการจะต้องใช้ค่า d.f. ประกอบด้วย

หรือหมายถึงการสูญเสียอิสระภาพของข้อมูล เมื่อเราสุ่มตัวอย่างมาจากประชากรทั้งหมด ตัวอย่างที่สุ่มมาเป็นตัวแทนของประชากรนี้ ถือว่าได้สูญเสียอิสระภาพไป

ในความหมายว่า ความเป็นอิสระของข้อมูลในตาราง หมายถึงว่า ถ้าเราหาค่าของจำนวนในตารางเท่ากับค่า degree of freedom ก็สามารถหาค่าอื่น ๆ ในตารางที่ยังไม่หรามได้ แต่จะต้องหราบจำนวนรวมทั้งหมด และจำนวนรวมในแต่ละrow และ column ด้วย

ตัวอย่าง d.f. = 1 ถ้าเราทราบค่าเพียง 1 cell ในตารางก็สามารถหาค่าอื่น ๆ ได้



**Degree of Freedom**

Degree of Freedom หมายถึง ความเป็นอิสระที่จะแปรปรวนไปได้ของข้อมูล เป็นจำนวนข้อมูลที่ต้องการรู้ใน cell หนึ่ง ๆ ว่า จะต้องรู้ค่าข้อมูลใน cell เป็นจำนวนกี่ cell แล้วจึงจะสามารถคำนวณหาค่าตัวเลขลงใน cell ที่เหลือได้ ตัวอย่างเช่น จากข้อมูลเรื่องรายได้กับการศึกษาของคนมีข้อมูลเมื่อลงในตาราง cross tabulation จะมี cell ที่จะลงข้อมูลจำนวน 4 cell

| รายได้ \ การศึกษา | สูง | ต่ำ | รวม |
|---|---|---|---|
| สูง | (20) | | 30 |
| ต่ำ | | | 70 |
| รวม | 60 | 40 | 100 |

จากตัวอย่างที่ให้ค่า total ของข้อมูลมา ถ้าใน cell เรารู้เพียง 1 cell คือ 20 (รู้ข้อมูลช่อง 1) แล้วอีก 3 cell ที่ไม่ทราบค่าก็สามารถคำนวณหาค่าออกมาได้

degree of freedom ของตารางนี้ = 1

ในกรณี R by C table คือมีจำนวน cell = 4 cell

จะหาค่า degree of freedom ได้จาก

$$d.f. = (r-1)(k-1)$$

(r = row, k = column)

เช่นมี row = 3, column = 4

$$d.f. = (3-1)(4-1) = 6$$

แสดงว่าเมื่อเรารู้ค่าตัวเลขใน 1 cell จาก 4 cell ก็สามารถคำนวณค่าที่เหลือลงในตารางได้ degree of freedom จะใช้ในการเปิดตารางหาค่า $x^2$, T-test เป็นต้น

**Test of Independent**

Test of Independent ใช้ทดสอบความเป็นอิสระแก่กันของตัวแปรที่มีระดับการวัด nominal ขึ้นไป (ใช้ได้ทุกชนิด) ขนาดที่จำนวนพอเหมาะคือตั้งแต่ 30 คนขึ้นไป และแต่ละ cell มีประชากรไม่น้อยกว่า 5

การทดสอบใช้ Chi-Square test($X^2$) เพียงบอกความสัมพันธ์กันในระหว่างตัวแปรโดย

- บอกว่ามีความสัมพันธ์กันหรือไม่
- บอกว่ามีความแตกต่างกันหรือไม่ แต่บอกว่าแตกต่างกันเท่าไหร่ไม่ได้ ถ้าแตกต่างกันก็แสดงว่ามีความสัมพันธ์กัน

การทดสอบความเป็นอิสระต่อกัน เป็นการเปรียบเทียบความแตกต่างของสิ่งที่เราได้สังเกต (observation) กับการคาดหวัง (expectation) ว่ามีความแตกต่างอย่างมีนัยสำคัญหรือไม่ (significance)

ในการทดสอบจะต้องทำเป็น Cross tabulation คือ เอาตัวแปร 2 ตัวมา cross กัน เป็นตาราง four - fold table

สูตร $X^2 = \sum \frac{(O - E)^2}{E}$

$E$ = Expected Values ($E = \frac{O}{N} \times 100$)

