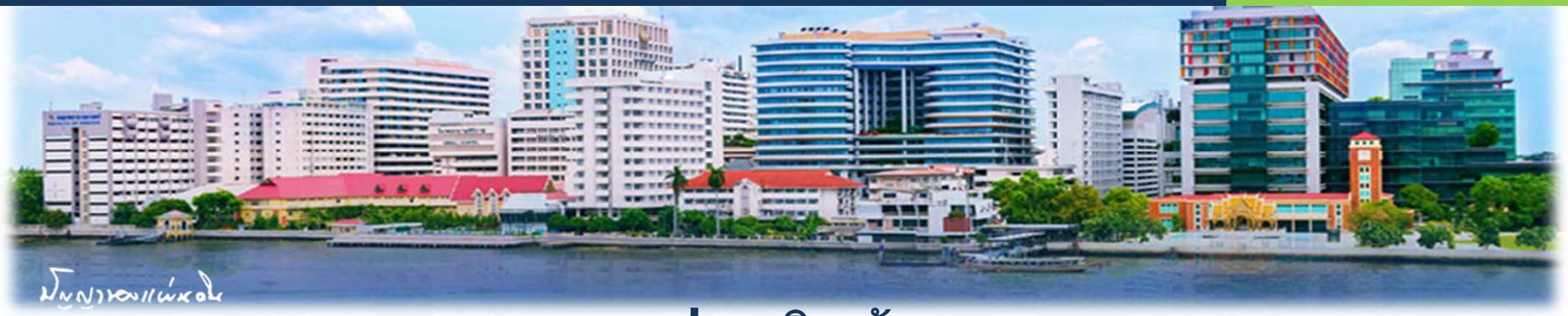
มหาวิทยาลัยมหิดล
คณะแพทยศาสตร์
ศิริราชพยาบาล

สถิติเบื้องต้น



งานประเมินต้นทุน
ฝ่ายการคลัง คณะแพทยศาสตร์ศิริราชพยาบาล

ความหมายของสถิติ

สถิติ หมายถึง ตัวเลขที่แสดงข้อเท็จจริงเกี่ยวกับเรื่องใดเรื่องหนึ่ง เช่น สถิติที่แสดงปริมาณน้ำฝน สถิติอุบัติเหตุ สถิตินักเรียน จำนวนผู้ป่วยเป็นเอดส์ของจังหวัดสุโขทัย

สถิติ หมายถึง ศาสตร์ หรือหลักการและระเบียบวิธีทางสถิติที่ว่าด้วย

1. การเก็บรวบรวมข้อมูล
2. การนำเสนอข้อมูล
3. การวิเคราะห์ข้อมูล
4. การตีความหมายข้อมูล

ประเภทของข้อมูล

1. ข้อมูลปฐมภูมิ เป็นข้อมูลที่เก็บรวบรวมจากแหล่งที่มาโดยตรง เช่น การสัมภาษณ์ การวัด การจดข้อมูลจากการทดลอง ฯลฯ ซึ่งทำได้ 2 วิธี คือ

1.1 การสำมะโน คือการเก็บรวบรวมข้อมูลจากทุกๆหน่วยของประชากรหรือเรื่องที่เราต้องการศึกษา

1.2 การสำรวจจากข้อมูลตัวอย่าง เป็นการเก็บข้อมูลจากกลุ่มตัวอย่าง เช่น การสำรวจความพึงพอใจในการทำงานของรัฐบาล การศึกษาพฤติกรรมของเด็กวัยรุ่นของไทย ฯลฯ เราเพียงสุ่มตัวอย่างให้มากพอในการศึกษาเท่านั้นไม่ได้ให้คนไทยทั้งประเทศเป็นคนตอบคำถาม

หมายเหตุ! การเก็บรวบรวมข้อมูลปฐมภูมิ นิยมใช้แบบสัมภาษณ์ การสอบถาม การทดลอง การสังเกตุจากแหล่งข้อมูลโดย ตรง โดยไม่มีผู้ใดรวบรวมไว้ก่อน

2. ข้อมูลทุติยภูมิ เป็นข้อมูลที่ถูกรวบรวมไว้เรียบร้อยแล้วตามแหล่งข้อมูลต่างๆ เช่น รายงานการส่งออก รายงานจำนวนนักเรียนของวิทยาลัยอาชีวศึกษาสุโขทัย รายงานอุบัติเหตุบนท้องถนนของปี 2553 เป็นต้น

ลักษณะของข้อมูล

ลักษณะของข้อมูล แบ่งเป็น 2 ลักษณะใหญ่ๆ คือ

1. ข้อมูลเชิงปริมาณ เป็นข้อมูลที่ใช้แทนขนาดหรือปริมาณ ซึ่งสามารถออกมาเป็นตัวเลขได้เลย เช่น จำนวนนักเรียนระดับ ปวช.1 - ปวช. 3 มีจำนวน 950 คน ปริมาณการผลิตมันสำปะหลังของปี 2549 คะแนนสูงสุดของการสอบวิชาสถิติของนักศึกษาระดับ ปวส.

2. ข้อมูลเชิงคุณภาพ เป็นข้อมูลที่ไม่สามารถวัดออกมาเป็นตัวเลขได้โดยตรง เช่น เพศ สถานภาพการสมรส วุฒิการศึกษา ความคิดเห็น เช่น ชอบมากที่สุด ชอบปานกลาง ไม่ชอบ เป็นต้น

ประเภทของสถิติ

นักคณิตศาสตร์ได้แบ่งสถิติในฐานะที่เป็นศาสตร์ออกเป็นสาขาใหญ่ ๆ 2 สาขาด้วยกัน ดังนี้

1.สถิติพรรณนา (Descriptive Statistics) หมายถึง การบรรยายลักษณะของข้อมูล (Data) ที่ผู้วิจัยเก็บรวบรวมจากประชากรหรือกลุ่มตัวอย่างที่สนใจ ซึ่งอาจจะแสดงในรูป ค่าเฉลี่ย มัธยฐาน ฐานนิยม ร้อยละ ส่วนเบี่ยงเบนมาตรฐาน ความแปรปรวน เป็นต้น

2.สถิติเชิงอนุมาน (Inferential Statistics) หมายถึง สถิติที่ว่าด้วยการวิเคราะห์ข้อมูลที่รวบรวมมาจากกลุ่มตัวอย่าง เพื่ออธิบายสรุปลักษณะบางประการของประชากร โดยมีการนำทฤษฎีความน่าจะเป็นมาประยุกต์ใช้ สถิติสาขานี้ ได้แก่ การประมาณค่าทางสถิติ การทดสอบสมมุติฐานทางสถิติ การวิเคราะห์การถดถอยและสหสัมพันธ์ เป็นต้น

