บทที่ 1

บทนำ

1.1 ความสำคัญและความเป็นมาของปัญหา

การเก็บรวบรวมข้อมูลปฐมภูมิเกี่ยวกับเรื่องที่เราสนใจ หรือเพื่อการวิจัยอาจกระทำโดยการสำรวจ ซึ่งถ้าเก็บจากทุกหน่วยในประชากรเรียกว่า การสำมะโน แต่ถ้าเก็บมาเพียงบางหน่วยก็เรียกว่าการสำรวจตัวอย่าง โดยทั่วไปแล้วประชากรที่สนใจศึกษามักจะมีขนาดใหญ่มาก ทำให้ไม่สามารถเก็บรวบรวมข้อมูลจากทุกหน่วยในประชากรได้ หรือหากจะทำก็ต้องใช้เวลาและทรัพยากรจำนวนมากซึ่งบ่อยครั้งเป็นสิ่งที่เป็นไปไม่ได้ที่จะทำภายใต้ข้อจำกัดทางทรัพยากรที่มีอยู่ ในทางปฏิบัติจึงมักจะใช้วิธีการสำรวจตัวอย่าง เพราะนอกจากจะเป็นการประหยัดค่าใช้จ่าย ประหยัดเวลาแล้วยังทำให้สามารถเก็บข้อมูลได้มากชนิดขึ้นด้วยทรัพยากรจำนวนเท่าๆ กัน และยังอาจใช้ทรัพยากรส่วนหนึ่งเพื่อเพิ่มคุณภาพของการประมาณค่าจากตัวอย่างได้ ในกรณีที่ประชากรมีค่าสังเกตบางค่าสูงมาก ตัวอย่างที่เลือกมาได้อาจมีหน่วยที่มีค่าสูงมากรวมอยู่ด้วย ซึ่งอาจจะมีผลกระทบต่อการประมาณค่าพารามิเตอร์ที่เราสนใจได้ ดังจะเห็นได้จากการประมาณค่ารวมประชากรจากการเลือกตัวอย่างแบบสุ่มอย่างง่ายชนิดไม่ใส่คืน (simple random sampling without replacement) ด้วยตัวประมาณ $\hat{Y}_o = N\bar{y}$ โดยที่ N แทนจำนวนหน่วยในประชากร และ $\bar{y}$ แทนค่าเฉลี่ยตัวอย่าง (sample mean) ถ้าเรานำค่าสังเกตที่เป็นค่าสูงมากมาใช้จะได้ว่า ค่าประมาณ $\hat{Y}_o$ จะมีค่ามากกว่าค่ารวมประชากร แต่ถ้าเราตัดค่าสังเกตเหล่านี้ทิ้งไป ค่าประมาณ $\hat{Y}_o$ ที่คำนวณได้ก็อาจจะต่ำกว่าค่ารวมประชากร เป็นต้น ทางแก้สำหรับปัญหาดังกล่าวอาจทำได้โดยการตัดค่าสังเกตที่เป็นค่าสูงมากทิ้งไป ถ้าการกระทำดังกล่าวไม่กระทบกระเทือนต่อผลการวิจัยมากนัก แต่ในบางกรณีจะเห็นได้ว่า เราไม่สามารถตัดค่าสังเกตที่เป็นค่าสูงมากเหล่านี้ทิ้งไปได้ เพราะบางครั้งจะทำให้การประมาณค่าคลาดเคลื่อนไปจากความเป็นจริงมาก และในทางตรงกันข้ามการนำค่าดังกล่าวมาใช้ก็มีผลทำให้ค่าประมาณที่ได้มากกว่าค่าที่แท้จริงด้วย

จากปัญหาดังกล่าว ได้มีนักสถิติหลายท่านได้เล็งเห็นความสำคัญและพยายามหาเทคนิคต่าง ๆ มาใช้เพื่อที่จะทำให้การประมาณค่ารวมประชากรจากการเลือกตัวอย่างแบบสุ่มอย่างง่ายชนิดไม่ใส่คืน มีความแม่นยำ หรือใกล้เคียงกับค่าจริงยิ่งขึ้น จึงได้มีการสร้างตัวประมาณขึ้นมาหลายรูปแบบ เช่น ตัวประมาณที่อยู่ในรูปกรณฑ์ (root estimator) ตัวประมาณแบบวิธี

วินซอไรเซชั่น (Winsorization) เป็นต้น ทั้งนี้เพื่อวัตถุประสงค์เดียวกันก็คือ เพื่อลดอิทธิพลในกรณีที่มีค่าสังเกตของหน่วยตัวอย่างบางหน่วยเป็นค่าสูงมาก

