บทที่ 1
ความหมายของสถิติ และการนำไปใช้ประโยชน์

ความหมายของสถิติ

สถิติเป็นกระบวนการที่เกิดขึ้นเพื่อแก้ปัญหาเกี่ยวกับปรากฏการณ์ธรรมชาติทางวิทยาศาสตร์ ทางสังคม การบริหารธุรกิจ และการปกครอง จัดเป็นเทคนิคที่ใช้แก้ปัญหาโดยการวินิจฉัยข้อมูล ช่วยให้เกิดการตัดสินใจจากพื้นฐานความไม่แน่นอน เป็นวิธีทางวิทยาศาสตร์ที่ช่วยในการตัดสินใจในรูปของตัวเลข การรวบรวมข้อความจริงมาศึกษา จึงกล่าวได้ว่าสถิติเป็นเทคนิคเกี่ยวกับตัวอย่าง ปัจจุบันวิทยาศาสตร์ทุกสาขาคือสถิติ นั่นคือต้องมีการทดลองทางวิทยาศาสตร์ทุกแขนง ผลที่ได้จัดเป็นเพียงสถิติที่ใช้อ้างอิง ปัจจุบันกระบวนการทางวิทยาศาสตร์ใช้วิธีการสรุปผลงานที่ได้จากการทดลองในรูปของตัวอย่างและนำมากล่าวโดยอ้างหลักของความน่าจะเป็นหรือโอกาส (probability) โดยใช้หลักการตัดสินใจทางสถิติ (statistical decision)

คำว่า สถิติ ตรงกับภาษาอังกฤษว่า Statistics ซึ่งมีรากศัพท์มาจากคำว่า State ความหมายเดิมจึงหมายถึง ข้อมูล (data) หรือข่าวสาร (information) ที่เป็นประโยชน์แก่รัฐหรือประเทศในด้านต่างๆ เช่น ข้อมูลในการบริหารงานหรือวางแผนเกี่ยวกับกำลังคน การเก็บภาษีอากรเพื่อเป็นรายได้ของรัฐ การเกณฑ์ทหารเพื่อเข้าประจำการรักษาความปลอดภัยและป้องกันประเทศ การจัดการศึกษา การประกันสังคม และการสาธารณสุข เป็นต้น ต่อมาความหมายของคำว่า สถิติ หมายรวมถึงการค้นคว้าและพัฒนาในด้านเนื้อหาและวิธีการของนักคณิตศาสตร์และนักสถิติจำนวนมาก จึงอาจสรุปความหมายของสถิติหมายถึง ตัวเลขหรือข้อความจริงต่างๆ ที่จดบันทึกไว้เป็นหลักฐานอาจเป็นตัวเลขที่ใช้บรรยายเหตุการณ์หรือข้อเท็จจริงของเรื่องต่างๆ ที่เราต้องการศึกษา เช่น สถิติจำนวนผู้ป่วย สถิติจำนวนคนเกิด สถิติจำนวนคนตาย เป็นต้น หรือกระบวนการที่เกี่ยวข้องกับตัวเลขและข้อความจริงที่ถูกสรุปและตีความโดยกระบวนการทางสถิติ

จากความหมายเหล่านี้สามารถอธิบายความหมายของสถิติศาสตร์ คือ ศาสตร์ที่ว่าด้วยระเบียบวิธีการทางสถิติ (statistical method) ซึ่งประกอบด้วย การเก็บรวบรวมข้อมูล (collection of data) การนำเสนอข้อมูล (presentation of data) การวิเคราะห์ข้อมูล (analysis of data) และการตีความหมายข้อมูล (interpretation of data) สถิติในความหมายนี้เป็นเครื่องมือ (tool) ที่สำคัญที่สุดของการวิจัย การประเมินผล และการบริหารของวิชาการทุกสาขาทำให้ขอบข่ายของสถิติมีความหมายกว้าง และเกี่ยวข้องกับศาสตร์แขนงอื่นๆ เช่น วิศวกรรมศาสตร์ เศรษฐศาสตร์ บริหารธุรกิจ จิตวิทยา ศึกษาศาสตร์ มนุษยศาสตร์ เป็นต้น

ประโยชน์ของสถิติ

สถิติเป็นศาสตร์ที่ใช้เป็นเครื่องมือช่วยในการตัดสินใจอย่างมีเหตุผล การแก้ปัญหาส่วนใหญ่มีความจำเป็นต้องใช้ข้อมูลสารสนเทศและกระบวนการทางสถิติมาช่วยในการสรุปผลปัญหาหรืองานต่างๆ ในชีวิตประจำวันทั้งในวงการธุรกิจและราชการต้องใช้สถิติมาช่วยในการตัดสินใจและวางแผนในกรณีต่อไปนี้

1) การพยากรณ์อากาศของกรมอุตุนิยมวิทยา ต้องอาศัยวิธีวิเคราะห์ทางสถิติช่วยในการสรุปผลการพยากรณ์อากาศแต่ละช่วงเวลาประชาชนทั่วไปนำประกาศของกรมอุตุนิยมวิทยา ใช้ประกอบการตัดสินใจก่อนออกเดินทางไกลหรือใกล้ในแต่ละครั้ง

2) การตรวจทดสอบประสิทธิผลของผลิตภัณฑ์ต่างๆ ที่ผลิตได้ เพื่อให้มีคุณภาพตามเกณฑ์มาตรฐานที่กำหนด

3) การทำโพลเพื่อสำรวจความคิดเห็นของประชาชนที่มีต่อเรื่องใดเรื่องหนึ่ง โดยเฉพาะอย่างยิ่งเป็นกรณีที่เป็นที่สนใจของสังคม

4) ข้อมูลทางสถิติทำให้ทราบสถานการณ์ต่างๆ ในปัจจุบัน

5) ข้อมูลทางสถิติทำให้ทราบจุดเด่น-จุดด้อยของงาน ทำให้สามารถปรับปรุงและพัฒนางานให้มีคุณภาพตรงตามวัตถุประสงค์ วัตถุประสงค์ที่สำคัญของการติดตามผลโครงการคือ

- เพื่อรายงานความก้าวหน้า ปัญหาและอุปสรรคในการดำเนินงาน

- เพื่อชี้ประเด็นของปัญหา ให้ข้อเสนอแนะและแนวทางในการแก้ไขปัญหา

- เพื่อนำข้อมูลไปปรับปรุงแผนการดำเนินงานของโครงการในระยะต่อไปหรือเพื่อเป็นแนวทางใน การจัดทำแผนปฏิบัติงานของโครงการอื่นๆ

- เพื่อให้ผู้ปฏิบัติงานของโครงการหรือคณะทำงานมีความกระตือรือร้นในการปฏิบัติงาน

6) ข้อมูลทางสถิติทำให้สามารถคาดคะเนเหตุการณ์ในอนาคตได้อย่างถูกต้อง หรือใกล้เคียงความเป็นจริงมากที่สุด

7) การจัดทำแผนพัฒนาเศรษฐกิจและสังคม โดยอาศัยข้อมูลสถิติเป็นพื้นฐานในการจัดทำแผน การกำหนดเป้าหมาย และทิศทางของการพัฒนา เช่น การกำหนดหรือการวางนโยบายเกี่ยวกับการศึกษาภาคบังคับ การวางนโยบายเกี่ยวกับงบประมาณแผ่นดิน การวางนโยบายเกี่ยวกับการค้าทั้งในประเทศและนอกประเทศ อัตราค่าจ้างแรงงาน การเก็บภาษีอากร เป็นต้น ในช่วงภาวะวิกฤติเศรษฐกิจเช่นในปัจจุบันนี้ ข้อมูลสถิติเป็นสิ่งที่มีความจำเป็นอย่างยิ่งต่อการกำหนดนโยบาย และแก้ไขปัญหาต่างๆ ของรัฐบาล โดยเฉพาะใช้เป็นเครื่องเตือนภัยล่วงหน้า เพื่อกำหนดนโยบายหรือแผนงานต่างๆ ให้สอดคล้องกับภาวะเศรษฐกิจ

ตัวอย่างการใช้ข้อมูลสถิติสำหรับการพัฒนาในด้านต่างๆ ที่สำคัญในภาครัฐ

- ด้านการศึกษา ในการกำหนดนโยบายและการวางแผนพัฒนาการศึกษาและการกระจายโอกาสทางการศึกษาของประชาชนในระดับการศึกษาต่างๆ ข้อมูลสำคัญที่ต้องการใช้ ได้แก่ ประชากรก่อนวัยเรียนและวัยเรียน บุคลากรทางการศึกษา ปริมาณการผลิตและพัฒนาครูในแต่ละสาขา จำนวนสถานศึกษา ค่าใช้จ่ายในแต่ละระดับการศึกษา เป็นต้น

- ด้านการเกษตร ในการกำหนดนโยบายและวางแผนพัฒนาทางการเกษตรของประเทศ ข้อมูลที่ต้องการใช้ ได้แก่ ครัวเรือนที่ทำการเกษตร เนื้อที่การเพาะปลูก ผลิตผลทางการเกษตร จำนวนปศุสัตว์ ราคาสินค้าเกษตรกรรม เครื่องมือเครื่องใช้ทางการเกษตร ภาวะเศรษฐกิจและสังคมของครัวเรือนเกษตร การประมง การป่าไม้ ข้อมูลเกี่ยวกับแหล่งน้ำ และการชลประทาน เป็นต้น

- ด้านอุตสาหกรรม ใช้จัดทำแผนงานหรือกำหนดนโยบายและส่งเสริมอุตสาหกรรม ส่งเสริมการลงทุนและพัฒนาเทคโนโลยีทางด้านอุตสาหกรรม ได้แก่ ข้อมูลเกี่ยวกับปริมาณการผลิตทางอุตสาหกรรม ต้นทุนการผลิต จำนวนแรงงาน ค่าใช้จ่ายของสถานประกอบการ มูลค่าเพิ่ม ฯลฯ

- ด้านรายรับ - รายจ่ายของครัวเรือน เป็นข้อมูลที่มีความสำคัญที่ใช้วัดความเจริญเติบโตทางเศรษฐกิจ การครองชีพและการกระจายรายได้ของประชากร ซึ่งเป็นเครื่องชี้วัดที่สำคัญของผลการพัฒนาประเทศ ข้อมูลสถิติที่สำคัญ ได้แก่ รายได้รายจ่ายของครัวเรือน ภาวะหนี้สิน สภาพความเป็นอยู่ ที่อยู่อาศัยของครัวเรือน เป็นต้น

- ด้านสาธารณสุข การจัดทำแผนพัฒนาด้านสาธารณสุข การพัฒนางานวิชาการทางการแพทย์/สาธารณสุข เพื่อให้ประชาชนมีสุขภาพอนามัยที่ดี จำเป็นต้องใช้สถิติเกี่ยวกับการเกิด การตาย การเจ็บป่วยของประชาชน การรักษาพยาบาล ความเป็นอยู่และสภาพทางสังคมของประชากร การอนามัยและสุขาภิบาล พฤติกรรมด้านการบริโภค การสูบบุหรี่และดื่มสุรา เป็นต้น

- ด้านคมนาคมและขนส่ง การปรับปรุงบริการและพัฒนาทางการคมนาคมขนส่ง และการสื่อสารของประเทศ เพื่อสนับสนุนการพัฒนาด้านต่างๆ และกระจายความเจริญไปสู่ภูมิภาค ข้อมูลที่ใช้ได้แก่ รายรับ - รายจ่ายของการประกอบการขนส่ง ปริมาณผู้ใช้บริการในแต่ละเส้นทาง ปริมาณการขนส่งทางถนน ทางน้ำ และทางอากาศ รายละเอียดเส้นทางคมนาคม ข้อมูลเกี่ยวกับการจัดสรรความถี่วิทยุ จำนวนครัวเรือนที่มีเครื่องรับวิทยุ โทรทัศน์ เป็นต้น

นิยามของคำศัพท์ต่างๆ

1) ประชากรและพารามิเตอร์

ประชากร (population) หมายถึง ทุกหน่วยในเรื่องที่สนใจศึกษา หน่วยต่างๆ ในประชากรอาจเป็นบุคคล กลุ่มบุคคล องค์กรต่างๆ สัตว์ สิ่งของ ฯลฯ เช่น ต้องการศึกษาเกี่ยวกับรายได้เฉลี่ยของคนขับรถแท็กซี่

ในกรุงเทพมหานคร คนขับรถแท็กซี่ในกรุงเทพฯ ทั้งหมดคือ ประชากร จำนวนคนขับรถแท็กซี่ทั้งหมดก็คือขนาดประชากร (population size) สนใจอายุเฉลี่ยของนักศึกษามหาวิทยาลัย ประชากรคือ นักศึกษามหาวิทยาลัยทุกคน

พารามิเตอร์ (parameters) หมายถึง ค่าคงที่ที่ใช้บรรยายลักษณะของประชากรที่ประมวลได้จากประชากรทั้งหมดโดยวิธีการทางสถิติ สัญลักษณ์ที่ใช้ คือ

μ แทนค่าเฉลี่ยของประชากร

σ แทนค่าส่วนเบี่ยงเบนมาตรฐานของประชากร

ρ แทนค่าสัมประสิทธิ์สหสัมพันธ์ของประชากร

τ แทนยอดรวมของประชากร

P แทนสัดส่วนของประชากร (หรืออาจใช้สัญลักษณ์ π)

2) ตัวอย่างและค่าสถิติ

ตัวอย่าง (sample) หมายถึง บางส่วนของประชากรที่ถูกเลือกมาศึกษาอย่างมีขั้นตอนและมีประสิทธิภาพด้วยทฤษฎีการเลือกตัวอย่าง จำนวนทั้งหมดที่เราเลือกเรียกว่า ขนาดตัวอย่าง (sample size) เช่น สนใจอายุการใช้งานเฉลี่ยของหลอดไฟยี่ห้อ A ประชากรคือ หลอดไฟยี่ห้อ A ทุกหลอด ตัวอย่างคือ หลอดไฟยี่ห้อ A บางหลอดที่ถูกเลือกเป็นตัวอย่าง

ค่าสถิติ หมายถึง ค่าที่ใช้บรรยายลักษณะของตัวอย่างที่ประมวลได้จากตัวอย่าง โดยวิธีทางสถิติ สัญลักษณ์ที่ใช้ คือ

$\bar{x}$ แทนค่าเฉลี่ยของตัวอย่าง x

S แทนค่าส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง

r แทนค่าสัมประสิทธิ์สหสัมพันธ์ของตัวอย่าง

3) ข้อมูล (data) หมายถึง ตัวเลขหรือข้อความที่ได้จากการรวบรวมขึ้นเพื่อศึกษาตามลักษณะของสิ่งที่สนใจ เช่น ปริมาณการส่งข้าวสารไปจำหน่ายยังต่างประเทศในรอบ 6 เดือนที่ผ่านมา เป็นข้อมูลที่เป็นตัวเลข อัตราดอกเบี้ยเงินกู้ของธนาคารพาณิชย์มีแนวโน้มว่าจะเพิ่มขึ้นในอนาคต คุณภาพของสินค้าดี มีตำหนิเล็กน้อย มีตำหนิมาก ซึ่งเป็นข้อมูลที่เป็นข้อความ เป็นต้น

4) ตัวแปร (variable) หมายถึง ลักษณะของสิ่งที่สนใจ เมื่อชั่ง ตวง วัดหรือตรวจสอบแล้วมีค่าต่างๆ กันในแต่ละหน่วยของประชากรหรือตัวอย่าง เช่น เงินเดือน อายุ คะแนนสอบ น้ำหนัก เพศ ศาสนา ความคิดเห็น เป็นต้น

5) ค่าสังเกต (observed value) เป็นข้อมูลหรือผลการทดลองต่างๆ ที่บันทึกด้วยตัวเลข เช่น เพศของทารกอาจบันทึกเป็น 0 = ชาย และ 2 = หญิง ผลผลิตข้าว 1 แปลงทดลองก็คือ 1 ค่าสังเกต ความสูงของนักศึกษาค่าสังเกตที่ได้ของแต่ละคนอาจวัดเป็นเซนติเมตร

บทที่ 2

ระเบียบวิธีทางสถิติ

ระเบียบวิธีทางสถิติแบ่งออกเป็น 4 ขั้นตอน คือ การเก็บและรวบรวมข้อมูล (collection of data) การนำเสนอข้อมูล (presentation of data) การวิเคราะห์ข้อมูล (analysis of data) และการตีความหมายหรือหาข้อสรุปของข้อมูล (interpretation of data) ซึ่งมีรายละเอียดดังต่อไปนี้

1. การเก็บและรวบรวมข้อมูล (collection of data) เป็นการรวบรวมข่าวสารข้อมูลหรือข้อเท็จจริงที่ต้องการจากประชากรที่มีคุณสมบัติที่สอดคล้องตามความต้องการ การเก็บและรวบรวมข้อมูลนี้จัดว่าเป็นขั้นตอนที่สำคัญที่สุดในระเบียบวิธีการทางสถิติ เพราะว่าการเก็บรวบรวมข้อมูลที่มีความเชื่อถือได้น้อยจะทำให้ผลที่ได้จากการวิเคราะห์และตีความออกมานั้นมีความเชื่อถือได้ในระดับต่ำ ดังนั้นขั้นตอนนี้ต้องมีการวางแผนในการรวบรวมข้อมูล มีการควบคุมขั้นตอนการเก็บและต้องมีการตรวจสอบข้อมูลให้ละเอียดก่อนว่าสามารถนำไปวิเคราะห์ได้หรือไม่ การเก็บและรวบรวมข้อมูลที่ใช้ในการศึกษาแบ่งออกเป็น 3 วิธี คือ

1. การเก็บรวบรวมข้อมูลจากงานทะเบียนหรือการบันทึก
2. การเก็บรวบรวมข้อมูลจากการสำรวจ ได้แก่

- การสำมะโน (census) เป็นการเก็บรวบรวมข้อมูลจากทุกๆ หน่วย ในประชากรที่ศึกษา
- การสำรวจตัวอย่าง (sample survey) เป็นการเก็บรวบรวมข้อมูลจากพียงบางหน่วยของประชากร เพื่อที่จะได้ตัวแทนที่ดีของประชากรเป็นการประหยัดเวลาและค่าใช้จ่าย แต่ถ้าตัวอย่างที่เลือกมาไม่เป็นตัวแทนที่ดีหรือมีขนาดน้อยเกินไปก็จะเกิดความคลาดเคลื่อนได้
- การลงทะเบียน (registration) เป็นวิธีหนึ่งของการเก็บข้อมูลที่มีลักษณะต่อเนื่องมีประโยชน์ คือ ทำให้ทราบถึงการเคลื่อนไหวและการเปลี่ยนแปลงของประชากรตลอดเวลา เช่น ทะเบียนบ้าน การลงทะเบียนของนักศึกษา การขึ้นทะเบียนทหารกองเกิน เวชระเบียนของโรงพยาบาล
- การทดลอง (experiment) เป็นการเก็บข้อมูลที่สามารถป้องกันควบคุมปัจจัยอื่นๆ ที่จะเข้ามาแทรกซ้อนต่อเรื่องที่ศึกษาได้ เช่น เราต้องการศึกษาปริมาณผลผลิตต่อไร่ของข้าว 4 พันธุ์ เราก็สามารถควบคุมปริมาณปุ๋ย ปริมาณน้ำ และสภาพแวดล้อมให้เท่าๆ กัน หรือไม่แตกต่างกันได้ เป็นต้น

3. การเก็บรวบรวมข้อมูลจากการทดลองสามารถทำได้หลายวิธี เช่น การสัมภาษณ์ (interview) การส่งไปรษณีย์ (mail) การทอดแบบ การตอบแบบสอบถาม โทรศัพท์ การชั่ง ตวง วัด นับ และการสังเกต

2. การนำเสนอข้อมูล (data presentation) เป็นการนำเสนอข้อมูลสถิติที่ได้รวบรวมไว้นำออกเผยแพร่ให้คนทั่วๆ ไปเข้าใจและเป็นการเตรียมพร้อมข้อมูลเพื่อการวิเคราะห์ต่อไป วิธีการนำเสนอข้อมูลมีหลายแบบแล้วแต่ความเหมาะสมกับชนิดของข้อมูลและปริมาณของข้อมูล โดยทั่วๆ ไปแบ่งเป็น 2 ลักษณะคือ

2.1 การนำเสนอข้อมูลแบบไม่มีแบบแผน การนำเสนอข้อมูลในรูปแบบนี้จะไม่มีกฎเกณฑ์ใดๆ อาจนำเสนอในรูปของบทความหรือบทความกึ่งตาราง

- การนำเสนอข้อมูลในรูปบทความ (textual presentation) เป็นการนำเสนอเกี่ยวกับรายงานต่างๆ โดยนำเอาสถิติแทรกลงไป การนำเสนอแบบนี้มักจะพบเห็นในรายการทางโทรทัศน์ วิทยุ หนังสือพิมพ์

- การนำเสนอข้อมูลในรูปบทความกึ่งตาราง (semi tabular presentation) มีลักษณะเป็นบทความแต่นำเอาตัวเลขต่างๆ มาอ้างอิงประกอบมากขึ้นและนำมาจัดเป็นตารางเพื่อให้เห็นการเปรียบเทียบอย่างชัดเจน

2.2 การนำเสนอข้อมูลแบบมีแบบแผน เป็นการนำเสนอข้อมูลที่มีระเบียบแบบแผนและกฎเกณฑ์ ในรูปของตารางหรือกราฟโดยมีจุดมุ่งหมายเพื่อให้การนำเสนอนั้นง่ายและรัดกุมขึ้น ตลอดจนผู้อ่านก็สามารถเข้าใจได้ง่าย

- การนำเสนอข้อมูลด้วยตาราง (tabular presentation) เป็นการจัดข้อมูลให้อยู่ในรูปของตารางซึ่งประกอบด้วยส่วนต่างๆ ดังนี้คือ หมายเลขตาราง (table number) ชื่อเรื่อง (title) หมายเหตุคำนำ (head note)

- การนำเสนอข้อมูลด้วยแผนภูมิ (chart Presentation) เป็นการนำเสนอข้อมูลที่จะช่วยให้สามารถเห็นลักษณะเด่นของข้อมูลอย่างรวดเร็ว และชัดเจนง่ายต่อการเข้าใจ ดึงดูดความสนใจได้มากกว่าตาราง ทั้งยังสามารถเปรียบเทียบข้อมูลได้ง่าย แผนภูมิที่ใช้ทางสถิติ ได้แก่ กราฟเส้น กราฟแท่ง แผนภูมิภาพ แผนภูมิเชิงประกอบ แผนภูมิพื้นที่หรือปริมาณ แผนที่สถิติ ส่วนประกอบของแผนภูมิ ดังนี้

กราฟเส้น (Line Graph) เหมาะสำหรับข้อมูลที่มีความต่อเนื่องซึ่งส่วนมากใช้กับข้อมูลที่เป็นอนุกรมเวลา (time series)

กราฟแท่ง (bar chart) เหมาะสำหรับข้อมูลที่เป็นความถี่ (จำนวน) ประกอบด้วยแท่งสี่เหลี่ยมผืนผ้าที่มีความกว้างเท่ากันแสดงปริมาณของข้อมูลโดยความสูงของแท่งกราฟ อาจจะนำเสนอในแนวนอนหรือแนวตั้งก็ได้ กราฟแท่งจะใช้เปรียบเทียบข้อมูลที่ไม่ต้องการรายละเอียดมากนัก

กราฟวงกลม (Pie Chart) เหมาะสำหรับเปรียบเทียบข้อมูลที่เน้นให้เห็นถึงความแตกต่างของสัดส่วนอย่างรวดเร็ว โดยแบ่งวงกลมซึ่งมีมุม 360° เทียบกับ 100% แล้วแบ่งวงกลมออกตามสัดส่วนของข้อมูล

แผนภูมิรูปภาพ (Pictogram) เป็นการนำเสนอข้อมูลที่ต้องการดึงดูดความสนใจ โดยนำเสนอเป็นรูปภาพซึ่งอาจจะใช้ภาพที่มีขนาดเท่าๆ กัน เปรียบเทียบปริมาณของข้อมูลโดยจำนวนภาพหรืออาจจะใช้ภาพ ภาพเดียวเปรียบเทียบปริมาณของข้อมูลโดยใช้ขนาดของภาพให้มีขนาดแตกต่างกันตามปริมาณของข้อมูล

3. การวิเคราะห์ข้อมูล (analysis of data) เป็นการนำเอาข้อมูลที่รวบรวมได้มาประมวลผลตามวัตถุประสงค์ สมมติฐาน และคำถามการวิจัยที่ตั้งไว้ เช่น เปรียบเทียบความแตกต่างค่าเฉลี่ยของประชากร 2 กลุ่มโดยใช้ Z หรือ t เปรียบเทียบความแตกต่างระหว่างค่าเฉลี่ยของประชากรที่มากกว่า 2 กลุ่ม โดยใช้การทดสอบความ

แปรปรวนสถิติทดสอบคือ F ทดสอบความสัมพันธ์ระหว่างตัวแปรที่เป็นข้อมูลเชิงคุณภาพโดยใช้ X^2 ทดสอบความสัมพันธ์ระหว่างข้อมูล 2 ชุดที่เป็นข้อมูลเชิงปริมาณหรือทวิ โดยใช้การวิเคราะห์สหสัมพันธ์ ทดสอบอิทธิพลและพยากรณ์ โดยใช้การวิเคราะห์ความถดถอย หรืออาจจะใช้สถิติชั้นสูงระดับมัลติวาริเอท เช่น MANOVA CANONICAL FACTOR-ANALYSIS DISCREMINAN ก็ได้ การประมวลผลอาจจะประมวลด้วยมือหรือเครื่องคอมพิวเตอร์ก็ได้ ปัจจุบันนี้มีโปรแกรมสำเร็จรูปทางสถิติที่สามารถนำมาช่วยการวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพรวดเร็วและใช้ได้ทุกขั้นตอน เช่น โปรแกรมสำเร็จ SPSS for Windows, MINITAB, SAS เป็นต้น

4. การตีความหมายหรือหาข้อสรุปของข้อมูล เป็นการนำผลที่ได้จากการวิเคราะห์มาตีความสรุป เขียนเป็นรายงานผล เช่น t = 3.1 หมายความว่าอย่างไร มีความแตกต่างกันระหว่างค่าเฉลี่ยของประชากรสองกลุ่มหรือไม่ ค่า R = -0.85 หมายความว่ามีความสัมพันธ์กันอย่างไร มาก น้อย มีทิศทางอย่างไร ซึ่งต้องอาศัยความรู้ที่ต้องศึกษาต่อไป

บทที่ 3

ประเภทของข้อมูล ระดับการวัดข้อมูล และชนิดของสถิติ

ประชากร กลุ่มตัวอย่าง ตัวแปร

ประชากร (Population) หมายถึง กลุ่มสมาชิกทั้งหมดที่ต้องการศึกษาอาจเป็นสิ่งมีชีวิตหรือไม่มีชีวิตก็ได้ และประชากรในทางสถิติ หมายถึง บุคคล กลุ่มบุคคล องค์กรต่างๆ สัตว์ สิ่งของ เช่น การศึกษาอายุเฉลี่ยของคนไทย ประชากรคือคนไทยทุกคน สำหรับการแบ่งประเภทของประชากรสามารถแบ่งออกได้เป็น 2 ประเภท คือ 1) ประชากรที่นับได้ (Finite population) หมายถึง ประชากรที่มีจำนวนจำกัด สามารถนับได้ เช่น จำนวนหนังสือในห้องสมุด

2) ประชากรที่นับไม่ได้ (Infinite population) หมายถึง ประชากรที่มีจำนวนมากไม่สามารถนับได้ครบถ้วน หรือมีเกิดขึ้นใหม่อยู่เรื่อยๆ จนไม่ทราบจำนวนที่แน่นอน เช่น จำนวนเมล็ดข้าวเปลือกที่เก็บเกี่ยว

กลุ่มตัวอย่าง (Sample) หมายถึง ส่วนหนึ่งของประชากรที่นำมาศึกษา ซึ่งเป็นตัวแทนของประชากรของกลุ่มตัวอย่างและควรได้มาจากกลุ่มตัวแทนที่ดีของประชากรที่ศึกษา มีการเลือกตัวอย่างและขนาดตัวอย่างที่เหมาะสม เพื่อการอ้างอิงไปยังกลุ่มประชากรซึ่งต้องอาศัยสถิติเข้ามาช่วยในการสุ่มตัวอย่างและการกำหนดขนาดของกลุ่มตัวอย่าง

ตัวแปร คือ สิ่งที่โดยสภาพทั่วไปแล้วสามารถแปรค่าได้ค่าที่แปรออกมาของตัวแปรย่อมมีคุณสมบัติแตกต่างกันออกไป ตัวแปร คือ สิ่งที่ผู้วิจัยสนใจที่จะวัดเพื่อให้ได้ข้อมูลออกมาในรูปใดรูปหนึ่ง

ประเภทของตัวแปร การจำแนกประเภทของตัวแปรมี 4 ลักษณะคือ

1. พิจารณาคุณสมบัติของค่าที่แปรออกมาแบ่งเป็น 2 ชนิด คือ

- ตัวแปรเชิงปริมาณ (Quantitative Variables) เป็นตัวแปรที่ถูกวัดมามีค่าเป็นตัวเลข เช่น ส่วนสูง รายได้ ราคา

- ตัวแปรเชิงคุณภาพ (Qualitative Variables) เป็นตัวแปรที่ข้อมูลไม่ใช่ตัวเลข แต่เป็นข้อมูลที่มีลักษณะเป็นการแบ่งประเภทให้เห็นถึงความแตกต่างของกลุ่มตัวอย่างแต่ละกลุ่ม เช่น ระดับการศึกษา อาชีพ

2. พิจารณาความต่อเนื่องตามธรรมชาติของตัวแปรแบ่งเป็น 2 ชนิด คือ

- ตัวแปรค่าต่อเนื่อง (Continuous Variables) เป็นตัวแปรที่มีค่าต่อเนื่องกันตลอด เช่น ส่วนสูง น้ำหนัก คะแนนสอบ เป็นต้น

- ตัวแปรค่าไม่ต่อเนื่อง (Discrete Variables) ตัวแปรประเภทนี้มีค่าเฉพาะตัวแยกออกจากกันเด็ดขาด วัดค่าเป็นจำนวนเต็ม เช่น จำนวนหนังสือ

3. พิจารณาความเป็นไปได้ของผู้วิจัยที่จัดกระทำกับตัวแปรแบ่งเป็น 2 ชนิดคือ

- ตัวแปรที่กำหนดได้ (Active Variables) เป็นตัวแปรที่ผู้วิจัยสามารถกำหนดให้กับการทดลองได้ เช่น วิธีสอน การจัดสอนซ่อมเสริม การจัดสภาพห้องเรียน เป็นต้น

- ตัวแปรที่จัดกระทำขึ้นไม่ได้ (Attribute of Organismic Variables) เป็นตัวแปรที่ยากจะกำหนดให้ผู้รับการทดลองได้ ตัวแปรเหล่านี้เป็นลักษณะของผู้รับการทดลอง เช่น เพศ สภาพเศรษฐกิจ ความถนัด เป็นต้น

4. พิจารณาถึงความสัมพันธ์ระหว่างตัวแปรในเชิงเหตุผล เป็นการแบ่งตามลักษณะการใช้เป็นวิธีแบ่งตัวแปรที่นิยมกันมากที่สุดแบ่งเป็น

- ตัวแปรอิสระหรือตัวแปรต้น (Independent Variables) เป็นตัวแปรที่จะทำให้เกิดสิ่งอื่นตามมา เป็นตัวแปรที่เป็นเหตุตัวแปรที่มาก่อน

- ตัวแปรตาม (Dependent Variables) เป็นตัวแปรที่เป็นผลมาจากตัวแปรต้น โดยอาจกำหนดหรือจัดตัวแปรให้แตกต่างกัน ไม่มีอิสระในตัวเอง ต้องแปรเปลี่ยนไปตามเหตุการณ์หรือการทดลอง

- ตัวแปรแทรกซ้อน (Extraneous Variables) เป็นตัวแปรอื่นๆ ที่อาจมีผลต่อตัวแปรตาม โดยผู้วิจัยต้องพยายามควบคุมตัวแปรแทรกซ้อน เช่น ควบคุมด้วยการเลือกกลุ่มตัวอย่างควบคุมโดยวิธีการทางสถิติ

ชนิดของตัวแปร

1) ตัวแปรต้นหรือตัวแปรอิสระ คือ สิ่งที่ต้องจัดให้แตกต่างกัน ไม่ขึ้นอยู่กับสิ่งใด มีความเป็นอิสระในตัวเองคือสิ่งที่ต้องจัดให้แตกต่างกัน

2) ตัวแปรตาม คือ สิ่งที่ต้องติดตามดูผลจากการจัดสิ่งที่แตกต่างกัน ไม่มีอิสระในตัวเอง ต้องแปรเปลี่ยนไปตามเหตุการณ์หรือการทดลอง

3) ตัวแปรควบคุม คือ สิ่งที่ต้องจัดให้เหมือนกัน เป็นการควบคุมเพื่อให้แน่ใจว่าผลการทดลองเกิดจากตัวแปรต้นอย่างแท้จริง

ระดับการวัด

ระดับการวัดของตัวแปร เป็นการจัดเรียงลำดับของตัวแปร โดยสามารถแบ่งระดับในการวัดได้เป็น 4 ระดับ ได้แก่

1. มาตรนามบัญญัติ (Nominal Scale) ลักษณะเด่นของมาตรานี้คือ เป็นตัวแปรที่ถูกจัดเป็นกลุ่มๆ โดยที่ตัวแปรนี้ไม่สามารถจัดลำดับก่อนหลัง หรือบอกระยะห่างได้ เช่น เพศ แบ่งออกเป็นกลุ่มเพศชาย และกลุ่มเพศหญิง โดยอาจให้เลข 1 แทนเพศชาย และเลข 2 แทนเพศหญิง เป็นต้น ซึ่งตัวเลข 1, 2 ที่ใช้แทนกลุ่มต่างๆ ถือเป็นตัวเลขในระดับนามบัญญัติไม่สามารถนำมาบวก ลบ คูณ หาร หรือหาสัดส่วนได้

2. มาตราจัดลำดับ (Ordinal Scale) ลักษณะของมาตรานี้ มีลักษณะคล้ายกับมาตรานามบัญญัติ คือ สามารถจัดเป็นกลุ่มๆ ได้ และไม่สามารถบอกระยะห่างระหว่างกลุ่มได้เช่นเดียวกับมาตรานามบัญญัติ แต่มาตราจัดลำดับสามารถจัดลำดับก่อนหลังของตัวแปรได้ เช่น ก สอบได้ที่ 1 ข สอบได้ที่ 2 ค สอบได้ที่ 3 ซึ่งตัวเลข 1, 2, 3 เป็นตัวเลขในระดับอันดับที่สามารถนำมาบวก ลบกันได้

3. มาตราอันตรภาคชั้น (Interval Scale) คุณลักษณะของมาตรานี้สามารถแบ่งตัวแปรออกเป็นกลุ่มๆ ได้ จัดลำดับก่อนหลังของตัวแปรได้ อีกทั้งมีระยะห่างของช่วงการวัดที่เท่ากัน และที่สำคัญมาตรานี้เป็นมาตรการวัดที่ไม่มีศูนย์แท้ (Absolute Zero) กล่าวคือ ศูนย์ของมาตรานี้ไม่ได้หมายความว่าไม่มี แต่เป็นศูนย์ที่เกิดจากการสมมติขึ้น เช่น ผลคะแนนสอบวิชาสถิติของนาย ก สอบได้ 0 คะแนน มิได้หมายความว่าเขาไม่มีความรู้ เพียงแต่เขาไม่สามารถทำข้อสอบซึ่งเป็นตัวแทนของความรู้ทั้งหมดได้ ระดับนี้สามารถนำตัวเลขมาบวก ลบ คูณ หาร กันได้

4. มาตราอัตราส่วน (Ratio Scale) มาตรานี้เป็นมาตราที่มีลักษณะเหมือนกับมาตราอันตรภาคชั้นทุกประการ แต่สิ่งที่แตกต่างกันในมาตรานี้คือเป็นมาตราที่มีศูนย์แท้ (Absolute Zero) นั่นหมายความว่า ผลที่ได้จากการวัดในมาตรานี้หากเท่ากับศูนย์แสดงว่าไม่มีอย่างแท้จริง และสามารถนำตัวเลขมาบวก ลบ คูณ หาร หรือหาอัตราส่วนกันได้เช่น น้ำหนัก ความสูง อายุ เป็นต้น

ประเภทของสถิติ

สถิติแบ่งออกเป็น 2 ประเภทคือ

1. สถิติพรรณนา (Descriptive Statistics) เป็นสถิติที่ใช้อธิบายคุณลักษณะของสิ่งที่ต้องการศึกษากลุ่มใดกลุ่มหนึ่ง ไม่สามารถอ้างอิงไปยังกลุ่มอื่นๆ ได้ สถิติที่อยู่ในประเภทนี้ เช่น ค่าเฉลี่ย ค่ามัธยฐาน ค่าฐานนิยม ส่วนเบี่ยงเบนมาตรฐาน พิสัย ฯลฯ

1.1 ค่าเฉลี่ย คือ ค่ากลางของการแจกแจงของค่าของข้อมูล เป็นค่าที่ใช้มากที่สุดและมีประโยชน์มากของการวัดแนวโน้มเข้าสู่ส่วนกลาง เหมาะสำหรับข้อมูลแบบต่อเนื่องที่มีมาตราวัดอัตรภาคชั้น (interval scale) และมาตราอัตราส่วน (ratio scale)

1.2 ค่ามัธยฐาน คือ ค่าของข้อมูลที่ตำแหน่งกลางของการแจกแจงที่มีจำนวนความถี่ของข้อมูลที่มีค่ามากกว่าหรือน้อยกว่าจุดนี้เป็นจำนวนความถี่เท่ากับครึ่งหนึ่งของจำนวนข้อมูลทั้งหมดเป็นจุดที่แบ่งการแจกแจงออกเป็น 2 ส่วน คือ ด้านซ้าย และด้านขวาเท่าๆ กัน ค่ามัธยฐานเหมาะมากกับข้อมูลที่เป็นตัวเลขเชิงปริมาณที่มีการแจกแจงแบบเบ้

1.3 ค่าฐานนิยม คือ ค่าของข้อมูลที่มีความถี่มากที่สุดในการแจกแจงหนึ่ง เหมาะสำหรับข้อมูลที่แบ่งเป็นชั้น (categorical data) ซึ่งตัวแปรเป็นแบบแบ่งประเภท มีมาตราการวัดนามบัญญัติ (norminal scale) เป็นค่าที่หยาบที่สุดของการวัดแนวโน้มเข้าสู่ส่วนกลาง ถ้าวัดค่าฐานนิยมของข้อมูลที่เป็นแบบต่อเนื่อง

(continuous data) โดยการจัดกลุ่มข้อมูลนั้นให้เป็นกลุ่ม (grouped data) แบ่งเป็นชั้นๆ ที่มีความกว้างของชั้นไม่เท่ากัน อาจทำให้รูปทรงของการแจกแจงบิดเบือนไปจากที่เป็นจริงได้ ทำให้เห็นได้ชัดว่าต้องระมัดระวังในการตีความหมายข้อมูล

2 สถิติอนุมาน (Inferential Statistics) เป็นสถิติที่ใช้อธิบายคุณลักษณะของสิ่งที่ต้องการศึกษากลุ่มใดกลุ่มหนึ่งหรือหลายกลุ่มแล้วสามารถอ้างอิงไปยังกลุ่มประชากรได้ แต่กลุ่มตัวอย่างที่นำมาศึกษาจะต้องเป็นตัวแทนที่ดีของประชากรและได้มาโดยวิธีการสุ่มตัวอย่าง ซึ่งสถิติอ้างอิงสามารถแบ่งออกเป็น 2 ประเภท คือ

2.1 สถิติมีพารามิเตอร์ (Parametric Statistics) เป็นวิธีการทางสถิติที่จะต้องเป็นไปตามข้อตกลงเบื้องต้นดังนี้