ประโยชน์ของสถิติ

ประโยชน์ของสถิติมิใช่เพียงแต่ใช้เป็นเครื่องมือในการช่วยตัดสินใจ ให้เป็นไปอย่างมีประสิทธิภาพเท่านั้น ยังเป็นเครื่องมือที่ทรงคุณประโยชน์ในการประเมินผลงานโครงการต่าง ๆ ที่จัดทำไปแล้ว ว่าได้ผลตามเป้าหมายที่วางไว้เพียงไร สมควรที่จะต้องปรับปรุงหรือแก้ไขโครงการนั้น ๆ หรือไม่อย่างไรอีกด้วย

เนื่องจากสถิติมีขอบข่ายกว้างขวาง ได้รับการนำไปใช้ประโยชน์แทบทุกแขนงวิชาการ ดังนั้น นักบริหาร นักวิชาการ หรือแม้แต่สามัญชนทั่วไป จึงควรมีความรู้ทางสถิติตามสมควร กล่าวคือ อย่างน้อยก็สามารถอ่านข้อมูลจากตาราง จากแผนภูมิ หรือจากแผนภาพต่าง ๆ ให้เข้าใจได้ถูกต้อง ประโยชน์ของสถิติสรุปได้ คือ

1. ด้านการวางแผนเพื่อพัฒนา เศรษฐกิจของประเทศ
2. ด้านธุรกิจ
3. ด้านการเกษตรกรรม
4. สถิติเป็นเครื่องมือที่สำคัญยิ่งสำหรับการวิจัย ทั้งนี้เพราะ

4.1. ข้อมูลที่รวบรวมมาจากการวิจัยมีตัวเลขจำนวนมาก การนำสถิติมาจัดตัวเลขเหล่านั้นให้เป็นระเบียบ จะทำให้ผู้อ่านเข้าใจได้ถูกต้องตรงความเป็นจริงในเวลาอันรวดเร็ว

4.2. การทำงานวิจัยเป็นการศึกษาเพื่อแก้ปัญหาข้อสงสัยด้วยกระบวนการวิทยาศาสตร์ ข้อมูลที่รวบรวมมาได้ เมื่อนำมาผ่านกระบวนการทางสถิติก็จะทำให้นักวิจัยมีข้อมูลที่น่าเชื่อถือได้ประกอบการตัดสินใจ

คำศัพท์ที่เกี่ยวข้องกับ สถิติ

1. ประชากร (population) หมายถึง กลุ่มที่มีลักษณะที่เราสนใจ หรือกลุ่มที่เราต้องการจะศึกษาหาข้อมูลที่เกี่ยวข้อง เปรียบเหมือนเอกภพสัมพัทธ์ในเรื่องเซต

2. กลุ่มตัวอย่าง (sample) หมายถึง ส่วนหนึ่งของกลุ่มประชากรที่เราสนใจ ในกรณีที่กลุ่มประชากรที่จะศึกษานั้นเป็นกลุ่มขนาดใหญ่ เกินความสามารถหรือความจำเป็นที่ต้องการ หรือเพื่อประหยัดในด้านงบประมาณและเวลา สามารถศึกษาข้อมูลเพียงบางส่วนของกลุ่มประชากรได้

3. ค่าพารามิเตอร์ หมายถึง ค่าต่างๆที่คำนวณมาจากกลุ่มประชากร จะถือเป็นค่าคงตัว กล่าวคือ คำนวณกี่ครั้งๆก็จะไม่เปลี่ยนแปลง

4. ค่าสถิต หมายถึง ค่าต่างๆที่คำนวณมาจากกลุ่มตัวอย่าง จะเป็นค่าที่เปลี่ยนแปลงได้ตามกลุ่มตัวอย่างที่เลือกสุ่มมา จึงถือว่าเป็นค่าตัวแปรสุ่ม

5. ตัวแปร ในทางสถิติ หมายถึง ลักษณะบางอย่างที่เราสนใจ ค่าของตัวแปร อาจอยู่ในรูปข้อความ หรือตัวเลขก็ได้

6. ค่าที่เป็นไปได้ หมายถึง ค่าของตัวแปรที่อาจจะเกิดขึ้นได้จริง

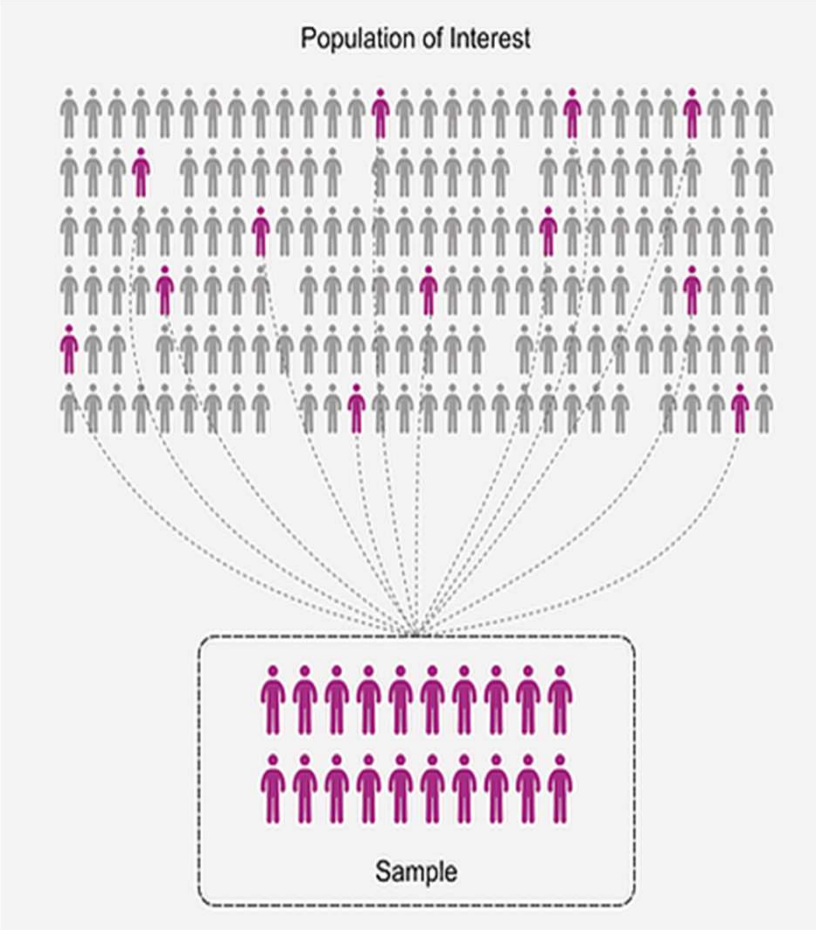
7. ค่าจากการสังเกต หมายถึง ค่าที่เก็บรวบรวมได้มาจริงๆ