โดยทั่วไปแล้วในการทำวิจัยเรื่องหนึ่ง ๆ มักจะมีการเก็บรวบรวมข้อมูลของตัวแปรต่าง ๆ หลายตัวแปรไปพร้อม ๆ กัน ทั้งนี้เพื่อตอบสนองต่อวัตถุประสงค์ของการวิจัยที่กำหนดไว้ อย่างไรก็ตามผู้วิจัยจะพบอยู่เสมอว่า ในบรรดาตัวแปรต่าง ๆที่เก็บรวบรวมมานั้น มีตัวแปรบางตัวที่มีความสัมพันธ์ต่อกัน ดังนั้นการเพิ่มประสิทธิภาพของตัวประมาณค่ารวมประชากรน่าจะกระทำได้โดยอาศัยข้อมูลจากตัวแปรอื่นที่มีความสัมพันธ์กับตัวแปรที่เราสนใจศึกษามาช่วยในการเพิ่มประสิทธิภาพของการประมาณค่าดังจะเห็นได้ว่า ตัวประมาณอัตราส่วน (ratio estimator) ตัวประมาณความถดถอย (regression estimator) และตัวประมาณความแตกต่าง (Difference estimator) ต่างก็ใช้ประโยชน์จากตัวแปรอื่นมาช่วยในการเพิ่มประสิทธิภาพของการประมาณค่าทั้งสิ้น จากหลักการดังกล่าวในทำนองเดียวกันการประมาณค่ารวมประชากรจากตัวอย่างที่มีค่าสังเกตบางค่าเป็นค่าสูงมาก ย่อมสามารถกระทำได้โดยอาศัยหลักการหรือเทคนิคอย่างเดียวกัน

วิทยานิพนธ์นี้ได้เสนอตัวประมาณ 3 รูปแบบ โดยอาศัยคุณสมบัติเกี่ยวกับตัวประมาณความถดถอย[1] และการแบ่งชั้นภูมิเมื่อเลือกตัวอย่างแล้ว[2] (poststratification) มาใช้เพื่อกำหนดรูปแบบและวิธีการประมาณค่ารวมประชากรเฉพาะกรณีเมื่อเลือกตัวอย่างแบบสุ่มอย่างง่ายชนิดไม่ใส่คืนเท่านั้น นอกจากนี้ยังได้ทำการศึกษาคุณสมบัติของตัวประมาณที่เสนอขึ้นมาทั้ง 3 รูปแบบพร้อมทั้งทำการเปรียบเทียบประสิทธิภาพของตัวประมาณที่เสนอแนะ กับตัวประมาณ $\hat{Y}_o$ และตัวประมาณที่เสนอโดยไมเคิลและคาดาบา (Michael and Kadaba) ในปี ค.ศ.1981 โดยพิจารณาจากค่าประสิทธิภาพสัมพัทธ์ (relative efficiency) ที่คำนวณได้จากการจำลองข้อมูลขึ้นในเครื่องคอมพิวเตอร์ IBM 370/3031 ด้วยวิธีมอนติคาร์โล (Monte Carlo method) เป็นหลัก

---

[1] อ่านรายละเอียดเพิ่มเติมได้ในภาคผนวก ก

[2] อ่านรายละเอียดเพิ่มเติมได้ในภาคผนวก ก

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาวิธีการประมาณค่ารวมประชากร ในกรณีที่ประชากรนั้นมีค่าลักเกตบางหน่วยที่เป็นค่าสูงมาก และทำการเลือกตัวอย่างแบบสุ่มอย่างง่าย ชนิดไม่ใส่คืนแล้วพบว่าตัวอย่างบางหน่วยมีค่าสูงมาก โดยใช้ตัวประมาณค่ารวมประชากรที่วิทยานิพนธ์เสนอ

2. เพื่อศึกษาดูว่าประสิทธิภาพของตัวประมาณในข้อหนึ่งเปลี่ยนแปลงอย่างไร เมื่อพบว่า ตัวแปรตาม (Y) ที่สนใจศึกษาและตัวแปรอิสระ (X) มีความสัมพันธ์เชิงเส้นต่อกันมากขึ้น

3. เพื่อศึกษาดูว่าประสิทธิภาพของตัวประมาณในข้อหนึ่งเปลี่ยนแปลงอย่างไร เมื่อพบว่าในประชากร หรือในตัวอย่างมีค่าร้อยละของจำนวนค่าลักเกตที่เป็นค่าสูงมากเพิ่มขึ้น

