7.การสุ่มตัวอย่างและการประมาณค่า (Sampling and Estimation)

7.1 ประชากร (Population)

ประชากรในทางสถิติ หมายถึง ตัวเลขหรือหน่วยของข้อมูล (Data) ที่จะใช้ในการศึกษาอาจเป็นตัวบุคคล ครอบครัว หรือกลุ่มต่างๆของสังคม ปริมาณสิ่งของต่างๆ ปริมาณน้ำฝน ปริมาณข้าวที่ผลิตได้ในแต่ละไร่ น้ำหนัก ส่วนสูง เราต้องกำหนดขอบเขตของการศึกษาว่าจะครอบคลุมแค่ไหน เราเรียกทุกๆหน่วยที่อยู่ในขอบเขตนั้นว่า ประชากร (populations)

ประชากรจำแนกได้เป็น 2 ประเภท

1. ประชากรที่มีจำนวนจำกัด (Finite Population) มีขนาดพอที่จะนับจำนวนได้แน่นอน เช่น จำนวนพลเมืองในประเทศไทย จำนวนรถยนต์ที่เข้า-ออกใน มศว ประสานมิตรใน 1 วัน เป็นต้น

2. ประชากรที่มีจำนวนอนันต์ (Infinite Population) หมายถึง ประชากรที่มีขนาดใหญ่จนไม่สามารถนับได้แน่นอน เช่น จำนวนเมล็ดข้าวในกระสอบหนึ่งๆ จำนวนน้ำฝนในเดือนหนึ่งๆ

7.2 ตัวอย่าง (Sample) ในการดำเนินงานเก็บข้อมูลทางสถิติ มักจะพบบ่อยๆว่าการเก็บข้อมูลจากทุกๆหน่วยในประชากรนั้นทำได้ยาก เสียค่าใช้จ่ายมาก และไม่จำเป็นต้องทำถึงขนาดนั้น ในทางสถิติเรานิยมใช้ตัวอย่างเป็นตัวแทนของประชากร ซึ่งตัวอย่างที่ใช้จะต้องเป็นตัวแทนที่ดีของประชากรได้เพียงไรนั้น ขึ้นอยู่กับทฤษฎีการสุ่มตัวอย่างต่อไป

7.3 ค่าพารามิเตอร์และค่าสถิติ (Parameter and Statistics)

ค่าพารามิเตอร์ในทางสถิติ หมายถึง ค่าอันที่จริง (True Value) ที่หาได้โดยวิธีทางสถิติเพื่ออธิบายถึงลักษณะต่างๆของประชากร เป็นค่าคงที่ เช่น มัชฌิมเลขคณิต (Arithmetic mean) สัดส่วน (Proportion) หรือค่าความแปรปรวน (Variance) ของประชากร

โดยปกติค่าพารามิเตอร์ไม่สามารถจะคำนวณหาออกมาได้โดยตรง เพราะประชากรมักมีขนาดใหญ่มาก หรือจำนวนอนันต์ ดังนั้น จึงจำเป็นต้องอาศัยทฤษฎีการสุ่มตัวอย่าง เพื่อที่จะสุ่มตัวอย่างมาชุดหนึ่งให้มีจำนวนมากพอถ้าตัวอย่างชุดนี้ได้มาอย่างถูกต้องตามทฤษฎีการสุ่มตัวอย่าง ตัวอย่างชุดนี้จะมีลักษณะคล้ายคลึงกับประชากรมาก

ลักษณะของข้อมูลที่ดี ที่จะใช้เป็นตัวแทนของประชากรทั้งหมดต้องไม่มีความเอนเอียง (Unbiased) และให้ความคลาดเคลื่อน (error) น้อยที่สุด

สาเหตุที่ทำให้เกิดความเอนเอียง (Biased) มี 2 ประการใหญ่ๆ คือ

1. เกิดจากการใช้สูตรในการประมาณค่าไม่ถูกต้อง คือ หาค่า parameters เพื่อคำนวณค่าสถิติไม่ถูกต้อง

2. เกิดจากการเลือกตัวอย่าง ตัวอย่างที่เลือกมานั้นไม่ได้เป็นตัวแทนที่ดีของประชากร

ลักษณะตัวอย่างที่ดี ตัวอย่างที่ดีควรได้มาจากการสุ่ม ตามทฤษฎีความน่าจะเป็น กล่าวคือ ทุกๆหน่วยในประชากรมีโอกาสที่จะได้รับเลือกมาเป็นตัวอย่าง เรียกว่า probability sample กล่าวคือ ตัวอย่างที่เลือกไม่ได้เฉพาะเจาะจง คือ ไม่มีความเอนเอียงในการเลือก