ประชากร กับ ตัวอย่าง

Population คือ ประชากร หรือ ข้อมูลทั้งหมด (หรือขนาดใหญ่) ของสิ่งที่เรากำลังสนใจ เช่น นักเรียนทั้งหมด ลูกค้าทั้งหมด สินค้าทั้งหมด เป็นต้น

Sample คือ ข้อมูลส่วนหนึ่งจากประชากรที่เรานำมาใช้วิเคราะห์

สิ่งที่สำคัญในส่วนนี้ คือ จำนวน Sample ที่เราดึงมาใช้นั้น มีจำนวนและความหลากหลายมากพอที่จะเป็นตัวแทนของประชากรทั้งหมดหรือไม่ เช่น หากจะวิเคราะห์ปัจจัยในการเกิดโรคไข้หวัดใหญ่ของคนไทย ที่มีประชากร 80 ล้านคน เราควรศึกษากลุ่มตัวอย่างปริมาณกี่คน จากภูมิภาคไหน อายุเท่าไหร่ อาชีพอะไร และมีกิจกรรมในชีวิตประจำวันอย่างไรบ้าง ซึ่งคำถามทั้งหมดนี้ ไม่มีคำตอบที่ชัดเจน เพียงแต่นักวิเคราะห์ต้องสามารถหาเหตุผลมาตอบเพื่อให้กลุ่มตัวอย่างที่นำมาวิเคราะห์มีน้ำหนักชัดเจนพอ



การวัดแนวโน้มเข้าสู่ส่วนกลาง (measures of central tendency)

Mean หรือ ค่าเฉลี่ย ทางสถิติ หมายถึง ค่ากลางของข้อมูลแต่ละชุด

เช่น **9, 3, 1, 8, 3, 6** ค่าเฉลี่ย = (9 + 3 + 1 + 8 + 3 + 6) / 6 = 5

Median หรือ มัธยฐาน หมายถึง ค่ากึ่งกลางของข้อมูลชุดนั้น หรือค่าที่อยู่ในตำแหน่งกึ่งกลางของข้อมูลชุดนั้น เมื่อได้จัดเรียงค่าของข้อมูลจากน้อยที่สุดไปหามากที่สุด หรือจากมากที่สุดไปหาน้อยที่สุด ค่ากึ่งกลางจะเป็นตัวแทนที่แสดงว่ามีข้อมูลที่มากกว่าและน้อยกว่านี้อยู่ 50%

เช่น **9, 3, 1, 8, 3, 6**

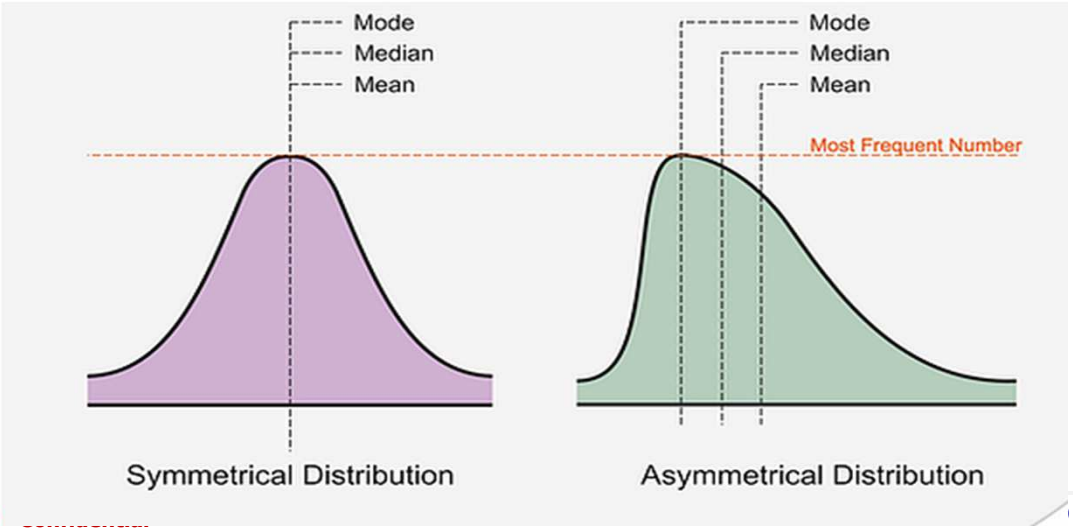
จัดเรียงเป็น 1, 3, 3, 6, 8, 9 มีมัธยฐานคือค่าระหว่าง 3 กับ 6 ได้แก่ 4.5

Mode หรือ ฐานนิยม ค่าของคะแนนที่ซ้ำกันมากที่สุดหรือ ค่าคะแนนที่มีความถี่สูงที่สุดในข้อมูลชุดนั้น

เช่น **9, 3, 1, 8, 3, 6** ฐานนิยม คือ **3**

ในกรณีที่หาค่ามัชฌิเลขคณิตและมัธยฐานได้แล้ว สามารถที่จะนำมาคำนวณหาฐานนิยมได้ โดยใช้สูตร

Mode = 3Median - 2Mean



การวัดการกระจาย (Measure of Dispersion)

Range หรือ พิสัย คือความแตกต่างระหว่างข้อมูลที่มีค่าสูงสุด และ ต่ำสุด
เช่น **9, 3, 1, 8, 3, 6** พิสัย คือ 9 – 1 =8

Variance หรือ ค่าแปรปวน
ใช้เพื่อวัดการกระจายของข้อมูลคิดจากค่าเฉลี่ยของความต่างจากค่าเฉลี่ยยกกำลัง 2

Standard deviation หรือ ค่าเบี่ยงเบนมาตรฐาน เพื่อดูการกระจายข้อมูลจากค่าเฉลี่ย คือ รากที่ 2 ของค่าแปรปวน