4. เพื่อเปรียบเทียบวิธีการประมาณค่ารวมประชากรระหว่างตัวประมาณที่กล่าวในข้อหนึ่งกับตัวแประมาณที่เสนอโดยไมเคิล และคาดาบา (1981) และตัวประมาณ $\hat{Y}_o$

1.3 สมมติฐานของการวิจัย

1. ในกรณีที่ตัวแปร Y และตัวแปร X มีความสัมพันธ์เชิงเส้นต่อกันมากขึ้นภายใต้การเลือกตัวอย่างแบบสุ่มอย่างง่าย ชนิดไม่ใส่คืน และได้ตัวอย่างบางหน่วยมีค่าสูงมากและเป็นค่ามีอยู่จริงในประชากรนั้น ประสิทธิภาพของตัวประมาณที่เสนอแนะทั้ง 3 รูปแบบจะสูงขึ้น

2. ในกรณีที่ประชากรหรือตัวอย่างมีร้อยละของจำนวนค่าลักเกตที่เป็นค่าสูงมากเพิ่มขึ้น ประสิทธิภาพของตัวประมาณค่ารวมประชากรที่เสนอ จะสูงขึ้น

3. ตัวประมาณค่ารวมประชากรที่เสนอ จะมีประสิทธิภาพสูงกว่าตัวประมาณ $\hat{Y}_o$ และตัวประมาณซึ่งเสนอโดยไมเคิลและคาดาบาในปี ค.ศ. 1981

1.4 ข้อตกลงเบื้องต้น

1. ตัวแปร Y มีค่าลักเกตที่เป็นค่าสูงมากอยู่ในประชากรจริง ๆ และมีโอกาสถูกเลือกมาเป็นตัวอย่างได้ ส่วนค่าลักเกตของตัวแปร Y ที่เป็นค่าที่ต่ำมากจะไม่มีอยู่ในประชากร

2. ตัวแปร Y และตัวแปร X มีความสัมพันธ์เชิงเส้นต่อกัน และทราบค่าเฉลี่ยประชากรของตัวแปร X ในแต่ละกลุ่ม กล่าวคือ ทั้งในกลุ่มที่ตัวแปร Y มีค่าลักเกตสูงมาก และกลุ่มที่ตัวแปร Y มีค่าลักเกตที่เป็นค่าไม่สูงมาก หรือเป็นค่าปกติ ทั้งนี้ในตัวแปร X ไม่มีค่าลักเกตที่เป็นค่าสูงมากหรือต่ำมากอยู่ด้วยในประชากร

3. วิธีการเลือกตัวอย่างเป็นแบบสุ่มตัวอย่างง่าย ชนิดไม่ใส่คืน

4. ในวิทยานิพนธ์นี้จะถือว่าค่าสังเกต Y_i ; i = 1,2..., N เป็นค่าสูงมากก็ต่อเมื่อค่าสังเกต Y_i มีค่ามากกว่าค่าขอบเขตบน (upper confidence interval) ของช่วงความเชื่อมั่น 99% ของ Y_i[1] กล่าวคือ มีค่ามากกว่า ค่าเฉลี่ยของ Y_i + 2.576 × $\sqrt{\text{ความความแปรปรวนของ } Y_i}$ ซึ่งจะเห็นได้ว่า จำนวนค่าสังเกต Y_i ที่มีค่ามากกว่าค่าขอบเขตบนของช่วงความเชื่อมั่น 99% ของ Y_i นั้นมีจำนวนน้อย เพราะฉะนั้นหลักเกณฑ์นี้จึงพออนุโลมได้ว่าค่าสังเกต Y_i นี้เป็นค่าสูงมากได้

1.5 ขอบเขตของการวิจัย

1. ในการศึกษาการเปรียบเทียบครั้งนี้ จะเป็นการเปรียบเทียบตัวประมาณที่เสนอแนะกับตัวประมาณ $\hat{Y}_o$ และตัวประมาณที่เสนอโดยไมเคิลและคาดาบาในปี ค.ศ. 1981 เท่านั้น