$O$ = Observed Values

สรุปแล้ว $X^2$ ใช้ทดสอบความสัมพันธ์และความแตกต่างแต่บอกขนาดความแตกต่างไม่ได้

เมื่อหาค่า $x^2$ ได้จากสูตรดังกล่าวข้างต้นแล้วให้นำค่าที่คำนวณได้ไปเปรียบเทียบกับตารางเพื่อเปรียบเทียบระดับความเชื่อมั่น

ถ้าค่า $x^2$ น้อยกว่าค่าในตารางแสดงว่า ตัวแปร 2 ตัว มีความสัมพันธ์กัน มีความแตกต่างกันอย่างมีนัยสำคัญ

ในการแปลความหมายของค่าไคสแควร์ เพื่อทราบว่าผลต่างจะมีนัยสำคัญหรือไม่ ต้องอาศัยค่าไคสแควร์ที่ระดับความมีนัยสำคัญต่าง ๆ และในการดูจากตารางนั้นจะต้องหาค่าขั้นแห่งความเป็นอิสระ (degree of freedom) ด้วย โดยการใช้สูตร

$$d.f. = (r-1)(k-1)$$

ข้อจำกัดในการใช้ไคสแควร์เพื่อทดสอบสมมติฐานมีดังนี้

1. จำนวนของประชากรหรือตัวอย่างในแต่ละเซล (cell) จะต้องมีค่าไม่น้อยกว่า 5 และค่า N ไม่ต่ำกว่า 30

2. ค่าไคสแควร์ จะไม่สามารถบอกทิศทางหรือระดับความสัมพันธ์ได้

<u>แกมมา</u> (Gamma) เป็นระเบียบวิธีการทางสถิติที่ใช้สำหรับหาค่าความสัมพันธ์ระหว่างตัวแปร 2 ตัว และสามารถบอกทิศทางขนาดความสัมพันธ์ของตัวแปรได้ว่าเป็นไปในทางบวกหรือลบ (Positive and Nagative assosiation) ใช้สำหรับตัวแปรที่มีระดับการวัดของตัวแปรเป็นอันดับสเกล (ordinal scale) ขึ้นไป และความถี่ของข้อมูลมีลักษณะการกระจายแบบปกติ (normal distribution)

สูตรทั่วไปของแกมมา

$$\text{Gamma} = \frac{CP - IP}{CP + IP}$$

เมื่อ CP = Consistance pair

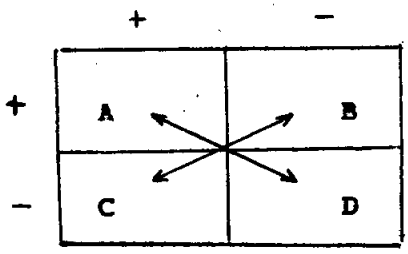
IP = Inconsistance pair

<u>ยูลส์คิว</u> (Yule's Q) เป็นระเบียบวิธีการทางสถิติที่บอกให้ทราบถึงทิศทางและระดับความสัมพันธ์ของตัวแปร โดยมีข้อแม้ว่าข้อมูลที่ได้จะต้องมีลักษณะเป็น Four-fold tables และอย่างน้อยระดับการวัดของตัวแปรเป็นอันดับสเกล (ordinal scale) และความถี่ของข้อมูลมีลักษณะการกระจายแบบปกติ (normal distribution)

สูตรทั่วไปของยูลส์คิว

$$\text{Yule's Q} = \frac{(B \times C) - (A \times D)}{(B \times C) + (A \times D)}$$

A, B, C, D เป็น four cells of the table



ค่าของกูลด์คิว และแกมมา มีเกณฑ์วัดระดับความสัมพันธ์ดังนี้

| | |
|---|---|
| .70 หรือมากกว่า | ถือว่ามีความสัมพันธ์เชิงปฏิภาคในระดับสูงมาก (a very strong positive association) |
| + .50 ถึง +.69 | ถือว่ามีความสัมพันธ์เชิงปฏิภาคในระดับค่อนข้างสูง (a substantial positive association) |
| + .30 ถึง + .49 | ถือว่ามีความสัมพันธ์เชิงปฏิภาคในระดับปานกลาง (a moderate positive association) |
| + .10 ถึง + .29 | ถือว่ามีความสัมพันธ์เชิงปฏิภาคในระดับค่อนข้างต่ำ (a low positive association) |
| + .01 ถึง + .09 | ถือว่ามีความสัมพันธ์เชิงปฏิภาคในระดับแทบจะไม่มีความสัมพันธ์ (a negligible positive association) |
| | ถือว่าไม่มีความสัมพันธ์ (no association) |
| -.01 ถึง -.09 | ถือว่ามีความสัมพันธ์เชิงนิเสธในระดับแทบจะไม่มีความสัมพันธ์ (a negligible negative association) |
| -.10 ถึง -.29 | ถือว่ามีความสัมพันธ์เชิงนิเสธในระดับค่อนข้างต่ำ (a low nagative association) |
| -.30 ถึง -.49 | ถือว่ามีความสัมพันธ์เชิงนิเสธในระดับปานกลาง (a moderate negative association) |
| -.50 ถึง -.69 | ถือว่ามีความสัมพันธ์เชิงนิเสธค่อนข้างสูง (a substantial negative association) |
| -.70 หรือมากกว่า | ถือว่ามีความสัมพันธ์เชิงนิเสธในระดับสูงมาก (a very strong negative association) |