- ข้อมูลต้องอยู่ในระดับช่วงขึ้นไป
- ข้อมูลต้องมีการแจกแจงเป็นโค้งปกติ
- กลุ่มประชากรแต่ละกลุ่มที่นำมาศึกษาต้องมีความแปรปรวนเท่ากัน

สถิติประเภทนี้เช่น t-test, Z-test, ANOVA, Regression ฯลฯ

2.2 สถิติไร้พารามิเตอร์ (Nonparametric Statistics) เป็นวิธีการทางสถิติที่สามารถนำมาใช้ได้โดยปราศจากข้อตกลงเบื้องต้น สถิติที่อยู่ในประเภทนี้ เช่น ไคสแควร์, Median test, Sign test ฯลฯ โดยปกติแล้วนักวิจัยนิยมใช้สถิติมีพารามิเตอร์ทั้งนี้เพราะผลลัพธ์ที่ได้มีอำนาจทดสอบ (Power of Test) สูงกว่าการใช้สถิติไร้พารามิเตอร์ ดังนั้นหากข้อมูลมีคุณสมบัติที่สอดคล้องกับข้อตกลงเบื้องต้นให้ใช้สถิติมีพารามิเตอร์

บทที่ 4

สถิติพรรณนา (Descriptive Statistics)

สถิติพรรณนา คือ สถิติที่เกี่ยวกับระเบียบวิธีหรือบรรยายถึงลักษณะของข้อมูลเฉพาะที่ได้มาจากการเก็บรวบรวมข้อมูล ซึ่งผลของการศึกษาจะบอกได้เฉพาะลักษณะของกลุ่มที่ศึกษาเท่านั้น ไม่สามารถนำผลไปอ้างอิงหรือพยากรณ์ค่าของกลุ่มอื่นๆ ได้ สถิติประเภทนี้ ส่วนใหญ่จะเป็นการคำนวณหาค่าการกระจายของข้อมูล การวัดแนวโน้มเข้าสู่ส่วนกลาง สถิติที่ใช้อธิบายคุณลักษณะของข้อมูลอาจจะเกี่ยวข้องกับวิธีการทางสถิติต่อไปนี้

1. การนำเสนอข้อมูล (Presentation)

- การนำเสนอในรูปบทความ

- การนำเสนอในรูปตารางเป็นร้อยละ (Percentage)

- การนำเสนอในรูปกราฟ เช่น กราฟแท่ง (Bar Graph) กราฟเส้น (Line Graph) กราฟวงกลม (Pie Graph)

2. การแจกแจงความถี่ (Frequency)

3. การวัดแนวโน้มเข้าสู่ส่วนกลาง ซึ่งประกอบด้วยสถิติต่อไปนี้

- ตัวกลางเลขคณิตหรือค่าเฉลี่ย (Arithmatic Mean or Average) การวัดค่าเฉลี่ยคือ ค่าที่ได้จากการนำข้อมูลทั้งหมดมารวมกัน แล้วหารด้วยจำนวนข้อมูลทั้งหมด

- ตัวกลางเรขาคณิต (Geometric Mean)

- ตัวกลางฮาร์โมนิก (Harmonic Mean)

- ฐานนิยม (Mode) คือ ค่าที่มีความถี่มากที่สุด

- มัธยฐาน (Median) คือ ค่าที่อยู่ตำแหน่งกลางของข้อมูล

- ควอไทล์ (Quartiles)

- เดไซล์ (Deciles)

- เปอร์เซ็นไทล์ (Percentiles)

4. การวัดการกระจายของข้อมูล ซึ่งประกอบด้วยสถิติดังต่อไปนี้

- พิสัย (Range)

- ส่วนเบี่ยงเบนควอไทล์ (Quartile Deviation)

- ส่วนเบี่ยงเบนเฉลี่ย (Mean Deviation)

- ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation)

- ค่าแปรปรวน (Variance)
- สัมประสิทธิ์การกระจาย (Coefficient of Variation)
- การวัดความเบ้ (Skewness)
- การวัดความโด่ง (Kurtosis)

5. การหาความสัมพันธ์ระหว่างตัวแปร มีค่าสถิติที่ใช้ดังนี้
   - สหสัมพันธ์ของเพียร์สัน (Pearson Correlation)
   - สหสัมพันธ์เชิงอันดับ (Spearman Rank Correlation)

บทที่ 5

การทดสอบสมมติฐานทางสถิติ

เป็นส่วนหนึ่งของสถิติอนุมาน ซึ่งเป็นการทดสอบเกี่ยวกับพารามิเตอร์ที่ไม่ทราบค่า โดยสุ่มตัวอย่างจากประชากรแล้ว อาศัยการแจกแจงของตัวสถิติ สร้างสถิติทดสอบเกี่ยวกับพารามิเตอร์นั้น ซึ่งเป็นการทดสอบค่าเฉลี่ยของประชากรหนึ่งชุด การทดสอบความแตกต่างของค่าเฉลี่ยของประชากรสองชุด การทดสอบความแปรปรวนของประชากร และการทดสอบความแปรปรวนในสองประชากร การทดสอบสมมติฐานจะเกี่ยวข้องกับ

1) สมมติฐาน คือ การตั้งข้อสมมติเกี่ยวกับค่าพารามิเตอร์ที่สงสัยว่าค่าพารามิเตอร์นั้น จะมีค่าตามที่ตั้งขึ้นจริงหรือไม่

2) สมมติฐานที่จะทดสอบ เรียกว่า สมมติฐานหลัก (Null Hypothesis) แทนด้วย H_0 สมมติฐานที่แย้งกับสมมติฐานหลัก และนำมาพิจารณาในการทดสอบด้วยเรียกว่า สมมติฐานแย้งหรือสมมติฐานรอง (Alternative Hypothesis) ซึ่งแทนด้วย H_1

3) บริเวณยอมรับ (Acceptance region) คือ บริเวณที่ทำให้เกิดการยอมรับ H_0 ส่วนบริเวณปฏิเสธ (Rejection region) หรือบริเวณวิกฤต (Critical region) คือ บริเวณที่ทำให้เกิดการปฏิเสธ H_0

4) ผลการตัดสินใจจากการทดสอบสมมติฐาน เนื่องจากสมมติฐานที่จะทดสอบ (H_0) เป็นความเชื่อหรือคำยืนยันเกี่ยวกับลักษณะของประชากรซึ่งยังไม่สามารถบอกได้ว่าเป็นจริงหรือเท็จ จนกว่าจะพิสูจน์โดยเก็บรวบรวมข้อมูลทั้งหมด มาวิเคราะห์ตามลักษณะของประชากรที่ต้องการพิสูจน์นั้น ซึ่งบางครั้งการเก็บรวมรวมข้อมูลทั้งหมดจากประชากรเป็นสิ่งที่ทำได้ยากเพราะต้องเสียค่าใช้จ่ายและเวลามาก จึงทำได้เพียงการสำรวจจากตัวอย่างเพื่อทดสอบเท่านั้นเอง ดังนั้นผลการตัดสินใจจากการทดสอบสมมติฐานใดๆ สามารถสรุปได้ดังตารางที่ 1

ตารางที่ 1 ผลการตัดสินใจจากการทดสอบสมมติฐาน

การตัดสินใจ	ข้อเท็จจริงของ H_0	
	H_0 เป็นจริง	H_0 ไม่เป็นจริง
ปฏิเสธ H_0	ความผิดพลาดประเภทที่ 1	ตัดสินใจถูก
ยอมรับ H_0	ตัดสินใจถูก	ความผิดพลาดประเภทที่ 2

ผลการทดสอบไม่ว่าจะยอมรับหรือปฏิเสธสมมติฐานหลัก ย่อมอาจมีความผิดพลาดเกิดขึ้นได้ 2 กรณีเสมอ คือ

1) การปฏิเสธ H_0 เมื่อ H_0 เป็นจริง เรียกว่า ความผิดพลาดประเภทที่ 1 (Type I error) ความน่าจะเป็นที่จะเกิดความผิดพลาดประเภทที่ 1 แทนด้วย α

2) การยอมรับ H_0 เมื่อ H_0 เป็นเท็จ เรียกว่า ความผิดพลาดประเภทที่ 2 (Type II error) ความน่าจะเป็นที่จะเกิดความผิดพลาดประเภทที่ 2 แทนด้วย β

การทดสอบสมมติฐานทางสถิติที่ดี คือ การทดสอบสมมติฐานที่ให้โอกาสที่จะเกิดความผิดพลาดประเภทที่ 1 และประเภทที่ 2 มีค่าต่ำสุด ดังนั้นต้องพยายามควบคุมความผิดพลาดทั้งสองประเภทให้มีโอกาสเกิดขึ้นน้อยที่สุด แต่ขนาดของความผิดพลาดสองประเภทนี้สวนทางกัน กล่าวคือ ถ้า α มีค่ามากแล้ว β จะมีค่าน้อย การควบคุมความผิดพลาดทั้งสองประเภทนี้สามารถลดลงได้ถ้าเพิ่มขนาดตัวอย่างให้มากขึ้น

การทดสอบทางเดียวและสองทาง

ในการทดสอบสมมติฐานใดๆ เราจะยอมรับว่าสมมติฐานหลักเป็นจริงก่อน แล้วจึงสุ่มตัวอย่างและคำนวณค่าสถิติที่ได้จากตัวอย่างสุ่ม ถ้าค่าสถิติที่ใช้ในการทดสอบนั้นแตกต่างจากพารามิเตอร์ที่กำหนดใน H_0 มากเพียงพอที่จะปฏิเสธ H_0 เราจึงจะปฏิเสธ H_0 หรือกล่าวว่าแตกต่างอย่างมีนัยสำคัญ เมื่อพิจารณาความแตกต่างดังกล่าว จะพบว่ามี 2 แบบคือ

1) แตกต่างอย่างมีทิศทาง คือ ค่าพารามิเตอร์ที่แท้จริงมากกว่าค่าพารามิเตอร์ที่กำหนดใน H_0 และอีกกรณีคือ ค่าพารามิเตอร์ที่แท้จริงน้อยกว่าค่าพารามิเตอร์ที่กำหนดใน H_0

2) แตกต่างแบบไม่มีทิศทาง คือ ค่าพารามิเตอร์ที่แท้จริงมีค่าไม่เท่ากับค่าพารามิเตอร์ที่กำหนดใน H_0 โดยความแตกต่างทั้ง 2 แบบนี้จะเขียนอยู่ในสมมติฐานแย้ง (H_1) ถ้าทดสอบสมมติฐานแบบมีทิศทางจะเรียกว่า การทดสอบแบบทางเดียว แต่ถ้าทดสอบสมมติฐานแบบไม่มีทิศทางจะเรียกว่า การทดสอบแบบสองทาง

1. การทดสอบแบบทางเดียว (One - Tailed Test)

ให้ θ เป็นพารามิเตอร์ที่ต้องการทดสอบ และให้ θ_0 เป็นค่าคงที่ที่ต้องการทดสอบหรือเป็นค่าพารามิเตอร์ที่คาดหวังไว้นั่นเอง สมมติฐานที่จะทดสอบอยู่ในลักษณะ

1) $H_0 : \theta = \theta_0$

$H_1 : \theta_0 > \theta$

เมื่อยอมรับว่า H_0 เป็นจริงก่อน บริเวณปฏิเสธ H_0 จะอยู่ปลายหางทางขวาของการแจกแจงของตัวสถิติที่ใช้ทดสอบ ($\hat{\theta}$)

รูปที่ 1 บริเวณวิกฤตของการทดสอบด้านขวา

2) $H_0: \theta_0 = \theta$

$H_1: \theta_0 < \theta$

เมื่อยอมรับว่า H_0 เป็นจริงก่อน บริเวณปฏิเสธ H_0 จะอยู่ปลายหางทางซ้ายของการแจกแจงของตัวสถิติที่ใช้ทดสอบ ($\hat{\theta}$)

รูปที่ 2 บริเวณวิกฤตของการทดสอบด้านซ้าย

2. การทดสอบแบบสองทาง (Two - Tailed Test)

ให้ θ เป็นพารามิเตอร์ที่ต้องการทดสอบ และให้ θ_0 เป็นค่าคงที่ที่ต้องการทดสอบหรือเป็นค่าพารามิเตอร์ ที่ คาดหวังไว้นั่นเอง สมมติฐานที่จะทดสอบอยู่ในลักษณะ

$H_0: \theta_0 = \theta$

$H_1: \theta_0 \neq \theta$

เมื่อยอมรับว่า H_0 เป็นจริงก่อน บริเวณปฏิเสธ H_0 จะอยู่ปลายหางทั้งสองข้างของการแจกแจงของตัวสถิติทดสอบ $\hat{\theta}$ ดังรูป

รูปที่ 3 บริเวณวิกฤตของการทดสอบแบบ 2 ทาง

ขั้นตอนของการทดสอบสมมติฐาน มีดังนี้

1. ตั้งสมมติฐานหลัก (H_0) และสมมติฐานทางเลือก (H_1) ให้มีความหมายตรงข้ามกันเสมอ

2. กำหนดระดับนัยสำคัญ α

3. เลือกตัวสถิติทดสอบที่เหมาะสมแล้วหาจุดวิกฤตเพื่อกำหนดบริเวณปฏิเสธ H_0 ให้สอดคล้องกับ H_1 และ α

4. คำนวณค่าสถิติที่ใช้ทดสอบจากตัวอย่างขนาด n ที่สุ่มมา

5. สรุปผลคือ ตัดสินใจยอมรับหรือปฏิเสธ H_0 โดยพิจารณาจากเงื่อนไขนี้ ถ้าค่าสถิติทดสอบที่คำนวณได้จากขั้นตอนที่ 4 ตกอยู่ในบริเวณยอมรับ เราจะตัดสินใจยอมรับ H_0 แต่หากตกอยู่ในเขตวิกฤตจะตัดสินใจปฏิเสธ H_0

การทดสอบสมมติฐานเกี่ยวกับค่าเฉลี่ยประชากรหนึ่งประชากร

เมื่อ μ คือค่าเฉลี่ยของประชากรและ μ_0 คือ ค่าคงที่ที่ต้องการทดสอบหรือเป็นค่าเฉลี่ยที่คาดว่าจะเป็น สมมติฐานที่จะทดสอบอยู่ในลักษณะ

1) $H_0 : \mu = \mu_0$ **แย้งกับ** $H_1 : \mu > \mu_0$ **หรือ**

2) $H_0 : \mu = \mu_0$ **แย้งกับ** $H_1 : \mu < \mu_0$ **หรือ**

3) $H_0 : \mu = \mu_0$ **แย้งกับ** $H_1 : \mu \neq \mu_0$

ตัวสถิติที่ใช้ในการทดสอบขึ้นอยู่กับลักษณะของประชากรและขนาดตัวอย่างสุ่ม ซึ่งแบ่งเป็น 3 กรณีคือ

1. ประชากรมีการแจกแจงแบบปกติ และทราบค่าความแปรปรวนภายใต้ H_0 **เป็นจริง**

ตัวสถิติที่ใช้ทดสอบคือ $Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

H_0	H_1	เขตวิกฤต
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$Z \geq z_\alpha$
$H_0 : \mu = \mu_0$	$H_1 : \mu < \mu_0$	$Z \leq -z_\alpha$
$H_0 : \mu = \mu_0$	$H1 : \mu \neq \mu_0$	$Z \geq z_{\frac{\alpha}{2}}$ หรือ $Z \leq -z_{\frac{\alpha}{2}}$

2. ประชากรมีการแจกแจงแบบใดๆ ไม่ทราบความแปรปรวนประชากรแต่ตัวอย่างมีขนาดใหญ่ (n ≥ 30)

เราประมาณความแปรปรวน σ^2 **ด้วยความแปรปรวนตัวอย่าง** S^2 **ดังนั้น ภายใต้** H_0 **เป็นจริง**

ตัวสถิติที่ใช้ทดสอบคือ $Z = \frac{\bar{x}-\mu_0}{\frac{S}{\sqrt{n}}}$

ซึ่งเกณฑ์ในการตัดสินใจใช้เกณฑ์เดียวกับกรณีที่ 1

3. ประชากรมีการแจกแจงแบบปกติ ไม่ทราบความแปรปรวนประชากร แต่ตัวอย่างมีขนาดเล็ก (n < 30) ภายใต้ H_0 **เป็นจริง**

ตัวสถิติที่ใช้ทดสอบคือ $T = \frac{\bar{x}-\mu_0}{\frac{S}{\sqrt{n}}}$ **เมื่อ** $df = n-1$

เกณฑ์การตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

H_0	H_1	เขตวิกฤต
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$T \geq t_\alpha$
$H_0 : \mu = \mu_0$	$H_1 : \mu < \mu_0$	
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$T \geq t_{\frac{\alpha}{2}}$ หรือ $T \leq -t_{\frac{\alpha}{2}}$

ตัวอย่างที่ 1 บริษัทผลิตเชือกแห่งหนึ่งประดิษฐ์เชือกใยสังเคราะห์ชนิดใหม่และอ้างอิงว่าทนแรงดึง เฉลี่ยได้ 15 ปอนด์ ค่าเบี่ยงเบนมาตรฐานการผลิตเท่ากับ 0.05 ปอนด์ ถ้าสุ่มเลือกเชือกใยสังเคราะห์ชนิดใหม่นี้มาทดสอบจำนวน 50 เส้น พบว่าทดแรงดึงเฉลี่ยได้ 14.8 ปอนด์ จงทดสอบสมมติฐานว่าเชือกใยสังเคราะห์ชนิดใหม่นี้ได้มาตรฐานดังกล่าวอ้างที่ระดับนัยสำคัญ 0.01

ให้ μ คือ ค่าเฉลี่ยที่เชือกทนแรงดึงได้ (ปอนด์)

1. ตั้งสมมติฐาน $H_0 : \mu = 15$

$H_1 : \mu \neq 15$

2. ระดับนัยสำคัญ (α) = 0.01 และ $\frac{\alpha}{2} = 0.005$

3. บริเวณปฏิเสธ H_0 คือ $Z \geq 2.58$ หรือ $Z \leq -2.58$

4. ตัวสถิติที่ใช้ทดสอบคือ $Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

$$= \frac{14.8-15}{\frac{0.05}{\sqrt{50}}} = -2.822$$

5. เพราะว่า $Z = -2.822$ ตกอยู่บริเวณวิกฤต ดังนั้นจึงตัดสินใจ ปฏิเสธ H_0 (ยอมรับ H_1) สรุปผลว่า เชือก ใยสังเคราะห์ที่บริษัทผลิตโดยเฉลี่ยไม่ทนแรงดึง 15 ปอนด์ได้ตามที่อ้าง

ตัวอย่างที่ 2 ในการทดลองอัตราการสิ้นเปลืองน้ำมันเชื้อเพลิงของรถยนต์ยี่ห้อหนึ่ง โดยการให้คนขับรถ 6 คน น้ำมันเชื้อเพลิง 1 ลิตร คนขับแต่ละคนขับได้ระยะทางเป็น 7.2, 9.3, 11.5, 8.7, 10.2 และ 9.6 กิโลเมตร ผู้ผลิตรถยนต์ต้องการโฆษณาว่ารถยี่ห้อนี้มีอัตราการสิ้นเปลืองเชื้อเพลิงได้อย่างน้อยที่สุด 10 กิโลเมตรต่อลิตร จงทดสอบว่าคำโฆษณานี้จริงหรือไม่ ที่ระดับนัยสำคัญ 0.05

1. ตั้งสมมติฐาน $H_0 : \mu = 10$

$H_1 : \mu < 10$

2. ระดับนัยสำคัญ (α) = 0.05

3. บริเวณปฏิเสธ H_0 คือ $T \leq -t_\alpha$ หรือ $T \leq -2.015$

4. ตัวสถิติที่ใช้ทดสอบคือ $T = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$

คำนวณค่า $\sum x_i = 56.5$ $\sum x_i^2 = 542.47$ $\bar{x} = 9.42$

$s^2 = \frac{542.47-(56.5)^2/6}{5} = 2.085$ $s = 1.44$

ดังนั้นค่า $T = \frac{9.42-10}{\frac{1.44}{\sqrt{6}}} = -0.99$

5. เพราะว่า T = -0.99 ตกอยู่บริเวณยอมรับ ดังนั้นจึงตัดสินใจยอมรับ H_0 สรุปว่า รถยนต์ยี่ห้อนี้มีอัตราสิ้นเปลืองน้ำมันเชื้อเพลิงไม่น้อยกว่า 10 กิโลเมตรต่อลิตร

การทดสอบสมมติฐานเกี่ยวกับผลต่างค่าเฉลี่ยของสองประชากรที่อิสระกัน

เมื่อ μ_1, μ_2 คือ ค่าเฉลี่ยของประชากรที่ 1 และ 2 ตามลำดับ d_0 คือ ค่าผลต่างของค่าเฉลี่ยของสองประชากรดังกล่าวสมมติฐานที่จะทดสอบอยู่ในลักษณะ

1) $H_0 : \mu_1 - \mu_2 = d_0$ แย้งกับ $H_1 : \mu_1 - \mu_2 > d_0$ หรือ

2) $H_0 : \mu_1 - \mu_2 = d_0$ แย้งกับ $H_1 : \mu_1 - \mu_2 < d_0$ หรือ

3) $H_0 : \mu_1 - \mu_2 = d_0$ แย้งกับ $H_1 : \mu_1 - \mu_2 \neq d_0$

ตัวสถิติที่ใช้ในการทดสอบขึ้นอยู่กับการแจกแจงของประชากร ความแปรปรวนของประชากรและขนาดของตัวอย่างที่สุ่มมา ซึ่งแบ่งได้ 3 กรณี คือ

1. ประชากรทั้งสองมีการแจกแจงแบบปกติ ทราบค่าความแปรปรวน σ_1^2 และ σ_2^2 เมื่อสุ่มตัวอย่างโดยอิสระจากประชากรที่ 1 และ 2 มาขนาด n_1 และ n_2 ตามลำดับ

ตัวสถิติที่ใช้ทดสอบคือ $Z = \frac{(\bar{x}_1-\bar{x}_2)-d_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

H_0	H_1	เขตวิกฤต
$H_0 : \mu_1 - \mu_2 = d_0$	$H_1 : \mu_1 - \mu_2 > d_0$	$Z \geq z_\alpha$
$H_0 : \mu_1 - \mu_2 = d_0$	$H_1 : \mu_1 - \mu_2 < d_0$	$Z \leq -z_\alpha$
$H_0 : \mu_1 - \mu_2 = d_0$	$H_1 : \mu_1 - \mu_2 \neq d_0$	$Z \geq z_{\frac{\alpha}{2}}$ หรือ $Z \leq -z_{\frac{\alpha}{2}}$

2. ประชากรที่ 1 และประชากรที่ 2 มีการแจกแจงแบบใดๆ ไม่ทราบความแปรปรวนของสองประชากร แต่สุ่มตัวอย่างขนาดใหญ่ (n_1 และ $n_2 \geq 30$)

ตัวสถิติที่ใช้ทดสอบคือ $Z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

เกณฑ์การตัดสินใจที่ระดับนัยสำคัญ α เหมือนกับ กรณีที่ 1

3. ประชากรทั้งสองมีการแจกแจงแบบปกติ ไม่ทราบความแปรปรวนของสองประชากร แต่สุ่มตัวอย่างขนาดเล็ก (n_1 หรือ $n_2 < 30$) ตัวสถิติที่ใช้ทดสอบขึ้นอยู่กับความแปรปรวนของประชากรทั้ง 2 กลุ่ม ดังนี้

3.1 เมื่อ $\sigma_1^2 = \sigma_2^2 = \sigma^2$

ตัวสถิติที่ใช้ทดสอบคือ $T = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$

เมื่อ d.f. = $n_1 + n_2 - 2$ และ $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

H_0	H_1	เขตวิกฤต
$H_0 : \mu_1 - \mu_2 = d_0$	$H1 : \mu_1 - \mu_2 > d_0$	$T \geq t_\alpha$
$H_0 : \mu_1 - \mu_2 = d_0$	$H1 : \mu_1 - \mu_2 < d_0$	$T \leq -t_\alpha$
$H_0 : \mu_1 - \mu_2 = d_0$	$H1 : \mu_1 - \mu_2 \neq d_0$	$T \geq t_{\frac{\alpha}{2}}$ หรือ $T \leq -t_{\frac{\alpha}{2}}$

3.2 เมื่อ $\sigma_1^2 \neq \sigma_1^2$

ตัวสถิติที่ใช้ทดสอบคือ $T = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ เมื่อ df = $\frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{(\frac{S_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{S_2^2}{n_2})^2}{n_2 - 1}}$

เกณฑ์การตัดสินใจที่ระดับนัยสำคัญ α เหมือนกับ กรณีที่ 3.1

ตัวอย่างที่ 3 ผู้ผลิตรายหนึ่งยืนยันว่า ความทนทานเฉลี่ยของเส้นด้ายยี่ห้อ A สูงกว่าความทนทานเฉลี่ยของยี่ห้อ B อย่างน้อยที่สุด 12 กิโลกรัม เพื่อทดสอบคำยืนยัน เขาได้สุ่มตัวอย่างของเส้นด้ายแต่ละยี่ห้อมา 50 เส้น แต่ละเส้นมีความยาวเท่ากัน นำมาทดสอบหาค่าความทนทานภายใต้เงื่อนไขเดียวกัน ปรากฏว่าเส้นด้ายตัวอย่างยี่ห้อ A มีความทนทานเฉลี่ย 86.7 กิโลกรัมด้วยส่วนเบี่ยงเบนมาตรฐาน 6.28 กิโลกรัม ขณะที่

เส้นด้ายตัวอย่างยี่ห้อ B มีความทนทานเฉลี่ย 77.8 กิโลกรัม ด้วยส่วนเบี่ยงเบนมาตรฐาน 5.61 กิโลกรัม จงทดสอบคำยืนยันของผู้ผลิตโดยใช้ระดับนัยสำคัญ 0.05

ค่าสถิติ $\bar{x}_1$ = 86.7 s_1 = 6.28 s_1^2 = 39.44 n_1 = 50

$\bar{x}_2$ = 77.8 s_2 = 5.61 s_2^2 = 31.47 n_1 = 50

ให้ μ_1 : ความทนทานเฉลี่ยของเส้นด้ายยี่ห้อ A

μ_2 : ความทนทานเฉลี่ยของเส้นด้ายยี่ห้อ B

1. ตั้งสมมติฐาน $H_0 : \mu_1 - \mu_2 = 12$

$H_1 : \mu_1 - \mu_2 < 12$

2. ระดับนัยสำคัญ (α) = 0.05

3. บริเวณปฏิเสธ H_0 คือ $Z \leq -Z_{0.05}$

$Z \leq -1.645$

4. ตัวสถิติที่ใช้ทดสอบคือ $Z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(86.7 - 77.8) - 12}{\sqrt{\frac{39.44}{50} + \frac{31.47}{50}}} = -2.063$

5. เพราะว่า Z = -2.063 ตกอยู่บริเวณวิกฤตจึงปฏิเสธ H_0 สรุปว่า ความทนทานเฉลี่ยของเส้นด้ายยี่ห้อ A สูงกว่าความทนทานเฉลี่ยของเส้นด้ายยี่ห้อ B น้อยกว่า 12 กิโลกรัม อย่างมีนัยสำคัญที่ระดับ 0.05

การทดสอบสมมติฐานเกี่ยวกับผลต่างของค่าเฉลี่ยของสองประชากร

กรณีที่ตัวอย่างมีความสัมพันธ์กัน

เมื่อตัวอย่างที่สุ่มมาจากประชากรที่ 1 และประชากรที่ 2 โดยไม่เป็นอิสระกัน ตัวสถิติที่ใช้ในการทดสอบผลต่างของค่าเฉลี่ยระหว่าง 2 ประชากรดังกล่าว จะอาศัยการแจกแจงของ $\bar{d}$ โดยกำหนดให้ μ_d แทนผลต่างของค่าเฉลี่ยประชากรโดยที่ $\mu_d = \mu_1 - \mu_2$ และให้ d_0 คือ ค่าคงที่ของผลต่างของค่าเฉลี่ยประชากรที่ต้องการทดสอบ ดังนั้นสมมติฐานที่ต้องการทดสอบจะอยู่ในลักษณะ

1) $H_0 : \mu_1 - \mu_2 = d_0$ แย้งกับ $H_1 : \mu_1 - \mu_2 > d_0$

$H_0 : \mu_d = d_0$ แย้งกับ $H_1 : \mu_d > d_0$ หรือ

2) $H_0 : \mu_1 - \mu_2 = d_0$ แย้งกับ $H_1 : \mu_1 - \mu_2 < d_0$

$H_0 : \mu_d = d_0$ แย้งกับ $H_1 : \mu_d < d_0$ หรือ

3) $H_0 : \mu_1 - \mu_2 = d_0$ แย้งกับ $H_1 : \mu_1 - \mu_2 \neq d_0$

$H_0 : \mu_d = d_0$ แย้งกับ $H_1 : \mu_d \neq d_0$

ตัวสถิติที่ใช้ทดสอบสมมติฐานขึ้นอยู่กับขนาดของตัวอย่างที่สุ่มมาดังนี้

1.1 เมื่อตัวอย่างมีขนาดเล็ก (n < 30) ภายใต้ H_0 เป็นจริง

ตัวสถิติที่ใช้ทดสอบคือ $T = \frac{\bar{D} - d_0}{\frac{S_d}{\sqrt{n}}}$ **เมื่อ** $df = n - 1$

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

H_0	H_1	เขตวิกฤต
$H_0 : \mu_d = d_0$	$H1 : \mu_d > d_0$	$T \geq t_\alpha$
$H_0 : \mu_d = d_0$	$H1 : \mu_d < d_0$	$T \leq -t_\alpha$
$H_0 : \mu_d = d_0$	$H1 : \mu_d \neq d_0$	$T \geq t_{\frac{\alpha}{2}}$ หรือ $T \leq -t_{\frac{\alpha}{2}}$

1.2 เมื่อตัวอย่างมีขนาดใหญ่ (n ≥ 30) ภายใต้ H_0 เป็นจริง

ตัวสถิติที่ใช้ทดสอบคือ $Z = \frac{\bar{D} - d_0}{\frac{S_d}{\sqrt{n}}}$

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

H_0	H_1	เขตวิกฤต
$H_0 : \mu_d = d_0$	$H_1 : \mu_d > d_0$	$Z \geq z_\alpha$
$H_0 : \mu_d = d_0$	$H_1 : \mu_d < d_0$	$Z \leq -z_\alpha$
$H_0 : \mu_d = d_0$	$H_1 : \mu_d \neq d_0$	$Z \geq z_{\frac{\alpha}{2}}$ หรือ $Z \leq -z_{\frac{\alpha}{2}}$

ตัวอย่างที่ 4 ปริมาณการขาย ก่อนและหลังการอบรมเกี่ยวกับเทคนิคการขายของพนักงานขาย 12 คน ที่สุ่มมาได้เป็นดังนี้ (1000 บาท/เดือน)

พนักงานคนที่	1	2	3	4	5	6	7	8	9	10	11	12
ก่อนอบรม	135	142	130	143	135	159	126	139	144	152	130	144
หลังอบรม	136	141	140	148	138	155	135	138	148	160	132	150

ที่ระดับนัยสำคัญ 0.05 จะสรุปได้หรือไม่ว่าปริมาณการขายหลังการอบรมสูงกว่าก่อนอบรม

ให้ μ_1 **: แทนปริมาณการขายเฉลี่ยก่อนอบรม**

μ_2 **: แทนปริมาณการขายเฉลี่ยหลังอบรม**

1. ตั้งสมมติฐาน $H_0 : \mu_2 = \mu_1$ **หรือ** $H_0 : \mu_2 - \mu_1 = 0$ **หรือ** $H_0 : \mu_d = 0$

$H_1 : \mu_2 > \mu_1$ หรือ $H_0 : \mu_2 - \mu_1 > 0$ หรือ $H_0 : \mu_d > 0$

2. ระดับนัยสำคัญ (α) = 0.05

3. บริเวณปฏิเสธ H_0 คือ $T \geq 2.72$

4. ตัวสถิติที่ใช้ทดสอบคือ $T = \frac{\bar{D} - d_0}{\frac{S_d}{\sqrt{n}}}$

พนักงานคนที่	1	2	3	4	5	6	7	8	9	10	11	12
ก่อนอบรม	135	142	130	143	135	159	126	139	144	152	130	144
หลังอบรม	136	141	140	148	138	155	135	138	148	160	132	150
d_i	1	-1	0	5	3	-4	9	-1	4	8	2	6

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{12} d_i = \frac{1-1+0+5+3-4+\cdots+6}{12} = 3.5$$

$$s_p^2 = \frac{n\sum d_i^2 - (\sum d_i)^2}{n(n-1)}$$

$$= \frac{12\left[1^2+(-1^2)+0^2+\cdots+6^2\right]-[1-1+0+\cdots+6]^2}{12(12-1)}$$

$$= 18.82$$

$$s_d = 4.34$$

ดังนั้น $T = \frac{3.5-0}{\frac{4.34}{\sqrt{12}}} = 2.795$

5. เพราะว่า T = 2.795 ตกอยู่บริเวณวิกฤตจึงปฏิเสธ H_0 ที่ระดับนัยสำคัญ 0.05 สรุปว่าปริมาณการขายหลังการอบรมสูงกว่าก่อนอบรม

การทดสอบสมมติฐานเกี่ยวกับสัดส่วนของหนึ่งประชากร

ให้ P แทนสัดส่วนของประชากร P_0 เป็นค่าคงที่ของสัดส่วนประชากรที่ต้องการเปรียบเทียบหรือเป็นค่าสัดส่วนที่คาดว่าจะเป็น สมมติฐานที่ต้องการทดสอบอยู่ในลักษณะ

1) $H_0 : P = P_0$ แย้งกับ $H_1 : P > P_0$ หรือ
2) $H_0 : P = P_0$ แย้งกับ $H_1 : P < P_0$ หรือ
3) $H_0 : P = P_0$ แย้งกับ $H_1 : P \neq P_0$

เมื่อขนาดตัวอย่างที่สุ่มมามีขนาดใหญ่ ($n \geq 30$) ภายใต้ H_0 เป็นจริง

ตัวสถิติที่ใช้ทดสอบคือ $Z = \frac{\hat{P} - P_0}{\sqrt{\frac{p_0 q_0}{n}}}$

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

H_0	H_1	เขตวิกฤต
$H_0 : P = P_0$	$H_1 : P > P_0$	$Z \geq z_\alpha$
$H_0 : P = P_0$	$H_1 : P < P_0$	$Z \leq -z_\alpha$
$H_0 : P = P_0$	$H_1 : P \neq P_0$	$Z \geq z_{\frac{\alpha}{2}}$ หรือ $Z \leq -z_{\frac{\alpha}{2}}$

ตัวอย่างที่ 5 บริษัทประกันภัยยืนยันว่า 20% ของบ้านเรือนในจังหวัดหนึ่งได้ทำประกันอัคคีภัยไว้ มีเหตุผลพอหรือไม่ ที่จะหักล้างคำยืนยัน ถ้าหากว่าสุ่มตัวอย่างบ้านเรือนในจังหวัดนี้มา 1000 หลัง แล้วพบว่ามี 236 หลังที่ทำประกันอัคคีภัยไว้ โดยใช้ระดับนัยสำคัญ 0.01

ให้ P คือ สัดส่วนของบ้านเรือนที่ทำประกันอัคคีภัยในจังหวัดดังกล่าว

1. ตั้งสมมติฐาน $H_0 : P = 0.20$

$H_1 : P \neq 0.20$

2. ระดับนัยสำคัญ $\alpha = 0.01$

3. บริเวณวิกฤต คือ $Z \geq 2.57$ หรือ $Z \leq -2.57$

4. ภายใต้ H_0 เป็นจริง ตัวสถิติที่ใช้ทดสอบคือ $Z = \frac{\hat{P} - P_0}{\sqrt{\frac{p_0 q_0}{n}}}$

เมื่อ $\hat{P} = \frac{236}{1000} = 0.236$ จะได้ $Z = \frac{0.236 - 0.2}{\sqrt{\frac{(0.2)(0.8)}{1000}}} = 2.846$

5. เพราะว่า Z = 2.846 ตกอยู่บริเวณวิกฤตจึงปฏิเสธ H_0 สรุปว่า สัดส่วนของครัวเรือนที่ทำประกันอัคคีภัยไว้แตกต่างจาก 20% อย่างมีนัยสำคัญที่ระดับนัยสำคัญ 0.01

การทดสอบสมมติฐานเกี่ยวกับผลต่างของสัดส่วนประชากร P_1-P_2

เมื่อ P_1 และ P_2 เป็นสัดส่วนประชากรของประชากรที่ 1 และ 2 ตามลำดับ P_0 คือ ค่าคงที่ของผลต่างของสัดส่วนประชากรที่ต้องการทดสอบหรือยืนยันกับพารามิเตอร์ สมมติฐานที่ใช้ในการทดสอบจะอยู่ในลักษณะ

1) $H_0 : P_1 - P_2 = P_0$ แย้งกับ $H_1 : P_1 - P_2 > P_0$ หรือ

2) $H_0 : P_1 - P_2 = P_0$ แย้งกับ $H_1 : P_1 - P_2 < P_0$ หรือ

3) $H_0 : P_1 - P_2 = P_0$ แย้งกับ $H_1 : P_1 - P_2 \neq P_0$

เมื่อ n_1 และ n_2 คือขนาดตัวอย่างที่สุ่มมาจากประชากรที่ 1 และประชากรที่ 2 ตามลำดับ ซึ่งมีขนาดใหญ่พอ (n_1 และ $n_2 \geq 30$) ตัวสถิติที่ใช้ทดสอบมี 2 กรณีคือ

1) เมื่อ $P_0 \neq 0$

ตัวสถิติที่ใช้ในการทดสอบ คือ $$Z = \frac{(\hat{p}_1 - \hat{p}_2) - P_0}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_1\hat{q}_2}{n_2}}}$$

2) เมื่อ $P_0 = 0$

ตัวสถิติที่ใช้ในการทดสอบ คือ $$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

H_0	H_1	เขตวิกฤต
$H_0 : P_1 - P_2 = P_0$	$H_1 : P_1 - P_2 > P$	$Z \geq z_\alpha$
$H_0 : P_1 - P_2 = P_0$	$H_1 : P_1 - P_2 < P$	$Z \leq -z_\alpha$
$H_0 : P_1 - P_2 = P_0$	$H_1 : P_1 - P_2 \neq P_0$	$Z \geq z_{\frac{\alpha}{2}}$ หรือ $Z \leq -z_{\frac{\alpha}{2}}$

ตัวอย่างที่ 6 บริษัทผลิตเตาไมโครเวฟแห่งหนึ่ง ได้เริ่มนำเตารุ่นใหม่ซึ่งมี 3 สี คือ สีขาว สีฟ้า และสีเขียว ออกจำหน่าย จากการสุ่ม 1,000 เตาแรกที่จำหน่ายได้ บริษัทพบว่าเตาสีขาวมียอดจำหน่ายมากที่สุดรวม 400 เตา ส่วนเตาสีฟ้าและสีเขียวมียอดจำหน่ายใกล้เคียงกัน ที่ระดับนัยสำคัญ 0.01 จงตรวจสอบว่าเตาสีขาวได้รับความนิยมสูงที่สุดหรือไม่

ให้ P คือ สัดส่วนของยอดจำหน่ายเตาสีขาว

1. ตั้งสมมติฐาน $H_0 : P = 0.33$

$H_1 : P > 0.33$

2. ระดับนัยสำคัญ $\alpha = 0.01$

3. บริเวณวิกฤต คือ $Z \geq 2.33$

4. ภายใต้ H_0 เป็นจริง ตัวสถิติที่ใช้ทดสอบคือ $$Z = \frac{\hat{P} - P_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

เมื่อ $\hat{P} = \frac{400}{1000} = 0.236$ จะได้ $$Z = \frac{0.4 - 0.33}{\sqrt{\frac{(0.33)(0.67)}{1000}}} = 4.7$$