ข้อสังเกต โดยทั่วไปการเก็บข้อมูลสถิติจากตัวอย่างจะมีความคลาดเคลื่อนในข้อมูลทุกครั้ง แต่หากเราใช้ probability sample เราสามารถวิเคราะห์ข้อมูลได้ว่าความคลาดเคลื่อน และความเอนเอียงจากข้อมูลเป็นเท่าไร

สำหรับความคลาดเคลื่อนจากข้อมูล (Sample error) นั้นเราวัดได้ด้วยค่าความแปรปรวน (Variance) ของตัวที่ใช้ประมาณค่า กรณีที่ความประมาณค่าไม่มีความเอนเอียง ในทางทฤษฎีเราสามารถควบคุม (control) ความคลาดเคลื่อนจากตัวอย่างได้ โดยการเพิ่มขนาดของตัวอย่าง (Sample size) ให้ใหญ่ขึ้น และเลือกใช้สูตรในการประมาณค่าให้สอดคล้องกับแบบแผนตัวอย่างที่กำหนดขึ้น ก็จะได้ค่าประมาณที่ไม่มีความเอนเอียง (Unbiased Estimator) ได้

ข้อดีของการสุ่มตัวอย่าง

1. ใช้ดีในกรณีที่ประชากรมีขนาดใหญ่เกินไป ไม่สามารถตรวจสอบได้ทั้งหมด
2. เป็นการสิ้นเปลืองค่าใช้จ่ายมากเกินไปหากต้องเก็บข้อมูลจากทุกหน่วยในประชากร
3. เป็นการประหยัดค่าใช้จ่ายโดยไม่ต้องเสียเวลาศึกษาประชากรโดยตรง
4. หากเลือกตัวอย่างได้ดีแล้ว การศึกษาตัวอย่างจะให้ผลที่เที่ยงตรงมากกว่า การศึกษาจากประชากรโดยตรง

7.4 วิธีการสุ่มตัวอย่าง (Sampling Methodology) ที่ใช้กันแพร่หลายมี 5 วิธี

1. การสุ่มตัวอย่างอย่างง่าย (Sample Random Sampling)
2. การสุ่มตัวอย่าง แบบขั้นภูมิ (Stratified Random Sampling)
3. การสุ่มตัวอย่าง แบบมีระบบ (Systematic Random Sampling)
4. การสุ่มตัวอย่าง แบบแบ่งเป็นกลุ่ม (Cluster Random Sampling)
5. การสุ่มตัวอย่างหลายขั้น (Multi-Stage Sampling)

7.5 การสุ่มตัวอย่างแบบแทนที่ และไม่แทนที่ (Sampling with and without replacement)

การสุ่มตัวอย่างแบบแทนที่ หมายความว่า เมื่อเลือกตัวอย่างขึ้นมาหนึ่งหน่วยจากประชากรแล้ว หลังจากนั้นจะเอาหน่วยตัวอย่างนี้ใส่คืนลงไปในประชากรอีกก่อนทำการเลือกตัวอย่างหน่วยต่อไป ซึ่งทำให้หน่วยตัวอย่างหน่วยนี้มีโอกาสได้รับเลือกขึ้นมาเป็นตัวอย่างอีก

การสุ่มตัวอย่างแบบไม่แทนที่ หมายความว่า หลังจากหน่วยตัวอย่างใดถูกเลือกขึ้นมาจากประชากรแล้ว จะไม่ใส่กลับคืนไปในประชากรอีกก่อน การสุ่มเลือกครั้งต่อไป หรือ แต่ละหน่วยตัวอย่างจะไม่มีโอกาสได้รับเลือกซ้ำอีก

ผลที่ได้จากแต่ละหน่วยตัวอย่างจะเป็นอิสระต่อกัน (Statistical independent) เมื่อเป็นการสุ่มตัวอย่างแบบแทนที่ แต่ผลที่ได้จากแต่ละหน่วย ตัวอย่างจะขึ้นต่อกัน (Statistical dependent) เมื่อเป็นการสุ่มตัวอย่างแบบไม่แทนที่ คือความน่าจะเป็นที่หน่วยตัวอย่างหน่วยใด จะถูกสุ่มในครั้งต่อไป จะไปอยู่กับผลของการสุ่มครั้งก่อนหน้า