2. ข้อมูลที่ใช้ในการวิจัยครั้งนี้ได้จากการจำลองขึ้นในเครื่องคอมพิวเตอร์ IBM 370/3031 เพื่อสร้างประชากรขนาดเท่ากับ 500 และ 1000 ซึ่งในแต่ละขนาดของประชากรมีร้อยละของจำนวนค่าสังเกตที่เป็นสูงมากคิดเป็น 1.8% 2.8% และ 3.2% ในขนาดประชากรเท่ากับ 500 และในกรณีขนาดประชากรเท่ากับ 1000 จะมีร้อยละของจำนวนค่าสังเกตที่เป็นค่าสูงมากคิดเป็น 1.8% 2.8% และ 3.3%[2] สำหรับขนาดตัวอย่างที่ใช้ศึกษามีขนาดเท่ากับ 50 100 และ 200 ในแต่ละสถานการณ์ที่ศึกษาจะกระทำซ้ำ ๆ กัน 100 ครั้งในขนาดประชากรเท่ากับ 500 และกระทำซ้ำ ๆ กัน 50 ครั้ง ในขนาดประชากรเท่ากับ 1000[3] และกำหนดให้ตัวแปร Y

---

[1] มนตรี พิริยะกุล, เทคนิคการวิเคราะห์ความถดถอย เล่ม 2 (กรุงเทพมหานคร : มหาวิทยาลัยรามคำแหง) หน้า 15

[2] ร้อยละของค่าสังเกตที่เป็นค่าสูงมากที่พบในประชากรแต่ละขนาดดังกล่าวเป็นร้อยละของค่าสังเกตที่เป็นค่าสูงมากที่มีโอกาสพบมากในแต่ละการแจกแจงของตัวแปร ที่มีค่าพารามิเตอร์ตามที่ศึกษา หลังจากที่ทดลองหาค่าเริ่มต้นที่เหมาะสมแล้ว

[3] เหตุที่ใช้จำนวนซ้ำ ๆ ในขนาดประชากร 1000 ไม่เท่ากับจำนวนซ้ำในขนาดประชากร 500 เพราะเวลาที่ใช้หาค่าเริ่มต้นในการทำ simulation เพื่อจำลองประชากรขนาด 1000 ใช้เวลามากและได้ลองศึกษาค่าเริ่มต้นอย่างสุ่ม 10 ค่า ดูผลลัพธ์ที่ได้ ปรากฏว่าให้ผลลัพธ์คล้องกันกับค่าเริ่มต้น 50 ค่า

และตัวแปร X มีความสัมพันธ์เชิงเส้นต่อกัน ณ ระดับค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient = ρ_{XY}) เท่ากับ 0.1 0.3 0.5 0.7 และ -0.1 -0.3 -0.5 -0.7 โดยให้

ก. ตัวแปร Y มีการแจกแจงแบบล็อกนอร์มอล (lognormal)[1] ที่มีพารามิเตอร์ $\mu = 2$ $\sigma^2 = 1$ และความคลาดเคลื่อน (ε) มีการแจกแจงแบบล็อกนอร์มอลที่มีพารามิเตอร์ $\mu = 0$ $\sigma^2 = 1$

ข. ตัวแปร Y มีการแจกแจงแบบแกมม่า (gamma)[2] ที่มีพารามิเตอร์ $\alpha = 2$ $\beta = 1$ และความคลาดเคลื่อนมีการแจกแจงแบบแกมม่าที่มีพารามิเตอร์ $\alpha = 0.02$, $\beta = 10$

และจำลองตัวแปร X จากสมการเชิงเส้น $Y_i = \beta_o + \beta_1 X_i + \varepsilon_i$; i = 1,2, ...,N เมื่อ $\beta_o = \mu_Y - \beta_1\mu_X$ และ β_1 เป็นจำนวนจริงใด ๆ ที่ทำให้ค่า ρ_{XY} เป็นไปตามที่ศึกษาจะโดยกำหนดให้ μ_X (ค่าเฉลี่ยของประชากร X) = 2

---

[1]lognormal distribution เป็น positively skewed distribution ที่มีฟังก์ชั่นความหนาแน่นคือ

$$f(y) = \begin{cases} \frac{1}{y\sqrt{2\pi}} \exp(-\frac{1}{2}(\log y - \mu))/\frac{2}{\sigma} & ; y > 0 \\ 0 & ; \text{อื่น ๆ} \end{cases}$$

เมื่อค่า μ และ σ^2 เป็นค่าเฉลี่ย และความแปรปรวนของ N = log Yซึ่ง N มีการแจกแจงแบบปกติ สำหรับค่าเฉลี่ยของ y เท่ากับ $E(y) = \exp(\mu + \frac{1}{2}\sigma^2)$ ความแปรปรวนของ y คือ $V(y) = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2)$