5. เพราะว่า $Z = 4.7$ ตกอยู่บริเวณวิกฤตจึงปฏิเสธ H_0 สรุปว่า สัดส่วนของเตาสีขาวที่ขายได้มากกว่า 33% อย่างมีนัยสำคัญที่ระดับ 0.01

การตรวจสอบสมมติฐานเกี่ยวกับความแปรปรวนของหนึ่งประชากร

ให้ σ^2 เป็นความแปรปรวนของประชากร และ σ_0^2 เป็นค่าคงที่ของความแปรปรวนประชากรที่ต้องการทดสอบหรือยืนยันกับพารามิเตอร์ สมมติฐานที่ใช้ในการทดสอบจะอยู่ในลักษณะ

1. $H_0 : \sigma^2 = \sigma_0^2$ แย้งกับ $H_1 : \sigma^2 > \sigma_0^2$
2. $H_0 : \sigma^2 = \sigma_0^2$ แย้งกับ $H_1 : \sigma^2 < \sigma_0^2$
3. $H_0 : \sigma^2 = \sigma_0^2$ แย้งกับ $H_1 : \sigma^2 \neq \sigma_0^2$

ภายใต้ H_0 เป็นจริง ตัวสถิติที่ใช้ทดสอบคือ $\chi^2 = (n-1)\frac{S^2}{\sigma_0^2}$ เมื่อ df = n – 1

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

H_0	H_1	เขตวิกฤต
$H_0 : \sigma^2 = \sigma_0^2$	$H_1 : \sigma^2 > \sigma_0^2$	$\chi^2 \geq \chi_\alpha^2$
$H_0 : \sigma^2 = \sigma_0^2$	$H_1 : \sigma^2 < \sigma_0^2$	$\chi^2 \leq -\chi_{1-\alpha}^2$
$H_0 : \sigma^2 = \sigma_0^2$	$H_1 : \sigma^2 \neq \sigma_0^2$	$\chi^2 \leq \chi_{1-\frac{\alpha}{2}}^2$ หรือ $\chi^2 \geq \chi_{\frac{\alpha}{2}}^2$

ตัวอย่างที่ 7 ในการศึกษาเกี่ยวกับจุดอ่อนตัวของสารน้ำมันชนิดหนึ่ง โดยมีการกำหนดไว้ว่าค่าเบี่ยงเบนมาตรฐานที่แท้จริงของจุดอ่อนตัวสูงที่สุดของสารชนิดนี้เป็น 0.5 ^{0}C จึงมีผู้นำสารตัวอย่างนี้มา 10 ตัวอย่าง วัดค่าเบี่ยงเบนมาตรฐานได้เป็น 0.58 ^{0}C จงทดสอบว่าสารน้ำมันชนิดนี้มีจุดอ่อนตัวเป็นไปตามที่กำหนดไว้หรือไม่ ที่ระดับนัยสำคัญ 0.01

ให้ σ^2 คือ ความแปรปรวนของจุดอ่อนตัวของสารน้ำมันชนิดนี้

1. ตั้งสมมติฐาน $H_0 : \sigma^2 = 0.25$

   $H_1 : \sigma^2 > 0.25$

2. ระดับนัยสำคัญ $\alpha = 0.01$

3. ค่าวิกฤต คือ $\chi^2_{0.01(9)} = 21.67$

   บริเวณวิกฤต คือ $\chi^2 \geq 21.67$

4. ภายใต้ H_0 เป็นจริง ตัวสถิติที่ใช้ทดสอบคือ $\chi^2 = (n-1)\frac{S^2}{\sigma_0^2}$

   $= (10\text{-}1)\ \frac{0.58^2}{0.25} = 12.11$

5. เพราะว่า $\chi^2 = 12.11$ ตกอยู่บริเวณยอมรับ H_0 สรุปได้ว่าค่าเบี่ยงเบนมาตรฐานของจุดอ่อนตัวของสารน้ำมันชนิดนี้เท่ากับ 0.5 ^{0}C ที่ระดับนัยสำคัญ 0.01

การทดสอบสมมติฐานเกี่ยวกับอัตราส่วนของความแปรปรวนสองประชากร

ให้ σ_1^2 และ σ_2^2 เป็นความแปรปรวนของประชากรที่ 1 และ 2 ตามลำดับ สมมติฐานที่จะทดสอบอยู่ในลักษณะ

1. $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ แย้งกับ $H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$ หรือ
2. $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ แย้งกับ $H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$ หรือ
3. $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ แย้งกับ $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

ภายใต้ H_0 เป็นจริง ตัวสถิติที่ใช้ทดสอบคือ $F = \frac{S_1^2}{S_2^2}$ เมื่อ $df = n_1 - 1, n_2 - 1$

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

H_0	H_1	เขตวิกฤต
$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$	$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$	$F \geq f_{\alpha}$
$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$	$H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$	$F \leq f_{1-\alpha}$
$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$	$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$	$F \leq f_{1-\alpha/2}$ หรือ $F \geq f_{\alpha/2}$

ตัวอย่างที่ 8 จงทดสอบความแปรปรวนของความยาวรากต้นถั่วเหลือง 2 พันธุ์ จำนวน 12 ต้น และ 10 ต้น ตาม ลำดับที่ระดับนัยสำคัญ 0.05 ว่ามีค่าเท่ากันหรือไม่

ให้ σ_1^2 คือ ความแปรปรวนของถั่วเหลืองพันธุ์ที่ 1 มีค่าเท่ากับ 0.177

σ_2^2 คือ ความแปรปรวนของถั่วเหลืองพันธุ์ที่ 2 มีค่าเท่ากับ 0.141

1. ตั้งสมมติฐาน $H_0 : \sigma_1^2 = \sigma_2^2$

   $H_1 : \sigma_1^2 \neq \sigma_2^2$

2. ระดับนัยสำคัญ $\alpha = 0.05$

3. ค่าวิกฤต คือ $F_{0.025(11,9)} = 3.87$ และ $F_{0.975(11,9)} = \frac{1}{F_{0.025(9,11)}} = \frac{1}{3.59} = 0.28$

   บริเวณวิกฤต คือ $F \geq 3.87$ และ $F \leq 0.28$

4. ภายใต้ H_0 เป็นจริง ตัวสถิติที่ใช้ทดสอบคือ $F = \frac{S_1^2}{S_2^2} = \frac{0.177}{0.141} = 1.255$

5. เพราะว่า F =1.255 ตกอยู่บริเวณยอมรับ H_0 สรุปว่าความแปรปรวนของความยาวรากต้นถั่วเหลือง 2 พันธุ์ไม่แตกต่างกันที่ระดับนัยสำคัญ 0.05

บทที่ 6

การทดสอบค่าไคสแควร์

ในเก็บรวบรวมข้อมูลที่นับในรูปความถี่ (Frequency Data) หรือข้อมูลจำแนกประเภท (Categorical Data) โดยข้อมูลที่นับนั้นเป็นข้อมูลเชิงคุณภาพ ไม่สามารถวัดเป็นตัวเลขได้ เช่น การสำรวจความคิดเห็นสิ่งที่ชอบ เช่น ลักษณะสินค้าหรือสีของสินค้า OTOP สีของสินค้าที่นักท่องเที่ยวชอบ เป็นต้น ซึ่งข้อมูลจากเก็บจากภาคสนามจริง และการนับจำนวนในรูปความถี่นั้น เรียกว่า ค่าความถี่จากการสังเกต (Observed frequency) แทนด้วย O_i และค่าความถี่ที่ผู้ศึกษาคาดหวังว่าเกิดเป็นไปตามหลักการทฤษฎี เรียกว่า ค่าความถี่คาดหวัง (Expected frequency) แทนด้วย E_i การทดสอบไคสแควร์จะแบ่งเป็น 3 การทดสอบคือ

1) การทดสอบภาวะสารูปสนิทดี (Goodness – of – Fit Test) มีวัตถุประสงค์เพื่อทดสอบเกี่ยวกับลักษณะต่างๆ ของประชากร ว่าเป็นไปตามที่คาดไว้หรือไม่อีกวัตถุประสงค์หนึ่งคือ เพื่อทดสอบเกี่ยวกับการแจกแจงของประชากร ข้อมูลมาจากตัวอย่าง 1 กลุ่ม โดยมีตัวแปร 1 ตัว และตัวแปรมีสเกลการวัดแบบแบ่งประเภทซึ่งมีข้อมูลเป็นจานวนนับ

2) การทดสอบความเป็นอิสระ (Test for Independence) มีวัตถุประสงค์เพื่อทดสอบความเป็นอิสระหรือความสัมพันธ์ระหว่างตัวแปร 2 ตัว และตัวแปรมีสเกลการวัดแบบแบ่งประเภทซึ่งมีข้อมูลเป็นจำนวนนับ

3) การทดสอบภาวะเอกพันธ์ (Test of Homogeneity) ในกรณีที่ตัวอย่างกลุ่มเดียวเรามักทดสอบภาวะสารูปสนิทดี ระหว่างการแจกแจงของตัวอย่างกับการแจกแจงที่กำหนด ส่วนกรณีที่มีตัวอย่าง 2 กลุ่มที่เป็นอิสระกัน เราสุ่มกลุ่มตัวอย่างจากประชากรแต่ละกลุ่ม และจัดข้อมูลของตัว 2 แปรตามที่เป็นแบบจำแนกประเภทให้อยู่ในชั้นต่างๆ ข้อมูลจะอยู่ในตาราง 2 ทาง เมื่อ ตัวแปรในทางหนึ่งของตารางอ้างถึงกลุ่มประชากร และตัวแปรที่อยู่อีกทางหนึ่งของตารางเป็นตัวแปรตามที่สนใจศึกษา มีสเกลการวัดแบบจำแนกประเภท หรือเป็นชั้นๆ วัตถุประสงค์เพื่อทดสอบเกี่ยวกับตัวแปรตามที่สนใจศึกษาของประชากรกลุ่มต่างๆ ว่ามาจากประชากรเดียวกันหรือไม่หรือมาจากประชากรที่มีการแจกแจงแบบเดียวกันหรือไม่

กลุ่มตัวอย่างที่ใช้ในการทดสอบ

1. กลุ่มตัวอย่างกลุ่มเดียว (Simple Classification)

การทดสอบ χ^2 กรณีกลุ่มตัวอย่างกลุ่มเดียวเป็นการทดสอบตัวแปรเพียงด้านเดียวเพื่อต้องการทราบว่า ความถี่ที่ได้จากการสังเกต (Observed Frequency) จากกลุ่มตัวอย่าง เป็นไปตามความถี่ที่คาดหวัง (Expected Frequency) ไว้ หรือไม่ตามนัยสำคัญที่กำหนด การทดสอบโดยการใช้สูตรคำนวณ χ^2test คือ

$$\chi^2 = \sum_{i=1}^{k}\left(\frac{Oi - Ei}{Ei}\right)^2, df = k - 1$$

χ^2 = ค่าสถิติไคสแควร์

O_i = ความถี่ที่ได้จากการสังเกต (Observed Frequency)

E_i = ความถี่ที่คาดหวัง (Expected Frequency) ซึ่งมีค่าเท่ากับจำนวน ข้อมูลคูณด้วย สัดส่วนที่คาดหวัง

K = จำนวนกลุ่มตัวแปร กรณี d.f. = K-1

สมมุติฐาน

H_0 = สัดส่วนของกลุ่มต่างๆ เป็นไปตามทฤษฎีที่คาดหวัง

H_1 = สัดส่วนของกลุ่มต่างๆ ไม่เป็นไปตามทฤษฎีที่คาดหวัง

กฎการตัดสินใจ จะปฏิเสธ H_0 ถ้าค่าสถิติ $\chi^2_{calc} \geq \chi^2_{(c-1),\,(1-\alpha)}$

เมื่อ c คือ จำนวนกลุ่ม นอกนั้นไม่ปฏิเสธ H_0

ตัวอย่างที่ 1 ทฤษฎีพันธุศาสตร์กล่าวว่า การสุ่มผสมพันธุ์ที่มีลักษณะพันธุกรรม A กับ a พบว่าในรุ่นลูกจะมีลักษณะพันธุกรรมเป็น AA, Aa, aa ในอัตราส่วน 1:2:1 ถ้าผสมพันธุ์ทั้งหมด 60 คู่ ปรากฏว่าได้ลูกที่มีลักษณะพันธุกรรม AA, Aa, aa จำนวน 14, 28 และ 18 ตามลำดับ ผู้ทดลองต้องการทราบว่า จำนวนลูกตามลักษณะพันธุกรรมทั้ง 3 แบบสอดคล้องกับทฤษฎีพันธุศาสตร์หรือไม่ ที่ระดับความสำคัญ 0.05

1. สมมติฐาน $H_0 : p_1 = \frac{1}{4}, \quad p_2 = \frac{2}{4}, \quad p_3 = \frac{1}{4}$

$H_1 : p_1 \neq \frac{1}{4}, \quad p_2 \neq \frac{2}{4}, \quad p_3 \neq \frac{1}{4}$

2. ระดับนัยสำคัญ (α) = 0.05

3. ใช้ตัวทดสอบ χ^2 มีเขตวิกฤตเป็น $\chi^2 \geq \chi^2$ (0.05),(3-1) = 5.99

4. คำนวณตัวทดสอบ χ^2 จากค่าสังเกตและค่าความถี่ที่คาดหวัง ดังนี้

จากค่าสังเกต $O_1 = 14, O_2 = 28, O_3 = 18$

คำนวณค่าที่คาดหวัง

$E_1 = 60(\frac{1}{4}) = 15$

$E_2 = 60(\frac{2}{4}) = 30$

$E_3 = 60(\frac{1}{4}) = 15$

ดังนั้น $\chi^2 = \frac{(14-15)^2}{15} + \frac{(28-30)^2}{30} + \frac{(18-15)^2}{15} = 0.80$

5. นำค่า χ^2 ที่คำนวณได้ไปเทียบกับค่าวิกฤติจากตาราง ซึ่งมีค่า $\chi^2_{0.05,2} = 5.99$

แสดงว่าค่า $\chi^2_{\text{คำนวณ}} < \chi^2_{\text{วิกฤต}}$ จึงยอมรับ H_0 สรุปได้ว่าค่าสังเกตที่ได้สอดคล้องกับทฤษฎีพันธุศาสตร์

2. กลุ่มตัวอย่างสองกลุ่ม (Two-way Classification)

การทดสอบในกรณีตัวแปรสองตัวนี้เป็นการทดสอบเพื่อดูว่าตัวแปรสองตัวนี้มีความเกี่ยวข้องหรือสัมพันธ์กันหรือไม่ ถ้าไม่สัมพันธ์กันหมายความว่าเป็นอิสระจากกัน ดังนั้นบางครั้งเราจึงเรียกว่า การทดสอบความเป็นอิสระ (The χ^2 – test for independence) ข้อมูลที่ได้จะอยู่ในมาตรานามบัญญัติ (Norminal scale) ซึ่งอาจเป็นจำนวนความถี่ สัดส่วน ร้อยละ ก็ได้ โดยแต่ละตัวแปรจะแบ่งเป็น 2 กลุ่มหรือประเภทขึ้นไป เช่น เพศ (ชาย – หญิง) กับวุฒิการศึกษา (ป.ตรี ป.โท ป.เอก) จะได้รูปแบบเป็น 2 x 3 ดังนั้นรูปแบบการวิเคราะห์อาจเป็นได้หลายรูปแบบขึ้นอยู่กับจำนวนกลุ่มของแต่ละตัวแปร

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

χ^2 = ค่าสถิติไคสแควร์

O_{ij} = ความถี่ที่ได้จากการสังเกต (Observed Frequency) ในแถวที่ I คอลัมน์ที่ j

E_{ij} = ความถี่ที่คาดหวัง (Expected Frequency) ในแถวที่ I คอลัมน์ที่ j

r = จำนวนแถว (row)

c = จำนวนคอลัมน์ (Column)

การหาค่าคาดหวัง $E_{ij} = \frac{r_i \times c_j}{N}$

เมื่อ r_i แทน ผลรวมของความถี่ในแถว i

c_j แทน ผลรวมของความถี่ในคอลัมน์ j

ตัวอย่างที่ 2 ในการศึกษาความพึงพอใจของผู้ปกครองนักเรียนที่มีต่อการบริหารโรงเรียนโดยเก็บข้อมูลกับผู้ปกครองอาชีพต่างๆ จำนวน 238 คน ได้ผลดังนี้

อาชีพ	ระดับความพึงพอใจ		
	มาก	ปานกลาง	น้อย
ข้าราชการ	30	20	7
เกษตรกร	40	30	12

ค้าขาย	47	33	19

ต้องการทดสอบว่า อาชีพของผู้ปกครองเกี่ยวข้องกับความพึงพอใจในการบริหารโรงเรียนหรือไม่

1. ตั้งสมมติฐานทางสถิติ

   H_0 : P = 0 (อาชีพไม่มีความสัมพันธ์กับความพึงพอใจ)

   H_1 : P ≠ 0 (อาชีพมีความสัมพันธ์กับความพึงพอใจ)

2. ระดับนัยสำคัญ (α) = 0.05

3. ใช้ตัวทดสอบ χ^2 มีเขตวิกฤตเป็น $\chi^2 \geq \chi^2$ (0.05),(3-1) (3-1) = 9.484

4. คำนวณตัวทดสอบ χ^2 โดยจะต้องคำนวณค่า E ก่อน ซึ่งสามารถคำนวณได้ดังนี้

อาชีพ	ระดับความพึงพอใจ			รวม
	มาก	ปานกลาง	น้อย	
ข้าราชการ	30	20	7	57
เกษตรกร	40	30	12	82
ค้าขาย	47	33	19	99
รวม	117	83	38	238

โดยที่ O_1 = 30, R = 57, C = 117 $\quad E = \frac{57 \times 117}{238} = 28.02$

O_2 = 20, R = 57, C = 83 $\quad E = \frac{57 \times 83}{238} = 19.88$

O_3 = 7, R = 57, C = 38 $\quad E = \frac{57 \times 38}{238} = 9.10$

O_4 = 40, R = 82, C = 117 $\quad E = \frac{82 \times 117}{238} = 40.31$

O_5 = 30, R = 82, C = 83 $\quad E = \frac{82 \times 38}{238} = 28.60$

O_6 = 12, R = 82, C = 38 $\quad E = \frac{82 \times 38}{238} = 13.09$

O_7 = 47, R = 99, C = 117 $\quad E = \frac{99 \times 117}{238} = 48.67$

O_8 = 35, R = 99, C = 83 $\quad E = \frac{99 \times 83}{238} = 15.81$

O_8 = 19, R = 99, C = 38 $\quad E = \frac{99 \times 38}{238} = 15.81$

นำค่า E ที่ได้เขียนลงในตารางโดยใส่เก็บเอาไว้

อาชีพ	ระดับความพึงพอใจ			รวม
	มาก	ปานกลาง	น้อย	
ข้าราชการ	30 (28.02)	20 (19.88)	7 (9.10)	57
เกษตรกร	40 (40.31)	30 (28.60)	12 (13.09)	82
ค้าขาย	47 (48.67)	33 (34.53)	19 (15.81)	99
รวม	117	83	38	238

คำนวณค่า χ^2 โดยแทนค่าลงในสูตร

$$\chi^2 = \frac{(30-28.02)^2}{28.02} + \frac{(20-19.88)^2}{19.88} + \frac{(7-9.10)^2}{9.10} + \frac{(40-40.31)^2}{40.31} + \frac{(30-28.60)^2}{28.60} + \frac{(12-13.09)^2}{13.09} + \frac{(47-48.67)^2}{48.67} + \frac{(33-34.52)^2}{34.53} + \frac{(12-13.09)^2}{13.09}$$

= 0.14 + 0.12 + 0.49 + .0024 + .049 + .076 + .034 + .068 + .091

= 1.0704

5. นำค่า χ^2 ที่คำนวณได้ไปเทียบกับค่าวิกฤติจากตาราง ซึ่งมีค่า $\chi^2_{.05,\,4}$ = 9.484 แสดงว่า ค่า $\chi^2_{\text{คำนวณ}} < \chi^2_{\text{วิกฤต}}$ จึงยอมรับ H_O สรุปได้ว่า อาชีพของผู้ปกครองไม่มีความสัมพันธ์กับความพึงพอใจในการบริหารอย่างไม่มีนัยสำคัญทางสถิติหรือกล่าวได้ว่า อาชีพกับความพึงพอใจไม่เกี่ยวข้องกัน

บทที่ 7

การทดสอบความแตกต่างของค่าเฉลี่ย

การทดสอบสมมติฐานเพื่อเปรียบเทียบความแตกต่างระหว่างค่าเฉลี่ยสองกลุ่มขึ้นไป ข้อมูลที่รวบรวมได้จากกลุ่มตัวอย่างแต่ละกลุ่มนั้นเป็นข้อมูลในมาตราอันตรภาคหรือมาตราอันตราส่วนโดยนำค่าเฉลี่ย ($\overline{X}$) ที่ได้จากกลุ่มตัวอย่างตั้งแต่ 2 กลุ่มขึ้นไปมาเปรียบเทียบกัน เพื่อนำไปสู่การสรุปว่าค่าเฉลี่ยของประชากรแต่ละกลุ่มแตกต่างกันหรือไม่ การทดสอบความแตกต่างระหว่างค่าเฉลี่ยจำแนกได้เป็น 2 กรณี คือ

1. การทดสอบความแตกต่างระหว่างค่าเฉลี่ยสองค่าที่ได้จากกลุ่มตัวอย่างที่เป็นอิสระจากกัน (Independent Sample)

2. การทดสอบความแตกต่างระหว่างค่าเฉลี่ยสองค่าที่ได้จากกลุ่มตัวอย่างสองกลุ่มที่ไม่เป็นอิสระจากกัน (Dependent Sample)

สถิติทดสอบ

การทดสอบสมมติฐานที่ใช้สถิติทดสอบมี 2 ประเภท คือ

1. การทดสอบสถิติที่ใช้พารามิเตอร์ (Parameter Statistics) ตัวอย่างสถิติทดสอบ เช่น T-test, Z-test, F-test เป็นต้น

2. การทดสอบสถิติที่ไม่ใช้พารามิเตอร์ (Non-Parameter Statistics) ตัวอย่างสถิติทดสอบ เช่น Chi-Square test, Binomial-test, Run-test เป็นต้น

การเลือกใช้สถิติทดสอบใดจะต้องคำนึงถึงข้อตกลงเบื้องต้น (Assumption) ของสถิติทดสอบแต่ละตัว เช่น ขนาดของข้อมูล จำนวนตัวแปร จำนวนกลุ่มตัวอย่าง ระดับข้อมูล ลักษณะการแจกแจงของข้อมูล เป็นต้น

การทดสอบค่าเฉลี่ยแยกการทดสอบได้เป็น 3 วิธี คือ

1. การทดสอบค่าเฉลี่ยของประชากรหนึ่งกลุ่ม

การทดสอบค่าเฉลี่ยของประชากรหนึ่งกลุ่ม เป็นการทดสอบว่าค่าเฉลี่ยที่ได้แตกต่างจากค่าที่กำหนดไว้หรือไม่ เช่น การทดสอบว่ารายได้เฉลี่ยของคนไทยเป็น 10,000 บาทหรือไม่ เป็นต้น ซึ่งก่อนการเลือกใช้สถิติทดสอบใดจะต้องพิจารณาข้อตกลงเบื้องต้นก่อนคือ ข้อมูลที่ได้ต้องมาจากการสุ่ม (Random) มีการแจกแจงเป็นโค้งปกติ (Normal Curve) และเลือกใช้สถิติทดสอบ ดังนี้

- สถิติทดสอบ Z-test กรณีขนาดกลุ่มตัวอย่างมากกว่าหรือเท่ากับ 30 หรือถ้าขนาดตัวอย่างน้อยกว่า 30 จะต้องทราบค่าความแปรปรวนของประชากร (σ^2)
- สถิติทดสอบ T-test กรณีขนาดกลุ่มตัวอย่างน้อยกว่า 30

2. การทดสอบผลต่างระหว่างค่าเฉลี่ยของประชากรสองกลุ่มที่เป็นอิสระกัน

การทดสอบค่าเฉลี่ยของประชากรสองกลุ่ม เป็นการทดสอบเพื่อต้องการทราบว่าค่าเฉลี่ยของสองกลุ่มที่ได้แตกต่างกันหรือไม่ เช่น การทดสอบว่ารายได้เฉลี่ยของคนไทยที่อาศัยในจังหวัดกรุงเทพฯแตกต่างจากจังหวัดใกล้เคียงหรือไม่ เป็นต้น ซึ่งก่อนการเลือกใช้สถิติทดสอบใดจะต้องพิจารณาข้อตกลงเบื้องต้นก่อนคือ ข้อมูลทั้งสองกลุ่มจะต้องมาจากการสุ่มและเป็นอิสระต่อกัน (Independent) และมีการแจกแจงเป็นโค้งปกติ ซึ่งจะเลือกใช้สถิติทดสอบดังนี้

- **สถิติทดสอบ Z-test กรณี ขนาดกลุ่มตัวอย่างมากกว่าหรือเท่ากับ 30 หรือถ้าขนาดตัวอย่างน้อยกว่า 30 จะต้องทราบค่าความแปรปรวนของประชาการ ($\sigma_1^2 = \sigma_2^2$)**
- **สถิติทดสอบ T-test กรณี ขนาดกลุ่มตัวอย่างน้อยกว่า 30**

กลุ่มตัวอย่างเป็นอิสระจากกันถ้าได้มาโดยวิธีใดวิธีหนึ่ง

วิธีที่ 1 มีกลุ่มใหญ่ที่ต้องการศึกษา (Subjects) กลุ่มใหญ่ 1 กลุ่มแล้วสุ่มแยกเป็น 2 กลุ่มย่อย (Subgroup) เช่น จากนักเรียนชั้น ป. 6 (ประชากร) ของโรงเรียนแห่งหนึ่งจำนวน 400 คน ผู้วิจัยสุ่มมาศึกษา 80 คน โดยสุ่ม เป็นกลุ่มทดลอง และกลุ่มควบคุมกลุ่มละ 40 คน นักเรียนสองกลุ่มนี้ถือว่าเป็นอิสระจากกัน

วิธีที่ 2 กลุ่มตัวอย่างแต่ละกลุ่ม ถูกสุ่มมาจากประชากรขนาดใหญ่แต่ละกลุ่มตัวอย่างมีประชากร 2 กลุ่มนี้ถือว่าเป็นอิสระจากกัน

<u>หมายเหตุ</u> 1. โดยทฤษฎี T-test ใช้เมื่อกลุ่มตัวอย่างมีขนาดเล็ก ($n_1 < 30, n_2 < 30$) แต่ในทางปฏิบัติ T-test ใช้กับกลุ่มตัวอย่างขนาดใดก็ได้ ขอเพียงแต่ให้ประชากรของกลุ่มตัวอย่างที่สุ่มมามีการแจกแจงปกติหรือเข้าใกล้การแจกแจงปกติ (Weiss, 1995)

2. T-test มีโอกาสใช้มากกว่า Z–Test ทั้งนี้เพราะการใช้ Z–Test เราไม่มีโอกาสรู้ค่าความแปรปรวนของประชากร (σ^2) จึงต้องประมาณ ด้วยความแปรปรวนของกลุ่มตัวอย่าง (S_1^2, S_2^2) ซึ่งเมื่อเป็นเช่นนี้ ค่าสถิติทดสอบจะมีการแจกแจงแบบ t (t–distribution) มากกว่าการแจกแจงแบบ Z นั่นคือ ถ้าแทนค่าความแปรปรวนด้วยควรใช้ T-test

ข้อตกลงเบื้องต้น

ในการทดสอบความมีนัยสำคัญระหว่างค่าเฉลี่ยสองค่าที่ได้จากกลุ่มตัวอย่างที่เป็นอิสระจากกันมีข้อตกลงเบื้องต้น (Assumtion) ที่สำคัญ 2 ประการ คือ

1. กลุ่มตัวอย่างทั้งสองกลุ่มมาจากประชากร 2 กลุ่มซึ่งแตกต่างกัน การกระจายเป็นโค้งปกติ (Normal Distribution)

2. กลุ่มตัวอย่างทั้งสองกลุ่มต้องเป็นอิสระจากกัน (Independent Sample)

การเลือกใช้ Z–test และ T-test

1. กลุ่มตัวอย่างแต่ละกลุ่มมีขนาดใหญ่ (n1 และ n2 แต่ละกลุ่มเท่ากับหรือมากกว่า 30) ใช้ Z–test

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

ในทางปฏิบัติอาจไม่สามารถหา σ^2 ได้ ซึ่งสามารถใช้ S_1^2, S_2^2 แทนได้

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

2. กลุ่มตัวอย่างแต่ละกลุ่มมีขนาดเล็ก (n ในแต่ละกลุ่มน้อยกว่า 30) ให้ใช้ t–test โดยต้องคำนึงถึงองศาอิสระ (degree of freedom : df) ในการใช้ t–test นี้มี 2 กรณี คือ

1) ไม่ทราบความแปรปรวนของประชากร ทั้ง 2 กลุ่ม และตั้งข้อตกลง (assume) ว่า ความแปรปรวนของประชากรทั้งสองกลุ่มเท่ากัน ($\sigma_1^2 = \sigma_2^2$)

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

$$df = n_1 + n_2 - 2$$

เรียก pooled t-test (Weiss, 1995)

2) ไม่ทราบความแปรปรวนของประชากรทั้ง 2 กลุ่ม และตกลงว่า (assume) ความแปรปรวนของประชากรทั้งสองกลุ่มไม่เท่ากัน ($\sigma_1^2 \neq \sigma_2^2$)

สูตร

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$df = \frac{\left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right]^2}{\frac{\left[\frac{S_1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{S_2^2}{n_2}\right]^2}{n_2 - 1}}$$

เรียก Nonpooled t–test

ในทางปฏิบัติ

1. ถ้า $n_1 = n_2$ สามารถใช้ pooled t–test ได้เลยโดยไม่ต้องทดสอบความแปรปรวน

2. ถ้า $n_1 \neq n_2$ ให้ทดสอบด้วย F-test ถ้าค่า F-test ไม่มีนัยสำคัญทางสถิติ ให้ใช้ poolet t–test แต่ถ้ามีนัยสำคัญทางสถิติตามระดับที่ตั้งไว้ ใช้ Nonpooled t–test

การทดสอบความแปรปรวน

ในกรณีที่ไม่สามารถตัดสินใจได้ว่าหรือไม่ สามารถทดสอบได้โดยใช้ F-test for homogeneity of variance (การใช้ F-test ทดสอบความเป็นเอกพันธ์ของความแปรปรวน)

$$F = \frac{S_1^2}{S_2^2} \qquad F = \frac{S_2^2}{S_1^2}$$

$$df_1 = n_1 - 1, df_2 = n_2 - 1 \quad \text{หรือ} \quad df_1 = n_2 - 1, df_2 = n_1 - 1$$

ในการทดสอบความเท่ากันของความแปรปรวนโดยใช้ F-test นี้ในการที่จะต้องนำเอาค่าความแปรปรวนที่มีค่ามากกว่า เป็นเศษ และค่าความแปรปรวนที่มีค่าน้อยกว่าเป็นตัวหาร

เมื่อคำนวณได้ค่า F แล้ว ให้นำไปเรียบเทียบกับค่าวิกฤต ถ้าค่าที่คำนวณได้มากกว่าหรือเท่ากับค่าวิกฤต แสดงว่าค่าความแปรปรวนของทั้งสองกลุ่มไม่เท่ากัน ($\sigma_1^2 \neq \sigma_2^2$) แต่ถ้าได้น้อยกว่าค่าวิกฤตแสดงว่าค่าความแปรปรวนของทั้งสองกลุ่มเท่ากัน ($\sigma_1^2 = \sigma_2^2$) หลังจากนั้นจึงเลือกใช้สูตร T-test ที่ถูกต้อง

3. การทดสอบผลต่างระหว่างค่าเฉลี่ยของประชากรสองกลุ่มที่ไม่เป็นอิสระจากกัน

การทดสอบผลต่างระหว่างค่าเฉลี่ยของประชากรสองกลุ่มที่เป็นอิสระกันนั้น ได้แก่ ข้อมูลที่วัดมาจากกลุ่ม 2 กลุ่มที่มีลักษณะเหมือนกันหรือใกล้เคียงกันมาก หรือการเลือกมาเป็นคู่ๆ เช่น ฝาแฝด สามีภรรยา เป็นต้น หรือการวัด 2 ครั้งจากกลุ่มๆ เดียวกัน เช่น การทดลองก่อน-หลัง เป็นต้น ซึ่งก่อนการเลือกใช้สถิติทดสอบใด จะต้องพิจารณาข้อตกลงเบื้องต้นก่อน ในที่นี้คือ ข้อมูลต้องมีการแจกแจงเป็นโค้งปกติ และสถิติที่ใช้ในการทดสอบคือ T-test

ลักษณะของกลุ่มตัวอย่างที่ไม่เป็นอิสระจากกันหรือกล่าวได้ว่ามีความสัมพันธ์กัน มีหลายลักษณะ คือ

1) มีเพียงกลุ่มตัวอย่างเดียวแต่เก็บข้อมูล 2 ครั้ง เช่น การ Test–retest หรือ Before and After (Kohout, 1974) เช่น การทดสอบก่อนและหลังการเรียน (Pretest–Posttest) การทดสอบซ้ำของกลุ่มตัวอย่างเดียว เพื่อต้องการพิสูจน์ว่าวิธีการสอน มีผลต่อพัฒนาการการเรียนรู้ของนักเรียนหรือไม่

2) กลุ่มตัวอย่าง 2 กลุ่ม มีคุณลักษณะที่สำคัญบางประการเหมือนกันเป็นคู่ๆ (Matched) เช่น คู่แฝด

3) กลุ่มตัวอย่าง 2 กลุ่ม มีความสัมพันธ์กันอย่างใกล้ชิด เช่น การเปรียบเทียบความคิดเห็นทางการเมืองของสามีภรรยา

สูตรที่ใช้ทดสอบ $t = \dfrac{\sum D}{\sqrt{\dfrac{n\sum D^2 - (\sum D)^2}{n-1}}}$

df = n -1

D **แทนค่าผลต่างระหว่างคู่คะแนน**

n **แทนจำนวนคู่**

บทที่ 8

การวิเคราะห์ความแปรปรวน

หลักการวิเคราะห์ความแปรปรวน

การวิเคราะห์ความแปรปรวน (Analysis Of Variance : ANOVA) เป็นวิธีการทางสถิติที่พัฒนาขึ้นโดย R.A. Fisher เพื่อใช้ในการวิเคราะห์ข้อมูลตัวอย่าง ซึ่งครั้งแรกได้ประยุกต์ใช้กับการทดลองทางการเกษตร เช่นการเปรียบเทียบผลผลิตข้าวสาลีจากการใช้ปุ๋ยชนิดต่างๆ หรือการเปรียบเทียบเมล็ดพันธุ์ต่างๆ เป็นต้น ต่อมาก็ได้มีการประยุกต์ใช้กับการวิจัยทางวิทยาศาสตร์ในหลากหลายสาขา

ในการวิเคราะห์ความแปรปรวนจะเป็นวิธีการทางสถิติที่แยกความแปรปรวนทั้งหมด (Total Variation) ของข้อมูลออกมาเป็นส่วนๆ ตามสาเหตุต่างๆ กัน คือ

ความแปรปรวนทั้งหลาย = ความผันแปรจาก Treatment + ความผันแปรระหว่างหน่วยทดลอง + ความผิดพลาดจากการทดลอง

หรือ Total = Treatment + Inherent + Extraneous Variate

หรือ Total = Treatment + Experiment + Error

นั่นคือการวิเคราะห์ความแปรปรวน เป็นวิธีการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของประชากรตั้งแต่สองชุดขึ้นไป โดยแยกความแปรปรวนของข้อมูลที่ได้มาทั้งหมดออกเป็นส่วนๆ แต่ละส่วนใช้วัดการกระจายเฉพาะอย่างสามารถกำหนดสมมติฐานหลักในการทดสอบ ดังนี้

$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

และสมมติฐานรอง คือ

H_1 : มีค่าเฉลี่ยของประชากรอย่างน้อยหนึ่งกลุ่มที่มีค่าไม่เท่ากับค่าเฉลี่ยของประชากรกลุ่มอื่น

หรืออาจกล่าวได้ว่าเป็นการเปรียบเทียบค่าเฉลี่ยของประชากร k กลุ่ม โดยที่ $k>2$ (มากกว่าสองกลุ่มประชากร) ซึ่งสะดวกกว่าที่จะทดสอบสมมติฐานเกี่ยวกับค่าเฉลี่ยของประชากรทีละสองกลุ่มประชากร คือ $H_0 : \mu_1 = \mu_2$, $H_0 : \mu_1 = \mu_3$,..., $H_0 : \mu_1 = \mu_k$, $H_0 : \mu_2 = \mu_3$,... $H_0 : \mu_2 = \mu_k$,..., $H_0 : \mu_{k-1} = \mu_{k-2}$,..., $H_0 : \mu_{k-1} = \mu_k$

1. ความแปรปรวนรวม (Total group Variance หรือ mean square total : MST) เป็นความแปรปรวนอันเกิดจากคะแนนแต่ละตัวเบี่ยงเบนจากค่าเฉลี่ยของกลุ่มรวม

2. ความแปรปรวนระหว่างกลุ่ม (Between-Treatment Variance หรือ mean square between group : MSB) เป็นความแปรปรวนของค่าตัวแปรตามระหว่างตัวแปรต้นแต่ละกลุ่ม ซึ่งความแปรปรวนที่เกิดขึ้นนี้อาจเนื่องมาจากอิทธิพลของตัวแปรต้น หรืออาจเกิดจากอิทธิพลของตัวแปรต้นเกิดขึ้นอย่างสุ่ม

(chance) ที่เนื่องมาจากความแตกต่างระหว่างตัวอย่างหรือเกิดจากความคลาดเคลื่อนในการทดลอง เช่น ใช้การวัดที่แตกต่างกัน

3. ความแปรปรวนภายในกลุ่ม (Within Treatment Variance หรือ mean square within group : MSW) เป็นความแปรปรวนอันเกิดจากคะแนนแต่ละตัวในแต่ละกลุ่มเบี่ยงเบนจากค่าเฉลี่ยของแต่ละกลุ่ม ซึ่งความแปรปรวนชนิดนี้ คือ ความแปรปรวนของความคลาดเคลื่อนนั่นเอง

การวิเคราะห์ความแปรปรวน เป็นการนำความแปรปรวนของตัวแปรหรือปรากฏการณ์มาแบ่งเป็นส่วนๆ ตามความต้องการของผู้วิจัย แล้วนำความแปรปรวนแต่ละส่วนมาเปรียบเทียบกันตามวัตถุประสงค์ของการวิจัย กล่าวคือ การวิเคราะห์ความแปรปรวนเป็นการเปรียบเทียบความแปรปรวนอันเกิดจากความแตกต่างของค่าเฉลี่ยระหว่างกลุ่ม ซึ่งเป็นผลของตัวแปรอิสระกับความแปรปรวนภายในกลุ่มหรือความคลาดเคลื่อนว่าตัวใดมีค่ามากกว่ากัน ในการทดสอบสมมติฐานจึงใช้อัตราส่วนของความแปรปรวนอันเกิดจากความแตกต่างของค่าเฉลี่ยระหว่างกลุ่ม (Between-group Variance) กับความแปรปรวนภายในกลุ่ม อันเกิดจากความคลาดเคลื่อน (Within-group หรือ Error Variance) เป็นหลักในการตัดสินใจ และเรียกอัตราส่วนนี้ว่า F-ratio

$$F = \frac{\text{ความแปรปรวนระหว่างกลุ่ม}}{\text{ความแปรปรวนภายในกลุ่ม}}$$

โดยมี df = (df ระหว่างกลุ่ม, df ภายในกลุ่ม)

$$F = \frac{MSB}{MSW}$$

ข้อตกลงเบื้องต้นของการวิเคราะห์ความแปรปรวน

เนื่องจากการวิเคราะห์ความแปรปรวนใช้ F เป็นสถิติทดสอบ โดยที่ F เป็นอัตราส่วนระหว่างความแปรปรวนระหว่างกลุ่ม (Mean Square Between) กับความแปรปรวนภายในกลุ่ม (Mean Square Within) ก่อนการวิเคราะห์ความแปรปรวน จำเป็นต้องพิจารณาข้อตกลงเบื้องต้นของการวิเคราะห์ความแปรปรวน (Assumption of Analysis of Variance) ดังนี้