ในกรณีที่ประชากรมีขนาดใหญ่มากเมื่อเทียบกับขนาดของตัวอย่าง ไม่ว่าจะเป็นสุ่มตัวอย่างแบบแทนที่หรือไม่แทนที่ ความแตกต่างของผลที่ได้จะมีไม่มากนัก เพราะว่าปัญหาที่เกี่ยวข้องในการสุ่มตัวอย่างแบบไม่แทนที่ คือ โอกาสหรือความน่าจะเป็นที่หน่วยตัวอย่างใดจะถูกเลือกในการสุ่มครั้งต่อไป จะเปลี่ยนแปลงไปเรื่อยๆ แต่การเปลี่ยนแปลงในการดำรงของโอกาสหรือความน่าจะเป็นเหล่านี้จะมีค่าน้อยมากจึงสามารถจะละเลยไปได้ ดังนั้นจึงอาจถือเหมือนว่าการสุ่มตัวอย่างทุกครั้งเป็นการสุ่มแบบแทนที่

โดยทั่วไป ความแตกต่างระหว่างการสุ่มตัวอย่างแบบแทนที่ กับการสุ่มตัวอย่างแบบไม่แทนที่จะน้อยมาก ถ้าขนาดของประชากรใหญ่กว่าของตัวอย่างไม่ต่ำกว่า 10 เท่า

7.6 การประมาณค่า (Estimation)

การอ้างอิงหรือการหาข้อสรุปในทางสถิติเกี่ยงกับเรื่องต่างๆ อาจแยกออกได้เป็น 2 แบบ คือ การประมาณค่าเกี่ยวกับลักษณะต่างๆของประชากร (parameters estimation) การทดสอบสมมติฐานเกี่ยวกับค่าต่างๆเหล่านั้น

การประมาณค่าต่างๆในประชากร (Points and interval estimators)

ความแตกต่างระหว่างค่าที่แท้จริงในประชากร (parameters) และค่าที่ประมาณได้จากตัวอย่าง (statistics) ยกตัวอย่างเช่น ค่าอาหารกลางวันของนิสิตในมหาวิทยาลัยแห่งหนึ่ง จำนวน 5,000 คน หากนำข้อมูลตัวเลขของนิสิตทั้งหมดรวมกันแล้วหารด้วย 5,000 จะได้ค่าเฉลี่ยประชากร (Population mean)

โดยทั่วไปในทางสถิติเมื่อหาค่าเฉลี่ยได้แล้ว เราก็ต้องการทราบว่า โดยเฉลี่ยแล้วค่าเหล่านั้นจะแตกต่างไปจากค่าเฉลี่ยที่หาได้มากน้อยเพียงใด ซึ่งตัวที่ใช้วัดการกระจายของค่าต่างๆ แต่ละค่าจากค่าเฉลี่ยของประชากร เรียกว่า ส่วนเบี่ยงเบนมาตรฐานของประชากร (Population Standard Deviation)

สมมติต่อว่า นิสิตจำนวน 100 คน ถูกเลือกมาเรียกว่า ขนาดของตัวอย่าง (Size of Sample) จากนั้นก็สามารถหาค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานได้เช่นเดียวกับประชากร ซึ่งค่าที่ได้มักเรียกว่า ค่าเฉลี่ยจากตัวอย่าง และส่วนเบี่ยงเบนมาตาฐานจากตัวอย่าง หรือที่เรียกว่า ค่าสถิติ โดยทั่วไปค่าสถิติเหล่านี้ใช้ในการประมาณค่าต่างๆ ที่ต้องการทราบในประชากร

ในการศึกษาเกี่ยวกับเรื่องต่างๆนั้น นอกจากจะศึกษาเกี่ยวกับค่าเฉลี่ยแล้ว อาจศึกษาเกี่ยวกับสัดส่วนต่างๆก็ได้ จะใช้สัญลักษณ์ต่างๆเหล่านี้แทนค่าต่างๆ คือ