<u>ตัวอย่าง</u> สมมติว่าท่านทำการวิจัยเรื่องหนึ่งและตั้งสมมติฐานว่า ประชากรที่มีการศึกษาสูง จะมีรายได้สูงกว่าประชากรที่มีการศึกษาต่ำ และท่านได้รวบรวมข้อมูลจากตัวอย่างของประชากรที่มีรายได้ในนิคมสร้างตนเองแก่งคอย จำนวน 407 ตัวอย่าง ดังต่อไปนี้

| การศึกษา \ รายได้ | สูง | กลาง | ต่ำ |
|---|---|---|---|
| สูง | 9 | 26 | 13 |
| กลาง | 19 | 75 | 83 |
| ต่ำ | 16 | 56 | 110 |

ก. ให้เลือกวิธีการทางสถิติที่เหมาะสมพิสูจน์สมมติฐานที่ตั้งไว้จากข้อมูลข้างบน พร้อมทั้งอธิบายเหตุผลที่เลือกวิธีการทางสถิติ

ข. ให้ตีความค่าที่ได้จากการพิสูจน์ทางสถิติ

สมมติฐาน : ประชากรที่มีการศึกษาสูงจะมีรายได้สูงกว่าประชากรที่มีการศึกษาต่ำ

| รายได้ / การศึกษา | ต่ำ | กลาง | สูง | รวม |
|---|---|---|---|---|
| ต่ำ | 110(92.12) | 56(70.21) | 16(19.68) | 182 |
| กลาง | 83(59.59) | 75(68.28) | 19(19.14) | 177 |
| สูง | 13(24.29) | 26(18.52) | 9( 5.19) | 48 |
| รวม | 206 | 157 | 44 | 407 |

ก. ในการพิสูจน์สมมติฐานที่ตั้งไว้จากข้อมูลข้างบน ในที่นี้เลือกใช้การทดสอบ Chi-Square Test ($X^2$) ซึ่งเป็นการทดสอบหาความสัมพันธ์ของตัวแปรทั้งสองว่า มีความสัมพันธ์เป็นอิสระต่อกันหรือไม่ หรือดูความแตกต่างระหว่างตัวแปร 2 ตัวว่าต่างกันหรือเหมือนกันเพียงใด ซึ่งเป็นวิธีการของ Non-Parametric Method และใช้ค่าแถมมา เพื่อหาความสัมพันธ์ระหว่างตัวแปร 2 ตัว และทิศทางของความสัมพันธ์นั้น เมื่อพบว่าตัวแปรทั้งสองมีความสัมพันธ์กันจริง เมื่อทดสอบ Chi-Square Test ($X^2$) แล้ว

เหตุที่เลือกใช้การทดสอบแบบ Chi-Square Test ($X^2$) และค่าแถมมา เพราะข้อมูลที่ได้มาเป็นข้อมูลที่มีลักษณะดังนี้คือ

(1) เป็นข้อมูลที่มีระดับการวัดในระดับ Ordinal Scale

เป็นที่ยอมรับกันว่า เมื่อใช้แกมมาวัดความสัมพันธ์ของตัวแปรอย่างน้อยที่สุด ตัวแปรทั้งสองต้องเป็นข้อมูลที่มีระดับการวัด Ordinal Scale ซึ่งเป็นระดับการวัดที่บอกเพียงความแตกต่างว่ามีมากน้อยเท่าใด

(2) มีจำนวนข้อมูล หรือ N มากกว่า 30 และในแต่ละ cell ของตารางมีค่ามากกว่า 5 ขึ้นไป