1. อิทธิพลจาก Treatment และสิ่งแวดล้อมอื่นๆ รวมกันได้วิธีบวก (Additive)

2. ตัวอย่างที่เลือกมาแต่ละประชากรที่นำมาทดสอบจะต้องเป็นตัวอย่างที่ได้มาอย่างไม่เจาะจง (แบบสุ่ม)

3. ตัวอย่างที่เลือกมาจากแต่ละประชากรนั้นมาจากประชากรที่มีการแจกแจงแบบปกติ (Normality)

4. ประชากรต่างๆ ที่นำมาทดสอบจะต้องมีค่าความแปรปรวนเท่ากัน $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 = \sigma^2$

ถ้าข้อมูลที่นำมาทดสอบมีความคลาดเคลื่อนไปจากข้อกำหนดในข้อใดข้อหนึ่งหรือหลายข้อก็จะทำให้ระดับนัยสำคัญ และความไวของ F หรือ T-test เปลี่ยนแปลงไป

ประเภทของของการวิเคราะห์ความแปรปรวน

การวิเคราะห์ความแปรปรวนแบ่งออกเป็น 2 ประเภทใหญ่ๆ ดังนี้

1. การวิเคราะห์ความแปรปรวนแบบทางเดียว (One way ANOVA)

การวิเคราะห์ความแปรปรวนแบบทางเดียวเป็นการวิเคราะห์ข้อมูลด้วยตัวแปรหรือหรือปัจจัยเดียว นั่นคือ พิจารณาความแตกต่างของข้อมูลจากปัจจัยที่มีผลต่อข้อมูลเพียงปัจจัยเดียวหรือวิเคราะห์ความแตกต่างของข้อมูลในระดับต่างๆ ของปัจจัย เช่น การทดสอบความแตกต่างระหว่างรายได้เฉลี่ยของพนักงานร้านขายอาหาร 3 แห่ง การเปรียบเทียบคะแนนสอบของนักศึกษา 4 กลุ่ม เป็นต้น

ความแปรปรวนทั้งหลาย = ความแปรผันระหว่างกลุ่มประชากร + ความแปรผันภายในกลุ่มระชากร

หรือ Total Variability = Between Groups Variability + Within Groups Variability

การวิเคราะห์ความแปรปรวนทางเดียว (Analysis in a One-Way classification problem หรือ Independent Group ANOVA หรือที่นิยมเรียกกันว่า One Way ANOVA) ใช้ในกรณีที่มีการจำแนกข้อมูลตามปัจจัยที่สนใจศึกษาเพียงปัจจัยเดียวเท่านั้น แต่มีหลายสิ่งทดลอง เช่น

ต้องทดสอบความแตกต่างของวิธีการเลี่ยงดูเด็กช่วงอายุ 1-2 ปี 5 วิธี

ต้องทดสอบความแตกต่างของผลผลิตทุเรียน 4 พันธุ์

ต้องทดสอบความแตกต่างของอายุการใช้งานของหลอดไฟ 4 ยี่ห้อ เป็นต้น

จากตัวอย่าง จะเห็นว่าวิธีการเลี้ยงดูเด็ก พันธุ์ทุเรียนและยี่ห้อของหลอดไฟ คือปัจจัยที่สนใจศึกษา ซึ่งผู้วิจัยสามารถกล่าวได้ว่า แต่ละกลุ่มของปัจจัยที่สนใจศึกษานั้น เป็นประชากรแต่ละประชากร และผู้วิจัยต้องการทดสอบว่าประชากรเหล่านั้น (ซึ่งมีผลมาจากปัจจัยต่างกันนั้น) มีค่าเฉลี่ยแตกต่างกันหรือไม่

ตัวอย่างสุ่มขนาด n ถูกเลือกมาจาก k ประชากร (ประชากร n_i โดยที่ i = 1, 2, ..., k) โดยที่ k ประชากรนี้เป็นอิสระจากกัน และตัวอย่างแต่ละกลุ่มมีการแจกแจงปกติ มีค่าเฉลี่ยเป็น $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$ และความแปรปรวนเป็น σ^2

ตารางที่ 2 แสดงข้อมูลที่จัดแบบจำแนกทางเดียว

ค่าสังเกต	ประชากร หรือสิ่งทดลอง (Treatment)						
	1	2	...	I	...	k	
1	X_{11}	X_{21}		X_{i1}		X_{k1}	
2	X_{12}	X_{22}		X_{i2}		X_{k2}	
.							

j	X_{1j}	X_{2j}		X_{ij}		X_{kj}	
.							
n	X_{1n_1}	X_{2n_2}		X_{in_i}		X_{kn_k}	
ผลรวม (Total)	$T_{1.}$	$T_{2.}$		$T_{i.}$		$T_{k.}$	$T_{..}$
ค่าเฉลี่ย (Mean)	$\overline{X}_{1.}$	$\overline{X}_{2.}$		$\overline{X}_{i.}$		$\overline{X}_{k.}$	$\overline{X}_{..}$

โดยที่ k คือ จำนวนสิ่งทดลอง (Treatment) หรือ จำนวนประชากร

n_i คือ จำนวนค่าสังเกตในสิ่งทดลอง ; $N = \sum_{i=1}^{k} n_i$

N คือ จำนวนค่าสังเกตทั้งหมด

X_{ij} คือ ค่าสังเกตที่ j จากสิ่งทดลองที่ i

$T_{i.}$ คือ ผลรวมของค่าสังเกตทุกๆ ค่าสังเกตจากสิ่งทดลองที่ i ; $T_i = \sum_{j=1}^{n_i} x_{ij}$

$\overline{X}_{i.}$ คือ ค่าเฉลี่ยของทุกๆ ค่าสังเกตจากสิ่งทดลองที่ i ; $\overline{X}_{i.} = \frac{\sum_{j=1}^{n} x_{ij}}{n_i} = \frac{T_{i.}}{n_i}$

$T_{..}$ คือ ผลรวมของค่าสังเกตทั้งหมด ; $T_{...} = \sum_{i=1}^{k} T_{i.} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$

$\overline{X}_{..}$ คือ ค่าเฉลี่ยของค่าสังเกตทั้งหมด ; $\overline{X}_{..} = \frac{T_{..}}{N} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}}{N}$

สิ่งที่ต้องการ คือต้องการทดสอบว่าค่าเฉลี่ยของประชากร k ประชากรเหล่านี้ แตกต่างกันอย่างมีนัยสำคัญหรือไม่

ตารางที่ 3 แสดงสูตรและตารางการวิเคราะห์ความแปรปรวนทางเดียว

Source of Variation	Degree of freedom	sum of squares	Mean square	F-statistic
Between groups (Treatment)	k-1	SSB= $\sum_{i=1}^{k} \frac{T_i^2}{n_i} - \frac{T^2}{n}$	MSB = $\frac{SSB}{k-1}$	F = $\frac{MSB}{MSW}$
Within groups (Error)	n-k	SSW = SST-SSA	MSW = $\frac{SSW}{n-k}$	
Total	n-1	SST= $\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2 - \frac{T^2}{n}$		

โดยที่ SST = SSB + SSW

จากตารางข้างต้นสามารถเขียนโดยสรุปได้ดังนี้

Source of Variance	Degree of freedom (df)	Sum Square (SS)	Mean Square (MS)	F-ratio
Between Groups	k-1	SSB	MSB	F= $\frac{MSB}{MSW}$
Within Groups	n-k	SSW	MSW	
Total	n-1	SST		

ตัวอย่างงานวิจัย

การเปรียบเทียบผลการเรียนรู้ด้วยบทเรียนบนระบบเครือข่ายรายวิชาการประยุกต์ใช้เทคโนโลยีสารสนเทศและการสื่อสาร ของนิสิตระดับบัณฑิตศึกษามหาวิทยาลัยมหาสารคามที่มีรูปแบบการเรียน (Learning Style) ต่างกัน

ตัวแปรอิสระหรือตัวแปรต้น ได้แก่ รูปแบบการเรียน (Learning Style) ซึ่งมี 4 แบบ คือ

1. แบบแอคคอมมอเดเตอร์ (Accommodator)

2. แบบไดเวอร์เจอร์ (Diverger)

3. แบบแอสซิมิเลเตอร์ (Assimilator)

4. แบบคอนเวอร์เจอร์ (Converger)

ตัวแปรตาม ได้แก่ ผลการเรียนรู้จากบทเรียนบนเครือข่าย รายวิชา 1601 505 การประยุกต์ใช้เทคโนโลยีสารสนเทศและการสื่อสาร

ตัวอย่างที่ 1 การเปรียบเทียบผลสัมฤทธิ์ทางการเรียนของนิสิตปริญญาโทที่เรียนด้วยบทเรียนบนเครือข่าย รายวิชา 1601 505 การประยุกต์ใช้เทคโนโลยีสารสนเทศและการสื่อสารที่มีรูปแบบการเรียนต่างกัน

แหล่งความแปรปรวน	SS	df	MS	F	P
ระหว่างกลุ่ม	1364.840	3	454.947	16.871	.000*
ภายในกลุ่ม	2588.800	96	26.967		
รวม	3953.640	99			

***มีนัยสำคัญทางสถิติที่ระดับ .05**

จากตารางดังกล่าวพบว่า นิสิตปริญญาโทที่เรียนด้วยบทเรียนบนเครือข่าย รายวิชา 1601 505 การประยุกต์ใช้เทคโนโลยีสารสนเทศและการสื่อสาร ที่มีรูปแบบการเรียนต่างกันมีผลสัมฤทธิ์ทางการเรียนแตกต่างกัน อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05

การทดสอบสมมติฐานและการสรุปผล โดยทั่วไปมีขั้นตอนการทดสอบ ดังนี้

1. กำหนดสมมติฐานที่ใช้ในการทดสอบ

$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

H_1 **: มีค่าเฉลี่ยของประชากรอย่างน้อยหนึ่งกลุ่มที่มีค่าไม่เท่ากับค่าเฉลี่ยของประชากรกลุ่มอื่น**

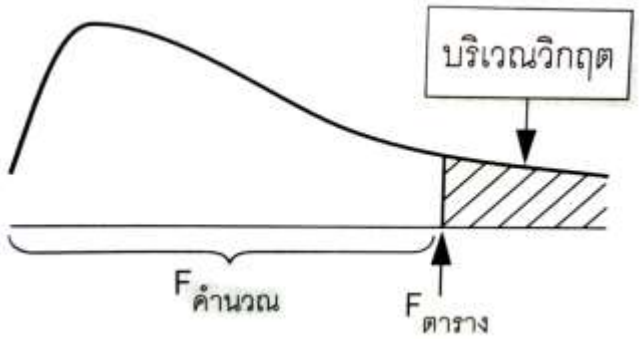
2. กำหนดระดับนัยสำคัญ (α)

3. สถิติที่ใช้ในการทดสอบ คือ F-Test และสร้างตาราง ANOVA (ได้ค่า $F_{\text{คำนวณ}}$**)**

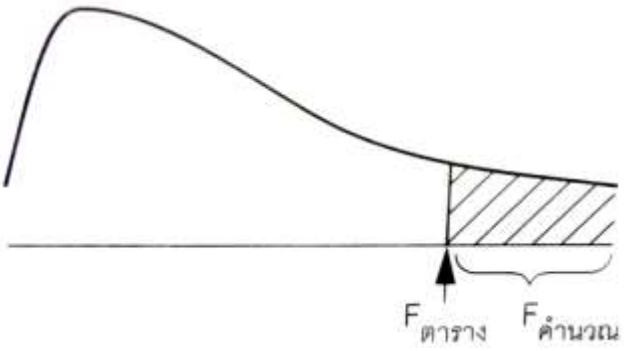
4. หาจุดวิกฤตและบริเวณวิกฤต โดยการเปิดตาราง F (ได้ค่า $F_{\text{ตาราง}}$**)**

5. สรุปผลการทดสอบ

- **ถ้า** $F_{\text{คำนวณ}} < F_{\text{ตาราง}}$ **จะ Accept** H_0 **นั่นคือ** $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$



- **ถ้า** $F_{\text{คำนวณ}} \geq F_{\text{ตาราง}}$ **จะ Reject** H_0 **นั่นคือ มีค่าเฉลี่ยของประชากรอย่างน้อยหนึ่งกลุ่มที่ต่างจากประชากรกลุ่มอื่น**



การเปรียบเทียบเชิงซ้อน

จากการทดสอบ ANOVA นี้ ถ้าผลที่ได้จากการทดสอบสมมติฐานคือ การปฏิเสธสมมติฐานหลักนั่นคือ ผู้วิจัยจะทราบแต่เพียงว่ามีอย่างน้อยหนึ่งกลุ่มที่แตกต่างจากกลุ่มอื่นๆ ซึ่งไม่ทราบว่าเป็นกลุ่มไหน ดังนั้นจึงมีวิธีการเปรียบเทียบความแตกต่างของค่าเฉลี่ยของประชากรต่างๆ เรียกว่า การเปรียบเทียบเชิงซ้อน (Multiple Comparison) ซึ่งสามารถบอกได้ว่าประชากรกลุ่มไหนที่มีความแตกต่างอย่างมีนัยสำคัญ วิธีที่ใช้ในการวิเคราะห์มีหลายวิธี เช่น

- **Least Square Difference (LSD)**
- **Duncan's new multiple range test**
- **Tukey's method**
- **Student – Newman – Keul (SNK)**
- **Scheffe's method**

นิยาม การเปรียบเทียบเชิงซ้อน หมายถึงการเปรียบเทียบระหว่างค่าเฉลี่ยของสิ่งทดลองต่างๆ เพื่อหาว่ามีสิ่งทดลองใดบ้างที่ส่งผลให้ H_0 ถูกปฏิเสธไป

สมมติฐานที่ใช้ในการวิเคราะห์

การเปรียบเทียบเชิงซ้อนนั้นเป็นการทดสอบสมมติฐานเกี่ยวกับความแตกต่างของค่าเฉลี่ยทีละคู่ ดังนี้

1. กำหนดสมมติฐานที่ใช้ในการทดสอบ

$H_0 : \mu_i = \mu_j \ ; i \neq j$

$H_1 : \mu_i \neq \mu_j \ ; i \neq j$

2. กำหนดระดับนัยสำคัญ (α)

3. สถิติที่ใช้ในการทดสอบ เช่น LSD

4. เปรียบเทียบผลต่างของค่าเฉลี่ยแต่ละคู่ กับค่าที่ได้ในข้อ 3

5. สรุปผลการทดสอบสมมติฐาน

ถ้าค่าสมบูรณ์ของผลต่างของค่าเฉลี่ย มีค่าน้อยกว่าหรือเท่ากับค่าในข้อ 3 แสดงว่ายอมรับ H_0

ถ้าค่าสมบูรณ์ของผลต่างของค่าเฉลี่ย มีค่ามากกว่าค่าในข้อ 3 แสดงว่าปฏิเสธ H_0

ตัวอย่างที่ 2 การเปรียบเทียบค่าเฉลี่ยเป็นคู่ๆ เทียบกับค่า LSD

$$\left| \overline{X}_D - \overline{X}_C \right| = \left| 2.8 - 4.0 \right| = 1.2 < LSD = 2.2336$$

$$\left| \overline{X}_A - \overline{X}_B \right| = \left| 2.8 - 5.2 \right| = 2.4 > LSD = 2.2336$$

$$\left| \overline{X}_D - \overline{X}_A \right| = \left| 2.8 - 7.8 \right| = 5.0 > LSD = 2.2336$$

$$|\overline{X}_D - \overline{X}_B| = |4.0 - 5.2| = 1.2 < LSD = 2.2336$$
$$|\overline{X}_C - \overline{X}_A| = |4.0 - 7.8| = 3.8 > LSD = 2.2336$$
$$|\overline{X}_C - \overline{X}_B| = |5.2 - 7.8| = 2.6 > LSD = 2.2336$$

$$\underline{\overline{X}_D \ \overline{X}_C} \quad \overline{\overline{X}_A \ \overline{X}_B}$$

หมายเหตุ การขีดเส้นใต้ค่าเฉลี่ย แสดงให้เห็นว่าค่าเฉลี่ยที่อยู่ภายใต้เส้นใต้เดียวกันจะมีค่าเฉลี่ยไม่แตกต่างกันหรือเท่ากันอย่างมีนัยสำคัญ

สรุปผล จากผลการวิเคราะห์ข้างต้น พบว่า

ค่าเฉลี่ยของ D ต่างจาก A, B อย่างมีนัยสำคัญ

ค่าเฉลี่ยของ C ต่างจาก B อย่างมีนัยสำคัญ

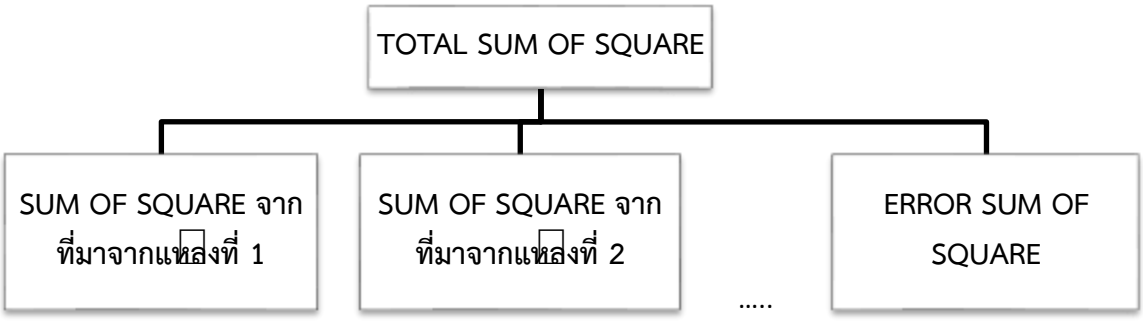
และ ค่าเฉลี่ยของ A ต่างจาก B อย่างมีนัยสำคัญ

หรือ อาจสรุปได้ว่า B มีค่าสูงสุด

2. การวิเคราะห์ความแปรปรวนแบบสองทาง (Multiple Factors ANOVA)

การวิเคราะห์ความแปรปรวนแบบหลายทางเป็นการศึกษาตัวแปรหรือปัจจัยที่มีผลกับข้อมูลมากกว่า 2 ปัจจัยขึ้นไป เช่น ระดับการศึกษาและเพศจะส่งผลร่วมกันต่อรายได้เฉลี่ยหรือไม่ การเปรียบเทียบนักศึกษา 4 กลุ่ม ที่มีการแยกกันสอบ 3 รอบ เป็นต้น

วิธีการวิเคราะห์ความแปรปรวน คือการแบ่ง SST ออกเป็นส่วนย่อยๆ ตามที่มาหรือแหล่งของข้อมูล และส่วนที่เนื่องจาก Random Errors จำนวนส่วนของความแปรปรวนที่จะแยกออกมานั้น ขึ้นอยู่กับแผนการทดลองที่ใช้ในการเก็บข้อมูล และตัวแบบทางสถิติ (Statistical Model) ที่เห็นว่าเหมาะสมกับการวิเคราะห์



การวิเคราะห์ความแปรปรวนแบบสองทาง (Two – way Analysis of Variance) เป็นการวิเคราะห์ข้อมูลจากหน่วยทดลองที่ได้จากการทดลองที่ประกอบด้วยปัจจัย 2 ปัจจัย โดยแต่ละปัจจัยยังแยกออกเป็นหลายระดับ หรือหลายประเภท หรือหลายชนิด ทั้งนี้ปัจจัยหนึ่งคือสิ่งทดลอง (Treatment) และอีกปัจจัยหนึ่งจะเรียกว่า กลุ่ม (Block) เช่น การเปรียบเทียบผลผลิตข้าวสาลี ดังตัวอย่าง

ตัวอย่างที่ 3 ถ้าต้องการเปรียบเทียบผลผลิตของแตงโม 4 พันธุ์ (ก, ข, ค, ง) โดยมีพื้นที่สำหรับปลูกแตงโม 5 แห่ง

พื้นที่ (Block)

1	2	3	4	5
ก	ก	ก	ก	ก
ข	ข	ข	ข	ข
ค	ค	ค	ค	ค
ง	ง	ง	ง	ง

ในที่นี้ถ้าผู้วิจัยต้องการศึกษาว่าพันธุ์แตงโมมีผลต่อผลผลิตหรือไม่ โดยเปรียบเทียบพันธุ์แตงโม 4 พันธุ์ (พันธุ์แตงโมเป็น Treatment) ปลูกแตงโมในพื้นที่ที่ต่างกัน 5 แห่ง (พื้นที่ที่ใช้ปลูกแตงโมเป็น Block) การปลูกแตงโมทั้ง 4 พันธุ์ ในแต่ละพื้นที่จะสุ่มปลูก ความแปรผันหรือความแปรปรวนทั้งหมดของผลผลิตแตงโม เกิดจาก 3 แหล่ง แหล่งแรกคือความแปรปรวนที่เกิดจากสิ่งทดลองหรือพันธุ์แตงโมที่แตกต่างกัน แหล่งที่ 2 คือ ความแปรปรวนที่เกิดจากกลุ่มหรือพื้นที่ที่ใช้ปลูกแตงโมที่แตกต่างกัน และแหล่งที่ 3 คือ ความแปรปรวนที่เกิดจากความคลาดเคลื่อนในการทดลอง

ในกรณีข้อมูลที่จัดแบบจำแนกสองทาง จากการทดลองหนึ่งซึ่งประกอบด้วย k สิ่งทดลอง และ b กลุ่ม ได้ตัวอย่างขนาด bk ตัวอย่าง ดังตารางต่อไปนี้

กลุ่ม (Block)	ประชากร หรือสิ่งทดลอง (Treatment)						รวม (Total)	ค่าเฉลี่ย (Mean)
	1	2	...	I	...	k		
1	X_{11}	X_{21}		X_{i1}		X_{k1}	$T_{.1}$	$\overline{X}_{.1}$
2	X_{12}	X_{22}		X_{i2}		X_{k2}	$T_{.2}$	$\overline{X}_{.2}$
.							.	
J	X_{1j}	X_{2j}		X_{ij}		X_{kj}	$T_{.j}$	$\overline{X}_{.j}$
.							.	
B	X_{1b}	X_{2b}		X_{ib}		X_{kb}	$T_{.b}$	$\overline{X}_{.b}$
รวม (Total)	$T_{1.}$	$T_{2.}$		$T_{i.}$		$T_{k.}$	$T_{..}$	
ค่าเฉลี่ย (Mean)	$\overline{X}_{1.}$	$\overline{X}_{2.}$		$\overline{X}_{i.}$		$\overline{X}_{k.}$		$\overline{X}_{..}$

โดยที่ k คือ จำนวนสิ่งทดลอง (Treatment)

b คือ จำนวนกลุ่ม (Block)

X_{ij} คือ ค่าสังเกตจากสิ่งทดลองที่ i ในกลุ่มที่ j

T_{ij} คือ ผลรวมของค่าสังเกตทุกๆ ค่าสังเกต จากสิ่งทดลองที่ i ; $T_{i.} = \sum_{j=1}^{b} X_{ij}$

$T_{i.}$ คือ ผลรวมของค่าสังเกตทุกๆ ค่าสังเกต จากกลุ่มที่ j ; $T_{.j} = \sum_{i=1}^{k} X_{ij}$

$T_{.j}$ คือ ค่าเฉลี่ยของทุกๆ ค่าสังเกต จากสิ่งทดลองที่ j ; $X_{i.} = \frac{\sum_{j=1}^{b} X_{ij}}{b} = \frac{T_{i.}}{b}$

$\overline{X}_{i.}$ คือ ค่าเฉลี่ยของทุกๆค่าสังเกต จากกลุ่มที่ i ; $\overline{X}_{.j} = \frac{\sum_{i=1}^{k} X_{ij}}{k} = \frac{T_{.j}}{k}$

$\overline{X}_{.j}$ คือ ผลรวมของค่าสังเกตทั้งหมด ; $T_{..} = \sum_{i=1}^{k} T_{i.} = \sum_{i=1}^{k} \sum_{j=1}^{b} X_{ij}$

$$\overline{X}_{..} = \frac{T_{..}}{N} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}}{N}$$

ตัวอย่างหัวข้อวิจัยที่เหมาะสมสำหรับการวิเคราะห์ความแปรปรวนสองทาง

การเปรียบเทียบสมรรถนะของไมโครโพรเซสเซอร์ของ Intel, Cyrix, AMD และ Motorola ภายใต้สภาพการใช้งานที่อุณหภูมิแตกต่างกันสองสภาวะ

ตัวแปรอิสระหรือตัวแปรต้น : ตัวที่ 1 : ไมโครโพรเซสเซอร์ จำแนกออกเป็น 4 ชนิด ได้แก่

- Intel
- Cyrix
- AMD
- Motorola

ตัวที่ 2 : อุณหภูมิ จำแนกออกเป็น 2 ระดับ ได้แก่

- เย็น
- ร้อน

ตัวแปรตาม : สมรรถนะ

ตัวอย่างงานวิจัยและการวิเคราะห์ Two-way ANOVA

สำหรับการวิเคราะห์ความแปรปรวนสองทางมีหลายวิธี เช่น วิธี CRF (Completed Randomized Factorials) และวิธี RMF เป็นต้น ในที่นี้จะขอกล่าวถึงวิธี CRF ซึ่งเป็นวิธีการวิเคราะห์ความแปรปรวนสองทางที่นิยมใช้กันโดยทั่วไปการวิเคราะห์ความแปรปรวนสองทางวิธี CRF

การวิเคราะห์ความแปรปรวนสองทางวิธี CRF หรือวิธี $CRF_{-a,b}$ เป็นวิธีการวิเคราะห์ความแปรปรวนสองทางทั่วๆ ไป

โดยที่ a เป็นจำนวนของตัวแปรอิสระหรือตัวแปรต้นตัวที่ 1

ส่วน b เป็นจำนวนของตัวแปรอิสระหรือตัวแปรต้นตัวที่ 2

จากตัวอย่างในการพัฒนาบทเรียนคอมพิวเตอร์ 2 แบบ ได้แก่ แบบปกติและแบบมัลติมีเดีย เพื่อนำไปทดลองใช้กับผู้เรียนที่มีความสามารถทางการเรียนแตกต่างกัน 3 กลุ่ม ได้แก่ กลุ่มเก่ง ปานกลางและอ่อน โดยที่ผู้วิจัยต้องการศึกษาผลสัมฤทธิ์ทางการเรียนที่เกิดขึ้นว่าแตกต่างกันหรือไม่อย่างไรซึ่งผลจากการทดลองใช้บทเรียนคอมพิวเตอร์ทั้ง 2 แบบ กับกลุ่มตัวอย่างที่มีจำนวนกลุ่มละ 4 คน รวมทั้งหมด 24 คน ได้ผลคะแนนปรากฏดังตารางต่อไปนี้

	B_1		B_2		B_3	
A_1	2	3	0	3	9	10
	4	5	6	4	12	15
A_2	2	1	3	6	8	9
	3	2	7	1	11	6

ทดสอบสมมติฐานระดับนัยสำคัญที่ 0.01

ตัวแปรอิสระหรือตัวแปรต้นที่ 1 จำนวน 2 ปัจจัย ได้แก่

A1 เป็นบทเรียนคอมพิวเตอร์แบบปกติ

A2 เป็นบทเรียนคอมพิวเตอร์แบบมัลติมีเดีย

ตัวแปรอิสระหรือตัวแปรต้นที่ 2 จำนวน 3 ปัจจัย ได้แก่

B_1 เป็นความสามารถทางการเรียนระดับเก่ง

B_2 เป็นความสามารถทางการเรียนระดับปานกลาง

B_3 เป็นความสามารถทางการเรียนระดับอ่อน

ตัวแปรตาม : ผลสัมฤทธิ์ทางการเรียน

สรุปผลการทดสอบสมมติฐานได้ทั้ง 3 กรณี ดังต่อไปนี้

1. บทเรียนคอมพิวเตอร์แบบปกติและบทเรียนคอมพิวเตอร์แบบมัลติมีเดีย ให้ผลสัมฤทธิ์ทางการเรียนที่ไม่แตกต่างกัน (หรือเท่ากัน)

2. กลุ่มผู้เรียนที่มีความสามารถทางการเรียนแตกต่างกัน ได้แก่ กลุ่มเก่ง ปานกลางและอ่อน จะมีผลสัมฤทธิ์ทางการเรียนแตกต่างกัน (หรือไม่เท่ากัน)

3. ไม่มีผลร่วมหรือปฏิสัมพันธ์ใดๆ เกิดขึ้นระหว่างบทเรียนคอมพิวเตอร์แบบปกติกับบทเรียนคอมพิวเตอร์แบบมัลติมีเดีย

ในกรณีนี้ สามารถสรุปเป็นตารางที่ระดับนัยสำคัญ 0.01 ได้ดังนี้

สาเหตุของความแปรปรวน	Sum of Squares	df	Mean Squares	F	ค่าวิกฤติของ F
A	8.16	1	8.16	1.78	8.29
B	247	2	123.50	26.94**	6.01
AB	16.34	2	8.17	1.78	6.01
W	82.50	18	4.583		
Total	354	23			

** มีนัยสำคัญที่ระดับ 0.01

2. สมมติฐานที่ใช้ในการทดสอบ

1. กำหนดสมมติฐานที่ใช้ในการทดสอบ

สำหรับการดสอบ Treatment

$H_0 : \mu_{1.} = \mu_{2.} = \mu_{3.} = \ldots = \mu_{k.}$

H_1 : มีค่าเฉลี่ยอย่างน้อยหนึ่งสิ่งทดลอง (Treatment) ที่มีค่าแตกต่างจากสิ่งทดลองกลุ่มอื่น

สำหรับการดสอบ Block

$H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3} = \ldots = \mu_{.b}$

H_1 : มีค่าเฉลี่ยอย่างน้อยหนึ่งกลุ่ม (Block) ที่มีค่าแตกต่างจากกลุ่มอื่น

2. กำหนดระดับนัยสำคัญ (α)

3. สถิติที่ใช้ในการทดสอบ คือ F-Test โดยการสร้างตาราง ANOVA

4. หาจุดวิกฤติและบริเวณวิกฤต

5. สรุปผลการทดสอบ

การวิเคราะห์เพิ่มเติม

ถ้าปฏิเสธสมมติฐานว่ามีอย่างน้อย 1 กลุ่มที่มีค่าเฉลี่ยต่างจากกลุ่มอื่น ผู้วิจัยยังสามารถวิเคราะห์ต่อว่ากลุ่มใดที่แตกต่างจากกลุ่มอื่น โดยวิธีเปรียบเทียบเชิงซ้อน การวิเคราะห์ทำเช่นเดียวกับกรณีการวิเคราะห์ความแปรปรวนทางเดียวซึ่งจะต้องทดสอบหลังการวิเคราะห์ (Post hoc test or postetior) โดยวิธีการเปรียบเทียบพหุคูณ (Multiple comparison) ซึ่งมีหลายวิธีในโปรแกรม SPSS ที่นิยมใช้ โดยแบ่งออกเป็น 2 กลุ่มใหญ่ๆ คือ

กลุ่มที่ 1 การเปรียบเทียบรายคู่ เมื่อค่าความแปรปรวนแต่ละกลุ่มไม่แตกต่างกัน ได้แก่

LSD (Least-significant), Boforroni, Sidak, Shceffe’s, RE–D–WF, R-E-G-WQ, S-N-K (Student-NBewman-Keuls), Turkey, Turkey’s-b, Ducan, Hochberg’s GT2, Gabriel, Waller–Duncan, Dunett

กลุ่มที่ 2 การเปรียบเทียบรายคู่ เมื่อค่าความแปรปรวนแต่ละกลุ่มแตกต่างกัน ได้แก่

Tamhane’s t2, Dunnett’s t3, Gamea-Howell, Dunnett’s C

สรุป

การวิเคราะห์ความแปรปรวน (Analysis of Variance) จะวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระตัวเดียวกับตัวแปรตามเพียงตัวเดียว โดยที่ตัวแปรอิสระหรือตัวแปรต้นอาจมีลักษณะเป็นตัวแปรเชิงคุณภาพ (Qualitative Variable) ที่จำแนกออกเป็นระดับหรือประเภทต่างๆ เช่น เก่ง-ปานกลาง-อ่อน ดีมาก-ดี-พอใช้-แย่ เป็นต้น ส่วนตัวแปรตามอาจมีลักษณะเป็นตัวแปรเชิงปริมาณ (Quantitative Variable) เพื่อศึกษาความสัมพันธ์ของตัวแปรอิสระหรือตัวแปรต้นว่าจะส่งผลอย่างไรกับตัวแปรตามตามสมมติฐานการวิจัยที่กำหนดไว้

บทที่ 9

การวิเคราะห์ความถดถอยและสหสัมพันธ์

การวิเคราะห์สหสัมพันธ์ของ Pearson

การวิเคราะห์สหสัมพันธ์ตามวิธีของ Pearson เป็นเทคนิคการวิเคราะห์ที่ให้ค่าระดับความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณทั้งสองว่ามีมากน้อยเพียงใดโดยไม่สนใจที่จะพยากรณ์ค่าของตัวแปร ตัวอย่างเช่น ความสัมพันธ์ระหว่างยอดขายกับค่าใช้จ่ายในการโฆษณาสินค้า ความสัมพันธ์ระหว่างรายได้และการประมาณการบริโภคสินค้าแต่ละชนิด หรือความสัมพันธ์ระหว่างคะแนนสอบความถนัดและปริมาณการผลิตของคนงานในโรงงานอุตสาหกรรม ระดับของความสัมพันธ์ที่ต้องการทราบนี้วัดได้จากค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient) กล่าวคือ

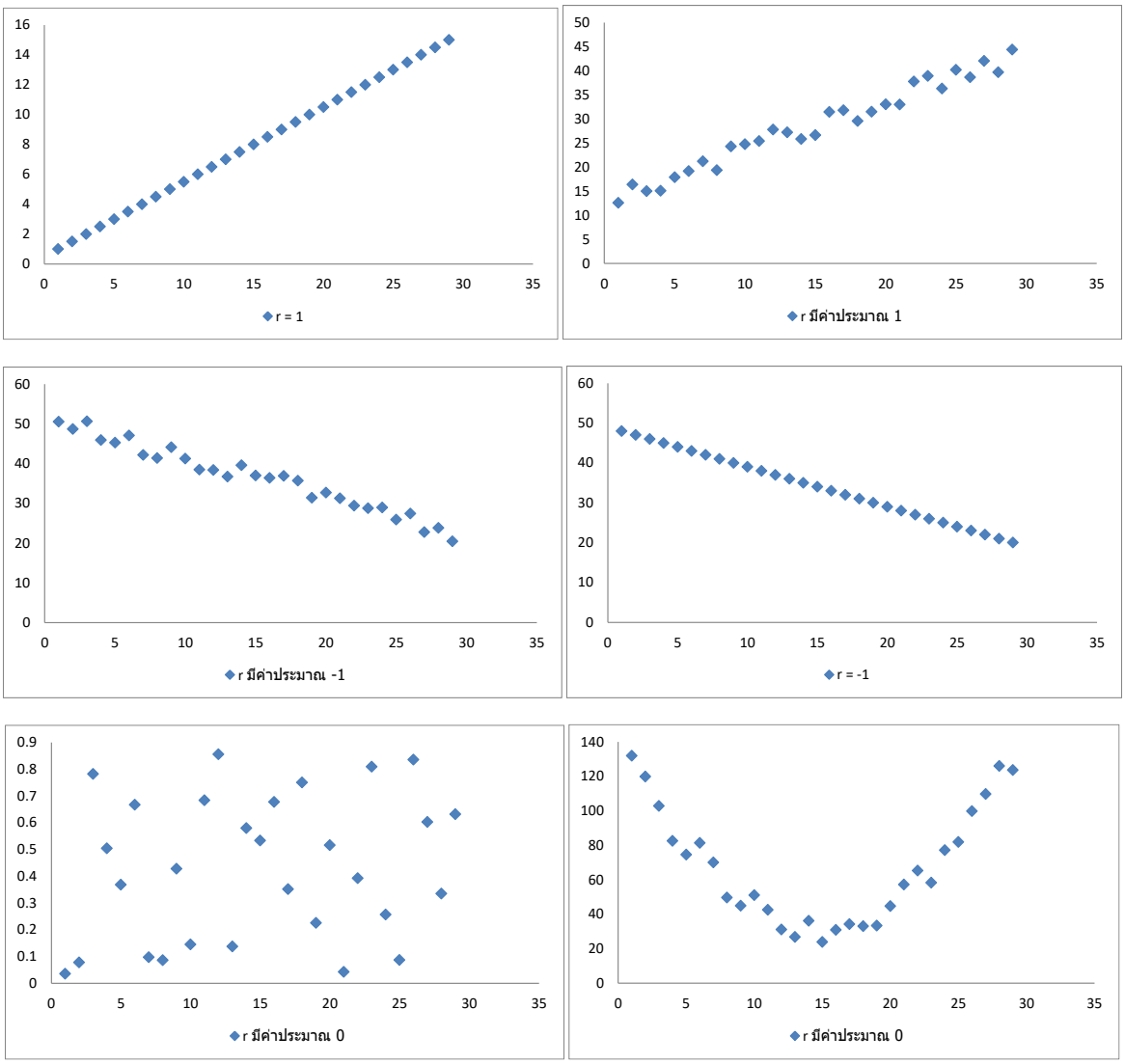
ถ้า X และ Y เป็นตัวแปรสองตัว n คือ จำนวนตัวอย่างที่นำมาวัดค่าตัวแปรทั้งสอง สัมประสิทธิ์สหสัมพันธ์ระหว่าง X และ Y เขียนแทนด้วยสัญลักษณ์ r ซึ่งมีสูตรที่ใช้ในการคำนวณหา ดังนี้

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

สัมประสิทธิ์สหสัมพันธ์มีค่าได้ตั้งแต่ -1 ถึง 1 ($-1 \le r \le 1$) ถ้าสัมประสิทธิ์สหสัมพันธ์มีค่าเป็นบวกแสดงว่าเมื่อตัวแปรตัวใดตัวหนึ่งมีค่าเพิ่มขึ้น ตัวแปรอีกตัวหนึ่งจะมีค่าเพิ่มขึ้นด้วย แต่ถ้าสัมประสิทธิ์สหสัมพันธ์มีค่าเป็นลบแสดงว่าเมื่อตัวแปรตัวใดตัวหนึ่งมีค่าเพิ่มขึ้น ตัวแปรอีกตัวหนึ่งจะมีค่าลดลงคือ เป็นไปในทางตรงกันข้าม

สำหรับขนาดหรือระดับของความสัมพันธ์นั้นพิจารณาจากค่าของสัมประสิทธิ์สหสัมพันธ์ ถ้าสัมประสิทธิ์สหสัมพันธ์มีค่าเข้าใกล้ +1 หรือ -1 แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันมาก แต่ถ้าสัมประสิทธิ์สหสัมพันธ์มีค่าเข้าใกล้ 0 แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันน้อย ในกรณีที่สัมประสิทธิ์สหสัมพันธ์มีค่าเท่ากับ 0 แสดงว่าตัวแปรทั้งสองไม่มีความสัมพันธ์กันเลย

ลักษณะของความสัมพันธ์ระหว่างตัวแปร X และ Y ที่มี r เท่ากับ +1, r มีค่าประมาณ -1 และ r มีค่าประมาณ 0 แสดงไว้ในภาพ