Variance

$$\sigma^2 = \frac{\sum (X - \bar{X})^2}{N}$$

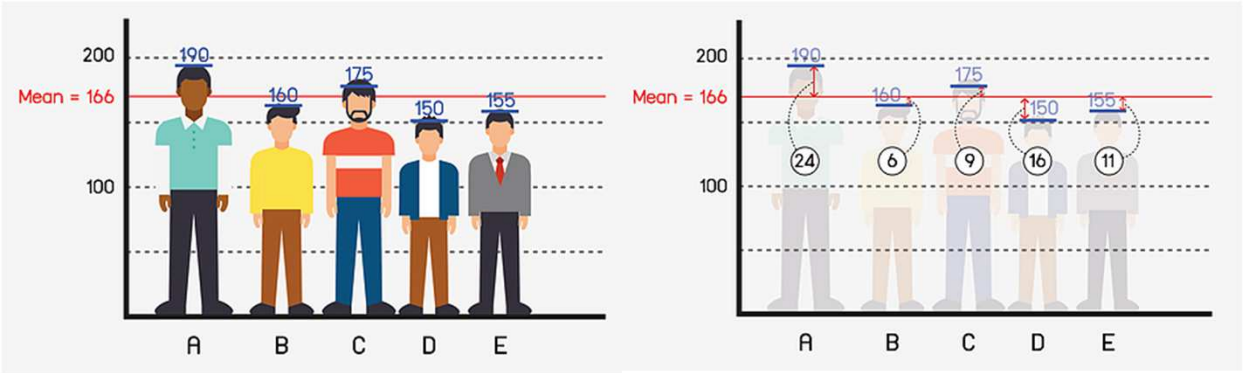
Standard deviation

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

σ = ค่าเบี่ยงเบนมาตรฐาน
σ^2 = ค่าแปรปวน
X = ค่าตัวแปร
$\bar{X}$ = ค่าเฉลี่ย
N = จำนวนตัวแปรทั้งหมด

Business Intelligence By Coraline

แล้ว Variance กับ Standard deviation ใช้งานต่างกันอย่างไร เพื่อให้เข้าใจง่ายๆ ขอเสนอตัวอย่างดังนี้



ค่าเฉลี่ย คือ Mean $\frac{190+160+175+150+155}{5} = 166$

ค่าแปรปวน คือ Variance $\frac{24^2+(-6)^2+9^2+(-16)^2+(-11)^2}{5} = 214$

ค่าเบี่ยงเบนมาตรฐาน คือ Standard Deviation $\sqrt{214} = 14.63$

ค่ามาตรฐานหรือคะแนนมาตรฐาน (Standard Scores)

ค่ามาตรฐาน เป็นค่าที่บอกให้ทราบความแตกต่างระหว่างค่าของข้อมูลนั้นกับค่าเฉลี่ยเลขคณิตของข้อมูลชุดนั้นเป็นกี่เท่า ของส่วนเบี่ยงเบนมาตรฐาน

สูตร $Z_i = \frac{Xi - \overline{X}}{S.D.}$

เมื่อ Z_i คือ คะแนนมาตรฐาน

XI คือ คะแนนดิบของข้อมูลที่มีค่าเฉลี่ยเลขคณิตที่จะแปลเป็นคะแนนมาตรฐาน

$\overline{X}$ คือ ค่าเฉลี่ยเลขคณิต

S.D. คือ ส่วนเบี่ยงเบนมาตรฐาน

ในการเปรียบเทียบค่าคะแนนของข้อมูลที่มาจากข้อมูลต่างชุดกัน ว่าจะมีความแตกต่างกันอย่างไร ซึ่งบางครั้งไม่สามารถเปรียบเทียบได้โดยตรง เพราะค่าเฉลี่ยเลขคณิตและส่วนเบี่ยงเบนมาตรฐานของข้อมูลมักจะไม่เท่ากัน ในการเปรียบเทียบให้มีความถูกต้องจึงมีความจำเป็นต้องมีการเปลี่ยน คะแนนของข้อมูลทั้งสองชุดให้เป็นค่ามาตรฐาน(ซึ่งมีค่าเฉลี่ยเลขคณิตและส่วนเบี่ยงเบนมาตรฐาน เท่ากันเสียก่อน) จึงจะเปรียบเทียบข้อมูล 2 ชุดนี้ได้ ในการเปลี่ยนค่าของข้อมูลของตัวแปรหรือข้อมูลแต่ละตัวให้เป็นค่ามาตรฐานที่นิยมใช้คือเปลี่ยนให้มีค่าเฉลียเลขคณิตเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1

ค่ามาตรฐานหรือคะแนนมาตรฐาน (Standard Scores)

ข้อสังเกต

1.คะแนนมาตรฐานเป็นตัวเลขไม่มีหน่วย

2.ส่วนเบี่ยงเบนมาตรฐานของคะแนนมาตรฐานทั้งหมดของชุดข้อมูล จะมีค่าเท่ากับ 1

3.คะแนนมาตรฐานของข้อมูลใดๆ จะเป็นบวก หรือลบก็ได้ขึ้นอยู่กับค่าของข้อมูลนั้นๆ กับมัชฌิมเลขของข้อมูลชุดนั้นว่าค่าใดมีค่ามากกว่ากัน

4.คะแนนมาตรฐานโดยทั่วไปจะมีค่า - 3 ถึง +3 แต่อาจจะมีบางข้อมูลที่มีคะแนนมาตรฐานสูงหรือต่ำกว่านี้เล็กน้อย

5.เมื่อแปลงข้อมูลทุกๆ ค่าในข้อมูลชุดใดชุดหนึ่งให้เป็นคะแนนมาตรฐานแล้วทำค่ามาตรฐานเหล่านั้นมาคำนวณหาค่ามัชฌิมเลขคณิตจะได้เท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานจะได้เท่ากับ 1 (คะแนนมาตรฐานจะมีมัชฌิมเลขคณิตเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1)

6. จะได้ $Z > 0$ และ จะได้ $Z < 0$

Ex1. สมศักดิ์สอบวิชาสถิติและภาษาอังกฤษ ซึ่งมีคะแนนเต็ม 100 คะแนนเท่ากัน ในการสอบสายใจได้คะแนน 75 และ 85 คะแนนตามลำดับ ถ้าค่ามัชฌิมเลขคณิตและส่วนเบี่ยงเบนมาาตรฐานของคะแนนสถิติของนักศึกษากลุ่มนี้คือ 60 และ 10 ค่าเลขคณิตและส่วนเฉลี่ยเบี่ยงเบนมาตรฐานของวิชาภาษาอังกฤษคือ 70 และ 12 ตามลำดับ แล้วจงเปรียบเทียบว่าสมศักดิ์เรียนวิชาไหนได้ดีกว่ากัน