Chi-square เป็นการทดสอบหาค่าความสัมพันธ์ของตัวแปร 2 ตัว โดยใช้หลักของความน่าจะเป็น ซึ่งเหมาะสมกับลักษณะของข้อมูลที่ได้มามาก เพราะข้อมูลที่ได้มาบอกค่าสังเกต ($O_{ij}$) ได้ เมื่อหาค่าของความน่าจะเป็นที่คำนวณได้ ($E_{ij}$) ก็สามารถจะใช้สูตร

$$x^2 = \frac{(O_{ij} - E_{ij})}{E_{ij}}$$

O = ค่าของข้อมูลที่รวบรวมได้

E = ค่าของข้อมูลที่คาดว่าจะเป็น

$$E = \frac{\text{row X Column}}{\text{total}}$$

คำนวณหาความสัมพันธ์ของตัวแปรทั้งสองได้ และเป็นวิธีการที่ไม่ยากเกินไป ส่วนแกมมานั้น เป็นการหาความสัมพันธ์ของตัวแปร จะบอกระดับและทิศทางความสัมพันธ์แก่ผู้วิจัยได้ จากสูตร

$$\text{Gamma} = \frac{CP - IP}{CP + IP}$$

ดังนั้นจึงเห็นว่า วิธีการเลือกใช้มีความเหมาะสมในการทดสอบสมมติฐานข้างต้น

การทดสอบสมมติฐาน

$$x^2 = \frac{(O_{ij} - E_{ij})}{E_{ij}}$$

$$= \frac{(110-92.12)^2}{92.12} + \frac{(56-70.21)^2}{70.21} + \frac{(16-19.68)^2}{19.68}$$
$$+ \frac{(83-59.59)^2}{59.59} + \frac{(75-68.28)}{68.28} + \frac{(19-19.14)^2}{19.14}$$
$$+ \frac{(13-24.29)^2}{24.29} + \frac{(26-18.52)^2}{18.52} + \frac{(9-5.19)^2}{5.19}$$

$$= 3.47 + 2.88 + 0.69 + 0.48 + 0.66 + 0.001 + 5.25 + 3.02 + 2.74$$

$$= 19.191$$

$$d.f. = (r-1)(k-1)$$

$$= (3-1)(3-1)$$

$$= 2 \times 2$$

$$= 4$$

ระดับนัยสำคัญ = 0.5

จาก d.f. = 4 ระดับนัยสำคัญ .05 เปิดค่า $x^2$ จากตาราง = 2.132 แสดงว่าค่า $x^2$ จากการคำนวณมากกว่า ค่า $x^2$ จากตารางใน d.f. = 4 ∴ ในการพิสูจน์สมมติฐานครั้งนี้ ผู้วิจัยยอมรับสมมติฐาน แสดงว่า การศึกษาและรายได้มีความสัมพันธ์กัน หรืออาจจะดีความว่า จากค่าที่คำนวณได้ โดย d.f. = 4 level of significance = 0.5 เมื่อเปิดตาราง Chi-square test ที่ d.f. = 4 level of significance = 0.5 แล้วได้ค่า = 2.132 ∴ ค่าที่คำนวณได้มากกว่าค่าในตาราง สรุปว่าตัวแปรทั้ง 2 ไม่เป็นอิสระต่อกัน แสดงว่า ทั้ง 2 ตัวแปรมีความสัมพันธ์กัน นั่นคือ การศึกษามีความสัมพันธ์กับรายได้

<u>จากตาราง</u>

| รายได้ / การศึกษา | สูง | ต่ำ |
|---|---|---|
| สูง | 119 | 96 |
| ต่ำ | 72 | 110 |

สูตร $\text{Gamma} = \frac{CP - IP}{CP + IP}$

CP = consistance pair

IP = inconsistance pair

$$\text{Gamma} = \frac{(199 \times 110) - (96 \times 72)}{(199 \times 110) + (96 \times 72)}$$

$$= \frac{13090 - 6912}{13090 + 6912}$$

$$= \frac{6178}{20002}$$

$$= 0.30$$

แสดงว่าตัวแปรทั้งสองตัว คือระดับการศึกษากับรายได้ มีความสัมพันธ์กันในทางบวก มีค่าแกมมาเท่ากับ 0.30 ซึ่งจัดว่ามีความสัมพันธ์ในระดับปานกลาง