การวัดความสัมพันธ์โดยใช้สัมประสิทธิ์สหสัมพันธ์ข้างต้นนี้เป็นการวัดความสัมพันธ์เชิงเส้นตรงเท่านั้น ถ้า r มีค่าเข้าใกล้ 0 แสดงว่าตัวแปรทั้งสองไม่มีความสัมพันธ์เชิงเส้นตรงเท่านั้น แต่ตัวแปรทั้งสองอาจมีความสัมพันธ์เชิงเส้นโค้งก็ได้

กำลังสองของสัมประสิทธิ์สหสัมพันธ์ r^2 เรียกว่า สัมประสิทธิ์การตัดสินใจ (coefficient of determination) ใช้วัดอิทธิพลของตัวแปรตัวหนึ่งว่ามีผลต่อตัวแปรอีกตัวหนึ่งมากน้อยเพียงใด สัมประสิทธิ์การตัดสินใจมีค่าได้ตั้งแต่ 0 ถึง 1

ตัวอย่างที่ 1 เกษตรกรต้องการศึกษาความสัมพันธ์ระหว่างความสูงของต้นกล้าไม้ชนิดหนึ่งและอายุ จึงสุ่มเก็บรวบรวมข้อมูลเกี่ยวกับความสูงและอายุของต้นกล้าจำนวน 8 ต้น ปรากฏดังนี้

ต้นกล้า	1	2	3	4	5	6	7	8
อายุ (สัปดาห์)	5	7	10	12	15	20	25	30
ความสูง (ซม.)	40	50	60	65	70	80	92	100

จงคำนวณหาสัมประสิทธิ์สหสัมพันธ์ (r) และสัมประสิทธิ์การตัดสินใจ (r^2) ของความสัมพันธ์ระหว่างความสูงของต้นกล้าไม้ชนิดนี้และอายุ

วิธีทำ

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

เมื่อ $n = 8$ **และ** $\sum_{i=1}^{n} x_i y_i$, $\bar{x}$, $\bar{y}$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i^2$ **หาได้จากตารางต่อไปนี้**

y_i	x_i	y_i^2	x_i^2	$x_i y_i$
40	5	1,600	25	200
50	7	2,500	49	350
60	10	3,600	100	600
65	12	4,225	144	780
70	15	4,900	225	1,050
80	20	6,400	400	1,600
92	25	8,464	625	2,300
100	30	10,000	900	3,000
รวม 557	รวม 124	รวม 41,689	รวม 2,468	รวม 9,880

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8}(124) = 15.50$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{8}(557) = 69.625$$

ดังนั้น
$$r = \frac{9,880 - 8(15.5)(69.625)}{\sqrt{2,468 - 8(15.5^2)}\sqrt{41,689 - 8(69.625^2)}}$$

$$= \frac{9,880 - 8,633.5}{\sqrt{546}\ \sqrt{2,907.875}}$$

$$= \frac{1,246.5}{(23.3666)\ (53.9247)}$$

$$= 0.9893$$

$$r^2 = (0.9893^2) = 0.9787$$

นั่นคือ ค่าใช้จ่ายในการโฆษณาที่เพิ่มขึ้นหรือลดลงมีผลทำให้ยอดขายสินค้าเพิ่มขึ้นหรือลดลงถึงร้อยละ 97.87 ที่เหลืออีกร้อยละ 2.13 (100 – 97.87) เป็นผลเนื่องมาจากสาเหตุอื่นๆ ที่ทำให้ยอดขายของสินค้าเพิ่มขึ้นหรือลดลง

เนื่องจาก r = 0.9893 มีค่าเป็นบวก ดังนั้นเมื่อค่าใช้จ่ายในการโฆษณาเพิ่มขึ้นจะมีผลทำให้ยอดขายสินค้าเพิ่มขึ้นตามหรือเมื่อค่าใช้จ่ายในการโฆษณาลดลงยอดขายสินค้าจะลดลงตามไปด้วย ในกรณีนี้เนื่องจาก r มีค่าเข้าใกล้ 1 มาก อาจสรุปได้ว่ายอดขายสินค้ามีความสัมพันธ์กับค่าใช้จ่ายในการโฆษณามาก

ตัวอย่างที่ 2 ในการวิเคราะห์ความสัมพันธ์ระหว่างความสูงกับเส้นผ่าศูนย์กลางลำต้นของต้นยางนา นักวิจัยได้เลือกตัวอย่างต้นยางนาโดยการสุ่มมาจำนวน 6 ต้น จากการเก็บข้อมูลความสูงและเส้นผ่าศูนย์กลางลำต้นปรากฏเป็นดังนี้

เส้นผ่าศูนย์กลางลำต้น (เซนติเมตร)	38	29	47	41	23	32
ความสูง (เซนติเมตร)	2,500	2,800	3,700	4,400	2,000	2,600

ความสูงกับเส้นผ่าศูนย์กลางลำต้นของต้นยางนามีความสัมพันธ์กันหรือไม่เพียงใด

วิธีทำ

ให้ X และ Y เป็นตัวแปรที่แทนเส้นผ่าศูนย์กลางลำต้น และความสูงของต้นยางนา ดังนั้นสัมประสิทธิ์สหสัมพันธ์ระหว่างเส้นผ่าศูนย์กลางลำต้นและความสูงของต้นยางนาคือ

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

เมื่อ n = 6 และ $\sum_{i=1}^{n} x_i y_i$, $\bar{x}$, $\bar{y}$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i^2$ หาได้จากตารางต่อไปนี้

y_i	x_i	y_i^2	x_i^2	x_iy_i
2,500	38	6,250,000	1,444	95,000
2,800	29	7,840,000	841	81,200
3,700	47	13,690,000	2,209	173,900
4,400	41	19,360,000	1,681	180,400
2,000	23	4,000,000	529	46,000
2,600	32	6,760,000	1,024	83,200
รวม 18,000	รวม 210	รวม 57,900,000	รวม 7,728	รวม 659,700

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}(210) = 35$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{6}(18,000) = 3,000$$

ดังนั้น
$$r = \frac{659,700 - 6(35)(3,000)}{\sqrt{7,728 - 6(35^2)}\sqrt{57,900,000 - 6(3,000^2)}}$$

$$= \frac{29,700}{\sqrt{378}\ \sqrt{3,900,000}}$$

$$= 0.77$$

$$r^2 = (0.77^2)$$

$$= 0.5929$$

นั่นคือ เส้นผ่าศูนย์กลางลำต้นมีความสัมพันธ์กับความสูงของต้นยางนาเพียงร้อยละ 59.29 ที่เหลืออีกร้อยละ 40.71 เป็นผลเนื่องมาจากอิทธิพลอื่นๆ เช่น สภาพดิน น้ำ ศัตรูตามธรรมชาติ

การวิเคราะห์สหสัมพันธ์เมื่อมีตัวแปรอิสระมากกว่าหนึ่งตัว

โดยทั่วไปการหาสัมประสิทธิ์สหสัมพันธ์ไม่ว่าตัวแปรอิสระจะมีจำนวนมากน้อยเพียงใดก็ตามหาได้จาก

$$r = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

เมื่อ $\hat{y}_i$ คือ ค่าประมาณจากตัวอย่างที่ i ซึ่งได้จากสมการเส้นถดถอย y_i คือ ค่าจากตัวอย่างที่ i ซึ่งเป็นข้อมูลที่เก็บรวบรวมได้และ $\bar{y}$ คือ ค่าเฉลี่ยของข้อมูลที่เก็บรวบรวมได้จากตัวอย่างทั้งหมด

ในกรณีที่มีตัวแปรอิสระ 2 ตัว คือ X_1 และ X_2 และมีสมการเส้นถดถอยเป็น

$$Y = b_0 + b_1X_1 + b_2X_2$$

อาจเขียนสัมประสิทธิ์สหสัมพันธ์ในเทอม x_{1i}, x_{2i}, y_i และ b_1, b_2 ได้ดังนี้

$$r = \sqrt{\frac{b_1 \sum_{i=1}^{n} x_{1i}y_i + b_2 \sum_{i=1}^{n} x_{2i}y_i}{\sum_{i=1}^{n} y_i^2}}$$

สัมประสิทธิ์การตัดสินใจสามารถหาได้โดยการยกกำลังสองค่า r เช่นเดียวกับการถดถอยเชิงเส้นตรง

ตัวอย่างที่ 3 ในการหาความสัมพันธ์ระหว่างราคาไม้ประดับกับจำนวนกิ่งและจำนวนดอก ได้เลือกไม้ประดับโดยการสุ่มมาจำนวน 8 ต้น ผลปรากฏดังนี้

จำนวนกิ่ง	จำนวนดอก	ราคา (บาท)
3	2	338
2	1	293
4	3	388
2	1	292
3	2	347
2	2	299
5	3	434
4	2	379

จงหาสัมประสิทธิ์สหสัมพันธ์ระหว่างราคาไม้ประดับกับจำนวนกิ่งและจำนวนดอก แล้วหาสัมประสิทธิ์การตัดสินใจ

วิธีทำ

เนื่องจากมีตัวแปรอิสระ 2 ตัว คือ X1 และ X2 จะได้

$$r = \sqrt{\frac{b_1 \sum_{i=1}^{n} x_{1i}y_i + b_2 \sum_{i=1}^{n} x_{2i}y_i}{\sum_{i=1}^{n} y_i^2}}$$

เมื่อ $b_1 = 41.49$, $b_2 = 7.31$, $\sum_{i=1}^{n} x_{1i}y_i = 9{,}061$, $\sum_{i=1}^{n} x_{2i}y_i = 5{,}777$, $\sum_{i=1}^{n} y_i^2 = 977{,}708$

ดังนั้น

$$r = \sqrt{\frac{41.49(9{,}061) + 7.31(5{,}777)}{977{,}708}}$$

$$= 0.6540$$

$$r^2 = 0.4277$$

นั่นคือ จำนวนกิ่งและจำนวนดอกของไม้ประดับมีผลทำให้ราคาของไม้ประดับเพิ่มขึ้นหรือลดลงเพียงร้อยละ 42.77 เท่านั้น ส่วนที่เหลืออีกร้อยละ (100 - 42.77) = 57.23 เป็นผลเนื่องมาจากสาเหตุอื่น

ตัวอย่างที่ 4 จากการศึกษาผลของฮอร์โมน (X_1) น้ำหนักเมล็ด (X_2) และอายุการเก็บรักษาภายหลังการเก็บเกี่ยว (X_3) ที่มีต่อระยะเวลาที่เมล็ดเริ่มงอก (Y) ของเมล็ดพันธุ์พืชหายากชนิดหนึ่งที่เลือกมาโดยการสุ่มจำนวน 25 เมล็ด โดยที่ $X_1 = 1$ เมื่อมีการใช้ฮอร์โมนกระตุ้น และ $X_1 = 0$ เมื่อไม่มีการใช้ฮอร์โมนกระตุ้น อายุการเก็บรักษาเมล็ดมีหน่วยเป็นเดือน ผลปรากฏดังนี้

เมล็ด	จำนวนชั่วโมงที่เมล็ดเริ่มงอก	ใช้ฮอร์โมน	น้ำหนักเมล็ด	อายุการเก็บรักษา
1	0.5	1	73	14
2	0.5	1	66	16
3	0.7	0	65	15
4	0.8	0	65	16
5	0.8	1	68	9
6	0.9	1	69	10
7	1.1	1	82	12
8	1.6	1	83	12
9	1.6	1	81	12
10	2.0	0	72	10
11	2.5	1	69	8
12	2.8	0	71	16
13	2.8	0	71	12
14	3.0	0	80	9
15	3.0	0	73	6
16	3.0	0	75	6
17	3.2	0	76	10
18	3.2	0	78	6
19	3.3	1	79	6
20	3.3	0	79	4
21	3.4	1	78	6
22	3.5	0	76	9

เมล็ด	จำนวนชั่วโมงที่เมล็ดเริ่มงอก	ใช้ฮอร์โมน	น้ำหนักเมล็ด	อายุการเก็บรักษา
23	3.6	0	65	12
24	3.7	0	72	12
25	3.7	0	80	6

จงหาสัมประสิทธิ์สหสัมพันธ์ระหว่างจำนวนชั่วโมงที่เมล็ดเริ่มงอกกับการใช้ฮอร์โมน น้ำหนักเมล็ด และอายุการเก็บรักษาเมล็ดภายหลังการเก็บเกี่ยวของพืชหายากชนิดนี้ แล้วหาสัมประสิทธิ์การตัดสินใจ

วิธีทำ

เนื่องจากมีตัวแปรอิสระ 3 ตัว คือ x_1, x_2 และ x_3 จะหาสัมประสิทธิ์สหสัมพันธ์ได้จากสมการ

$$r = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

แทนค่า x_1, x_2 และ x_3 เมื่อ i = 1, 2, 3, ..., 25 ลงในสมการเส้นถดถอยเชิงซ้อนที่ได้ คือ

$$\hat{Y} = 1.41411 - 1.17396\, X_1 + 0.03971\, X_2 - 0.15106\, X_3$$

จะได้ค่า $\hat{y}_1, \hat{y}_2, ..., \hat{y}_{25}$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{25}(58.5) = 2.34$$

เมื่อแทนค่าเหล่านี้ลงในสูตรคำนวณหาค่า r ข้างต้น จะได้

$$r = \sqrt{\frac{19.972}{31.860}} = 0.7918$$

$$r^2 = 0.6269 = 62.69\%$$

ดังนั้น จำนวนชั่วโมงที่เมล็ดเริ่มงอกจะมากหรือน้อยขึ้นอยู่กับการใช้ฮอร์โมน, น้ำหนักเมล็ดและอายุเมล็ด 62.69% ที่เหลืออีกประมาณ (100 - 62.69) = 37.31% ขึ้นอยู่กับสาเหตุหรือปัจจัยอื่นๆ เช่น การให้น้ำ อุณหภูมิ เป็นต้น

การทดสอบสมมติฐานเกี่ยวกับค่าสหสัมพันธ์

การพยากรณ์ค่า y_i เป็นช่วงหรือการหาช่วงความเชื่อมั่นของ y_i ณ ระดับความเชื่อมั่น $(1 - \alpha)$ 100% ของความสัมพันธ์ที่อยู่ในรูปเส้นตรง

$$\hat{y}_i \pm t_{(\frac{\alpha}{2}, n-2)} s_{\hat{Y}}$$

เมื่อ $s^2_{\hat{Y}}$ = ความแปรปรวนจากตัวอย่างของ $\hat{Y}$

$$= s^2_{Y \cdot X} \left[\frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

โดยที่ $s^2_{Y \cdot X} = \frac{1}{n-2} \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 - b \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \right]$

การทดสอบสมมติฐานที่ว่าตัวแปร X และ Y มีความสัมพันธ์เชิงเส้นตรงตามที่คาดไว้หรือไม่ หรือตัวแปรทั้งสองไม่มีความสัมพันธ์กันกล่าวคือ ทดสอบว่าสัมประสิทธิ์การถดถอยของประชากร B เท่ากับ b_0 ตามที่คาดไว้หรือเท่ากับ 0 ซึ่งแสดงว่าไม่มีความสัมพันธ์ระหว่างตัวแปร X และ Y

สมมติฐานเพื่อการทดสอบคือ

H_0 : $B = b_0$

H_a : $B > b_0$ หรือ $B < b_0$ หรือ $B \neq b_0$

ในกรณีที่สัมประสิทธิ์การถดถอย b ที่หาได้จากข้อมูลที่มีค่าเข้าใกล้ 0 ควรทดสอบดูว่าตัวแปรทั้งสองมีความสัมพันธ์กันจริงหรือไม่ เพราะความสัมพันธ์ที่เกิดขึ้นและมีขนาดน้อยมากอาจจะเป็นผลเนื่องมาจากความคลาดเคลื่อนในการเลือกตัวอย่างก็ได้ ดังนั้นควรทดสอบสมมติฐาน

H_0 : $B = 0$

H_a : $B \neq 0$

เพื่อเป็นการยืนยันว่ามีความสัมพันธ์กันหรือไม่จึงมีความจำเป็น ตัวสถิติเพื่อทดสอบสมมติฐานดังกล่าว คือ

$$t = \frac{b}{s_b}$$

เมื่อ $s^2_b = \frac{s^2_{Y \cdot X}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

จะปฏิเสธ H_0 ถ้าค่าสถิติที่คำนวณได้มากกว่าค่า $t_{(1 - \frac{\alpha}{2}, n-2)}$ หรือน้อยกว่าค่า $t_{(\frac{\alpha}{2}, n-2)}$ เมื่อ α คือ ระดับนัยสำคัญของการทดสอบ

ตัวอย่างที่ 5 เกษตรกรต้องการศึกษาความสัมพันธ์ระหว่างความสูงของต้นกล้าไม้ชนิดหนึ่งและอายุ จึงสุ่มเก็บรวบรวมข้อมูลเกี่ยวกับความสูงและอายุของต้นกล้าจำนวน 8 ต้น ปรากฏดังนี้

ต้นกล้า	1	2	3	4	5	6	7	8

อายุ (สัปดาห์)	5	7	10	12	15	20	25	30
ความสูง (ซม.)	40	50	60	65	70	80	92	100

จงหาช่วงความเชื่อมั่นของความสูง Y ณ ระดับความเชื่อมั่น 95% เมื่อต้นกล้ามีอายุ 50 สัปดาห์

วิธีทำ

ในการพยากรณ์ความสูง Y แบบช่วง เมื่อกำหนดอายุต้นกล้า X เป็น 50 สัปดาห์ โดยใช้ระดับความเชื่อมั่น 95% หาได้จากการแทนค่า $\hat{Y}$ และ $s_{\hat{Y}}$ ลงในช่วงความเชื่อมั่น เมื่อ X = 50 (หน่วยของ X เป็นสัปดาห์)

$$\hat{Y} = t_{(\frac{\alpha}{2},\, n-2)} s_{\hat{Y}}$$

จาก $\hat{Y} = 34.2385 + 2.283\, X$

$= 34.2385 + 2.283\,(50)$

$= 148.3885$ ซม.

$$s^2_{\hat{Y}} = s^2_{Y\cdot X}\left[\frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$= s^2_{Y\cdot X}\left[\frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}\right]$$

$$s^2_{Y\cdot X} = \frac{1}{n-2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 - b\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]$$

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

แต่ $n = 8$, $\sum_{i=1}^{n} y_i^2 = 41{,}689$, $\bar{y} = \frac{1}{8}(557) = 69.625$, $b = 2.283$, $\sum_{i=1}^{n} x_i y_i = 9{,}880$, $\bar{x} = \frac{1}{8}(124) = 15.5$, $\sum_{i=1}^{n} x_i^2 = 2{,}468$, $x_i = 50$

จะได้ $s^2_{Y\cdot X} = \frac{1}{8-2}[(41{,}689 - 8(69.625^2)) - 2.283(9{,}880 - 8(15.5)(69.625))]$

$= \frac{1}{6}[2{,}907.875 - 2.283\,(1{,}246.5)]$

$= 10.35258$

$$s^2_{\hat{Y}} = 10.35258\left[\frac{1}{8} + \frac{(50^2)}{2{,}468 - 8(15.5^2)}\right]$$

$= 10.35258\,(4.7038)$

$= 48.6965$

$s_{\hat{Y}}$ = 6.9783

ดังนั้น ช่วงความเชื่อมั่นของ Y ณ ระดับความเชื่อมั่น 95% คือ $148.3885 \pm 2.4469\ (6.9783)$

= (131.3133 , 165.4637) ซม.

นั่นคือ ความสูงของต้นกล้าเมื่อมีอายุ 50 สัปดาห์ อยู่ระหว่าง 131.3133 และ 165.4637 ซม.

ตัวอย่างที่ 6 เกษตรกรต้องการศึกษาความสัมพันธ์ระหว่างความสูงของต้นกล้าไม้ชนิดหนึ่งและอายุจึงสุ่มเก็บรวบรวมข้อมูลเกี่ยวกับความสูงและอายุของต้นกล้าจำนวน 8 ต้น ปรากฏดังนี้

ต้นกล้า	1	2	3	4	5	6	7	8
อายุ (สัปดาห์)	5	7	10	12	15	20	25	30
ความสูง (ซม.)	40	50	60	65	70	80	92	100

จงทดสอบความเชื่อที่ว่า $B \neq 0$ ณ ระดับนัยสำคัญ 0.05

วิธีทำ

สมมติฐานเพื่อการทดสอบ $H_0 : B = 0$

$H_a : B \neq 0$

ตัวสถิติเพื่อการทดสอบคือ $t = \frac{b}{s_b}$

โดยที่ $b = 2.283$, $s^2_{Y \cdot X} = 10.35258$ $\sum_{i=1}^{n} x_i^2 = 2{,}468$, $\bar{x} = 15.5$, $n = 8$

จาก $s_b^2 = \frac{s^2_{Y \cdot X}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

$= \frac{s^2_{Y \cdot X}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}$

$= \frac{10.35258}{2{,}468 - 8(15.5^2)}$

$= 0.01896$

$s_b = 0.1377$

จะได้ $t = \frac{2.283}{0.1377} = 16.58$

จะปฏิเสธ H_0 เมื่อตัวสถิติที่คำนวณได้มากกว่า $t_{(0.95,\ 6)} = 2.4469$ หรือน้อยกว่า $t_{(0.05,\ 6)} = -2.4469$

ดังนั้นต้องปฏิเสธ H_0 หรือยอมรับ H_a แสดงว่า $B \neq 0$

นั่นคือ ความสูงและอายุของต้นกล้าไม้ชนิดนี้มีความสัมพันธ์กัน

การวิเคราะห์สหสัมพันธ์ของ Spearman

ในการหาความสัมพันธ์ระหว่างตัวแปรหรือลักษณะที่สนใจศึกษาหากตัวแปรเหล่านั้นสามารถวัดออกมาเป็นค่าที่แทนด้วยตัวเลขได้ (ข้อมูลเชิงปริมาณ - มาตรอันตรภาคชั้นหรือมาตราอัตราส่วน)

การหาความสัมพันธ์ดังกล่าวสามารถใช้วิธีการของ Pearson ดังที่ได้กล่าวมาแล้ว แต่ถ้าตัวแปรหรือลักษณะที่สนใจศึกษาแทนด้วยตำแหน่งที่ (มาตราจัดลำดับ) การวิเคราะห์ด้วยวิธีข้างต้นไม่สามารถทำได้ต้องใช้วิธีการของสเปียร์แมน (Spearman) ซึ่งมีขั้นตอนในการหาความสัมพันธ์ดังนี้

1) ตั้งสมมติฐานเพื่อการทดสอบ ซึ่งโดยทั่วไปเป็นดังนี้

H_0 : ไม่มีความสัมพันธ์ระหว่างลักษณะที่สนใจศึกษา

H_1 : มีความสัมพันธ์ระหว่างลักษณะที่สนใจศึกษา

2) เขียนตำแหน่งที่เป็นคู่ๆ โดยที่จำนวนคู่ที่ใช้ไม่ควรน้อยกว่า 6 คู่

3) หากำลังสองของความแตกต่างระหว่างตำแหน่งที่ของแต่ละคู่ (d_i)

4) หาผลรวมของกำลังสองของความแตกต่างระหว่างตำแหน่งที่ของแต่ละคู่ เขียนแทนด้วย s

5) คำนวณหาสัมประสิทธิ์สหสัมพันธ์ เมื่อข้อมูลอยู่ในรูปตำแหน่งที่ (r_s) ซึ่งเป็นสูตรของสเปียร์แมน

$$r_s = 1 - \frac{6\,s}{n\,(n^2 - 1)}$$

เมื่อ n แทนจำนวนคู่

6) เปรียบเทียบค่าของ r_s ที่คำนวณได้กับค่าวิกฤติของสัมประสิทธิ์สหสัมพันธ์เมื่อข้อมูลอยู่ในรูปตำแหน่งที่ (rank correlation coefficient) จากตารางต่อไปนี้

ตารางที่ 4 ค่าวิกฤติของสัมประสิทธิ์สหสัมพันธ์เมื่อข้อมูลอยู่ในรูปตำแหน่งที่ ณ ระดับนัยสำคัญ α = 0.05

จำนวนคู่ (n)	ค่าวิกฤติ
6	-0.77 , +0.83
7	-0.71 , +0.75
8	-0.69 , +0.71

9	-0.67 , +0.68
10	-0.62 , +0.64

ในกรณีที่จำนวนคู่ (n) มากกว่า 10 การแจกแจงของ r สามารถอนุโลมได้ว่าเป็นการแจกแจงปกติ และขอบเขตในการยอมรับ H_0 หรือช่วงค่าวิกฤตคำนวณได้จาก

$$r_c = \pm \frac{z}{\sqrt{n-1}}$$

ถ้าค่า r_s ที่คำนวณได้ตกอยู่ภายในช่วงวิกฤตของสัมประสิทธิ์สหสัมพันธ์ต้องยอมรับ H_0 แต่ถ้าค่า r_s ตกนอกช่วงค่าวิกฤตต้องปฏิเสธ H_0

ตัวอย่างที่ 7 ในการทดสอบคุณภาพของน้ำข้าวกล้องงอก 6 ยี่ห้อ คือ A, B, C, D, E และ F ผู้ทดสอบได้แบ่งผู้ประเมินคุณภาพน้ำข้าวกล้องงอกดังกล่าวออกเป็น 2 กลุ่ม คือ กลุ่มวัยรุ่นและกลุ่มผู้ใหญ่ ถ้าผลการประเมินคุณภาพน้ำข้าวกล้องงอกทั้ง 6 ยี่ห้อ ซึ่งแทนด้วยตำแหน่งที่เป็นดังนี้

ยี่ห้อน้ำข้าวกล้องงอก	กลุ่มวัยรุ่น	กลุ่มผู้ใหญ่
A	1	5
B	3	3
C	2	6
D	5	1
E	4	4
F	6	2

ผู้ทดสอบสามารถสรุปได้หรือไม่ว่าไม่มีความสัมพันธ์ระหว่างคุณภาพของน้ำข้าวกล้องงอกกับกลุ่มของผู้บริโภค ณ ระดับนัยสำคัญ 0.05

วิธีทำ

สมมติฐานเพื่อการทดสอบ

H_0 : ไม่มีความสัมพันธ์ระหว่างคุณภาพของน้ำข้าวกล้องงอกกับกลุ่มของผู้บริโภค

H_a : มีความสัมพันธ์ระหว่างคุณภาพของน้ำข้าวกล้องงอกกับกลุ่มของผู้บริโภค

ยี่ห้อน้ำข้าวกล้องงอก	กลุ่มวัยรุ่น	กลุ่มผู้ใหญ่	d	d^2
A	1	5	-4	16

B	3	3	0	0
C	2	6	-4	16
D	5	1	4	16
E	4	4	0	0
F	6	2	4	16
				s = 64

$$r_s = 1 - \frac{6\,s}{n\,(n^2 - 1)}$$

$$= 1 - \frac{6\,(64)}{6\,(36 - 1)}$$

$$= -0.828$$

เนื่องจากค่าวิกฤติของสัมประสิทธิ์สหสัมพันธ์เมื่อข้อมูลอยู่ในรูปตำหน่งที่ ณ ระดับนัยสำคัญ 0.05 และจำนวนคู่เป็น 6 เท่ากับ -0.77, +0.83

ดังนั้น r_s ตกอยู่นอกช่วงค่าวิกฤติหรือช่วงที่จะยอมรับ H_0

นั่นคือ ต้องปฏิเสธ H_0 ที่ว่าไม่มีความสัมพันธ์ระหว่างคุณภาพของน้ำข้าวกล้องงอกกับกลุ่มของผู้บริโภค แสดงว่ากลุ่มวัยรุ่นมีความเห็นเกี่ยวกับคุณภาพของน้ำข้าวกล้องงอกทั้ง 6 ยี่ห้อแตกต่างจากกลุ่มผู้ใหญ่ และเนื่องจากสัมประสิทธิ์สหสัมพันธ์มีเครื่องหมายเป็น “ลบ” แสดงว่า ความสัมพันธ์ระหว่างคุณภาพของน้ำข้าวกล้องงอกกับกลุ่มของผู้บริโภคเป็นไปในทางตรงกันข้าม

ตัวอย่างที่ 8 จากการศึกษาถึงความสัมพันธ์ระหว่างประสบการณ์ในการทำอาชีพเกษตรกับรายได้เฉลี่ยต่อเดือนของเกษตรกรในพื้นที่แห่งหนึ่ง ปรากฏผลดังนี้

ชื่อเกษตรกร	ประสบการณ์ในการทำอาชีพเกษตร (ปี)	รายได้ต่อเดือน (บาท)
วิจิตร	12	7,000
ไพฑูรย์	13	7,800
สุนทร	16	4,000
สนอง	12	8,000
ภักดี	16	70,000
สมนึก	12	8,000
วิชิต	18	10,400
อเนก	20	13,000
สามารถ	20	25,000
สงัด	16	10,500
ถาวร	14	15,000
เถลิง	14	14,000
สนั่น	13	11,700

เราสามารถสรุปได้หรือไม่ว่าประสบการณ์ในการทำอาชีพเกษตรกับรายได้เฉลี่ยต่อเดือนของเกษตรกรในพื้นที่นี้มีความสัมพันธ์กัน ณ ระดับนัยสำคัญ 0.05

วิธีทำ

สมมติฐานเพื่อการทดสอบ

H_0 : ประสบการณ์ในการทำอาชีพเกษตรกับรายได้เฉลี่ยต่อเดือนของเกษตรกรไม่มีความสัมพันธ์กัน

H_a : ประสบการณ์ในการทำอาชีพเกษตรกับรายได้เฉลี่ยต่อเดือนของเกษตรกรมีความสัมพันธ์กัน

เมื่อเปลี่ยนค่าของข้อมูลให้อยู่ในรูปตำแหน่งที่ จะเป็นดังนี้

ชื่อเกษตรกร	ตำแหน่งที่ของจำนวนปีที่มีประสบการณ์	ตำแหน่งที่ของรายได้	d	d^2
วิจิตร	2	2	0	0
ไพฑูรย์	4.5	3	1.5	2.25
สุนทร	9	1	8	64.00

ชื่อเกษตรกร	ตำแหน่งที่ของจำนวนปีที่มีประสบการณ์	ตำแหน่งที่ของรายได้	d	d^2
สนอง	2	4.5	-2.5	6.25
ภักดี	9	13	-4	16.00
สมนึก	2	4.5	-2.5	6.25
วิชิต	11	6	5	25.00
อเนก	12.5	9	3.5	12.25
สามารถ	12.5	12	0.5	0.25
สงัด	9	7	2	4.00
ถาวร	6.5	11	-4.5	20.25
เถลิง	6.5	10	-3.5	12.25
สนั่น	4.5	8	-3.5	12.25
				s = 181.00

$$r_s = 1 - \frac{6\,s}{n\,(n^2 - 1)}$$

$$= 1 - \frac{6\,(181)}{13\,(168)}$$

$$= 0.503$$

เนื่องจากจำนวนคู่ (n) มากกว่า 10 ช่วงค่าวิกฤตคำนวณจาก

$$r_c = \pm \frac{Z}{\sqrt{n - 1}}$$

$$= \pm \frac{1.96}{\sqrt{13 - 1}} = \pm 0.566$$

ดังนั้นยอมรับ H_0 ที่ว่าประสบการณ์ในการทำอาชีพเกษตรกับรายได้เฉลี่ยต่อเดือนของเกษตรกรไม่มีความสัมพันธ์กัน

การวิเคราะห์การถดถอยอย่างง่าย

การวิเคราะห์การถดถอยอย่างง่าย (Simple regression) หรือการวิเคราะห์การถดถอยเมื่อมีตัวแปรอิสระตัวแปรเดียว เป็นการศึกษาถึงความสัมพันธ์ระหว่างตัวแปร 2 ตัว หรือลักษณะที่สนใจศึกษา 2 ลักษณะ เช่น ในการหาความสัมพันธ์ระหว่างยอดขายและค่าใช้จ่ายในการโฆษณา เพื่อนำไปใช้ในการพยากรณ์ยอดขายสินค้าเมื่อทราบค่าใช้จ่ายในการโฆษณา

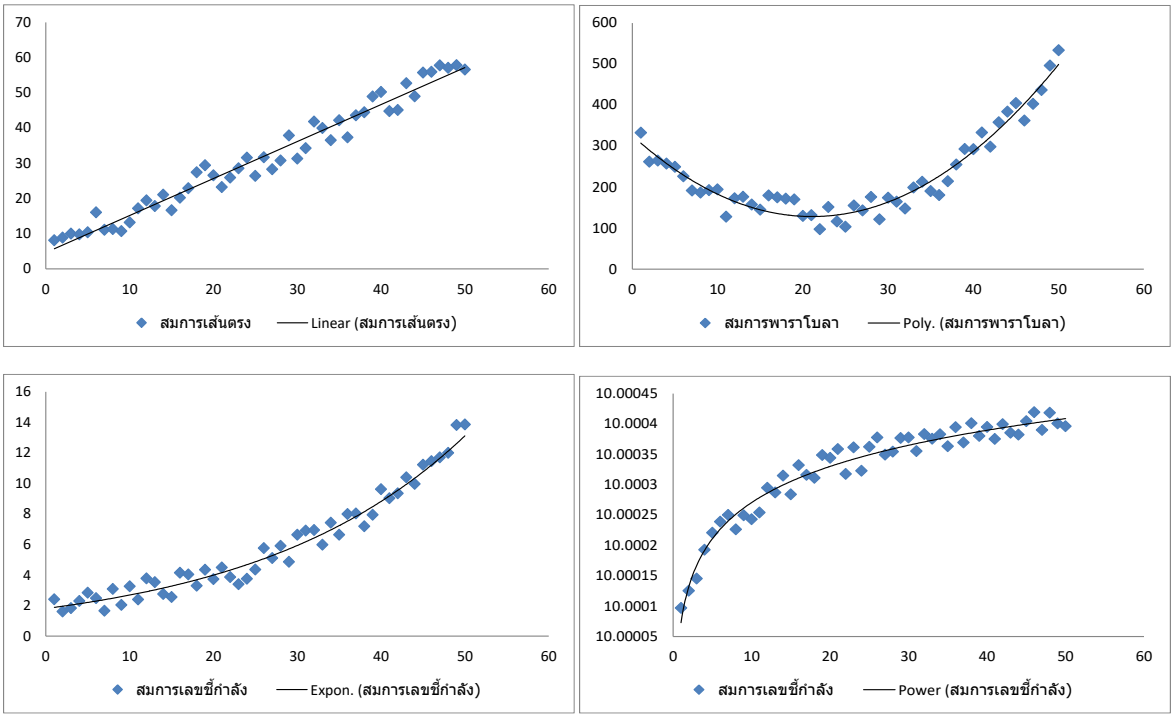
ในการหาความสัมพันธ์ระหว่างตัวแปรหรือลักษณะที่สนใจศึกษา 2 ลักษณะนี้มีวิธีหาได้หลายวิธีเช่น วิธีลากด้วยมือ (freehand method) วิธีเลือกจุด (selected point method) เป็นต้น แต่วิธีที่ให้ความถูกต้องและเชื่อถือได้มาก คือ วิธีกำลังสองน้อยสุด (least square method)

การหาความสัมพันธ์ระหว่างตัวแปรสองตัวหรือลักษณะสองลักษณะของประชากรที่สนใจศึกษาเป็นวิธีที่นิยมใช้กันทั่วไปเนื่องจากมีความถูกต้องและเชื่อถือได้มาก ซึ่งมีขั้นตอนการวิเคราะห์ดังต่อไปนี้

1) นำข้อมูลที่จะหาความสัมพันธ์ที่ประกอบด้วยตัวแปร 2 ตัว ซึ่งเรียกว่า ตัวแปรอิสระ (independent variable) และตัวแปรตาม (dependent variable) มาลงจุดแสดงความสัมพันธ์จะได้แผนภูมิที่เรียกว่า แผนภาพการกระจาย (scatter plot) โดยตัวแปรตามเป็นตัวแปรที่ผู้วิเคราะห์ต้องการพยากรณ์ค่าตัวแปรอิสระเป็นตัวแปรที่นำมาสร้างความสัมพันธ์

2) พิจารณาความสัมพันธ์ระหว่างตัวแปรทั้งสอง จากแผนภาพการกระจายว่ามีแนวโน้มที่จะแทนได้ด้วยรูปของสมการทางคณิตศาสตร์ใด เช่น เส้นตรง พาราโบลา สมการเลขชี้กำลังหรือรูปสมการอื่นที่สามารถเขียนแทนความสัมพันธ์ทางคณิตศาสตร์ได้

แผนภาพการกระจายของข้อมูล 3 ชุด ที่ความสัมพันธ์แทนได้ด้วยสมการเส้นตรง สมการพาราโบลา และสมการเลขชี้กำลัง ดังภาพ



3) หาค่าคงที่ที่ไม่ทราบค่า (unknown constant) ของสมการทางคณิตสาสตร์ที่ได้กำหนดไว้ใน (2) โดยใช้หลักการของวิธีกำลังสองน้อยที่สุด กล่าวคือ พยายามทำให้ผลรวมของส่วนเบี่ยงเบนระหว่างค่าจริงและค่าประมาณของข้อมูลที่นำมาสร้างความสัมพันธ์ยกกำลังสองมีค่าน้อยสุด นั่นคือ

ถ้าให้ y เป็นค่าจริงของข้อมูลที่เก็บรวบรวมมาได้

$\hat{y}$ เป็นค่าประมาณที่หาได้จากความสัมพันธ์ที่สร้างขึ้น

$\sum_{i=1}^{n}(y_i - \hat{y}_i)$ จะมีค่าน้อยที่สุด เมื่อ n แทนจำนวนตัวอย่างที่ใช้ในการเก็บรวบรวมข้อมูลเพื่อนำมาสร้างความสัมพันธ์

การหาค่าคงที่ที่ไม่ทราบค่าโดยวิธีกำลังสองน้อยสุด หาได้จากสมการปกติ (normal equation) ที่สร้างขึ้นมาจากรูปสมการทั่วไปของความสัมพันธ์ระหว่างตัวแปรอิสระ X และตัวแปรตาม Y ดังนี้

3.1) สมการเส้นตรง ซึ่งมีรูปสมการทั่วไปเป็น $Y = a + bX$

3.2) สมการเส้นโค้ง

สมการพาราโบลา ซึ่งมีรูปสมการทั่วไปเป็น $Y = a + bX + cX^2$ มีสมการปกติเป็น

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4$$

สมการไฮเพอร์โบลาซึ่งมีรูปสมการทั่วไปเป็น $Y = 1/(a + bX)$ สมการนี้สามารถเปลี่ยนให้อยู่ในรูปสมการเส้นตรงได้ กล่าวคือ $1/Y = a + bX$ ดังนั้น ความสัมพันธ์ระหว่าง X และ Y อยู่ในรูปไฮเพอร์โบลาแต่ความสัมพันธ์ระหว่าง X และ 1/Y อยู่ในรูปเส้นตรง ซึ่งมีสมการปกติเป็น

$$\sum_{i=1}^{n} (1/y_i) = an + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (x_i/y_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

สมการเลขชี้กำลังซึ่งมีรูปสมการทั่วไปเป็น $Y = ab^X$ สมการนี้สามารถเปลี่ยนให้อยู่ในรูปสมการเส้นตรงได้เช่นเดียวกันกล่าวคือ

$$\log Y = \log a + (\log b) X$$

ดังนั้นความสัมพันธ์ระหว่าง X และ Y อยู่ในรูปเลขชี้กำลังแต่ความสัมพันธ์ระหว่าง X และ log Y อยู่ในรูปเส้นตรงซึ่งมีสมการปกติเป็น

$$\sum_{i=1}^{n} (\log y_i) = (\log a)\, n + (\log b) \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i (\log y_i) = (\log a) \sum_{i=1}^{n} x_i + (\log b) \sum_{i=1}^{n} x_i^2$$

สมการเส้นตรงและเส้นโค้งบางชนิดที่ยกมาเป็นตัวอย่างในการหาสมการปกตินี้ใช้กันมากในทางปฏิบัติ ถ้าผู้วิเคราะห์ไม่ต้องการความละเอียดถูกต้องในการสร้างความสัมพันธ์ระหว่างตัวแปรสองตัวมากนัก ส่วนสมการเส้นโค้งอื่นที่ไม่ได้นำมากล่าวในที่นี้สามารถนำมาหาสมการปกติโดยวิธีกำลังสองน้อยสุดเพื่อใช้ในการหาค่าคงที่ที่ไม่ทราบค่าได้เช่นเดียวกัน