วิธีทำ

สูตร $Z_i = \frac{Xi - \overline{X}}{S.D.}$

ค่ามาตรฐานของคะแนนวิชาสถิติ คือ $\frac{75-60}{10} = 1.5$

ค่ามาตรฐานของคะแนนวิชาภาษาอังกฤษ คือ $\frac{85-70}{12} = 1.25$

ค่ามาตรฐานของคะแนนวิชาสถิติของสมศักดิ์สูงกว่าค่ามาตรฐานของคะแนนวิชาภาษาอังกฤษ แสดงว่าสมศักดิ์เรียนวิชาสถิติได้ดีกว่าวิชาภาษาอังกฤษ

การประมาณค่า (Estimation)

การประมาณค่า เป็นวิธีการใช้ค่าสถิติที่ได้จากตัวอย่างไปประมาณค่าพารามิเตอร์ เป็นการหาข้อสรุปที่เกี่ยวกับพารามิเตอร์ ในลักษณะของการประมาณ ซึ่งมักแสดงในรูปตัวเลข เช่น ประมาณค่าเฉลี่ยของประชากร ประมาณค่าสัดส่วนของประชากร เป็นต้น อาจกล่าวได้ว่ากระบวนการในการประมาณค่า เป็นการนำตัวเลข ค่าสถิติที่ได้มาจากกลุ่มตัวอย่าง ไปประมาณหาค่าความจริงระดับประชากร ในเรื่องเดียวกันนั้น การประมาณค่ามี 2 แบบ คือ

1. การประมาณค่าแบบจุด (Point Estimation)

เป็นการประมาณค่าพารามิเตอร์ของประชากรด้วยค่าเพียงค่าเดียว (Single valued Estimation หรือ Point Estimation) ซึ่งการประมาณค่าแบบนี้อาจจะมีค่าเท่ากับค่าพารามิเตอร์หรืออาจมีโอกาสที่จะได้ค่าที่คาดเคลื่อนไปจากค่าพารามิเตอร์ได้มาก ทั้งนี้ขึ้นอยู่กับหน่วยตัวอย่างที่นำมาวิเคราะห์ (ถ้าหน่วยตัวอย่างนั้นได้มาจากการสุ่มตัวอย่าง ก็จะสามารถควบคุมความคลาดเคลื่อนได้ระดับหนึ่ง)

หมายเหตุ การประมาณค่าพารามิเตอร์ซึ่งเป็นลักษณะของประชากรโดยใช้ข้อมูลตัวอย่าง หรือทำการประมาณค่าพารามิเตอร์ด้วยค่าสถิติ สัญลักษณ์ที่ใช้แทนค่าพารามิเตอร์และค่าสถิติ เช่น

ลักษณะที่ต้องการทราบ	ค่าสถิติ	ค่าพารามิเตอร์
ค่าเฉลี่ย	$\bar{x}$	μ
ค่าสัดส่วน	$\hat{p}$	p

ค่าที่ประมาณได้จากการประมาณค่าแบบจุด จะมีลักษณะเป็นตัวเลขประมาณค่าเดียว เรียกเป็นค่าประมาณ (estimate) เช่น ประมาณค่าเฉลี่ยค่าใช้จ่ายในการรักษาพยาบาลของผู้ป่วยโรคหอบที่มารับการรักษาในโรงพยาบาลสุโขทัย เท่ากับ 4,950.-บาท

การประมาณค่า (Estimation)

2. การประมาณค่าแบบช่วง (Interval Estimation)

ค่าที่ประมาณได้จากการประมาณค่าแบบช่วง จะได้ช่วงของตัวเลขที่ประมาณ เรียก ช่วงการประมาณ เช่น ค่าเฉลี่ยค่าใช้จ่ายในการรักษาพยาบาลของผู้ป่วยโรคหอบที่มารับการรักษา ในโรงพยาบาลสุโขทัยอยู่ระหว่าง 3,370 - 6,480. บาท ในการประมาณค่าแบบช่วง นิยมเขียนเป็นสัญญลักษณ์ทางคณิตศาสตร์ แทนค่าที่ทำการประมาณโดยครอบคลุมค่าต่ำสุด - สูงสุด เช่น หากเป็นการประมาณค่าเฉลี่ยของประชากร คือ μ จะเขียนเป็น $a < \mu < b$ เรียกค่า a และ b ว่า ค่าต่ำสุด และค่าสูงสุดของช่วงประมาณ μ ในการประมาณค่าแบบช่วง นอกจาก ขึ้นอยู่กับค่าที่ต้องการประมาณ และการแจกแจงความน่าจะเป็นของค่าที่ต้องการประมาณแล้ว ยังขึ้นกับระดับความเชื่อมั่น (confidence level) อีกด้วย ระดับความเชื่อมั่นนี้ จะเป็นค่าที่บอกเราว่า ช่วงประมาณที่สร้างขึ้นจะครอบคลุมค่าพารามิเตอร์ด้วยความน่าจะเป็นมากน้อยเพียงใด

ช่วงความเชื่อมั่น (confidence interval) หมายถึง ช่วงของค่าประมาณที่ประกอบไปด้วยค่าต่ำสุด (a) และค่าสูงสุด (b) ที่คำนวณขึ้นมา ช่วงดังกล่าวจะคลุมค่าของพารามิเตอร์ ด้วยความน่าจะเป็นตามที่กำหนด ตัวอย่างเช่น ช่วงความเชื่อมั่น 90% ของค่าใช้จ่ายโดยเฉลี่ยของผู้ป่วยโรคหอบที่มารับการรักษา ในโรงพยาบาลสุโขทัย อยู่ระหว่าง 3,370 - 6,480.-บาท หมายถึงว่า “มีความมั่นใจ 90% ที่ช่วงของการประมาณค่าใช้จ่ายโดยเฉลี่ยที่ได้ (3,370 - 6,480.-บาท) จะครอบคลุมค่าใช้จ่ายที่เป็นค่าเฉลี่ยจริงของผู้ป่วย” ที่ระดับบความเชื่อมั่นนี้ จะเขียนแทนด้วยสัญญลักษณ์ $(1-\alpha)100\%$ โดยที่ $0 < \alpha < 1$ และเรียกค่า $1-\alpha$ ว่า สัมประสิทธิ์ของความเชื่อมั่น (confidence coefficient)

การประมาณค่าสัดส่วน

การประมาณค่าสัดส่วนประชากรกลุ่มเดียว

ในกรณีที่ข้อมูลเป็นข้อมูลเชิงคุณภาพไม่สามารถหาค่าเฉลี่ยได้ เช่น ลักษณะของสินค้า คุณภาพของสินค้า ดังนั้น จึงสนใจการประมาณสัดส่วนของประชากร โดยประมาณค่าของลักษณะ ต่าง ๆ ที่สนใจ