ข. <u>การตีความจากสมมติฐาน</u>

จากการทดสอบสมมติฐานพบว่า ประชากรที่มีการศึกษาสูงจะมีรายได้สูงกว่าประชากรที่มีการศึกษาต่ำ ผลการทดสอบพบว่า Chi-square เท่ากับ 19.191 ซึ่งในระดับนัยสำคัญเท่ากับ 0.05 และมีค่ามากกว่าค่า Chi-square จากตาราง แสดงว่า การศึกษาของประชากรมีความสัมพันธ์กับรายได้ของประชากร และค่าแกมมาเท่ากับ 0.30 แสดงถึงความสัมพันธ์ในทางบวก หมายถึง การศึกษาของประชากรสูงจะทำให้รายได้ของประชากรสูงตามไปด้วย แต่มีความสัมพันธ์กันในระดับปานกลาง แสดงว่า การเปลี่ยนแปลงการศึกษาของประชากรมีผลต่อการเปลี่ยนแปลงรายได้ของประชากรต่ำ

## การวิเคราะห์ข้อมูลด้วยวิธีการสถิติที่ใช้พารามิเตอร์ (Parametric Method)

เทคนิคในการวิเคราะห์ข้อมูลที่สูงขึ้นไป ได้แก่ การวิเคราะห์ข้อมูลโดยการวิเคราะห์ความสัมพันธ์เชิงฟังชั่นของข้อมูล ที่ใช้กันมาก คือ.-

1. Pearson Product - Moment Correlation
2. Partial Correlation
3. Regression Analysis

### Pearson Product-Moment Correlation

The Pearson correlation coefficient (r) is used to measure the strength of a relationship between two variables, it also indicates the direction of the relationship by a plus or minus sign. The strength of the relationship from this statistical test indicates both the goodness of fit of a linear regression line to the data and, when r is squared, the proportion of variance in one variable explained by the other.

Mathematically, r is defined as the ratio of covariation to square root of the product of the variation in X (independent variable) and the variation in Y (dependent variable). The formula is:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\ (y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n} (x_i - \bar{x}^2)\right]\left[\sum_{i=1}^{n} (y_i - \bar{y}^2)\right]}}$$

where

$x_i$ = the observation of variable x

$y_i$ = the observation of variable y

n = number of observations

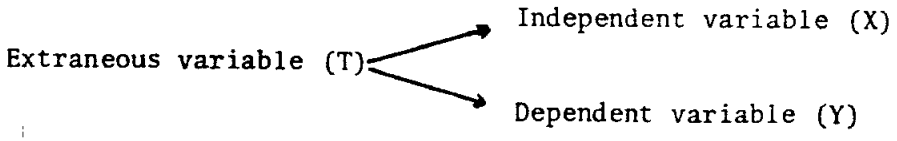
$\bar{x}$ = mean of variable x

$\bar{y}$ = mean of variable y

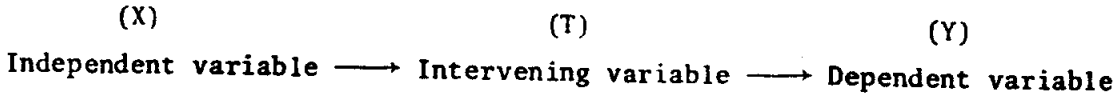
Partial Correlation Analysis

Partial correlation provides the researcher with a single measure of association describing the relationship between two variables while adjusting for the effects of one or more additional variables. Generally, partial correlation is a statistical method used to locate spurious relationships and to determine the existence of an intervening variable (Blalock, 1957).

A spurious correlation is defined as a relationship between two variables, independent (X) and dependent (Y), in which X's correlation with Y is solely the result of the fact that X varies along with some other variable (T) which is indeed the true predictor of X. In this case, when the effects of T are controlled or held constant Y no longer varies with X.




The distinction between the application of partial correlation in determining the spurious correlation and intervening variable is a logical and theoretical issue, not a statistical one. Instead of causing both X and Y, T may be an intervening causal link between them. That is, X may cause T, which in turn may cause Y. T is not limited to only one variable; in fact, there might be a group of variables intervening between X and Y. When the intervening variable is controlled the existing relationship between X and Y vanishes or is substantially reduced.