สำหรับประชากร

μ แทน ค่าเฉลี่ย หรือมัชฌิมเลขคณิต

σ แทน ส่วนเบี่ยงเบนมาตรฐาน

N แทน ขนาดของประชากร

π แทน สัดส่วน

σ_p แทน ความคลาดเคลื่อนมาตรฐานของสัดส่วน

สำหรับตัวอย่าง

$\bar{x}$ แทน ค่าเฉลี่ย หรือมัชฌิมเลขคณิต

s แทน ส่วนเบี่ยงเบนมาตรฐาน

n แทน ขนาดของประชากร

p แทน สัดส่วน

s_p แทน ความคลาดเคลื่อนมาตรฐานของสัดส่วน

ถ้าสมมติว่า ตัวอย่างจำนวน n และ $x_1, x_2, ..., x_n$ คือค่าที่วัดได้จาก n หน่วย นั้นถูกเลือกมาโดยสุ่มจากประชากรซึ่งมีการแจกแจงแบบปกติ มีมัชฌิมเลขคณิตเป็น μ และส่วนเบี่ยงเบนมาตรฐาน σ ซึ่งยังไม่ทราบค่า โดยทั่วไปมักใช้ค่าเฉลี่ยจากตัวอย่าง $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ เป็นตัวประมาณค่าเฉลี่ยของงประชากร (μ)

และใช้ส่วนเบี่ยงเบนมาตรฐานจากตัวอย่าง $s = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$ เป็นตัวประมาณค่า ส่วนเบี่ยงเบนมาตรฐานของประชากร (σ)

ตัวประมาณค่านี้นิยมใช้กันอยู่ทั่วไป มี 2 แบบคือ การประมาณค่าแบบจุด (point estimate) เป็นการประมาณค่าลักษณะต่างๆของประชากรด้วยตัวเลขเพียงค่าเดียว ซึ่งเป็นการประมาณค่าพารามิเตอร์ของประชากรโดยใช้ค่าสถิติ

เช่น จากตัวอย่างจำนวน n สามารถคำนวณค่าเฉลี่ย $(\bar{x})$ และสัดส่วน (p) ได้ และถ้าเราใช้ค่านี้เป็นตัวประมาณค่า μ และ π การประมาณค่าในลักษณะนี้เรียกว่า ตัวประมาณค่าแบบจุด แต่การประมาณค่าแบบนี้เราไม่สามารถทราบได้เลยว่าค่าประมาณที่ได้มีค่าใกล้เคียงกับค่าจริงมากน้อยเพียงใด

แม้การประมาณค่าโดยการกำหนดออกมาในเทอมของค่าสองค่า ซึ่งเป็นช่วงที่คาดว่าค่าที่เราต้องการทราบในประชากรจะอยู่ในช่วงนั้นเรียกว่า interval estimate จะเป็นที่นิยมมากกว่า การประมาณค่าแบบช่วงมักอยู่ในรูปของ $a < \theta < b$ ซึ่งค่า a และ b จะขึ้นอยู่กับตัวประมาณแบบจุด $\hat{\theta}$ ของพารามิเตอร์ θ และการแจกแจงความน่าจะเป็นของ $\hat{\theta}$

โดยทั่วไปหากสมมติให้ข้อมูลชุดหนึ่งมีค่าเป็น $x_1, x_2, ..., x_n$ และสามารถคำนวณค่าสถิติต่างๆ เช่น $\bar{x}$ และ s^2 ได้ ค่า $\bar{x}, s^2$ จากตัวอย่างชุดนี้ อาจแตกต่างไปจากตัวอย่างชุดแรก ตามความจริงแล้วถ้าตัวอย่างหลายๆชุดถูกเลือกออกมาจากประชากรเดียวกันแล้ว อาจถือว่าตัวสถิติที่เราสนใจก็เป็นตัวแปรอีกชุดหนึ่ง ซึ่งการแจกแจงเหล่านั้นเรียกว่า การแจกแจงของตัวอย่าง (Sampling distribution)

7.7 การแจกแจงของค่าเฉลี่ยที่ได้จากตัวอย่าง (The Sampling Distribution of $\bar{x}$)

การแจกแจงของค่าเฉลี่ยจากตัวอย่างที่สำคัญที่สุดคือ mean $(\bar{x})$ ถ้าเราทำการสุ่มตัวอย่างขนาดเท่าๆกันมาหลายๆครั้งจากประชากรเดียวกัน จะเห็นว่าค่าเฉลี่ยจากตัวอย่างก็จะเปลี่ยนไปจากตัวอย่างชุดหนึ่งไปยังตัวอย่างอีกชุดหนึ่ง ซึ่งค่าเหล่านี้จะเป็นตัวกำหนดการแจกแจง $\bar{x}$