การประมาณค่าสัดส่วนของประชากรกลุ่มเดียวมี 2 แบบ คือการประมาณค่าสัดส่วนประชากรกลุ่มเดียวแบบจุด และ การประมาณค่าสัดส่วนประชากรกลุ่มเดียวแบบช่วง

การประมาณค่าสัดส่วน

1. การประมาณค่าสัดส่วนประชากรกลุ่มเดียวแบบจุด

ในกรณีที่ข้อมูลเป็นข้อมูลเชิงคุณภาพ และต้องการประมาณสัดส่วนของประชากรของลักษณะที่สนใจ จึงแบ่งประชากรได้ 2 ลักษณะคือ ลักษณะที่สนใจกับลักษณะที่ไม่สนใจ นั่นคือ ประชากรมีการแจกแจงแบบทวินาม

ให้ p เป็นสัดส่วนของประชากรที่สนใจ

q เป็นสัดส่วนของประชากรที่ไม่สนใจ

โดยที่ $q = 1 - p$

ในการประมาณค่า p จะใช้ $\hat{p}$ ซึ่งเป็นสัดส่วนของตัวอย่างเป็นตัวประมาณค่าสัดส่วน ของประชากร p โดยที่

$$\hat{p} = \frac{\text{จำนวนตัวอย่างที่มีลักษณะสนใจ}}{\text{จำนวนตัวอย่างทั้งหมด}}$$

Ex.1 จากการสอบถามนักศึกษาหญิงจำนวน 120 คน พบว่าใช้น้ำหอมยี่ห้อ A จำนวน 64 คน จงประมาณสัดส่วนของผู้ใช้น้ำหอมยี่ห้อนี้

วิธีทำ

ให้ $\hat{p}$ เป็นตัวประมาณสัดส่วนของผู้ใช้น้ำหอมยี่ห้อ A

$$\hat{p} = \frac{\text{จำนวนตัวอย่างที่มีลักษณะสนใจ}}{\text{จำนวนตัวอย่างทั้งหมด}}$$

$$\hat{p} = \frac{64}{120}$$

$$\hat{p} = 0.5333$$

นั่นคือ ค่าประมาณสัดส่วนแบบจุดของผู้ใช้น้ำหอมยี่ห้อ A ในกลุ่มนักศึกษาหญิงเท่ากับ 0.5333 หรือ 53.33%

การประมาณค่าสัดส่วน

2. การประมาณค่าสัดส่วนประชากรกลุ่มเดียวแบบช่วง

การสร้างช่วงความเชื่อมั่น p ขึ้นอยู่กับการแจกแจง ตัวประมาณ $\hat{p}$ มีค่าไม่ใกล้ 0 หรือ 1 มากเกินไป $\hat{p}$ จะมีการแจกแจงปกติ จะใช้ $\hat{p}$ หาค่าความแปรปรวน

ค่าความแปรปรวน = $\frac{\hat{p}(1-\hat{p})}{n}$

เมื่อ n มีขนาดใหญ่ p จะมีการแจกแจงปกติ จะได้ค่าประมาณสัดส่วนแบบช่วง คือ

$$\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Ex.2 จากการสุ่มนักศึกษาหญิงวิทยาลัยอาชีวศึกษาสุโขทัย จำนวน 100 คน พบว่ามาเรียนที่วิทยาลัยโดยการอยู่หอพัก จำนวน 60 คน จงประมาณสัดส่วนของนักศึกษาหญิงที่มาเรียนโดยการอยู่หอพักที่ระดับความเชื่อมั่น 95 %

วิธีทำ

$$\hat{p} = \frac{\text{จำนวนตัวอย่างที่มีลักษณะสนใจ}}{\text{จำนวนตัวอย่างทั้งหมด}}$$

$$\hat{p} = \frac{60}{100}$$

$$\hat{p} = .60$$

$$1 - \hat{p} = 0.4$$

ค่าประมาณแบบช่วงของ p คือ

$$\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

ที่ระดับความเชื่อมั่น 95 % ค่า $Z_{\frac{\alpha}{2}} = 1.96$

ค่าประมาณแบบช่วง คือ

$$0.60 - 1.96\sqrt{\frac{0.6(0.4)}{100}} < p < 0.60 + 1.96\sqrt{\frac{0.6(0.4)}{100}}$$

$$0.60 - 1.96(0.049) < p < 0.60 + 1.96(0.049)$$

$$0.60 - 0.096 < p < 0.60 + 0.096$$

$$0.504 < p < 0.696$$

ที่ระดับความเชื่อมั่น 90 % ค่าประมาณของนักศึกษาหญิงวิทยาลัยอาชีวศึกษาสุโขทัยที่มาเรียนโดยอยู่หอพักอยู่ระหว่าง 0.504 ถึง 0.696

การทดสอบสมมติฐาน (Hypothesis Testing)

การทดสอบสมมติฐาน (Hypothesis Testing) เป็นส่วนหนึ่งของสถิติเชิงอนุมาน (Statistical Inference) ซึ่งเป็นการทดสอบเกี่ยวกับพารามิเตอร์ที่ไม่ทราบค่า โดยสุ่มตัวอย่างจากประชากรแล้ว อาศัยการแจกแจงของตัวสถิติ สร้างสถิติทดสอบเกี่ยวกับพารามิเตอร์นั้นๆ

ศัพท์ที่ควรรู้ในการทดสอบสมมติฐาน

ในการทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ที่ไม่ทราบค่าใดๆ จึงควรรู้จัก ความหมายหรือนิยามของคำศัพท์ต่างๆ ดังต่อไปนี้

1. สมมติฐาน คือ ความเชื่อหรือคำกล่าวอ้างยืนยันเกี่ยวกับลักษณะของประชากร ซึ่งอาจมีเพียงประชากรเดียวหรือหลายประชากรก็ได้
2. สมมติฐานที่จะทดสอบ เรียกว่า สมมติฐานหลัก (Null Hypothesis) เขียนแทนด้วย H_0 สมมติฐานที่แย้งกับสมมติฐานหลัก และนำมาพิจารณาในการทดสอบด้วย เรียกว่า สมมติฐานแย้ง หรือสมมติฐานรอง (Alternative Hypothesis) ซึ่งแทนด้วย H_1
3. บริเวณยอมรับ (Acceptance region) คือบริเวณที่ทำให้ เกิดการยอมรับ H_0 ส่วน บริเวณปฏิเสธ (Rejection region) หรือบริเวณวิกฤต (Critical region) คือบริเวณที่ทำให้เกิดการ ปฏิเสธ H_0
4. ผลการตัดสินใจจากการทดสอบสมมติฐาน เนื่องจากสมมติฐานที่จะทดสอบ (H_0) เป็นความเชื่อ หรือคำยืนยันเกี่ยวกับลักษณะของ ประชากรซึ่งยังไม่สามารถบอกได้วาเป็นจริงหรือเท็จ จนกว่าจะทำการพิสูจน์โดยเก็บรวบรวม ข้อมูลทั้งหมด มาวิเคราะห์ตามลักษณะของประชากรที่ต้องการพิสูจน์นั้น ซึ่งบางครั้งการเก็บรวมรวมข้อมูลทั้งหมดจากประชากรเป็นสิ่งที่ทำได้ยากเพราะต้องเสียค่าใช้จ่าย และเวลามาก จึงทำได้เพียงการสำรวจจากตัวอย่าง เพื่อทำการทดสอบเท่านั้นเอง ดังนั้นผลการตัดสินใจจากการทดสอบ สมมติฐานใดๆ สามารถสรุปได้ ดังตาราง