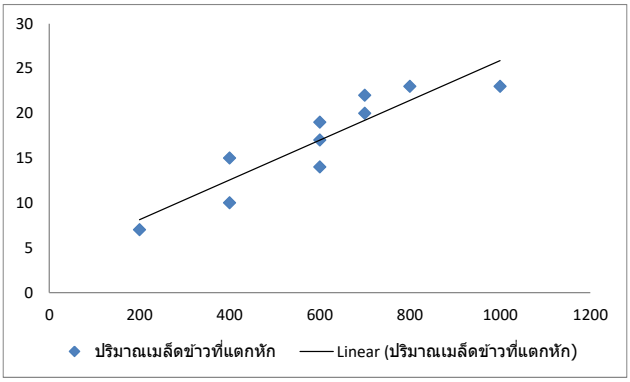
ตัวอย่างที่ 9 เจ้าของโรงสีข้าวแห่งหนึ่งต้องการศึกษาความสัมพันธ์ระหว่างปริมาณข้าวที่เครื่องสีข้าวแต่ละเครื่องผลิตได้ในแต่ละวันกับปริมาณเมล็ดข้าวแตกหักที่เกิดจากกระบวนการสีข้าว เขาจึงเลือกเครื่องสีข้าวมาเป็นตัวอย่างจำนวน 10 เครื่อง จากการเก็บรวบรวมข้อมูลจากเครื่องสีข้าวทั้ง 10 เครื่อง เกี่ยวกับปริมาณข้าวที่ผลิตได้ในวันที่กำหนดไว้และปริมาณเมล็ดข้าวที่แตกหักปรากฏดังนี้

เครื่องสีข้าว	1	2	3	4	5	6	7	8	9	10
ปริมาณข้าวที่ผลิตได้ (กก.)	600	700	600	400	700	800	600	400	200	1,000
ปริมาณข้าวที่แตกหัก (กก.)	19	20	14	10	22	23	17	15	7	23

จงสร้างความสัมพันธ์ระหว่างปริมาณข้าวที่ผลิตได้กับปริมาณเมล็ดข้าวที่แตกหักจากกระบวนการสีข้าว เพื่อนำมาใช้ในการพยากรณ์ปริมาณเมล็ดข้าวที่แตกหักเมื่อทราบปริมาณข้าวที่เครื่องสีข้าวแต่ละเครื่องผลิตได้

วิธีทำ

นำข้อมูลที่เก็บรวบรวมได้มาลงจุด เพื่อสร้างแผนภาพการกระจายไว้ใช้ในการพิจารณากำหนดรูปแบบของความสัมพันธ์ เนื่องจากต้องการพยากรณ์ปริมาณเมล็ดข้าวแตกหักเมื่อทราบปริมาณข้าวที่เครื่องสีข้าวแต่ละเครื่องผลิตได้ ดังนั้นจะให้ Y ซึ่งเป็นตัวแปรตามแทนปริมาณเมล็ดข้าวแตกหัก และ X ซึ่งเป็นตัวแปรอิสระแทนปริมาณข้าวที่เครื่องสีข้าวแต่ละเครื่องผลิตได้



จากแผนภาพการกระจายจะเห็นได้ว่าความสัมพันธ์ระหว่างปริมาณข้าวที่ผลิตได้กับปริมาณข้าวแตกหักจากกระบวนการผลิต พอจะอนุโลมให้อยู่ในรูปเส้นตรงซึ่งมีสมการทั่วไปเป็น $Y = a + bX$ ได้ โดยมีสมการปกติเป็น

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i \quad = \quad a\sum_{i=1}^{n} x_i \quad + \quad b\sum_{i=1}^{n} x_i^2$$

เมื่อ n = 10, a และ b เป็นค่าคงที่ที่ไม่ทราบค่าซึ่งจะต้องคำนวณหา และเพื่อสะดวกต่อการคำนวณหาค่า a และ b จากสมการปกติข้างต้น มักนิยมสร้างตารางเพื่อหาค่า $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i y_i$ และ $\sum_{i=1}^{n} x_i^2$ ดังนี้

y_i	x_i	x_i^2	$x_i y_i$
19	600	360,000	11,400
20	700	490,000	14,000
14	600	360,000	8,400
10	400	160,000	4,000
22	700	490,000	15,400
23	800	640,000	18,400
17	600	360,000	10,200
15	400	160,000	6,000
7	200	40,000	1,400
23	1,000	1,000,000	23,000
รวม 170	รวม 6,000	รวม 4,060,000	รวม 112,200

เมื่อแทนค่า n = 10, $\sum_{i=1}^{n} y_i$ = 170, $\sum_{i=1}^{n} x_i$ = 6,000, $\sum_{i=1}^{n} x_i y_i$ = 112,200 และ $\sum_{i=1}^{n} x_i^2$ = 4,060,000 ลงในสมการข้างต้นจะได้

$$170 \quad = \quad 10\,a \quad + \quad 6{,}000\,b$$

$$112{,}200 \quad = \quad 6{,}000\,a \quad + \quad 4{,}060{,}000\,b$$

แก้สมการทั้งสอง จะได้ a = 3.68, b = 0.022

ดังนั้นสมการแสดงความสัมพันธ์ซึ่งโดยทั่วไปเรียกว่า สมการเส้นถดถอย (regression line) ระหว่างปริมาณข้าวที่ผลิตได้กับปริมาณข้าวแตกหักจากระบวนการสีข้าว คือ

$$\hat{Y} \quad = \quad 3.68 \quad + \quad 0.022\,X$$

จากสมการเส้นถดถอยเชิงเส้นตรงที่ได้ จะเห็นได้ว่าค่า a คือ ปริมาณเมล็ดข้าวเสียหายที่เกิดจากการทำงานของเครื่องสีข้าว แต่ค่า b เป็นปริมาณเมล็ดข้าวที่เสียหายเพิ่มขึ้นเมื่อจำนวนข้าวที่ผลิตเพิ่มขึ้น 1 หน่วย ซึ่งในที่นี้ถ้าเครื่องสีข้าวผลิตสินค้าได้เพิ่มขึ้นหนึ่งหน่วย จะมีร้อยละของเมล็ดข้าวแตกหักเพิ่มขึ้น 0.022 หน่วย หรือถ้าเครื่องสีข้าวผลิตได้เพิ่มขึ้น 1,000 หน่วย จะมีเมล็ดข้าวแตกหักเพิ่มขึ้น 22 หน่วย นั่นคือ ถ้าผู้

วิเคราะห์ทราบปริมาณข้าวที่เครื่องสีข้าวเครื่องใดเครื่องหนึ่งผลิตได้ จะสามารถพยากรณ์ปริมาณเมล็ดข้าวแตกหักเสียหายที่เครื่องสีข้าวนั้นผลิตจากสมการเส้นถดถอยที่สร้างขึ้น เช่น ถ้าผู้วิเคราะห์ต้องการทราบว่าเครื่องสีข้าวที่ผลิตได้ 1,200 กก. จะมีปริมาณเมล็ดข้าวเสียหายเท่าใด ถ้าแทนค่า X = 1,200 ลงในสมการเส้นถดถอย

$$\hat{Y} = 3.68 + 0.022\,X$$

จะได้ $$\hat{Y} = 3.68 + 0.022\,(1{,}200)$$

$$= 30.08 \text{ กก.}$$

นั่นคือ เครื่องสีข้าวที่ผลิตได้ 1,200 กก. จะมีปริมาณเมล็ดข้าวแตกหัก 30.08 กก.

ตัวอย่างที่ 10 เกษตรกรต้องการศึกษาความสัมพันธ์ระหว่างความสูงของต้นกล้าไม้ชนิดหนึ่งและอายุ จึงสุ่มเก็บรวบรวมข้อมูลเกี่ยวกับความสูงและอายุของต้นกล้าจำนวน 8 ต้น ปรากฏดังนี้

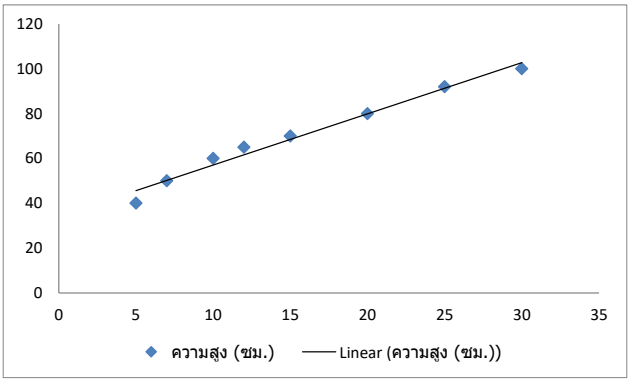
ต้นกล้า	1	2	3	4	5	6	7	8
อายุ (สัปดาห์)	5	7	10	12	15	20	25	30
ความสูง (ซม.)	40	50	60	65	70	80	92	100

ถ้าเกษตรกรต้องการพยากรณ์ความสูงของต้นกล้าเมื่อทราบอายุ จงสร้างสมการเส้นถดถอย

วิธีทำ

นำข้อมูลที่จะใช้ในการวิเคราะห์คือ อายุและความสูง มาลงจุดเพื่อสร้างแผนภาพการกระจายไว้ใช้ในการพิจารณากำหนดรูปแบบของความสัมพันธ์

เนื่องจากต้องการพยากรณ์ความสูงเมื่อกำหนดอายุ ดังนั้นจะให้ Y ซึ่งเป็นตัวแปรตามแทนความสูงต้นกล้าและ X ซึ่งเป็นตัวแปรอิสระแทนอายุต้นกล้า



จากแผนภาพการกระจายจะเห็นได้ว่าความสัมพันธ์ระหว่างความสูงต้นกล้าและอายุต้นกล้า สามารถอนุโลมให้อยู่ในรูปเส้นตรง ซึ่งมีสมการทั่วไปเป็น $Y = a + bX$ ได้โดยที่มีสมการปกติเป็น

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

เมื่อ n = 8, a และ b เป็นค่าคงที่ที่ไม่ทราบค่าซึ่งจะต้องคำนวณหา สำหรับค่า $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i y_i$ และ $\sum_{i=1}^{n} x_i^2$ หาได้จากตารางต่อไปนี้

y_i	x_i	x_i^2	$x_i y_i$
40	5	25	200
50	7	49	350
60	10	100	600
65	12	144	780
70	15	225	1,050
80	20	400	1,600
92	25	625	2,300
100	30	900	3,000
รวม 557	รวม 124	รวม 2,468	รวม 9,880

เมื่อแทนค่า n = 8, $\sum_{i=1}^{n} y_i$ = 557, $\sum_{i=1}^{n} x_i$ = 124, $\sum_{i=1}^{n} x_i y_i$ = 9,880 และ $\sum_{i=1}^{n} x_i^2$ = 2,468 ลงในสมการข้างต้นจะได้

$$557 = 8a + 124b$$

$$9{,}880 = 124a + 2{,}468b$$

แก้สมการทั้งสอง จะได้ a = 34.2385, b = 2.283

ดังนั้นสมการแสดงความสัมพันธ์ระหว่างความสูงและอายุของต้นกล้า คือ

$$\hat{Y} = 34.2385 + 2.283X$$

นั่นคือ a = 34.2385 ซม. เป็นความสูงของต้นกล้าเมื่อต้นกล้ามีอายุยังไม่ถึง 1 สัปดาห์ แต่ถ้าต้นกล้ามีอายุเพิ่มขึ้นทุกๆ 1 สัปดาห์ จะทำให้ความสูงเพิ่มขึ้นอีก 2.283 ซม.

ถ้าต้นกล้ามีอายุ 30 สัปดาห์ เป็น 50 สัปดาห์ ความสูงของต้นกล้าหาได้จากการแทนค่า X = 50 ลงในสมการเส้นถดถอย

$$\hat{Y} = 34.2385 + 2.283X$$

จะได้ $\hat{Y} = 34.2385 + 2.283\,(50)$

= 148.3885 ซม.

ดังนั้นถ้าต้นกล้ามีอายุ 50 สัปดาห์ ต้นกล้าจะมีความสูง 148.3885 ซม.

ถ้าแทนค่า X = 5, 7, 10, 12, 15, 20, 25 และ 30 ลงในสมการเส้นถดถอยที่ได้ จะได้ค่า $\hat{Y}$ เท่ากับ 45.6535, 50.2195, 57.0685, 61.6345, 68.4835, 79.8985, 91.3135 และ 102.7285 ตามลำดับ ซึ่งเมื่อนำไปเปรียบเทียบกับยอดขายจริง คือ 40, 50, 60, 65, 70, 80, 92 และ 100 จะเห็นได้ว่าแตกต่างกันไม่มากนัก แสดงว่าสมการเส้นถดถอยเชิงเส้นตรงสามารถใช้ประมาณความสัมพันธ์ความสูงและอายุของต้นกล้าไม้ชนิดนี้ได้เป็นอย่างดี ซึ่งมีผลให้ค่าพยากรณ์มีความถูกต้องและเชื่อถือได้มาก

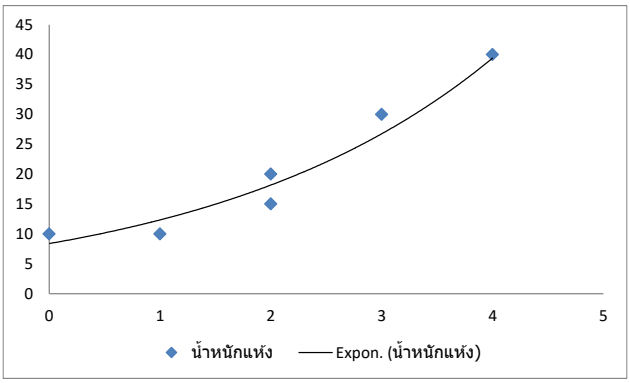
ตัวอย่างที่ 11 นักวิจัยเชื่อว่าน้ำหนักแห้งของต้นกล้าพืชชนิดหนึ่งมีความสัมพันธ์กับอายุการปลูก จากการเก็บรวบรวมข้อมูลโดยสุ่มต้นกล้ามาจำนวน 6 ต้น ปรากฏว่าน้ำหนักแห้งและอายุเป็นดังนี้

ต้นที่	1	2	3	4	5	6
น้ำหนักแห้ง (กรัม)	10	10	15	20	30	40
อายุ (วัน)	0	1	2	2	3	4

จงหาสมการเส้นถดถอยเพื่อใช้ในการพยากรณ์น้ำหนักแห้งของต้นกล้าพืชชนิดนี้

วิธีทำ

แผนภาพการกระจายแสดงความสัมพันธ์ระหว่างน้ำหนักแห้งของต้นกล้าพืชและอายุการปลูกโดยที่มี Y แทนน้ำหนักแห้งของต้นกล้าพืช และ X แทนอายุ เป็นดังภาพ



จากแผนภาพการกระจายความสัมพันธ์ระหว่างน้ำหนักแห้งของต้นกล้าพืชและอายุ สามารถอนุโลมให้อยู่ในรูปสมการเลขชี้กำลัง ซึ่งมีสมการทั่วไปเป็น $Y = ab^X$ หรือ $\log Y = \log a + (\log b) X$ และมีสมการปกติเป็น

$$\sum_{i=1}^{n} (\log y_i) = (\log a)\, n + (\log b) \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i (\log y_i) = (\log a) \sum_{i=1}^{n} x_i + (\log b) \sum_{i=1}^{n} x_i^2$$

เมื่อ n = 6, a และ b เป็นค่าคงที่ที่ไม่ทราบค่าซึ่งจะต้องคำนวณหา สำหรับค่า $\sum_{i=1}^{n} (\log y_i)$, $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i (\log y_i)$ และ $\sum_{i=1}^{n} x_i^2$ หาได้จากตารางต่อไปนี้

y_i	x_i	$\log y_i$	x_i^2	$x_i \log y_i$
10	0	1.0000	0	0.0000
10	1	1.0000	1	1.0000
15	2	1.1761	4	2.3522
20	2	1.3010	4	2.6020
30	3	1.4771	9	4.4313
40	4	1.6021	16	6.4084
	รวม 12	รวม 7.5563	รวม 34	รวม 16.7939

เมื่อแทนค่า n = 6, $\sum_{i=1}^{n} (\log y_i) = 7.5563$, $\sum_{i=1}^{n} x_i = 12$, $\sum_{i=1}^{n} x_i (\log y_i) = 16.7939$ และ $\sum_{i=1}^{n} x_i^2 = 34$ ลงในสมการข้างต้นจะได้

$$7.5563 = 6 \log a + 12 \log b$$

$$16.7939 = 12 \log a + 34 \log b$$

แก้สมการทั้งสอง จะได้ $\log a = 0.9231$ และ $\log b = 0.1681$ หรือ $a = 8.377$ และ $b = 1.473$

ดังนั้น สมการแสดงความสัมพันธ์ระหว่างน้ำหนักแห้งของต้นกล้าพืชและอายุของต้นกล้าชนิดนี้ คือ

$$\log \hat{Y} = 0.9231 + 0.1681 X$$

หรือ $\hat{Y} = (8.377)(1.473)^X$

นั่นคือ ถ้าต้นกล้ามีอายุ 5 วัน หรือ X = 5 จะได้

$$\log \hat{Y} = 0.9231 + 0.1681 (5)$$

$$= 1.7636$$

$$\hat{Y} = 58.02$$

ดังนั้น น้ำหนักแห้งของต้นกล้าพืช 58 กรัม

ถ้าต้นกล้ามีอายุ 8 วัน หรือ X = 8

จะได้ $\log \hat{Y} = 0.9231 + 0.1681 (8)$

$$= 2.2679$$

$\hat{Y}$ = 185.3

ดังนั้น น้ำหนักแห้งของต้นกล้าพืช 185 กรัม

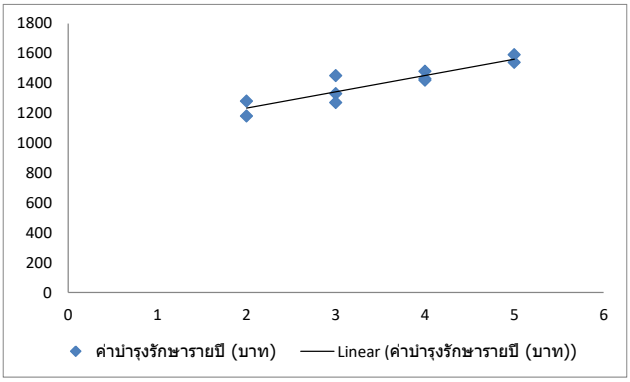
ตัวอย่างที่ 12 จากการศึกษาค่าบำรุงรักษารายปี (Y) ของรถไถนาที่มีอายุการใช้งาน (X) ดังต่อไปนี้

อายุการใช้งาน (ปี)	4	2	3	5	2	3	4	5	4	3
ค่าบำรุงรักษารายปี (บาท)	1,480	1,280	1,330	1,540	1,180	1,450	1,430	1,590	1,420	1,270

จงหาสมการเส้นถดถอยเพื่อแสดงความสัมพันธ์ระหว่างค่าบำรุงรักษารายปีและอายุการใช้งานของรถไถนา

วิธีทำ

แผนภาพการกระจายแสดงความสัมพันธ์ระหว่างค่าบำรุงรักษารายปี (Y) และอายุการใช้งาน (X) ของรถไถนา เป็นดังนี้



จากแผนภาพการกระจายความสัมพันธ์ระหว่างค่าบำรุงรักษารายปีและอายุการใชงานของรถไถนา สามารถอนุโลมให้อยู่ในรูปเส้นตรงซึ่งมีสมการทั่วไปเป็น $Y = a + bX$ และมีสมการปกติเป็น

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

เมื่อ n = 10, a และ b เป็นค่าคงที่ที่ไม่ทราบค่าซึ่งจะต้องคำนวณหาสำหรับค่า $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i y_i$ และ $\sum_{i=1}^{n} x_i^2$ หาได้จากตารางต่อไปนี้

y_i	x_i	x_i^2	$x_i y_i$

1480	4	16	5,920
1280	2	4	2,560
1330	3	9	3,990
1540	5	25	7,700
1180	2	4	2,360
1450	3	9	4,350
1430	4	16	5,720
1590	5	25	7,950
1420	4	16	5,680
1270	3	9	3,810
รวม 13,970	รวม 35	รวม 133	รวม 50,040

เมื่อแทนค่า n = 10, $\sum_{i=1}^{n} y_i$ = 13,970, $\sum_{i=1}^{n} x_i$ = 35, $\sum_{i=1}^{n} x_i y_i$ = 50,040 และ $\sum_{i=1}^{n} x_i^2$ = 133 ลงในสมการปกติข้างต้นจะได้

$$13{,}970 = 10a + 35b$$

$$50{,}040 = 35a + 133b$$

แก้สมการทั้งสอง จะได้ a = 1,015.33, b = 109.05

ดังนั้นสมการเส้นถดถอยเชิงเส้นตรง คือ

$$\hat{Y} = 1{,}015.33 + 109.05X$$

นั่นคือ ค่าบำรุงรักษารถไถนาที่ไม่ได้ใช้งานเลย (X=0) เท่ากับ 1,015.33 บาทต่อปี และเมื่อรถไถนาที่อายุการใช้งานเพิ่มขึ้น 1 ปี จะต้องเสียค่าบำรุงรักษาเพิ่มขึ้นจากเดิมอีก 109.05 บาท

จากสมการแสดงความสัมพันธ์ระหว่างค่าบำรุงรักษารายปีและอายุการใช้งานของรถไถนาที่ได้ สามารถนำมาใช้ในการพยากรณ์ค่าบำรุงรักษารายปีเมื่อรถไถนามีอายุการใช้งานต่างๆ ได้ เช่น ให้อายุการใช้งานของรถไถนาเท่ากับ 10 ปี ค่าบำรุงรักษารายปีจะเท่ากับ

$$\hat{Y} = 1{,}015.33 + 109.05(10)$$

$$= 2{,}105.83 \text{ บาท}$$

การวิเคราะห์การถดถอยเมื่อมีตัวแปรอิสระหลายตัวแปร (Multiple regression)

เป็นการหาความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระมากกว่า 1 ตัว เช่น ราคาขายของสินค้าขึ้นอยู่กับต้นทุนการผลิต ค่าขนส่ง และกำไรที่ต้องการได้รับ ยอดขายสินค้าขึ้นอยู่กับค่าใช้จ่ายในการโฆษณา

ทางหนังสือพิมพ์ ค่าใช้จ่ายในการโฆษณาทางวิทยุ และค่าใช้จ่ายในการโฆษณาทางโทรทัศน์หรือรายได้ของพนักงานขายขึ้นอยู่กับอายุและระดับการศึกษา ความสัมพันธ์ดังกล่าวนี้อาจเขียนในรูปของสมการถดถอยได้ดังนี้

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m$$

เมื่อ m คือ จำนวนตัวแปรอิสระที่ใช้ในการสร้างความสัมพันธ์ และ $X_1, X_2, \ldots, X_m$ คือ ตัวแปรอิสระที่ 1, 2, ..., m ตามลำดับ $b_1, b_2, \ldots, b_m$ คือ สัมประสิทธิ์การถดถอย (coefficient of regression) โดยที่ค่า b แต่ละค่าแสดงให้เห็นถึงการเปลี่ยนแปลงของค่าตัวแปรตาม Y เมื่อค่า X นั้นๆ เปลี่ยนไป 1 หน่วย โดยที่ค่า X ตัวอื่นๆคงที่ เช่น ค่า Y จะเปลี่ยนไป b_1 หน่วยถ้าค่า X_i เปลี่ยนไป 1 หน่วย โดยที่ค่า X ตัวอื่นๆ อีก m-1 ตัว คงที่ b_0 คือ ค่าของ Y เมื่อตัวแปรอิสระทุกๆ ตัวมีค่าเท่ากับ 0

เมื่อประมาณค่า $b_0, b_1, b_2, \ldots, b_m$ โดยใช้วิธีกำลังสองน้อยสุดจะสามารถหาค่า $b_0, b_1, b_2, \ldots, b_m$ ได้จากสมการปกติต่อไปนี้

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i} + \ldots + b_m \sum_{i=1}^{n} x_{mi}$$

$$\sum_{i=1}^{n} x_{1i}y_i = b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{1i}x_{2i} + \ldots + b_m \sum_{i=1}^{n} x_{1i}x_{mi}$$

$$\sum_{i=1}^{n} x_{2i}y_i = b_0 \sum_{i=1}^{n} x_{2i} + b_1 \sum_{i=1}^{n} x_{1i}x_{2i} + b_2 \sum_{i=1}^{n} x_{2i}^2 + \ldots + b_m \sum_{i=1}^{n} x_{2i}x_{mi}$$

$$\sum_{i=1}^{n} x_{mi}y_i = b_0 \sum_{i=1}^{n} x_{mi} + b_1 \sum_{i=1}^{n} x_{1i}x_{mi} + b_2 \sum_{i=1}^{n} x_{2i}x_{mi} + \ldots + b_m \sum_{i=1}^{n} x_{mi}^2$$

เมื่อ n แทนจำนวนตัวอย่างที่ใช้ในการเก็บรวบรวมข้อมูล

ประโยชน์ของการวิเคราะห์การถดถอยพหุคูณ การวิเคราะห์การถดถอยพหุคูณมีประโยชน์ดังต่อไปนี้

- ตรวจสอบได้ว่ามีปัจจัยอะไรบ้างที่สงสัยว่าจะมีผลต่อการเปลี่ยนแปลงของตัวแปรหรือลักษณะที่สนใจศึกษา

- บอกได้ว่าการเปลี่ยนแปลงของปัจจัยเหล่านั้นมีผลกระทบต่อการเปลี่ยนแปลงของตัวแปรหรือลักษณะที่สนใจศึกษาเป็นอย่างไร

- บอกได้ว่าลำดับความสำคัญของแต่ละปัจจัยที่มีผลกระทบต่อการเปลี่ยนแปลงของตัวแปรหรือลักษณะที่สนใจศึกษาเป็นอย่างไร

- บอกได้ว่าผลกระทบของแต่ละปัจจัยที่มีต่อตัวแปรรหรือลักษณะที่สนใจศึกษามีในทางบวกหรือ

ในทางลบ

- พยากรณ์ตัวแปรหรือลักษณะที่สนใจศึกษาได้เพื่อทราบค่าของปัจจัยทั้งหมดที่มีผลกระทบต่อตัวแปรหรือลักษณะที่สนใจศึกษานั้น

ตัวอย่างที่ 13 ในการหาความสัมพันธ์ระหว่างราคาไม้ประดับกับจำนวนกิ่งและจำนวนดอก ได้เลือกไม้ประดับโดยการสุ่มมาจำนวน 8 ต้น ผลปรากฏดังนี้

จำนวนกิ่ง	จำนวนดอก	ราคา (บาท)
3	2	338
2	1	293
4	3	388
2	1	292
3	2	347
2	2	299
5	3	434
4	2	379

จงหาสมการเส้นถดถอยที่ใช้แสดงความสัมพันธ์ระหว่างราคาไม้ประดับกับจำนวนกิ่งและจำนวนดอก และถ้าไม้ประดับมีกิ่งจำนวน 1 กิ่ง และดอกจำนวน 1 ดอก ราคาของไม้ประดับนี้ควรเป็นเท่าไร และถ้าไม้ประดับมี 5 กิ่ง 2 ดอก จะมีราคาเท่าไร

วิธีทำ

ให้ X_1 และ X_2 เป็นตัวแปรอิสระที่ใช้แทนจำนวนกิ่งและจำนวนดอกของไม้ประดับ ตามลำดับ และ Y เป็นตัวแปรตามที่ใช้แทนราคาไม้ประดับ

จะได้สมการที่ใช้แสดงความสัมพันธ์ระหว่างราคาไม้ประดับกับจำนวนกิ่งและจำนวนดอก ดังนี้

$$Y = b_0 + b_1X_1 + b_2X_2$$

ซึ่งมีสมการปกติคือ

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i}$$

$$\sum_{i=1}^{n} x_{1i}y_i = b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{1i}x_{2i}$$

$$\sum_{i=1}^{n} x_{2i}y_i = b_0 \sum_{i=1}^{n} x_{2i} + b_1 \sum_{i=1}^{n} x_{1i}x_{2i} + b_2 \sum_{i=1}^{n} x_{2i}^2$$

ค่า $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_{1i}$, $\sum_{i=1}^{n} x_{2i}$, $\sum_{i=1}^{n} x_{1i}y_i$, $\sum_{i=1}^{n} x_{1i}^2$, $\sum_{i=1}^{n} x_{1i}x_{2i}$, $\sum_{i=1}^{n} x_{2i}y_i$ และ $\sum_{i=1}^{n} x_{2i}^2$ หาได้จากตารางต่อไปนี้

y_i	x_{1i}	x_{2i}	x_{1i}^2	x_{2i}^2	$x_{1i}\,x_{2i}$	$x_{1i}\,y_i$	$x_{2i}\,y_i$	y_i^2
338	3	2	9	4	6	1,014	676	114,244
293	2	1	4	1	2	586	293	85,849
388	4	3	16	9	12	1,552	1,164	150,544
292	2	1	4	1	2	584	292	85,264
347	3	2	9	4	6	1,041	694	120,409
299	2	2	4	4	4	598	598	89,401
434	5	3	25	9	15	2,170	1,302	188,356
379	4	2	16	4	8	1,516	758	143,641
2,770	25	16	87	36	55	9,061	5,777	977,708

เมื่อแทนค่าที่ได้จากตารางข้างต้นลงในสมการปกติจะได้

$$2{,}770 = 8b_0 + 25b_1 + 16b_2$$

$$9{,}061 = 25b_0 + 87b_1 + 55b_2$$

$$5{,}777 = 16b_0 + 55b_1 + 36b_2$$

จากการแก้สมการทั้งสามจะได้ $b_0 = 201.97$, $b_1 = 41.49$ และ $b_2 = 7.31$ ดังนั้นสมการถดถอยพหุคูณคือ

$$\hat{Y} = 201.97 + 41.49\,X_1 + 7.31\,X_2$$

$$= 201.97 + 41.49\,(1) + 7.31\,(1)$$

$$= 250.77 \text{ บาท}$$

ดังนั้นไม้ประดับที่มี 1 กิ่งและ 1 ดอก จะมีราคา 250.77 บาท

ถ้าไม้ประดับมี 5 กิ่ง 2 ดอก ราคาไม้ประดับหาได้จากการแทนค่า $X_1 = 5$ และ $X_2 = 2$ ลงในสมการเส้นถดถอยที่ได้

$$\hat{Y} = 201.97 + 41.49 (5) + 7.31 (2)$$

$$= 201.97 + 207.45 + 14.62$$

$$= 424.04 \text{ บาท}$$

ดังนั้นไม้ประดับที่มี 5 กิ่ง 2 ดอก จะมีราคา 424.04 บาท

ตัวอย่างที่ 14 จากการเก็บรวบรวมข้อมูลเกี่ยวกับค่าใช้จ่ายต่อเดือน รายได้ต่อเดือน และขนาดครอบครัวของเกษตรกรในพื้นที่แห่งหนึ่งจำนวน 20 ครอบครัวที่เลือกมาเป็นตัวอย่างโดยการสุ่ม ผลเป็นดังนี้

ครอบครัว	ค่าใช้จ่าย (Y) (ร้อยบาท)	รายได้ (X_1) (ร้อยบาท)	ขนาดครอบครัว (X_2)
1	156	163	4
2	64	68	2
3	92	86	4
4	149	153	2
5	72	87	2
6	76	78	3
7	72	73	2
8	72	83	1
9	79	94	1
10	88	108	2
11	154	186	3
12	41	51	2
13	111	116	4
14	24	27	1
15	115	118	4
16	42	46	2
17	67	54	3

ครอบครัว	ค่าใช้จ่าย (Y) (ร้อยบาท)	รายได้ (X_1) (ร้อยบาท)	ขนาดครอบครัว (X_2)
18	121	129	3
19	111	133	2
20	47	59	1

สามารถสรุปได้หรือไม่ว่าขนาดครอบครัวมีผลทำให้ค่าใช้จ่ายของครอบครัวสูงขึ้นมากกว่ารายได้ของครอบครัว และถ้าครอบครัวหนึ่งของเกษตรกรในพื้นที่แห่งนี้มีรายได้ 10,000 บาทต่อเดือน และขนาดครอบครัวเท่ากับ 4 จงหาค่าใช้จ่ายของครอบครัวนี้

วิธีทำ

สมการที่ใช้แสดงความสัมพันธ์ระหว่างค่าใช้จ่ายกับรายได้ และขนาดครอบครัวของเกษตรกรในพื้นที่แห่งนี้คือ

$$Y = b_0 + b_1X_1 + b_2X_2$$

ซึ่งมีสมการปกติดังนี้

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i}$$

$$\sum_{i=1}^{n} x_{1i}y_i = b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{1i}x_{2i}$$

$$\sum_{i=1}^{n} x_{2i}y_i = b_0 \sum_{i=1}^{n} x_{2i} + b_1 \sum_{i=1}^{n} x_{1i}x_{2i} + b_2 \sum_{i=1}^{n} x_{2i}^2$$

ค่า $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_{1i}$, $\sum_{i=1}^{n} x_{2i}$, $\sum_{i=1}^{n} x_{1i}y_i$, $\sum_{i=1}^{n} x_{1i}^2$, $\sum_{i=1}^{n} x_{1i}x_{2i}$, $\sum_{i=1}^{n} x_{2i}y_i$ และ $\sum_{i=1}^{n} x_{2i}^2$ หาได้จากตารางต่อไปนี้

y_i	x_{1i}	x_{2i}	x_{1i}^2	x_{2i}^2	$x_{1i}\,x_{2i}$	$x_{1i}\,y_i$	$x_{2i}\,y_i$
156	163	4	26,569	16	652	25,428	624
64	68	2	4,624	4	136	4,352	128
92	86	4	7,396	16	344	7,912	368
149	153	2	23,409	4	306	22,797	298
72	87	2	7,569	4	174	6,264	144
76	78	3	6,084	9	234	5,928	228
72	73	2	5,329	4	146	5,256	144
72	83	1	6,889	1	83	5,976	73
79	94	1	8,836	1	94	7,426	79
88	108	2	11,664	4	216	9,504	176

y_i	x_{1i}	x_{2i}	x_{1i}^2	x_{2i}^2	$x_{1i}\ x_{2i}$	$x_{1i}\ y_i$	$x_{2i}\ y_i$
154	186	3	34,596	9	558	28,644	462
41	51	2	2,601	4	102	2,091	82
111	116	4	13,456	16	464	12,876	444
24	27	1	729	1	27	648	24
115	118	4	13,924	16	472	13,570	460
42	46	2	2,116	4	92	1,932	84
67	54	3	2,916	9	162	3,618	201
121	129	3	16,641	9	387	15,609	363
111	133	2	17,689	4	266	14,763	222
47	59	1	3,481	1	59	2,773	47
รวม 1,753	1,912	48	216,518	136	4,974	197,367	4,650

เมื่อแทนค่าที่ได้จากตารางข้างต้นลงในสมการปกติจะได้

$$1{,}753 = 20\ b_0 + 1{,}912\ b_1 + 48\ b_2$$

$$197{,}367 = 1{,}912\ b_0 + 216{,}518\ b_1 + 4{,}974\ b_2$$

$$4{,}650 = 48\ b_0 + 4{,}974\ b_1 + 136\ b_2$$

จากการแก้สมการทั้งสามจะได้ $b_0 = -4.95$, $b_1 = 0.8114$ และ $b_2 = 6.26$ ดังนั้นสมการเส้นถดถอยพหุคูณคือ

$$\hat{Y} = -4.95 + 0.8114\ X_1 + 6.26\ X_2$$

แสดงว่าครอบครัวที่มีรายได้เพิ่มขึ้นทุกๆ 1 ร้อยบาทต่อเดือน จะทำให้รายจ่ายสูงขึ้น 0.8114 ร้อยบาท หรือ 81.14 บาทต่อเดือน และครอบครัวที่มีขนาดครอบครัวเพิ่มขึ้นทุกๆ 1 คน จะทำให้รายจ่ายสูงขึ้น 6.26 ร้อยบาท หรือ 626 บาทต่อเดือน

กล่าวคือ ขนาดของครอบครัวมีผลทำให้ค่าใช้จ่ายของครอบครัวสูงขึ้นมากกว่ารายได้ของครอบครัว ทั้งนี้ เนื่องจากเมื่อขนาดของครอบครัวเพิ่มขึ้น 1 หน่วย จะมีผลทำให้รายจ่ายสูงขึ้นมากกว่าเมื่อรายได้เพิ่มขึ้น 1 หน่วย

การพยากรณ์รายจ่ายของครอบครัวที่มีรายได้ 10,000 บาทต่อเดือน และขนาดครอบครัวเท่ากับ 4 หาได้จากการแทนค่า $X_1 = 100$ (เนื่องจาก X_1 มีหน่วยเป็นร้อยบาท) และ $X_2 = 4$ ลงในสมการเส้นถดถอยพหุคูณที่ได้

$$\hat{Y} = -4.95 + 0.8114\ X_1 + 6.26\ X_2$$

= -4.95 + 0.8114 (100) + 6.26 (4)

= 101.23 ร้อยบาท

= 10,123 บาท

ดังนั้นครอบครัวของเกษตรกรที่มีรายได้ 10,000 บาทต่อเดือน และมีขนาดครอบครัวเท่ากับ 4 คน จะมีรายจ่ายเดือนละ 10,123 บาท

ตัวอย่างที่ 15 จากการศึกษาผลของฮอร์โมน (X_1) น้ำหนักเมล็ด (X_2) และอายุการเก็บรักษาภายหลังการเก็บเกี่ยว (X_3) ที่มีต่อระยะเวลาที่เมล็ดเริ่มงอก (Y) ของเมล็ดพันธุ์พืชหายากชนิดหนึ่งที่เลือกมาโดยการสุ่มจำนวน 25 เมล็ด โดยที่ $X_1 = 1$ เมื่อมีการใช้ฮอร์โมนกระตุ้น และ $X_1 = 0$ เมื่อไม่มีการใช้ฮอร์โมนกระตุ้น อายุการเก็บรักษาเมล็ดมีหน่วยเป็นเดือน ผลปรากฏดังนี้

เมล็ด	จำนวนชั่วโมงที่เมล็ดเริ่มงอก	ใช้ฮอร์โมน	น้ำหนักเมล็ด	อายุการเก็บรักษา
1	0.5	1	73	14
2	0.5	1	66	16
3	0.7	0	65	15
4	0.8	0	65	16
5	0.8	1	68	9
6	0.9	1	69	10
7	1.1	1	82	12
8	1.6	1	83	12
9	1.6	1	81	12
10	2.0	0	72	10
11	2.5	1	69	8
12	2.8	0	71	16
13	2.8	0	71	12
14	3.0	0	80	9
15	3.0	0	73	6
16	3.0	0	75	6
17	3.2	0	76	10

เมล็ด	จำนวนชั่วโมงที่เมล็ดเริ่มงอก	ใช้ฮอร์โมน	น้ำหนักเมล็ด	อายุการเก็บรักษา
18	3.2	0	78	6
19	3.3	1	79	6
20	3.3	0	79	4
21	3.4	1	78	6
22	3.5	0	76	9
23	3.6	0	65	12
24	3.7	0	72	12
25	3.7	0	80	6

จงหาสมการเส้นถดถอยพหุคูณเพื่อแสดงความสัมพันธ์ระหว่างจำนวนชั่วโมงที่เมล็ดเริ่มงอกกับการใช้ฮอร์โมน น้ำหนักเมล็ด และอายุการเก็บรักษาเมล็ดภายหลังการเก็บเกี่ยวของพืชหายากชนิดนี้ และจงพยากรณ์จำนวนชั่วโมงที่เมล็ดเริ่มงอกเมื่อมีการใช้ฮอร์โมน น้ำหนักเมล็ด 60 กรัม และมีอายุการเก็บรักษาภายหลังการเก็บเกี่ยว 5 เดือน