ตัวอย่าง สมมติว่ามีตัวเลขอยู่ 18 ตัว คือ

3	6	9	6	4	7
4	2	8	1	1	4
5	6	5	0	0	6

จากค่าเหล่านี้จะคำนวณค่าเฉลี่ย (μ) ได้ $= 4.39$, ค่าความแปรปรวน $(\sigma^2) = 6.01$ ถ้าแบ่งข้อมูลทั้งหมดออกเป็น 6 กลุ่ม กลุ่มละ 3 จำนวน จากนั้นก็เอาจำนวนค่าเฉลี่ยของแต่ละกลุ่มซึ่งจะได้ค่าเฉลี่ยดังนี้ คือ 4 , 4.67 , 7.33 , 2.33 และ 5.67 ซึ่งค่าทั้ง 6 ค่านี้ เป็นค่าเฉลี่ยของตัวอย่างแต่ละชุด ซึ่งมีขนาดเท่ากับ 3 และค่าเฉลี่ยของค่าเฉลี่ยที่ได้จากตัวอย่างก็ยังคงมีค่าเป็น 4.39 แต่ค่าความแปรปรวนของค่าเฉลี่ยที่ได้จากตัวอย่างเพียง 3.21 ซึ่งมีค่าน้อยกว่าความแปรปรวนของข้อมูลเดิมมาก

อันนี้แสดงให้เห็นว่า ค่าเฉลี่ยที่มาจากแต่ละกลุ่มจะใกล้เคียงกันมากกว่าข้อมูลเดิม และถ้าขนาดของตัวอย่างยิ่งใหญ่มากเท่าไร ค่าเฉลี่ยที่ได้จากตัวอย่างก็จะมีค่าใกล้เคียงกับค่าเฉลี่ยจริงๆมากขึ้นเท่านั้น ในกรณีนี้เนื่องจาก $\bar{x}$ เป็นตัวแปรสุ่ม จึงจำเป็นต้องทราบลักษณะการกระจายของค่าเฉลี่ยจากตัวอย่าง การกระจายของตัวแปรสุ่มจัดด้วยค่าของความแปรปรวน หรือส่วนเบี่ยงเบนมาตรฐาน ความแปรปรวนของค่าเฉลี่ย จากตัวอย่างขึ้นอยู่กับความแปรปรวนในประชากรของตัวแปร (x) เดิม

โดยทั่วไปใช้ σ_x^2 หรือใช้ σ^2 แทนความแปรปรวนในประชากร ดังนั้นความแปรปรวนของ $\bar{x}$ จึงเขียนแทนด้วย $\sigma_{\bar{x}}^2$

ความแปรปรวนของ $\bar{x}$ ส่วนหนึ่งขึ้นอยู่กับความแปรปรวนของลักษณะที่ต้องการจาก แต่ละหน่วยตัวอย่างในปรระชากร (σ^2) และยังขึ้นกับขนาดของตัวอย่าง (n) ด้วย

7.8 ทฤษฎีที่สำคัญเกี่ยวกับการสุ่มตัวอย่าง

ทฤษฎีที่ 1 : ถ้าตัวอย่างขนาด n ถูกสุ่มมาจากประชากรที่มีการแจกแจงซึ่งมีค่าเฉลี่ย μ และส่วนเบี่ยงเบนมาตรฐาน σ แล้ว ค่าเฉลี่ยของตัวอย่างก็จะมีการแจกแจงซึ่งมีค่าเฉลี่ยเท่ากับ μ เหมือนกัน แต่ส่วนเบี่ยงเบนมาตรฐานมีค่าเพียง $\frac{\sigma}{\sqrt{n}}$ ถ้าเป็นการสุ่มแบบแทนที่ และจะมีค่าเป็น $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ เมื่อเป็นการสุ่มตัวอย่างแบบไม่แทนที่

ทฤษฎีที่2: ตัวอย่างขนาด n ถูกสุ่มมาจากประชากรที่มีการแจกแจงแบบปกติ ซึ่งมีค่าเฉลี่ย μ และความแปรปรวน σ^2 การแจกแจงของ $\bar{x}$ ก็จะเป็นแบบปกติเหมือนกัน ซึ่งมีค่าเฉลี่ยเท่าเดิม และค่าความแปรปรวนคลาดเคลื่อนมาตรฐานเป็น $\frac{\sigma}{\sqrt{n}}$