การทดสอบสมมติฐาน (Hypothesis Testing)

ตารางผลการตัดสินใจจากการทดสอบสมมติฐาน

การตัดสินใจ	ข้อเท็จจริง	
	H_0 เป็นจริง	H_0 เป็นเท็จ
ปฏิเสธ H_0	ความผิดพลาดประเภทที่ 1	ตัดสินใจถูก
ยอมรับ H_0	ตัดสินใจถูก	ความผิดพลาดประเภทที่ 2

ผลการทดสอบไม่ว่าจะยอมรับหรือปฏิเสธสมมติฐานหลัก ย่อมอาจมีความผิดพลาด เกิดขึ้นได้ 2 กรณีเสมอ คือ

1) การปฏิเสธ H_0 เมื่อ H_0 เป็นจริง เรียกว่า ความผิดพลาดประเภทที่ 1
2) การยอมรับ H_0 เมื่อ H_0 เป็นเท็จ เรียกว่า ความผิดพลาดประเภทที่ 2

5. ขนาดของความผิดพลาดประเภทที่1 (Size of a type I error) คือ ความน่าจะเป็นที่จะเกิดความผิดพลาดประเภทที่ 1 เขียนแทนด้วย α เราเรียกว่า ระดับนัยสำคัญ (Level of significant) และขนาดของความผิดพลาดประเภทที่ 2 (Size of a type II error) คือ ความน่าจะเป็นที่จะเกิดความผิดพลาดประเภทที่ 2 แทนด้วย β และเรียก $1-\beta$ ว่า กำลังของการทดสอบ (Power of the test) ในการทดสอบสมมติฐานผู้ทดสอบต้องพยายามควบคุมความผิดพลาดทั้งสองประเภทให้มีโอกาสเกิดขึ้นน้อยที่สุดแต่ขนาดของความผิดพลาดสองประเภทนี้สวนทางกัน กล่าวคือ ถ้า α มีค่ามากแล้ว β จะมีค่าน้อย การควบคุมความผิดพลาดทั้งสองประเภทนี้สามารถลดลงได้ถ้าเพิ่มขนาดตัวอย่างให้มากขึ้น

ประเภทของการทดสอบสมมติฐาน

ในการทดสอบสมมติฐานใดๆ เราจะยอมรับว่าสมมติฐานหลักเป็นจริงก่อน จึงทำการสุ่ม ตัวอย่าง และคำนวณค่าสถิติที่ได้จากตัวอย่างที่สุ่ม ถ้าค่าสถิติที่ใช้ในการทดสอบนั้นแตกต่างจาก พารามิเตอร์ที่กำหนดใน H_0 มากเพียงพอที่จะปฏิเสธ H_0 เราจึงจะปฏิเสธ H_0 หรือกล่าวว่าแตกต่าง อย่างมีนัยสำคัญ 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) ถ้าทดสอบสมมติฐาน แบบมีทิศทางจะเรียกว่าการทดสอบแบบทางเดียว แต่ถ้าทดสอบสมมติฐาน แบบไม่มีทิศทางจะ เรียกว่า การทดสอบแบบสองทาง

ประเภทของการทดสอบสมมติฐาน

การทดสอบแบบทางเดียว (One - Tailed Test)

ให้ θ เป็นพารามิเตอร์ ที่ต้องการทดสอบ และให้ θ_0 เป็นค่าคงที่ที่ต้องการทดสอบหรือเป็นค่าพารามิเตอร์ที่คาดหวังไว้ นั้นเอง สมมติฐานที่จะทดสอบอยู่ในลักษณะ

1. $H_0 : \theta = \theta_0$
   $H_1 : \theta > \theta_0$

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

บริเวณวิกฤตของการทดสอบจะอยู่ด้านขวาและมีค่าเป็นบวก

2. $H_0 : \theta = \theta_0$
   $H_1 : \theta < \theta_0$

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

$$\frac{3(4.4721)}{5}$$

บริเวณวิกฤตของการทดสอบจะอยู่ด้านซ้ายและมีค่าเป็นลบ

ประเภทของการทดสอบสมมติฐาน

การทดสอบแบบทางสองทาง (Two - Tailed Test)

ให้ θ เป็นพารามิเตอร์ที่ต้องการทดสอบ และให้ $\theta < \theta_0$ เป็นค่าคงที่ที่ต้องการทดสอบหรือเป็นค่าพารามิเตอร์ที่คาดหวังไว้ สมมติฐานที่จะทดสอบจะอยู่ในลักษณะ

$H_0 : \theta = \theta_0$
$H_1 : \theta \neq \theta_0$

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

บริเวณวิกฤตของการทดสอบจะอยู่ด้านซ้ายและขวามีค่าเป็นได้ทั้งบวกและลบ

ขั้นตอนการทดสอบสมมติฐาน

ขั้นตอนการทดสอบสมมติฐานทางสถิติมีดังนี้

1. ตั้งสมมติฐานหลัก (H_0) และสมมติฐานทางเลือก (H_1) ให้มีความหมายตรงข้ามกันเสมอ
2. กำหนดระดับนัยสำคัญ α
3. เลือกตัวสถิติทดสอบที่เหมาะสม แล้วหาจุดวิกฤตเพื่อกำหนดบริเวณปฏิเสธ H_0 ให้ สอดคล้องกับ H_0 และ α
4. คำนวณค่าสถิติที่ใช้ทดสอบจากตัวอย่างขนาด n ที่สุ่มมา
5. ตัดสินใจยอมรับหรือปฏิเสธ H_0 โดยพิจารณาจากเงื่อนไขนี้ ถ้าค่าสถิติทดสอบที่คำนวณได้จากขั้นตอนที่ 4 ตกอยู่ในบริเวณยอมรับ เราจะตัดสินใจยอมรับ H_0 แต่หากตกอยู่บริเวณปฏิเสธ จะตัดสินใจปฏิเสธ H_0
6. สรุปผล