วิธีทำ

สมการที่ใช้แสดงความสัมพันธ์ระหว่างระหว่างจำนวนชั่วโมงที่เมล็ดเริ่มงอกกับการใช้ฮอร์โมน น้ำหนักเมล็ด และอายุการเก็บรักษาเมล็ดภายหลังการเก็บเกี่ยวของพืชหายากชนิดนี้ คือ

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

ซึ่งมีสมการปกติดังนี้

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i} + b_2 \sum_{i=1}^{n} x_{3i}$$

$$\sum_{i=1}^{n} x_{1i}y_i = b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{1i}x_{2i} + b_3 \sum_{i=1}^{n} x_{1i}x_{3i}$$

$$\sum_{i=1}^{n} x_{2i}y_i = b_0 \sum_{i=1}^{n} x_{2i} + b_1 \sum_{i=1}^{n} x_{1i}x_{2i} + b_2 \sum_{i=1}^{n} x_{2i}^2 + b_3 \sum_{i=1}^{n} x_{2i}x_{3i}$$

$$\sum_{i=1}^{n} x_{3i}y_i = b_0 \sum_{i=1}^{n} x_{3i} + b_1 \sum_{i=1}^{n} x_{1i}x_{3i} + b_2 \sum_{i=1}^{n} x_{2i}x_{3i} + b_3 \sum_{i=1}^{n} x_{3i}^2$$

จากข้อมูลข้างต้นจะได้

$$\sum_{i=1}^{n} y_i = 58.5, \quad \sum_{i=1}^{n} x_{1i} = 10, \quad \sum_{i=1}^{n} x_{2i} = 1{,}846,$$

$$\sum_{i=1}^{n} x_{3i} = 254, \quad \sum_{i=1}^{n} x_{1i}y_i = 16.2, \quad \sum_{i=1}^{n} x_{1i}^2 = 10,$$

$\sum_{i=1}^{n} x_{1i}x_{2i} = 748$, $\sum_{i=1}^{n} x_{1i}x_{3i} = 105$, $\sum_{i=1}^{n} x_{2i}y_i = 4{,}376$,

$\sum_{i=1}^{n} x_{2i}^2 = 137{,}086$, $\sum_{i=1}^{n} x_{2i}x_{3i} = 18{,}509$, $\sum_{i=1}^{n} x_{3i}y_i = 533.4$,

$\sum_{i=1}^{n} x_{3i}^2 = 2{,}892$

เมื่อแทนค่าข้างต้นลงในสมการปกติจะได้

$58.5 = 24\ b_0 + 10\ b_1 + 1{,}846\ b_2 + 254\ b_3$

$16.2 = 10\ b_0 + 10\ b_1 + 748\ b_2 + 105\ b_3$

$4{,}376 = 1{,}846\ b_0 + 748\ b_1 + 137{,}086\ b_2 + 18{,}509\ b_3$

$533.4 = 254\ b_0 + 105\ b_1 + 18{,}509\ b_2 + 2{,}892\ b_3$

จากการแก้สมการทั้งสี่จะได้ $b_0 = 1.41411$, $b_1 = -1.17396$, $b_2 = 0.03971$ และ $b_3 = -0.15106$ ดังนั้นสมการเส้นถดถอยพหุคูณคือ

$$\hat{Y} = 1.41411 - 1.17396\ X_1 + 0.03971\ X_2 - 0.15106\ X_3$$

การพยากรณ์จำนวนชั่วโมงที่เมล็ดเริ่มงอกเมื่อมีการใช้ฮอร์โมน น้ำหนักเมล็ด 60 กรัม และมีอายุการเก็บรักษาภายหลังการเก็บเกี่ยว 5 เดือน หาได้จากการแทนค่า $X_1 = 1$, $X_2 = 60$ และ $X_3 = 5$ ลงในสมการเส้นถดถอยพหุคูณที่ได้

$$\hat{Y} = 1.41411 - 1.17396\ X_1 + 0.03971\ X_2 - 0.15106\ X_3$$

$$= 1.41411 - 1.17396\ (1) + 0.03971\ (60) - 0.15106\ (5)$$

$$= 1.3018$$

ดังนั้นเมื่อมีการใช้ฮอร์โมน น้ำหนักเมล็ด 60 กรัม และมีอายุการเก็บรักษาภายหลังการเก็บเกี่ยว 5 เดือน เมล็ดพันธุ์พืชหายากชนิดนี้จะใช้เวลา 1.3018 ชั่วโมงที่จะเริ่มงอก

สมการเส้นถดถอยกรณีที่มีตัวแปรดัมมี่

การเพิ่มตัวแปรดัมมี่ (dummy variable) เข้าไปในสมการที่ใช้แทนความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม เป็นเทคนิควิธีหนึ่งในการวิเคราะห์สมการถดถอย

ความสัมพันธ์ที่พิจารณาจากแผนภาพการกระจายแทนได้ด้วยสมการเส้นตรง รูปสมการทั่วไปจะเขียนได้เป็น

$$Y = a + bX + cD$$

เมื่อ Y แทนตัวแปรตาม X แทนตัวแปรอิสระ และ D แทนตัวแปรดัมมี่ สมการปกติของสมการข้างต้นจะเป็นดังนี้

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} d_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} d_i x_i$$

$$\sum_{i=1}^{n} d_i y_i = a\sum_{i=1}^{n} d_i + b\sum_{i=1}^{n} d_i x_i + c\sum_{i=1}^{n} d_i^2$$

ตัวแปรดัมมี่นี้โดยทั่วไปจะมีค่าอยู่สองค่า โดยปกตินิยมแทนค่าด้วยค่า 0 หรือ 1 เช่น ใช้ค่า 0 แทนตัวแปรดัมมี่ของเหตุการณ์ปกติ และใช้ค่า 1 แทนตัวแปรดัมมี่ของเหตุการณ์ผิดปกติ เป็นต้น

หมายเหตุ จากรูปสมการทั่วไปที่มีความสัมพันธ์ $Y = a + bX + cD$ หากไม่มีเหตุการณ์ผิดปกติเกิดขึ้นเลย รูปสมการทั่วไปดังกล่าวจะอยู่ในรูป $Y = a + bX$ เนื่องจากค่า D เป็น 0 หมดทุกค่า

สำหรับการหาสมการเส้นถดถอยในกรณีที่มีตัวแปรดัมมี่นี้ ใช้วิธีการเช่นเดียวกันกับการหาสมการเส้นถดถอยพหุคูณ เพียงแต่แทนตัวแปรอิสระตัวหนึ่งด้วยตัวแปรดัมมี่ซึ่งมีเพียงสองค่า คือ 0 และ 1 เท่านั้น

ตัวอย่างที่ 16 ปริมาณการใช้ยางธรรมชาติในประเทศระหว่าง พ.ศ. 2520 ถึง พ.ศ. 2532 เป็นดังนี้

พ.ศ.	ปริมาณการใช้ยางธรรมชาติในประเทศ (หน่วย ตัน)
2520	91,562
2521	79,406
2522	91,469
2523	143,305
2524	172,746
2525	226,410
2526	274,548
2527	300,945
2528	329,616
2529	220,914
2530	297,626
2531	331,321
2532	360,635

จงหาสมการเส้นถดถอยเพื่อใช้พยากรณ์ปริมาณการใช้ยางธรรมชาติในอนาคต

วิธีทำ

จากข้อมูลข้างต้นจะสังเกตเห็นว่าปริมาณการใช้ยางธรรมชาติใน พ.ศ. 2521, 2522, 2529 และ 2530 ต่ำกว่าที่ควรจะเป็นและต้องการใช้ตัวแปรดัมมี่ในการแยกปีที่มีเหตุการณ์ปกติและปีที่เหตุการณ์ผิดปกติออกจากกัน กล่าวคือ แทนค่าตัวแปรดัมมี่ด้วย 0 สำหรับปีที่มีเหตุการณ์ปกติ และแทนค่าตัวแปรดัมมี่ด้วย 1 สำหรับปีที่มีเหตุการณ์ผิดปกติ

ดังนั้นสมการเส้นถดถอยที่ใช้แสดงความสัมพันธ์ระหว่างปริมาณการใช้ยางธรรมชาติและปี พ.ศ. ต่างๆ คือ

$$Y = a + bX + cD$$

และมีสมการปกติเป็น

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} d_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} d_i x_i$$

$$\sum_{i=1}^{n} d_i y_i = a\sum_{i=1}^{n} d_i + b\sum_{i=1}^{n} d_i x_i + c\sum_{i=1}^{n} d_i^2$$

ค่าต่างๆ ที่จะใช้แทนในสมการปกติทั้งสามหาได้จากตารางต่อไปนี้

พ.ศ.	x_i	y_i	$x_i y_i$	x_i^2	d_i	$d_i x_i$	$d_i y_i$	d_i^2
2520	-6	91,562	-549,372	36	0	0	0	0
2521	-5	79,406	-397,030	25	1	-5	79,406	1
2522	-4	91,469	-365,876	16	1	-4	91,469	1
2523	-3	143,305	-429,915	9	0	0	0	0
2524	-2	172,746	-345,492	4	0	0	0	0
2525	-1	226,410	-226,410	1	0	0	0	0
2526	0	274,548	0	0	0	0	0	0
2527	1	300,945	300,945	1	0	0	0	0
2528	2	329,616	659,232	4	0	0	0	0
2529	3	220,914	662,742	9	1	3	220,914	1
2530	4	297,626	1,190,504	16	1	4	297,626	1
2531	5	331,321	1,656,605	25	0	0	0	0
2532	6	360,635	2,163,810	36	0	0	0	0
รวม	0	2,920,503	4,319,743	182	4	-2	689,415	4

เมื่อแทนค่าต่างๆ ลงในสมการปกติจะได้

$$2{,}920{,}503 = 13a + 4c$$

$$4{,}319{,}743 = 182b - 2c$$

$$689{,}415 = 4a - 2b + 4c$$

จากการแก้สมการทั้งสามจะได้ $a = 206{,}540.13$, $b = 23{,}087.92$ และ $c = -58{,}870.30$ ดังนั้นสมการเส้นถดถอย คือ

$$\hat{Y} = 206{,}540.13 + 23{,}087.92\,X - 58{,}870.30\,D$$

(ช่วงเวลาเริ่มต้นอยู่ที่ พ.ศ. 2526, X มีหน่วยเป็น 1 ปี)

นั่นคือ เมื่อเวลาเพิ่มขึ้น 1 ปี ปริมาณการใช้ยางธรรมชาติจะเพิ่มขึ้นจากเดิมประมาณ 23,088 ตัน และในปีใดที่มีเหตุการณ์ผิดปกติเกิดขึ้นปริมาณการใช้ยางธรรมชาติจะลดลงจากเดิมประมาณ 58,870 ตัน

การพยากรณ์ค่าตัวแปรตาม (Y) เมื่อทราบค่าตัวแปรอิสระ (X) จากสมการเส้นถดถอยที่มีตัวแปรดัมมี่อยู่ด้วย ผู้พยากรณ์จะต้องคาดการณ์ล่วงหน้าได้ว่าจะมีเหตุการณ์ผิดปกติเกิดขึ้นในช่วงเวลาที่ต้องการพยากรณ์หรือไม่ ถ้าคาดว่าไม่มีเหตุการณ์ผิดปกติเกิดขึ้นต้องแทน D ด้วย 0 แต่ถ้าคาดว่าจะมีเหตุการณ์ผิดปกติเกิดขึ้นจะต้องแทน D ด้วย 1 จากตัวอย่างข้างต้น สัมประสิทธิ์ของ D มีค่าติดลบแสดงว่า ถ้ามีเหตุการณ์ผิดปกติเกิดขึ้นปริมาณการใช้ยางธรรมชาติจะลดลง การลดลงนี้จะลดเป็นจำนวนคงที่คือ 58,870 ตัน สำหรับช่วงเวลาใดๆ ที่เหตุการณ์ผิดปกติเกิดขึ้น เช่น ถ้าต้องการพยากรณ์ปริมาณการใช้ยางธรรมชาติในปี พ.ศ. 2537

ถ้าผู้พยากรณ์คาดว่าในปี พ.ศ. 2537 จะไม่มีเหตุการณ์ผิดปกติเกิดขึ้น แทนค่า $X = 11$ และ $D = 0$ ลงในสมการเส้นถดถอยจะได้

$$\hat{Y} = 206{,}540.13 + 23{,}087.92\,(11) - 58{,}870.30\,(0)$$

$$= 460{,}507.25 \text{ ตัน}$$

ถ้าผู้พยากรณ์คาดว่าในปี พ.ศ. 2537 จะมีเหตุการณ์ผิดปกติเกิดขึ้น แทนค่า $X = 11$ และ $D = 1$ ลงในสมการเส้นถดถอยจะได้

$$\hat{Y} = 206{,}540.13 + 23{,}087.92\,(11) - 58{,}870.30\,(1)$$

$$= 401{,}636.95 \text{ ตัน}$$

ตัวแปรดัมมี่ไม่จำเป็นต้องมีตัวเดียวเสมอไป อาจจะมีหลายตัวก็ได้ เช่น ให้ D_1 เป็นตัวแปรดัมมี่ที่ใช้แทนเหตุการณ์เกี่ยวกับการขึ้นราคาน้ำมันจากประเทศผู้ผลิต และ D_2 เป็นตัวแปรดัมมี่ที่ใช้แทนเหตุการณ์เกี่ยวกับความสงบเรียบร้อยตามชายแดนของประเทศไทยหรือในการศึกษาเกี่ยวกับปริมาณการผลิตสินค้าของโรงงานอุตสาหกรรมในแต่ละเดือน D_1 อาจเป็นตัวแปรดัมมี่เกี่ยวกับการขึ้นอัตราค่าแรงงานขั้นต่ำ และ D_2 อาจเป็นตัวแปรดัมมี่เกี่ยวกับการขาดแคลนวัตถุดิบที่ใช้ในการผลิต เป็นต้น

การใช้การวิเคราะห์การถดถอยอย่างมีประสิทธิภาพ

ในการนำการวิเคราะห์การถดถอยไปใช้ได้อย่างมีประสิทธิภาพนั้น ผู้วิเคราะห์ควรต้องคำนึงถึงปัจจัยดังต่อไปนี้

1) รู้วัตถุประสงค์ที่ชัดเจนในการนำการวิเคราะห์การถดถอยเชิงซ้อนมาใช้ เนื่องจากการใช้การถดถอยพหุคูณสำหรับแต่ละวัตถุประสงค์อาจจะมีวิธีการใช้ที่แตกต่างกัน หรือได้รับคำตอบที่มีความเชื่อถือได้ในระดับที่ต่างกัน

1.1) การตรวจสอบว่ามีปัจจัยอะไรบ้างที่สงสัยว่าจะมีผลต่อการเปลี่ยนแปลงของตัวแปรที่สนใจศึกษา พิจารณาจากการทดสอบสมมติฐานที่ว่า $b_i = 0$ หรือไม่ ถ้าค่า b ของปัจจัยซึ่งเป็นเหตุหรือตัวแปรอิสระ (X) ตัวใดแตกต่างจาก 0 อย่างมีนัยสำคัญ แสดงว่าปัจจัยซึ่งเป็นเหตุตัวนั้นมีผลต่อการเปลี่ยนแปลงของตัวแปรที่สนใจศึกษาหรือตัวแปรตาม ข้อมูลที่นำมาใช้วิเคราะห์ไม่ต้องทำให้เป็นค่ามาตรฐาน (standardized data)

$$Z_{ij} = \frac{x_{ij} - \bar{x}_i}{s_x} \qquad i = 1, 2, ..., k$$

$$Z_{ij} = \frac{y_{ij} - \bar{y}_i}{s_y} \qquad i = 1, 2, ..., n$$

1.2) การดูว่าการเปลี่ยนแปลงของปัจจัยที่เป็นเหตุมีผลกระทบต่อการเปลี่ยนแปลงของตัวแปรที่สนใจหรือไม่ พิจารณาจากค่า b_i (สัมประสิทธิ์การถดถอยของปัจจัยที่เป็นเหตุตัวที่ i) ถ้าค่า x_i เปลี่ยนไป 1 หน่วย โดยที่ค่า x ตัวอื่นๆ ที่เหลือทั้งหมดไม่เปลี่ยนแปลงค่า y จะเปลี่ยนแปลงไป b_i หน่วย โดยข้อมูลที่นำมาใช้วิเคราะห์ไม่ต้องทำให้เป็นค่ามาตรฐาน

1.3) การดูลำดับความสำคัญของแต่ละปัจจัยซึ่งมีผลกระทบต่อตัวแปรที่สนใจ พิจารณาจากค่า b_i เมื่อข้อมูลที่นำมาใช้วิเคราะห์ทั้งตัวแปรอิสระและตัวแปรตาม (x_i, y) เป็นค่ามาตรฐาน รูปแบบความสัมพันธ์ระหว่าง y และ x_i เพื่อข้อมูลเป็นค่ามาตรฐานคือ

$$\hat{Y} = b_1 X_1 + b_2 X_2 + ... + b_k X_k$$

x_i ที่มีค่า b_i สูงสุดจะมีผลกระทบต่อ y มากที่สุด และ x_i ที่มีค่า b_i สูงรองลงมาตามลำดับจะมีผลกระทบต่อ y รองลงมาตามลำดับ

1.4) การดูทิศทางของความสัมพันธ์ระหว่างปัจจัยแต่ละตัวกับตัวแปรตาม พิจารณาจากเครื่องหมายของ b_i ถ้ามีเครื่องหมายบวกแสดงว่าทิศทางของความสัมพันธ์เป็นบวก หรือถ้า x_i เพิ่มขึ้นหรือลดลง 1 หน่วย โดยที่ X ตัวอื่นๆ คงที่ ค่า $\hat{Y}$ จะเพิ่มขึ้นหรือลดลง b_i หน่วย แต่ถ้า b_i มีเครื่องหมายลบแสดงว่าถ้า x_i เพิ่มขึ้นหรือลดลง 1 หน่วยโดยที่ X ตัวอื่นๆ คงที่ ค่า $\hat{Y}$ จะลดลงหรือเพิ่มขึ้น b_i หน่วย ส่วนข้อมูลที่นำมาวิเคราะห์จะเป็นค่ามาตรฐานหรือไม่เป็นค่ามาตรฐานก็ได้

1.5) การพยากรณ์ตัวแปรหรือลักษณะที่สนใจศึกษาพิจารณาจากค่า $\hat{Y}$ ซึ่งได้จากการแทนค่า x_i ทุกๆ ค่าของสิ่งที่สนใจลงในสมการแสดงความสัมพันธ์ระหว่าง $\hat{Y}$ และ x_i ที่หาได้ข้อมูลที่นำมาวิเคราะห์ควรเป็นข้อมูลที่ไม่เป็นค่ามาตรฐาน เพื่อสะดวกต่อการแทนค่าในการพยากรณ์

2) รู้เกี่ยวกับเรื่องที่จะนำการวิเคราะห์การถดถอยพหุคูณมาใช้เป็นอย่างดี โดยเฉพาะอย่างยิ่งปัจจัยที่น่าจะมีผลต่อตัวแปรหรือลักษณะที่สนใจ การขาดความรู้เกี่ยวกับปัจจัยที่น่าจะมีผลต่อตัวแปรที่สนใจจะทำให้ไม่สามารถสรุปได้ว่ามีปัจจัยใดบ้างซึ่งมีผลกระทบต่อตัวแปรที่สนใจ และค่าพยากรณ์อาจคลาดเคลื่อนจากค่าจริงได้ ทั้งนี้ค่าพยากรณ์จะคลาดเคลื่อนมากหากขาดปัจจัยสำคัญๆ ที่มีผลต่อการเปลี่ยนแปลงของตัวแปรตาม

3) ตัวแปรที่นำมาวิเคราะห์ทั้งตัวแปรที่เป็นเหตุและตัวแปรที่เป็นผลจะต้องเป็นตัวแปรเชิงปริมาณชนิดต่อเนื่อง และตัวแปรที่เป็นเหตุทุกตัวจะต้องเป็นอิสระต่อกัน (independent) ในทางปฏิบัติตัวแปรที่เป็นเหตุมักจะมีตัวแปรเชิงคุณภาพหรือตัวแปรชนิดไม่ต่อเนื่องปนอยู่ด้วยเสมออาจอนุโลมใช้ได้ ถ้า

3.1) มีตัวแปรเชิงคุณภาพ ตัวแปรที่เป็นสัดส่วนไม่เกินร้อยละ 20 ของจำนวนตัวแปรทั้งหมดที่ใช้

3.2) ตัวแปรชนิดไม่ต่อเนื่อง มีค่าที่เป็นไปได้ทั้งหมดจำนวนมากพอสมควร

3.3) ในกรณีที่ตัวแปรอิสระซึ่งเป็นเหตุ คู่ใดคู่หนึ่งหรือหลายคู่มีความสัมพันธ์กันเอง (multicollinearity) ไม่สูงมากคือ มีค่าสัมประสิทธิ์สหสัมพันธ์ต่ำกว่า 0.80 ลงมา ให้อนุโลมใช้ตัวแปรอิสระที่เป็นเหตุทั้งสองได้ แต่ถ้าค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรคู่ใดคู่หนึ่งหรือหลายคู่สูงกว่า 0.80 ให้ตัดตัวแปรอิสระที่เป็นเหตุตัวที่มีความสัมพันธ์กับตัวแปรตามซึ่งเป็นผลต่ำทิ้งหรือตัดตัวแปรอิสระตัวที่วัดได้ยากกว่าหรือมีความเชื่อถือได้น้อยกว่าทิ้ง

4) จำนวนตัวอย่างที่นำมาใช้ในการวิเคราะห์ควรจะเป็นอย่างน้อยเท่ากับจำนวนตัวแปรซึ่งเป็นเหตุบวก 12 มิฉะนั้นจะมีผลทำให้เกิดความคลาดเคลื่อนสูงในการวิเคราะห์ หากมีความจำเป็นต้องใช้จำนวนตัวอย่างน้อยไม่ควรให้จำนวนตัวอย่างน้อยกว่าจำนวนตัวแปรที่เป็นเหตุบวก 4

5) ค่าของตัวแปรแต่ละตัวที่นำมาวิเคราะห์จะต้องเป็นค่าที่เปรียบเทียบขนาดกันได้อย่างแท้จริง ค่าของตัวแปรบางตัวอาจใช้ค่าสัมบูรณ์ (absolute value) ได้ แต่ค่าของตัวแปรบางตัวอาจต้องใช้ค่าสัมพัทธ์ (relative value)

ข้อควรระวังเกี่ยวกับการวิเคราะห์ถดถอยและสหสัมพันธ์

1) การวิเคราะห์ถดถอยและสหสัมพันธ์เป็นวิธีที่ใช้วัดความสัมพันธ์ระหว่างตัวแปรทั้งสองวิธี แต่การวิเคราะห์การถดถอยนอกจากจะบอกความสัมพันธ์ระหว่างตัวแปรอย่างคร่าวๆ โดยบอกว่าเมื่อค่า X ใดๆ เปลี่ยนไป 1 หน่วย ค่า Y จะเปลี่ยนไปเท่าไรเมื่อพิจารณาจากสัมประสิทธิ์การถดถอยแล้วยังสามารถใช้สมการเส้นถดถอยทำนายค่าตัวแปรตามเมื่อทราบค่าตัวแปรอิสระได้อีกด้วย ส่วนการวิเคราะห์สหสัมพันธ์ ค่า

สัมประสิทธิ์สหสัมพันธ์ที่คำนวณได้บอกแต่เพียงว่าตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันมากน้อยเพียงใดเท่านั้น แต่ความสัมพันธ์ที่หาได้นี้จะบอกถึงขนาดของความสัมพันธ์ระหว่างตัวแปรทั้งสองได้ดีกว่าความสัมพันธ์ที่วัดได้จากสัมประสิทธิ์การถดถอย ซึ่งไม่สามารถทราบค่าสูงสุดหรือต่ำสุดของความสัมพันธ์ที่อาจเกิดขึ้น

2) ในทางทฤษฎีการวิเคราะห์การถดถอยเป็นการศึกษาถึงความสัมพันธ์ระหว่างตัวแปรโดยที่ตัวแปรอิสระถูกกำหนดค่าไว้ล่วงหน้า และการวิเคราะห์สหสัมพันธ์เป็นการศึกษาถึงความสัมพันธ์ระหว่างตัวแปรโดยที่ไม่มีตัวแปรใดถูกกำหนดค่าไว้ล่วงหน้า แต่ในทางปฏิบัติทั่วๆ ไป ถึงแม้ว่าจะไม่มีตัวแปรอิสระถูกกำหนดค่าไว้ล่วงหน้าก็มักจะอนุโลมให้ใช้การวิเคราะห์การถดถอยได้

3) ในการพยากรณ์ค่าของตัวแปร (Y) เมื่อทราบค่าของตัวแปรอิสระ (X) ความเชื่อถือได้ของค่าพยากรณ์จะมีมากน้อยเพียงใดขึ้นอยู่กับขนาดของตัวอย่างที่นำมาสร้างความสัมพันธ์ โดยทั่วไปหากใช้ตัวอย่างมากจะทำให้ค่าพยากรณ์มีความเชื่อถือได้มาก แต่อย่างไรก็ตาม ความเชื่อถือได้ของค่าพยากรณ์นี้ยังขึ้นอยู่กับรูปแบบของความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามในขณะที่พยากรณ์ด้วยว่าเป็นไปเช่นเดียวกันกับรูปแบบของความสัมพันธ์เดิมที่นำมาสร้างความสัมพันธ์เพื่อนำไปใช้ในการพยากรณ์หรือไม่ ถ้ารูปแบบของความสัมพันธ์เปลี่ยนไปแล้วผลที่ได้จากการพยากรณ์จะเชื่อถือไม่ได้เลย แต่ถ้าเปลี่ยนไปบ้างเพียงเล็กน้อยผลจากการพยากรณ์จะเชื่อถือได้น้อยลง ในกรณีที่มีจำนวนตัวแปรอิสระมากขนาดตัวอย่างที่ใช้จะต้องมากตามไปด้วย กล่าวคือ จำนวนตัวอย่างจะต้องมากกว่าจำนวนตัวแปรอิสระมากพอสมควร

4) ตัวแปรอิสระและตัวแปรตามที่นำมาสร้างความสัมพันธ์จะต้องมีความสัมพันธ์กันจริงๆ หรือคาดว่าควรจะมีความสัมพันธ์กัน สำหรับตัวแปรที่ไม่มีความสัมพันธ์กันจริงๆ แล้วนำมาสร้างความสัมพันธ์สมการแสดงความสัมพันธ์ที่ได้จะไม่สามารถนำไปใช้ประโยชน์ได้เลย นอกจากนี้ผู้ใช้ผลการวิเคราะห์ยังอาจจะตัดสินใจดำเนินการต่างๆ ที่เกี่ยวข้องกับข้อมูลของตัวแปรทั้งสองนั้นไปอย่างผิดๆ ทำให้เกิดความเสียหายอย่างมากได้ เช่น การหาความสัมพันธ์ระหว่างจำนวนเด็กที่เกิดกับจำนวนอุบัติเหตุทางรถยนต์ของแต่ละสัปดาห์ซึ่งจริงๆ แล้วไม่ควรมีความสัมพันธ์กัน เป็นต้น

5) ถ้าผู้วิเคราะห์ต้องการพยากรณ์ค่าของตัวแปรใดจะต้องกำหนดให้ตัวแปรนั้นเป็นตัวแปรตามเสมอ และตัวแปรที่เหลือให้เป็นตัวแปรอิสระ สมการเส้นถดถอยที่หาได้จากการกำหนดตัวแปรดังกล่าว ไม่สามารถนำไปใช้ในการพยากรณ์ค่าตัวแปรอิสระเมื่อทราบค่าตัวแปรตามได้ ในกรณีนี้ต้องกำหนดให้ตัวแปรอิสระเดิมทำหน้าที่เป็นตัวแปรตาม และตัวแปรตามเดิมทำหน้าที่เป็นตัวแปรอิสระแล้วหาความสัมพันธ์ขึ้นมาใหม่จึงจะสามารถพยากรณ์ตามที่ต้องการได้ เช่น กรณีการหาความสัมพันธ์ระหว่างรายได้และรายจ่ายของร้านค้า ถ้าต้องการพยากรณ์รายจ่ายเมื่อทราบรายได้จะต้องกำหนดให้รายจ่ายเป็นตัวแปรตามและรายได้เป็นตัวแปรอิสระ แต่ถ้าต้องการพยากรณ์รายได้เมื่อทราบรายจ่ายจะต้องกำหนดให้รายได้เป็นตัวแปรตามและรายจ่ายเป็นตัวแปรอิสระ

6) ตัวแปรอิสระของการวิเคราะห์การถดถอยอาจเป็นตัวแปรเชิงคุณภาพได้ เช่น เพศ สถานภาพสมรส ศาสนา แต่ในการวิเคราะห์จะต้องแปลงข้อมูลเชิงคุณภาพเหล่านี้เป็นข้อมูลเชิงปริมาณเสียก่อน เช่น เพศชาย แทนด้วย 1 เพศหญิงแทนด้วย 0 ศาสนาพุทธแทนด้วย 1 ศาสนาอื่นๆ แทนด้วย 0 เป็นต้น

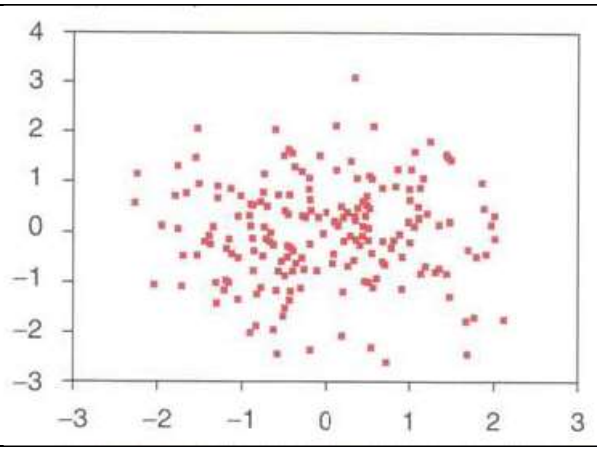
7) ตัวแปรอิสระที่เป็นข้อมูลเชิงคุณภาพหรือเป็นตัวแปรดัมมี่จะต้องไม่ให้มีจำนวนมากเกินไป เพราะจะมีผลทำให้ความถูกต้องเชื่อถือได้ของค่าประมาณของตัวแปรตามลดลง

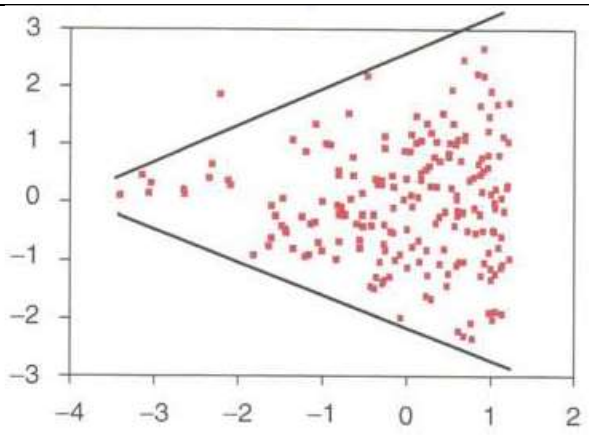
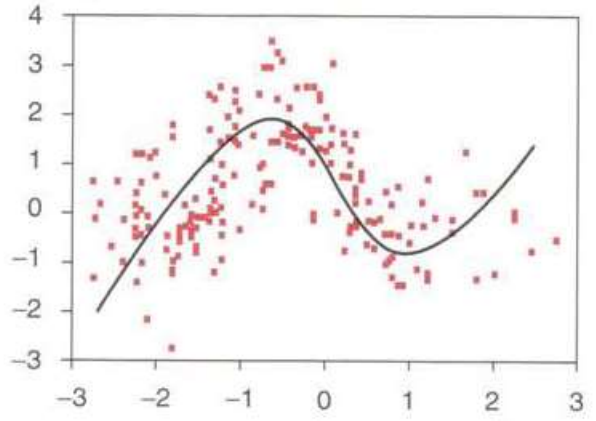
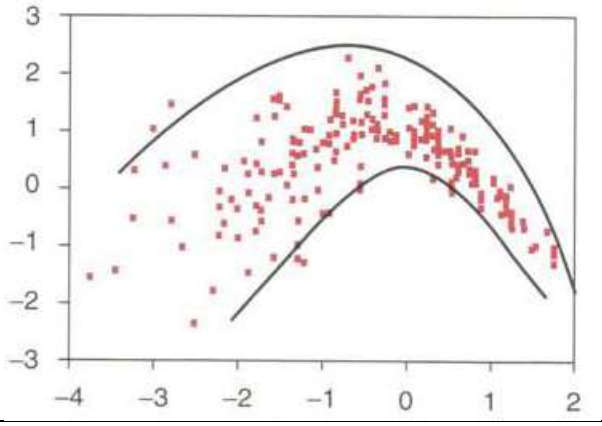
8) ตัวแปรอิสระที่เป็นข้อมูลซึ่งอยู่ในรูปสัดส่วนควรให้มีจำนวนน้อยที่สุด เพราะถ้ามีเป็นจำนวนมากแล้วจะทำให้ความถูกต้องเชื่อถือได้ของค่าประมาณของตัวแปรตามลดลงเช่นเดียวกัน

9) ตัวแปรตามไม่ควรเป็นตัวแปรเชิงคุณภาพที่มีค่าเพียง 2 ค่า (dichotomous variable) เท่านั้น เช่น (ชอบ, ไม่ชอบ) (เป็นโรคเบาหวาน, ไม่เป็นโรคเบาหวาน) (คุมกำเนิด, ไม่คุมกำเนิด) เพราะอาจจะทำให้ค่าประมาณของตัวแปรตามผิดพลาดได้มาก

10) สมการถดถอยเชิงเส้นตรงที่ได้จากการคำนวณจะไม่เหมาะสมพอที่จะนำไปใช้งานเมื่อ

- รูปแบบความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามไม่เป็นแบบเส้นตรง
- ข้อมูลที่ใช้ในการคำนวณมีค่าสูงหรือต่ำผิดปกติ (outliers) เกิดขึ้น
- ค่าส่วนต่าง $y_i - \hat{y}_i$ หรือค่าความคลาดเคลื่อนจากการทำนายด้วยสมการถดถอยที่ได้ควรมีการแจกแจงแบบปกติและมีค่าเฉลี่ยเท่ากับ 0 หรือใกล้เคียง 0
- Residual plot หรือแผนภาพการกระจายระหว่างค่าทำนาย $\hat{y}_i$ กับค่าส่วนต่าง $y_i - \hat{y}_i$ ควรมีรูปแบบเป็นไปโดยสุ่ม (random pattern) ไม่เช่นนั้นอาจแสดงให้เห็นว่ารูปแบบสมการถดถอยเชิงเส้นตรงที่ได้ไม่เหมาะสมในการใช้เพื่อทำนายค่าตัวแปรตามที่สนใจ

สมการถดถอยที่ได้ ไม่เกิดปัญหาใดๆ ค่าส่วนต่าง $y_i - \hat{y}_i$ หรือค่าความคลาดเคลื่อนจากการทำนายด้วยสมการถดถอยที่ได้เป็นไปโดยสุ่ม	

เกิดปัญหา Heteroscedasticity ความแปรปรวนของค่าส่วนต่าง $y_i - \hat{y}_i$ จากการทำนายด้วยสมการถดถอยที่ได้ไม่เท่ากันในแต่ละระดับของตัวแปรอิสระซึ่งกำหนดในสมการถดถอย แสดงให้เห็นว่าค่าส่วนต่าง $y_i - \hat{y}_i$ หรือค่าความความคลาดเคลื่อนจากการทำนายไม่เป็นไปโดยสุ่ม แต่กลับเกี่ยวข้องกับตัวแปรในสมการถดถอย ดังนั้นสมการถดถอยที่ได้จึงไม่เหมาะสมในการนำไปใช้งาน	
เกิดปัญหา Non-linearity ตัวแบบความสัมพันธ์เชิงเส้นตรงไม่เหมาะสมกับรูปแบบความสัมพันธ์ของข้อมูลการนำสมการถดถอยที่ได้ไปใช้งานจึงไม่เหมาะสม	
เกิดปัญหา Heteroscedasticity and non-linearity สมการถดถอยที่ได้ไม่เหมาะสมในการนำไปใช้งาน	

บทที่ 10

โปรแกรมวิเคราะห์สถิติ

โปรแกรมวิเคราะห์ทางสถิติมีความหลากหลายเช่นเดียวกันกับเทคนิคการวิเคราะห์ทางสถิติ โปรแกรมวิเคราะห์ทางสถิติเหล่านี้มีความแตกต่างกันไปตามแหล่งที่มาหรือวัตถุประสงค์ในการพัฒนาของผู้พัฒนาซึ่งอยู่ในหลากหลายวงการ หลายโปรแกรมถูกพัฒนาขึ้นเพื่อวัตถุประสงค์เฉพาะแต่ก็สามารถนำมาใช้งานในการวิเคราะห์โดยทั่วไปได้ อาทิเช่น

- ทางด้านการเกษตร โปรแกรม IRRISTAT ถูกพัฒนาขึ้นโดย International Rice Research Institute (IRRI) เป็นโปรแกรมที่ใช้สำหรับบริหารจัดการและวิเคราะห์ข้อมูลทางด้านการวางแผนการทดลอง (experimental data) ทำงานบนระบบปฏิบัติการ Windows โปรแกรม IRRISTAT เป็นโปรแกรมที่มีวัตถุประสงค์หลักในการพัฒนาเพื่อใช้งานทางด้านการเกษตรเป็นหลัก แต่คุณสมบัติหลายอย่างสามารถนำมาประยุกต์ใช้ในการวิเคราะห์ข้อมูลโดยทั่วไปได้ คุณสมบัติโดยทั่วไปที่โปรแกรมมี เช่น แผ่นงานสำหรับการจัดการข้อมูล, ค่าสถิติพื้นฐาน, Scatterplot, การวิเคราะห์ความแปรปรวน, การวิเคราะห์ถดถอยและสหสัมพันธ์, Experimental designs เป็นต้น

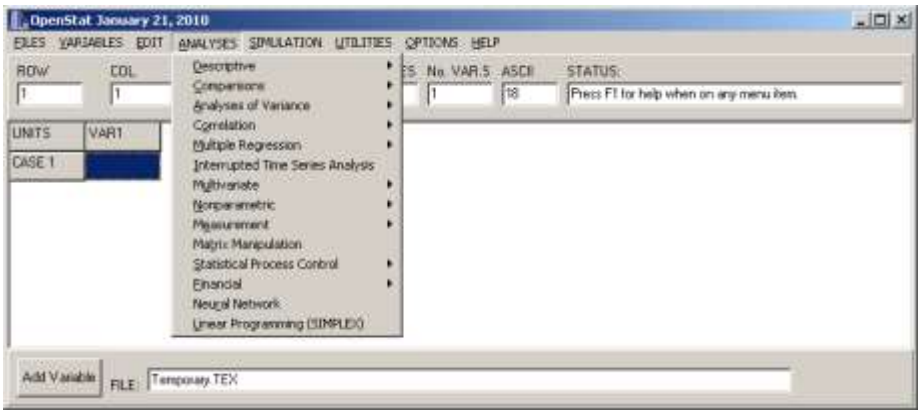
- ทางด้านการแพทย์และการสาธารณสุข เช่น โปรแกรม Epi Info (จากเวบไซต์ http://wwwn.cdc.gov/epiinfo/) เป็นโปรแกรมที่พัฒนาขึ้นโดย Department of Health and Human Service (Atlanta/Georgia/USA) เพื่องานทางด้านระบาดวิทยาและภูมิสารสนเทศ แต่ก็สามารถนำมาใช้ในการวิเคราะห์สถิติทั่วไปได้ ทำงานบนระบบปฏิบัติการ Windows

- ทางด้านประวัติศาสตร์และโบราณคดี เช่น โปรแกรม PAST พัฒนาขึ้นโดย University of Oslo สำหรับใช้ในการศึกษาเกี่ยวกับซากดึกดำบรรพ์ (Paleontology) มีความสามารถเด่นทางด้านการวิเคราะห์เชิงภูมิศาสตร์ (Geometrical analysis)

- ทางด้านการสำมะโนและการสำรวจ เช่น โปรแกรม CSPro (Census and Survey Processing System) จาก http://www.census.gov/population/international/software/cspro/ ซึ่งเป็นชุดโปรแกรมสำหรับการวิเคราะห์ข้อมูลสำมะโนและข้อมูลเชิงสำรวจ (census and survey data) พัฒนาขึ้นโดย MEASURE ซึ่งเป็นส่วนหนึ่งของ U.S. Bureau of the Census มีเครื่องมือหลายอย่างที่ช่วยในการนำเข้าข้อมูล ประมวลผลข้อมูล และนำเสนอข้อมูลจากแบบสอบถาม

- ทางด้านการศึกษา ส่วนมากเป็นโปรแกรมที่พัฒนาขึ้นโดยคณาจารย์ในแต่ละมหาวิทยาลัย และมักใช้ในจุดประสงค์เพื่อการเรียนการสอนเป็นหลัก ตัวอย่างเช่น โปรแกรม OpenStat ทำงานได้บนระบบปฏิบัติการ Windows และ Linux เป็นโปรแกรมสำหรับการวิเคราะห์ทางสถิติทั่วไป พัฒนาขึ้นโดย Bill

Miller แห่ง Iowa State University โปรแกรม OpenStat มีวัตถุประสงค์ในการพัฒนาเริ่มต้นเพื่อใช้ในการเรียนการสอนวิชาทางสถิติมีความสามารถในการคำนวณทางด้านสถิติหลากหลายดังภาพ



รูปที่ 4. OpenStat user interface.