The basic formula for the computation of partial correlation is:

$$r_{ij \cdot k} = \frac{r_{ij} - (r_{ik})\,(r_{jk})}{\sqrt{1 - r_{ik}^2}\,\sqrt{1 - r_{jk}^2}}$$

where

k = control variable

i = independent variable

j = dependent variable

Regression Analysis

Regression analysis เป็นการวิเคราะห์ regression ซึ่งเป็นหลักเกณฑ์หรือแนวความคิดพื้นฐานของการหาค่าสถิติในแบบ parametric statistic ซึ่งใช้หลักการ Regression Analysis เป็นเกณฑ์โดยใช้การวิเคราะห์ จากค่า Parameter ซึ่งถือว่าเป็นค่าแท้จริงของประชากรทั้งหมดที่ทำการศึกษา และค่าสถิติที่ได้จะถือว่ามี power efficiency สูงมาก สามารถยืนยันได้หนักแน่นกว่าวิธีการหาค่าสถิติแบบ Non-Parametric Statistics ซึ่งใช้หลักเกณฑ์ของความน่าจะเป็น (Probability) ในการคำนวณหาค่าทางสถิติต่าง ๆ

การวิเคราะห์ regression จะวิเคราะห์ความสัมพันธ์ของข้อมูล โดยใช้สมการเส้นตรงเป็นหลัก อาศัยสูตร คือ

$$Y = a + bx$$

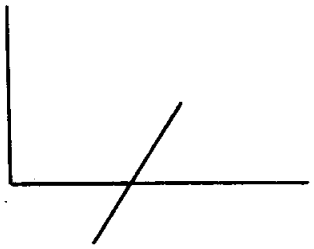
ซึ่งเมื่อรู้ค่า X จะสามารถคำนวณหาค่า Y ได้

เมื่อคำนวณได้ค่า X และ Y แล้วก็สามารถคำนวณหาค่าความสัมพันธ์กันของข้อมูลตามวิธี Parametric Statistics โดยใช้สูตรหาค่า correlation (r)

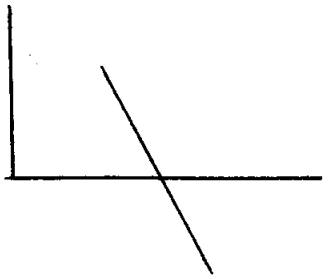
สูตร $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$r = \frac{\Sigma xy - \frac{(\Sigma x - \Sigma y)}{N}}{\sqrt{\left[\Sigma x^2 - \frac{(\Sigma x^2)}{N}\right]\left[y^2 - \frac{(\Sigma y^2)}{N}\right]}}$$

ถ้าค่า regression เป็น + (Positive)
เส้น regression จะเป็น



ถ้าค่า regression เป็น - (Negative)
เส้น regression จะเป็น



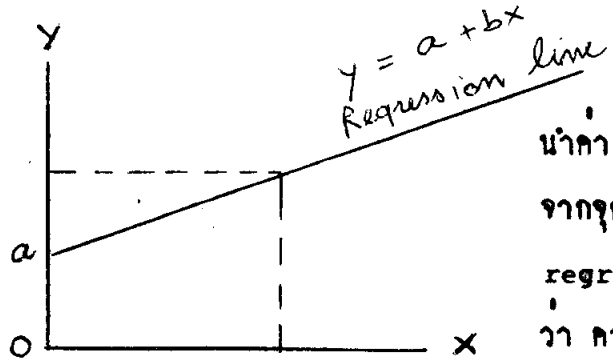
## Regression Line

Regression line $y = a+bx$

ซึ่งมีลักษณะเป็นสมการเส้นตรง

เป็นเส้นตรงแสดงความสัมพันธ์ระหว่างตัวแปร X และ Y

ซึ่งอาจมีทิศทางบวกหรือลบก็ได้จากสมการนี้จะใช้คาดคะเนหรือทำนาย (Predict) ค่าของ Y (dependent variable) จากค่าของX (independent variable) กล่าวคือใช้ทำนายค่าของตัวแปรตาม โดยดูจากตัวแปรอิสระ




นำค่า Y มา plot graph จากจุดที่ plot อยู่ใกล้ ๆ regression line แสดงว่า คาดคะเนค่า Y ได้ถูกต้อง

จาก $y = a + bx$

a คือจุด intercept จุดตัดบนแกน y

b คือ slope (ความเอียงของเส้น,ความลาด)

$a = \bar{y} - b\bar{x}$

$b = \frac{\Sigma xy}{\Sigma x^2}$

$x^2$ = ค่า $(x - \bar{x})^2$

กรณี $b = 0$ การทำนายใช้ไม่ได้

b ใกล้ 1 การทำนายใกล้ความถูกต้อง

$b = 1$ การทำนายถูกต้องสมบูรณ์

(ถ้า $b = 0$, $y$ = mean)