การทดสอบสมมติฐานค่าเฉลี่ยประชากร

การทดสอบสมมติฐานค่าเฉลี่ยประชากรกลุ่มเดียว

เมื่อ μ คือค่าเฉลี่ยของประชากรและ μ_0 คือค่าคงที่ที่ต้องการทดสอบ หรือเป็น ค่าเฉลี่ยที่คาดว่าจะเป็น สมมติฐานที่จะทดสอบอยู่ในลักษณะ

1. $H_0 : \mu = \mu_0$ แย้งกับ $H_1 : \mu > \mu_0$ หรือ
2. $H_0 : \mu = \mu_0$ แย้งกับ $H_0 : \mu < \mu_0$ หรือ
3. $H_0 : \mu = \mu_0$ แย้งกับ $H_1 : \mu \neq \mu_0$

ตัวสถิติที่ใช้ในการทดสอบขึ้นอยู่กับลักษณะของประชากรและขนาดตัวอย่างสุ่ม ซึ่ง แบ่งเป็น 3 กรณีคือ

ประชากรมีการแจกแจงแบบปกติและทราบค่าความแปรปรวน

ภายใต้ H_0 เป็นจริง

ค่าสถิติที่ใช้ในการทดสอบ คือ $$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

H_0	H_1	บริเวณวิกฤต
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$Z \geq Z_\alpha$
$H_0 : \mu = \mu_0$	$H_1 : \mu < \mu_0$	$Z \leq Z_\alpha$
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$Z \geq Z_{\frac{\alpha}{2}}$ หรือ $Z \leq Z_{\frac{\alpha}{2}}$

การทดสอบสมมติฐานค่าเฉลี่ยประชากร

<u>ประชากรมีการแจกแจงแบบใด ๆ และไม่ทราบค่าความแปรปรวน</u>

แต่ตัวอย่างมีขนาดใหญ่ ($n \geq 30$)

ภายใต้ H_0 เป็นจริง

ค่าสถิติที่ใช้ในการทดสอบ คือ $Z = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α เหมือนกับกรณีประชากรแจกแจงปกติทราบค่าความแปรปรวน

<u>ประชากรมีการแจกแจงแบบใด ๆ และไม่ทราบค่าความแปรปรวน</u>

แต่ตัวอย่างมีขนาดเล็ก ($n < 30$)

ภายใต้ H_0 เป็นจริง

ค่าสถิติที่ใช้ในการทดสอบ คือ $t = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$ โดยมีค่า $df = n - 1$

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

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

การทดสอบเกี่ยวกับผลต่างระหว่างค่าเฉลี่ยของสองประชากร

เมื่อ μ_1, μ_2 คือ ค่าเฉลี่ยของประชากรที่ 1 และ 2 ตามลำดับ

เมื่อ $\mu_1 - \mu_2$ คือ ค่าผลต่างของคาเฉลี่ยของสองประชากรดังกล่าว

สมมติฐานที่จะทดสอบอยู่ในลักษณะ

1. $H_0 : \mu_1 - \mu_2$ แย้งกับ $H_1 : \mu_1 > \mu_2$
2. $H_0 : \mu_1 - \mu_2$ แย้งกับ $H_1 : \mu_1 < \mu_2$
3. $H_0 : \mu_1 - \mu_2$ แย้งกับ $H_1 : \mu_1 \neq \mu_2$

ตัวสถิติที่ใช้ในการทดสอบขึ้นอยู่กับการแจกแจงของประชากร ความแปรปรวนของประชากรและขนาดของตัวอย่างที่สุ่มมา ซึ่งแบ่งได้ 3 กรณี คือ

<u>ประชากรทั้งสองมีการแจกแจงแบบปกติทราบค่าความแปรปรวน σ_1^2 และ σ_2^2</u>

เมื่อสุ่มตัวอย่างโดยอิสระจากประชากรที่ 1 และ 2 มาขนาด n_1 และ n_2 ตามลำดับ

สถิติที่ใช้ทดสอบคือ $Z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

เกณฑ์การตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

H_0	H_1	บริเวณวิกฤต
$H_0 : \mu_1 - \mu_2$	$H_1 : \mu_1 > \mu_2$	$Z \geq Z_\alpha$
$H_0 : \mu_1 - \mu_2$	$H_1 : \mu_1 < \mu_2$	$Z \leq Z_\alpha$
$H_0 : \mu_1 - \mu_2$	$H_1 : \mu_1 \neq \mu_2$	$Z \geq Z_{\frac{\alpha}{2}}$ หรือ $Z \leq Z_{\frac{\alpha}{2}}$

การทดสอบเกี่ยวกับผลต่างระหว่างค่าเฉลี่ยของสองประชากร

ประชากรทั้งสองมีการแจกแจงแบบใด ๆ ไม่ทราบค่าความแปรปรวน σ_1^2 และ σ_2^2 แต่สุ่มตัวอย่างขนาดใหญ่ (n_1 และ $n_2 \geq 30$)

สถิติที่ใช้ทดสอบคือ $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

เกณฑ์การตัดสินใจที่ระดับนัยสำคัญ α เช่นเดียวกับกรณีแรก ตารางข้างต้น

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

ประชากรทั้งสองมีการแจกแจงแบบใด ๆ ไม่ทราบค่าความแปรปรวน σ_1^2 และ σ_2^2 แต่สุ่มตัวอย่างขนาดเล็ก (n_1 และ $n_2 < 30$)

สถิติที่ใช้ทดสอบขึ้นอยู่กับความแปรปรวนของประชากรทั้ง 2 กลุ่ม ดังนี้

1. เมื่อ $\sigma_1 = \sigma_2$

สถิติที่ใช้ทดสอบ $T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}}$

เมื่อ $df = n_1 + n_1$

และ $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

เกณฑ์การตัดสินใจที่ระดับนัยสำคัญ α เป็นดังนี้

2. เมื่อ $\sigma_1 \neq \sigma_2$

สถิติที่ใช้ทดสอบ $T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

เมื่อ $df = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$

เกณฑ์ในการตัดสินใจที่ระดับนัยสำคัญ α ใช้เกณฑ์เดียวกับกรณีที่ 1 เมื่อ $\sigma_1 = \sigma_2$

อ้างอิง

- http://www.stvc.ac.th/elearning/stat/csu1.html
- https://www.coraline.co.th/single-post/basic-statistic-1
- https://www.coraline.co.th/single-post/basic-statistic-2