นอกจากนั้นยังมีโปรแกรมอื่นๆ อีก เช่น

โปรแกรม Instat และ SSC-Stat พัฒนาขึ้นโดย Statistical Services Centre แห่ง University of Reading

โปรแกรม AM พัฒนาโดย American Institutes for Research, Washington/DC.

โปรแกรม BrightStat โดย Dr. Daniel Stricker แห่ง University of Bern

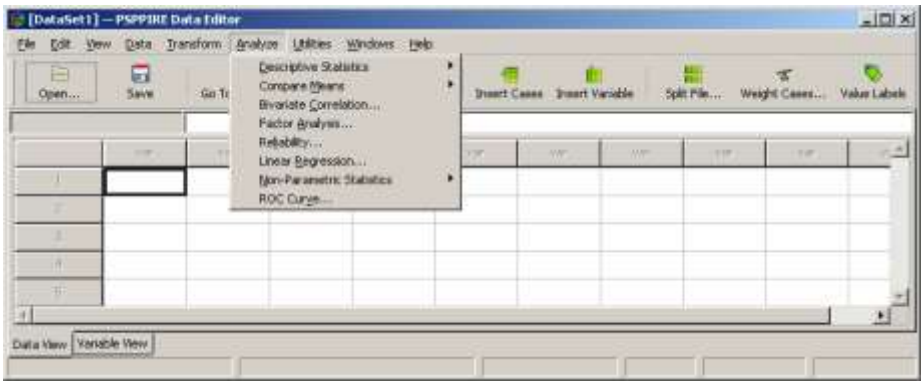
โปรแกรม Xtremes โดย Prof. R.D. Reiss (University of Siegen/Germany)

- ทางด้านประชากรศาสตร์และสังคมศาสตร์ เช่น โปรแกรม WinIDAMS พัฒนาขึ้นโดยองค์การความร่วมมือทางการศึกษาวิทยาศาสตร์ และวัฒนธรรมระหว่างประเทศ UNESCO โปรแกรมมีส่วนติดต่อผู้ใช้งานแบบกราฟฟิค ประกอบด้วยคำสั่งการคำนวณที่หลากหลาย เช่น การวิเคราะห์ถดถอย การวิเคราะห์อนุกรมเวลา การวิเคราะห์ปัจจัย การวิเคราะห์จำแนกประเภท กราฟต่างๆ หรือตัวอย่างเช่น โปรแกรม IVEware จากสถาบัน Social Research แห่ง University of Michigan

นอกจากโปรแกรมต่างๆ ดังที่กล่าวมาข้างต้นแล้ว ยังมีโปรแกรมจำนวนหนึ่งที่ถูกพัฒนาขึ้นเพื่อการวิเคราะห์โดยพื้นฐานทั่วๆ ไป และโปรแกรมทางสถิติซึ่งถูกพัฒนาขึ้นเพื่อวัตถุประสงค์จำเพาะตามเทคนิควิธีวิเคราะห์ทางสถิติ เช่น การวิเคราะห์อนุกรมเวลา การวิเคราะห์พหุตัวแปรหรืออื่นๆ ดังตัวอย่างโปรแกรมต่อไปนี้

โปรแกรม PSPP เป็นโปรแกรมแบบเปิดเผยรหัสที่มีความคล้ายคลึงกับโปรแกรม SPSS มาก เหมาะสำหรับการใช้เพื่อวิเคราะห์ข้อมูลทั่วๆ ไป โปรแกรม PSPP สามารถอ่านไฟล์ข้อมูลของ SPSS ได้ จุดประสงค์ในการพัฒนา PSPP เพื่อต้องการสร้างโปรแกรมที่สามารถใช้วิเคราะห์ข้อมูลได้เหมือน SPSS โปรแกรม PSPP

สามารถบันทึกผลลัพธ์การวิเคราะห์เป็นไฟล์ข้อความ ไฟล์ HTML หรือไฟล์ PostScript ข้อจำกัดของโปรแกรมคือยังมีคำสั่งการใช้งานที่น้อยเมื่อเปรียบเทียบกับโปรแกรม SPSS ดังภาพ



รูปที่ 5. PSPP user interface.

โปรแกรม BUGS/WinBUGS พัฒนาขึ้นโดย Alastair Stevens แห่ง Cambridge University สำหรับใช้ในเทคนิคการวิเคราะห์แบบเบย์เซียน (Bayesian analysis)

โปรแกรม VisiCube โปรแกรมสำหรับการสร้างกราฟ เพื่อใช้นำเสนอข้อมูล เหมาะกับผู้ที่ไม่มีทักษะความรู้ทางคณิตศาสตร์

โปรแกรม ADE-4 (2004) โปรแกรมสถิติสำหรับเทคนิคการวิเคราะห์พหุตัวแปร (Multivariate Analysis)

โปรแกรม EasyReg พัฒนาโดย Herman J. Bierens ศาสตราจารย์ด้านเศรษฐศาสตร์แห่ง Penn State University สำหรับการวิเคราะห์ทางเศรษฐศาสตร์ เป็นคู่แข่งขันและมีความคล้ายคลึงกับโปรแกรม E-Views ซึ่งเป็นโปรแกรมเชิงพาณิชย์

โปรแกรม Regress+ เป็นโปรแกรมฟรีสำหรับการวิเคราะห์โมเดลแบบไม่เป็นเชิงเส้นตรง

โปรแกรม NORM โดย Joe Schafer และ Maren Olsen แห่งภาควิชาสถิติ, Pennsylvania State University สำหรับการวิเคราะห์ข้อมูลเชิงปริมาณในแบบพหุตัวแปร

ฟรีแวร์หรือโอเพนซอร์ส

ในท้องตลาดโปรแกรมเชิงพาณิชย์ที่ใช้วิเคราะห์ทางสถิติเป็นโปรแกรมที่มีราคาสูง อาทิเช่น โปรแกรม SAS, SPSS เป็นต้น จึงเป็นสาเหตุหนึ่งที่ทำให้มีการพัฒนาโปรแกรมวิเคราะห์ทางสถิติในรูปแบบเปิดเผยรหัสหรือโปรแกรมฟรี

โปรแกรมทางสถิติฟรีเป็นทางเลือกในการศึกษาทดลอง ฝึกหัด และนำไปใช้ เมื่อเทียบกับโปรแกรมเชิงพาณิชย์ โดยทั่วไปโปรแกรมทางสถิติแบบฟรีให้ผลลัพธ์เช่นเดียวกับโปรแกรมเชิงพาณิชย์ และโปรแกรม

จำนวนมากยังง่ายในการเรียนรู้สามารถใช้งานด้วยระบบเมนูได้ อย่างไรก็ดีมีส่วนน้อยที่บังคับด้วยการพิมพ์คำสั่ง โปรแกรมทางสถิติฟรีเหล่านี้มาจากหลากหลายแหล่งไม่ว่าจะเป็นภาครัฐ หรือองค์กรต่างๆ หรือแม้กระทั่งนักพัฒนาอิสระ

โปรแกรมทางสถิติฟรีจำนวนหนึ่งพัฒนามาจากภาครัฐหรือองค์กรเอกชน ตัวอย่างเช่น Epi Info จาก CDC (ศูนย์ควบคุมและป้องกันโรคภัยไข้เจ็บ), IDAMS จาก UNESCO เป็นต้น โปรแกรมอีกจำนวนหนึ่งพัฒนามาจากองค์กรขนาดเล็กกว่าหรือองค์กรอิสระหรือมหาวิทยาลัย ตัวอย่างเช่น Instat หรือ Irristat โปรแกรมนอกเหนือจากนั้นเช่น R ถูกริเริ่มพัฒนาขึ้นโดยกลุ่มอาสาสมัครจำนวนมากจากทั่วโลก โดยโปรแกรม R นี้ไม่เพียงแค่เปิดเผยรหัสต้นฉบับแต่ยังเป็นฟรีโปรแกรม ตามความหมาย : สามารถแก้ไข ใช้งาน แจกจ่ายได้ตามต้องการโปรแกรมวิเคราะห์ทางสถิติในรูปแบบเปิดเผยรหัสที่นิยมใช้ในปัจจุบันได้แก่โปรแกรม R เนื่องจากมีฟังก์ชันการคำนวณที่ครอบคลุมการประยุกต์ใช้งานในแทบทุกวงการ

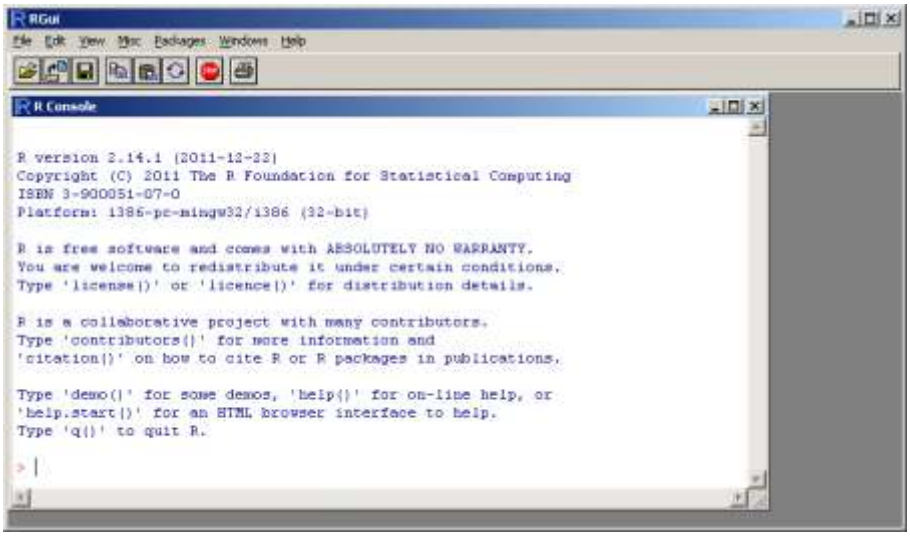
R เป็นโปรแกรมและเป็นภาษาโปรแกรมบนคอมพิวเตอร์สำหรับการวิเคราะห์ข้อมูล การประมวลผลทางด้านสถิติและกราฟ ปัจจุบันภาษา R เป็นที่นิยมใช้ในหมู่นักสถิติและนักวิเคราะห์ข้อมูล ภาษา R ได้ถูกนำมาเปรียบเทียบกับซอฟต์แวร์สถิติตัวอื่น อาทิ SAS, SPSS และ Stata โดยในปี 2552 นิวยอร์กไทมส์ได้มีบทความเกี่ยวกับภาษา R กล่าวถึงการยอมรับซอฟต์แวร์ตัวนี้ในหมู่นักสถิติ และการนำมาประยุกต์ในงานสถิติซึ่งมีผลต่อยอดขายกับซอฟต์แวร์ตัวอื่น อาทิ SAS

R เป็นซอฟต์แวร์ฟรีในรูปแบบของโอเพนซอร์ซซึ่งสามารถนำไปใช้งานและแจกจ่ายได้โดยอิสระ (ในที่นี้ ฟรี หมายถึง ไม่มีค่าใช้จ่ายโดยถูกต้องตามกฎหมาย) โปรแกรม R นอกจากจะฟรีแล้วยังเปิดเผยรหัสต้นฉบับ (open source) และยังเป็นฟรีโปรแกรม (free software) ในความหมายคือ รหัสต้นฉบับของโปรแกรมให้อิสระในการดัดแปลงแก้ไข และสามารถแจกจ่ายให้แก่ผู้อื่น โดยที่การแจกจ่ายซ้ำนั้นยังคงรักษาความเป็นฟรีโปรแกรมไว้เช่นเดิม

ภาษา R เป็นภาษาที่เหมือนกับภาษา S ซึ่งใช้ในโปรแกรม S-plus ผู้ใช้งานสามารถนำ code จากโปรแกรม S-plus มาประมวลผลในโปรแกรม R ได้ โดยแทบไม่ต้องปรับแก้โปรแกรมภาษา S เดิม โปรแกรม R สามารถทำงานได้บนหลากหลายระบบปฏิบัติการทั้ง Linux, Unix, BSD, Mac, Windows และที่สำคัญคือ R มีหลากหลายชุดคำสั่งสำหรับเทคนิคการประมวลผลทางสถิติและการสร้างกราฟแบบต่างๆ ที่มีความละเอียดสูงซึ่งเหมาะกับการนำไปใช้ตีพิมพ์ในเอกสารทางวิชาการ ชุดคำสั่งของโปรแกรม R มีเป็นจำนวนมากสำหรับใช้ในการวิเคราะห์ข้อมูล

โปรแกรม R ถูกเขียนและใช้งานโดยผู้คนจำนวนมากทั่วโลก เครื่องมือสิ่งอำนวยความสะดวก การสนับสนุนจากกลุ่มผู้ใช้คนอื่นๆ จึงมีเป็นจำนวนมากให้เรียกใช้งานได้จากบนอินเตอร์เน็ต ถึงแม้ว่าโปรแกรม R มีประสิทธิภาพสูง แต่การเรียนรู้ทำได้ยากสำหรับผู้ที่ไม่เคยหรือไม่คุ้นเคยกับศาสตร์การเขียนโปรแกรมคอมพิวเตอร์มาก่อน

ส่วนติดต่อผู้ใช้งาน (User interface) มาตรฐานของโปรแกรม R เมื่อเรียกใช้งานปรากฏดังภาพ



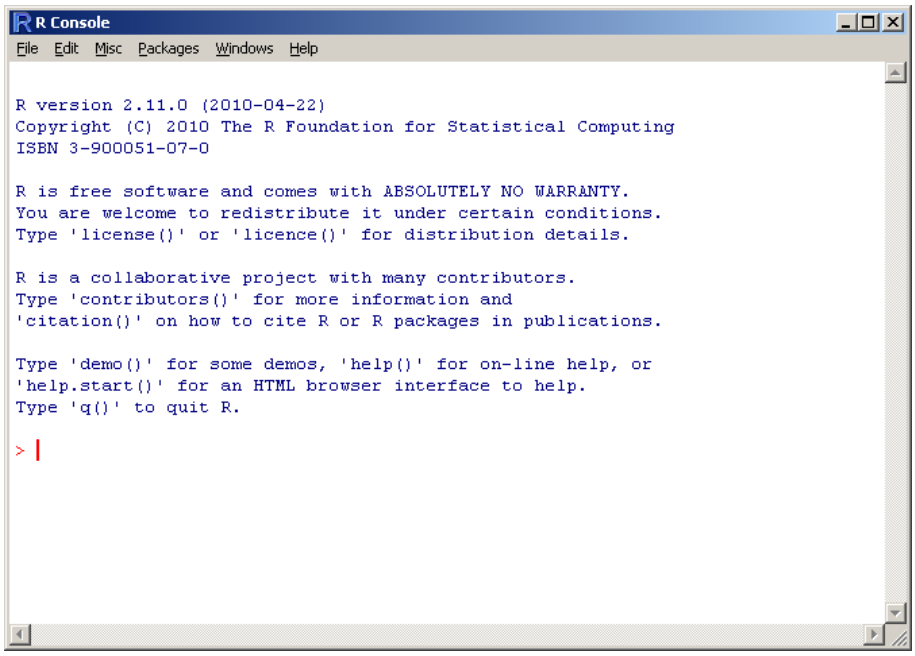
รูปที่ 6. User interface of R on Windows.

ผู้ใช้งานโปรแกรม R พึงตระหนักและยอมรับเงื่อนไขการใช้งานโปรแกรมว่าตนเองไม่สามารถฟ้องร้องเอาผิดใดๆ ทางกฎหมายกับผู้พัฒนาโปรแกรม หากตนเองได้รับความเสียหายจากการใช้งานโปรแกรม (ABSOLUTELY NO WARRANT) แต่ทั้งนี้ ผู้ใช้งานโปรแกรม R สามารถเข้ามามีส่วนร่วมในการพัฒนาโปรแกรม R ให้สมบูรณ์ยิ่งขึ้นได้ โดยสามารถแจ้งความคิดเห็น ข้อเสนอแนะ ปัญหาการใช้งานหรือข้อผิดพลาดของโปรแกรมที่พบไปยังทีมงานผู้พัฒนาโปรแกรม R-windows@R-project.org โดยสามารถตรวจสอบข้อผิดพลาดของโปรแกรม R ที่ยังไม่ได้รับการแก้ไขได้ที่ http://bugs.R-project.org/

โปรแกรม R และส่วนประกอบหลักของโปรแกรม รวมถึงรหัสต้นฉบับของโปรแกรม (Source Code) สามารถดาวน์โหลดได้จากเว็บไซต์ของโครงการ R ที่ http://www.r-project.org/ สำหรับโปรแกรมเสริม (Package) ต่างๆ ที่ถูกพัฒนาเพิ่มเติมขึ้นมาเพื่อให้ใช้งานร่วมกับโปรแกรม R โดยนักพัฒนาโปรแกรมคอมพิวเตอร์จากทั่วโลก สามารถดาวน์โหลดเพิ่มเติมได้ที่ http://cran.r-project.org/

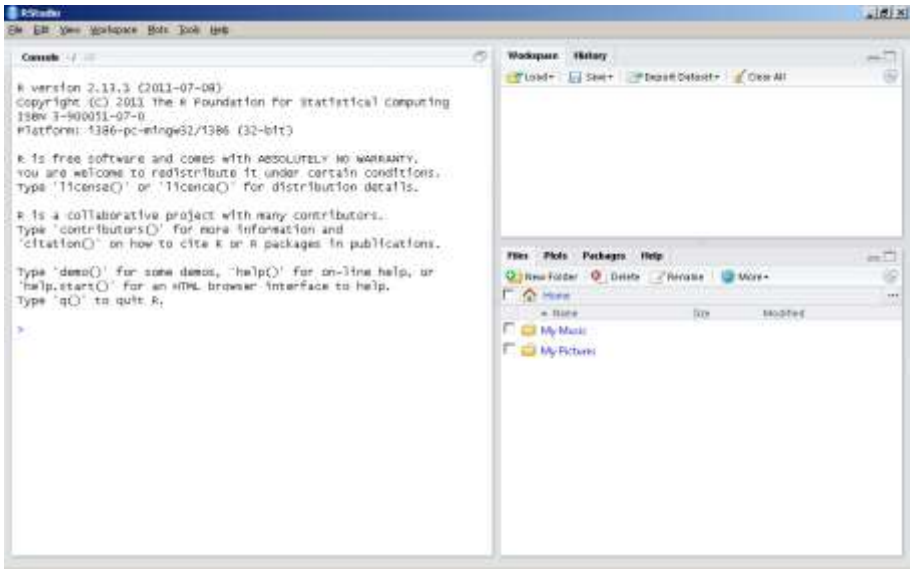
เนื่องจากโปรแกรม R เป็นโปรแกรมแบบเปิดเผยรหัสจึงมีการพัฒนาโปรแกรมต่อยอดจากโปรแกรม R โดยนักพัฒนาโปรแกรมจากทั่วโลกได้มีการปรับปรุงคุณสมบัติของโปรแกรม R เพื่อเป็นการเพิ่มเติมความสามารถและเพื่อตอบสนองความต้องการในการใช้งานเฉพาะด้าน ตัวอย่างของโปรแกรมที่ได้มีการปรับปรุงและพัฒนาขึ้นเช่น

R-Portable เป็นโปรแกรมที่พัฒนาขึ้นเพื่ออำนวยความสะดวกให้ผู้ใช้โปรแกรมสามารถเรียกใช้โปรแกรม R ได้โดยตรงจากสื่อบันทึกข้อมูลเช่น Flash Drive โดยที่ไม่ต้องเสียเวลาติดตั้งโปรแกรม R ลงในเครื่องคอมพิวเตอร์ การใช้งานโปรแกรมและส่วนติดต่อผู้ใช้งานของโปรแกรมยังคงคล้ายเดิม ดังภาพ



รูปที่ 7. R-Portable user interface.

RStudio เป็น โปรแกรมที่ปรับปรุงส่วนติดต่อผู้ใช้งานของโปรแกรม R และเพิ่มเติมคุณสมบัติของโปรแกรม R ให้สามารถทำงานในลักษณะของเครื่องแม่ข่าย เพื่อให้บริการสำหรับการวิเคราะห์ข้อมูลแก่ผู้ใช้งานได้พร้อมกันหลายคนในเวลาเดียวกัน



รูปที่ 8. RStudio user interface.

R Commander หรือเรียกในอีกชื่อ Rcmdr-package เป็นโปรแกรมที่ปรับปรุงส่วนติดต่อผู้ใช้งานของโปรแกรม R เพื่อให้สะดวกแก่การเรียกใช้คำสั่งต่างๆ จากเดิมที่ต้องพิมพ์คำสั่งเพื่อสั่งงานโปรแกรม เป็นการคลิกเลือกคำสั่งจากเมนูคำสั่งที่ได้มีการสร้างเตรียมไว้แล้ว เหมาะสำหรับผู้ที่เริ่มใช้งานหรือยังไม่คุ้นเคยกับการใช้งานคำสั่งโปรแกรมภาษา R

รูปที่ 9. R Commander user interface.

โปรแกรม R สามารถทำงานได้บนหลากหลายระบบปฏิบัติการ เมื่อเปรียบเทียบกับโปรแกรมอื่นๆ นอกจากนั้นเหตุผลสำคัญที่ทำให้โปรแกรม R ได้รับความนิยมในปัจจุบันเนื่องจากโปรแกรม R มีคำสั่งรองรับการทำงานสำหรับการวิเคราะห์ทางด้านสถิติที่หลากหลาย ครอบคลุมแทบทุกวงการ รายละเอียดดังปรากฏในตาราง

ตารางที่ 5 Operating system support

ชื่อโปรแกรม	Windows	Mac OS	Linux	BSD	Unix
ADaMSoft	+	+	+	+	+
Analyse-it	+				
BMDP	+				
Dataplot	+	+	+	+	+
Epi Info	+				
EViews	+	+			
GAUSS	+	+	+		+

ชื่อโปรแกรม	Windows	Mac OS	Linux	BSD	Unix
GraphPad Prism	+	+			
gretl	+	+	+		
JMP	+	+	+		
JHepWork	+	+	+	+	+
Maple	+	+	+		+
Matlab	+	+	+		+
Mathematica	+	+	+		+
MedCalc	+				
Minitab	+	ยกเลิก			
NCSS	+				
NMath Stats	+				
NumXL	+				
OpenEpi	+	+	+	+	+
Origin	+				
Primer	+				
PSPP	+	+	+	+	+
R Commander	+	+	+	+	+
R	+	+	+	+	+
RATS	+	+	+		+
RKWard	+		+		+
ROOT	+	+	+	+	+
Sage	บางส่วน	+	+		+
Salstat	+	+	+	+	+
SAS	+	ยกเลิก	+		+
SHAZAM	+				
SigmaXL	+				
SimStat	+				
SOCR	+	+	+	+	+
SOFA Statistics	+	+	+	+	+

ชื่อโปรแกรม	Windows	Mac OS	Linux	BSD	Unix
SPlus	+		+		+
SPSS	+	+	+		
Stata	+	+	+		+
Statgraphics	+				
STATISTICA	+				
StatPlus	+	+			
SYSTAT	+				
TSP	+	+	+	+	+
UNISTAT	+				
The Unscrambler	+				
Winpepi	+				
WPS	+	+	+		+
XLSTAT	+	+			
XploRe	+		+		+

ตารางที่ 6 โปรแกรม R เมื่อเปรียบเทียบกับโปรแกรมอื่นๆ ในด้านเทคนิคการวิเคราะห์ความแปรปรวน

ชื่อโปรแกรม	One-way	Two-way	MANOVA	GLM	Mixed model	Post-hoc	Latin squares
ADaMSoft	+	+					
Analyse-it	+	+			+		
BMDP	+	+	+	+	+	+	
Epi Info	+	+					
EViews	+						
GAUSS							
GenStat	+	+	+	+		+	+
GraphPad Prism	+	+				+	
gretl							
JMP	+	+	+	+		+	+
Mathematica	+	+	+	+		+	

ชื่อโปรแกรม	One-way	Two-way	MANOVA	GLM	Mixed model	Post-hoc	Latin squares
MedCalc	+	+		+		+	
Minitab	+	+	+	+		+	+
NCSS	+	+	+	+	+	+	+
NMath Stats	+	+					
Origin	+	+				+	
PSPP	+	+	+	+		+	+
R	+	+	+	+	+	+	+
R Commander	+	+	+	+		+	+
Sage	+	+	+		+	+	
Salstat	+						
SAS	+	+	+	+	+	+	+
SHAZAM	+	+		+		+	
SigmaXL	+	+					
SimStat	+	+		+	+	+	
SOCR	+	+				+	+
SOFA Statistics	+						
Stata	+	+	+	+	+		+
Statgraphics	+	+	+	+		+	+
STATISTICA	+	+	+	+		+	+
StatPlus	+	+	+	+		+	+
SPlus	+	+	+	+		+	+
SPSS	+	+	+	+		+	+
SYSTAT	+	+	+	+		+	+
TSP							
UNISTAT	+	+		+		+	+
The Unscrambler	+						
Winpepi	+	+					

ชื่อโปรแกรม	One-way	Two-way	MANOVA	GLM	Mixed model	Post-hoc	Latin squares
WPS	+			+		+	+
XLSTAT	+	+	+	+		+	

ตารางที่ 7 คุณสมบัติของโปรแกรม R เมื่อเปรียบเทียบกับโปรแกรมอื่นๆ ในด้านการวิเคราะห์ถดถอย

ชื่อโปรแกรม	OLS	WLS	2SLS	NLLS	Logistic	GLM	LAD	Stepwise	Quantile	Probit	Cox	Poisson	MLR
ADaMSoft	+	+		+	+			+					
Analyse-it	+												
BMDP	+				+			+			+		
Epi Info	+				+								
EViews	+	+	+	+	+	+	+	+	+	+		+	+
GAUSS													
GenStat	+	+	+	+	+	+	+	+	+	+		+	+
GraphPad Prism	+	+		+									+
gretl	+	+	+	+	+		+	+	+	+		+	
JMP	+	+		+	+	+		+		+		+	+
Mathematica	+	+		+	+	+	+		+	+		+	+
MedCalc	+			+	+			+					
Minitab	+	+		+	+			+					
NCSS	+	+		+	+	+	+	+		+	+	+	+
NMath Stats	+	+		+									+
Origin													
PSPP	+												
R	+	+	+	+	+	+	+	+	+	+	+	+	+

ชื่อโปรแกรม	OLS	WLS	2SLS	NLLS	Logistic	GLM	LAD	Stepwise	Quantile	Probit	Cox	Poisson	MLR
R Commander	+					+							
RATS	+	+	+	+	+	+	+	+	+	+		+	+
Sage	+	+	+	+	+	+	+	+	+	+		+	
Salstat													
SAS	+	+	+	+	+	+	+	+	+	+	+	+	+
SHAZAM	+	+	+	+	+	+	+	+	+	+		+	+
SigmaXL	+				+								+
SimStat	+			+	+	+		+					+
SOCR	+				+								
SPlus	+			+	+	+		+	+				
SPSS	+	+	+	+	+	+		+					
Stata	+	+	+	+	+	+	+	+	+	+	+	+	+
Statgraphics	+	+		+	+	+		+					
STATISTICA	+	+	+	+	+	+	+	+					
StatPlus	+		+	+	+	+		+					
SYSTAT	+	+	+	+	+	+	+	+					
TSP	+	+	+	+	+		+		+	+		+	+
UNISTAT	+	+	+	+	+	+		+		+		+	
The													

ชื่อโปรแกรม	OLS	WLS	2SLS	NLLS	Logistic	GLM	LAD	Stepwise	Quantile	Probit	Cox	Poisson	MLR
Unscrambler													
Winpepi	+				+					+		+	+
WPS	+	+			+	+		+					+
XLSTAT	+	+		+	+	+		+		+			

คุณสมบัติของโปรแกรม R เปรียบเทียบกับโปรแกรมอื่นๆ ในด้านการวิเคราะห์อนุกรมเวลา

ตารางที่ 8 Support for various time series analysis methods.

ชื่อโปรแกรม	ARIMA	GARCH	Unit root test	Cointegration test	VAR	Multivariate GARCH
Analyse-it						
BMDP	+					
EViews	+	+	+	+	+	+
GAUSS	+	+			+	+
GraphPad Prism						
gretl	+	+	+	+	+	
JMP	+					
Mathematica	+	+		+		
MedCalc						
Minitab	+					
NCSS	+					
NumXL	+	+				
NMath Stats						
Origin						
PSPP						
R	+	+	+	+	+	+
R Commander						
RATS	+	+	+	+	+	+
Sage	+	+	+	+	+	+
Salstat						
SAS	+	+	+	+	+	+
SHAZAM	+	+	+	+	+	
SimStat						
SOCR						
Stata	+	+	+	+	+	+

Statgraphics	+					
STATISTICA	+					
StatPlus	+					
SPlus		+			+	
SPSS	+					
SYSTAT	+					
TSP	+	+	+	+	+	
UNISTAT	+					
Winpepi						
WPS						
XLSTAT						

คุณสมบัติของโปรแกรม R เปรียบเทียบกับโปรแกรมอื่นๆ ในด้านการสร้างกราฟและแผนภูมิ

ตารางที่ 9 Support for various statistical charts and diagrams.

ชื่อโปรแกรม	กราฟแท่ง	Box plot	Correlogram	Histogram	กราฟเส้น	Scatterplot
ADaMSoft	+	+	+	+	+	+
Analyse-it						
BMDP				+		+
Epi Info	+			+	+	+
EViews	+	+	+	+	+	+
GAUSS	+	+		+	+	+
GenStat	+	+	+	+	+	+
GraphPad Prism	+	+	+	+	+	+
gretl	+	+	+	+	+	+
JMP	+	+	+	+	+	+
Mathematica	+	+		+	+	+
MedCalc	+	+		+	+	+
Minitab	+	+	+	+	+	+
NCSS	+	+	+	+	+	+
NMath Stats						

Origin	+	+	+	+	+	+
PSPP						
R	+	+	+	+	+	+
R Commander						
RATS	+	+	+	+	+	+
Sage	+	+	+	+	+	+
SAS	+	+	+	+	+	+
SHAZAM	+	+	+	+	+	+
SigmaXL	+	+		+	+	+
SimStat	+	+	+	+	+	+
SOCR	+	+	+	+	+	+
Stata	+	+	+	+	+	+
Statgraphics						
STATISTICA	+	+	+	+	+	+
StatPlus	+	+	+	+	+	+
SPlus						
SPSS	+	+	+	+	+	+
SYSTAT						
TSP			+	+	+	+
UNISTAT	+	+	+	+	+	+
The Unscrambler	+			+	+	+
Winpepi					+	+
WPS	+				+	+
XLSTAT	+	+	+	+	+	+

คุณสมบัติของโปรแกรม R เปรียบเทียบกับโปรแกรมอื่น ในเทคนิคการวิเคราะห์ด้านต่างๆ

ตารางที่ 10 Support for various statistical charts and diagrams.

ชื่อโปรแกรม	Descriptive statistics		Nonparametric statistics		Quality control	Survival analysis	Cluster analysis	Discriminant analysis
	Base statistics	Normality tests	Contingency tables analysis	Nonparametric comparison, ANOVA				
ADaMSoft	+	+	+	+			+	+
Analyse-it	+	+	+	+				
BMDP	+	+	+	+		+	+	+
Epi Info	+		+	+		+		
Gauss	+	+						
GenStat	+	+	+	+	+	+	+	+
GraphPad Prism	+	+		+		+		
JMP	+	+	+	+	+	+	+	+
Mathematica	+	+	+	+		+	+	
MedCalc	+	+	+	+	+	+		
Minitab	+	+	+	+	+	+	+	+
NCSS	+	+	+	+	+	+	+	+
NMath Stats	+	+					+	
OpenEpi	+		+					
Origin	+	+		+	+	+	+	+
PSPP	+	+						

ชื่อโปรแกรม	Descriptive statistics		Nonparametric statistics		Quality control	Survival analysis	Cluster analysis	Discriminant analysis
	Base statistics	Normality tests	Contingency tables analysis	Nonparametric comparison, ANOVA				
R	+	+	+	+	+	+	+	+
RATS	+	+						
SAS	+	+	+	+	+	+	+	+
SHAZAM	+	+						
SigmaXL	+	+	+	+	+	+		
SimStat	+	+	+	+	+			
SOCR	+	+	+	+		+	+	
SOFA Statistics			+	+				
Stata	+	+	+	+	+	+	+	+
Statgraphics	+	+	+	+	+	+	+	+
STATISTICA	+	+	+	+	+	+	+	+
StatPlus	+	+	+	+	+	+		
SPlus	+	+	+	+	+	+	+	+
SPSS	+	+	+	+	+	+	+	+
SYSTAT	+	+	+	+	+	+	+	+
TSP	+	+						
UNISTAT	+	+	+	+	+	+	+	+
The Unscrambler	+	+						

ชื่อโปรแกรม	Descriptive statistics		Nonparametric statistics		Quality control	Survival analysis	Cluster analysis	Discriminant analysis
	Base statistics	Normality tests	Contingency tables analysis	Nonparametric comparison, ANOVA				
Winpepi	+	+	+	+		+	+	
WPS	+			+			+	
XLSTAT	+	+	+	+		+	+	+

บรรณานุกรม

Andy Field. 2005. Discovering statistics using SPSS 2nd Ed. SAGE Publication, London. 779 p.

Andrea Corsini. 2012. Free Statistical Software. Available Source: http://en.freestatistics.info/stat.php, August 20, 2012.

Anonymous. 2012. What Analytics, Data mining, Big Data software you used in the past 12 months for a real project?. Available Source: http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html, August 20, 2012.

Anonymous. 2012. What programming/statistics languages you used for analytics/data mining?. Available Source: http://www.kdnuggets.com/polls/2012/analytics-data-mining-programming-languages.html, August 20, 2012.

Ashlee Vance. 2009. Data Analysts Captivated by R's Power. Available Source: http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?pagewanted=all, August 20, 2012.

Ashlee Vance. 2009. R You Ready for R?. Available Source: http://bits.blogs.nytimes.com/2009/01/08/r-you-ready-for-r/, August 20, 2012.

Ashlee Vance. 2009. SAS Warms to Open-Source One Letter at a Time. Available Source: http://bits.blogs.nytimes.com/2009/02/16/sas-warms-to-open-source-one-letter-at-a-time/, August 20, 2012.

Gordon Smyth. 2012. Statistical Packages: Free. Available Source: http://www.statsci.org/free.html, August 20, 2012.

Huck, S. W., W. H. Cormier and W. G. Bounds. 1974. Reading Statistics and Research. Harper & Row New York. Pages 52–53.

John C. Pezzullo. 2012. Free Statistical Software. Available Source: http://statpages.org/javasta2.html, August 20, 2012.

Kohout, F. J. 1974. Statistics for Social Scientists: A Coordinated Learning System. John Wiley and Sons, Inc., New York. Page 351.

Myra L. Samuels, Jeffrey A.Witmer and Andrew A. Schaffner. 2012. Statistics for the life sciences 4th Ed. Prentice Hall Boston, MA. 654 p.

Sarah Boslaugh and Paul Andrew Watters. 2008. Statistics in a Nutshell. O'Reilly Media, Sebastopol, CA. 452 p.

Weiss, R. E. 1995. The Influence of Variable Selection: A Bayesian Diagnostic Perspective. Journal of the American Statistical Association 90 (430): 619-625.

Wikipedia. 2012. Category:Statistical software. Available Source: http://en.wikipedia.org/wiki/Category:Statistical_software, August 20, 2012.

Wikipedia. 2012. Comparison of statistical packages. Available Source: http://en.wikipedia.org/wiki/Comparison_of_statistical_packages, August 20, 2012.

Wikipedia. 2012. Free statistical software. Available Source: http://en.wikipedia.org/wiki/Free_statistical_software, August 20, 2012.

Wikipedia. 2012. List of numerical analysis software. Available Source:
http://en.wikipedia.org/wiki/List_of_numerical_analysis_software, August 20, 2012.

Wikipedia. 2012. List of software categories. Available Source:
http://en.wikipedia.org/wiki/List_of_software_categories, August 20, 2012.

Wikipedia. 2012. List of statistical packages. Available Source:
http://en.wikipedia.org/wiki/List_of_statistical_packages, August 20, 2012.

Wikipedia. 2012. Outline of software. Available Source:
http://en.wikipedia.org/wiki/Lists_of_software, August 20, 2012.

ชนันกาญจน์ แสงประสาน. 2555. การทดสอบสมมติฐาน. แหล่งที่มา:
http://teacher.snru.ac.th/chanankarn/admin/document/userfiles/ch5.pdf, 17 สิงหาคม 2555

กัลยา วานิชย์บัญชา. 2552. หลักสถิติ. โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย. กรุงเทพฯ

คณาจารย์ภาควิชาสถิติ คณะวิทยาศาสตร์และเทคโนโลยี. 2554. สถิติเบื้องต้น. สำนักพิมพ์มหาวิทยาลัย กรุงเทพ. ปทุมธานี. 316 หน้า

ทรงศิริ แต้สมบัติ เปรมใจ ตรีสรานุวัฒนา สมบูรณ์ สุขพงษ์ และสายสุดา สมชิต. 2527. หลักสถิติ. ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์. 344 หน้า.

ประกายรัตน์ สุวรรณ. 2548. คู่มือการใช้โปรแกรม SPSS เวอร์ชัน 12 สำหรับ Windows. ซีเอ็ดยูเคชั่น กรุงเทพ. 330 หน้า.

พิชญ์สินี ชมภูคา. เอกสารเพิ่มเติมประกอบการสอน สถิติพื้นฐานเพื่อผู้บริหารท้องถิ่น. [ออนไลน์]. แหล่งที่มา: *http://www.hosting.cmru.ac.th/phitsinee/admin/blog/file/240811114508.pdf*, 1 7 สิงหาคม 2555.

วนิตา นุ่นเกลี้ยง, วันดี เอียดแก้ว และวิไลวรรณ หิตโกเมท. การทดสอบไคสแควร์. แหล่งที่มา: http:*//www.edu.tsu.ac.th/.../บทที่%2019%20การทดสอบไค%20-%20สแคว...,* 17 สิงหาคม 2555.

สรชัย พิศาลบุตร. 2541. สถิติธุรกิจ. วิทยพัฒน์ กรุงเทพ. 407 หน้า.