[2]gamma distribution เป็นอีก positively skewed distribution หนึ่งที่มีฟังก์ชั่นความหนาแน่นคือ

$$f(y) = \begin{cases} \frac{y^{\alpha-1}\exp^{(-y/\beta)}}{\Gamma(\alpha)\beta^2} & ; y > 0 , \alpha > 0, \beta > 0 \\ 0 & ; \text{อื่น ๆ} \end{cases}$$

มี $E(y) = \alpha\beta$ $V(y) = \alpha\beta^2$

3. กำหนดให้จำนวนหน่วยตัวอย่างที่มีค่าสูงมาก มีค่าเป็นไปได้ตั้งแต่ 2 จนถึงจำนวนค่าสังเกตที่เป็นค่าสูงมากที่พบในประชากร ทั้งนี้เพื่อให้สามารถคำนวณค่าความแปรปรวนของตัวประมาณค่ารวมประชากรที่เสนอขึ้นมาได้

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. ทำให้ทราบถึงวิธีการที่เหมาะสม และเป็นการเพิ่มประสิทธิภาพของการประมาณค่ารวมประชากรในกรณีที่ตัวอย่างสุ่มอย่างง่าย มีค่าของตัวอย่างบางหน่วยเป็นค่าสูงมาก

2. สามารถนำวิธีการประมาณค่าที่ได้จากข้อ 1 ไปประยุกต์ใช้กับข้อมูลจริงได้

1.7 ความหมายของคำต่าง ๆ ที่ใช้ในการวิจัย

ร้อยละของจำนวนค่าสังเกตที่เป็นค่าสูงมากที่พบในตัวอย่าง (หรือในประชากร) มีค่าเท่ากับสัดส่วนระหว่างจำนวนค่าสังเกตที่เป็นค่าสูงมากที่พบในตัวอย่าง (หรือในประชากร) กับขนาดตัวอย่าง (หรือขนาดประชากร) ที่ใช้ศึกษา คูณด้วย 100

N = จำนวนหน่วยตัวอย่างทั้งหมดในประชากรหรือขนาดประชากร (population size)

$Y_1, Y_2, \ldots, Y_N$ เป็นค่าสังเกตจากหน่วยที่ $1, 2, \ldots, N$ ตามลำดับ

Y = ค่ารวมประชากร (population total) = $\sum_{i=1}^{N} Y_i$

$\bar{Y}$ = ค่าเฉลี่ยประชากร (population mean) = $\sum_{i=1}^{N} Y_i/N = Y/N$

S_Y^2 = ค่าความแปรปรวนของประชากร (population variance) = $\sum_{i=1}^{N} \frac{(Y_i-\bar{Y})^2}{N-1}$

n = จำนวนหน่วยตัวอย่างทั้งหมดในตัวอย่างหรือขนาดตัวอย่าง (sample size)

$y_1, y_2, \ldots, y_n$ เป็นค่าสังเกตจากหน่วยตัวอย่างที่ $1, 2, \ldots, n$ ในตัวอย่างตามลำดับ

y = ค่ารวมตัวอย่าง (sample total) = $\sum_{i=1}^{n} y_i$

$\bar{y}$ = ค่าเฉลี่ยตัวอย่าง (sample mean) = $\sum_{i=1}^{n} y_i/n = y/n$

s_y^2 = ค่าความแปรปรวนของตัวอย่าง (sample variance) = $\sum_{i=1}^{n} (y_i-\bar{y})^2/n-1$

ρ_{XY} (correlation coefficient) หรือสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร Y และตัวแปร X เป็นค่าที่บ่งบอกถึงระดับความสัมพันธ์เชิงเส้นระหว่างตัวแปร Y และ X โดย

มีค่าเท่ากับ $\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y}) / \sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N} (Y_i - \bar{Y})^2}$ และอยู่ในช่วง

$-1 \leq \rho_{XY} \leq 1$

MSE = ความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean square error) มีค่าเท่ากับ $E(\hat{\theta}-\theta)^2$ หรือเท่ากับความแปรปรวนของ $\hat{\theta}$ + ความเอนเอียง (bias)ของ $\hat{\theta}$ ยกกำลังสองเมื่อความเอนเอียงของ $(B(\hat{\theta}))$ มีค่าเท่ากับ $E(\hat{\theta}) - \theta$ และถ้า $B(\hat{\theta}) = 0$ หรือ $E(\hat{\theta}) = \theta$ แล้ว $\hat{\theta}$ จะมีคุณสมบัติเป็นตัวประมาณที่ไม่เอนเอียงของ θ