บทที่ ๒

ทฤษฎีพื้นฐานเกี่ยวกับการประมาณผลในการสำรวจด้วยตัวอย่าง

วิธีการสำรวจด้วยตัวอย่างซึ่งเป็นการสำรวจเพียงส่วนหนึ่งของประชากร และถือว่าตัวอย่างที่เลือกขึ้นมานั้นเป็นตัวแทนของประชากรทั้งหมด

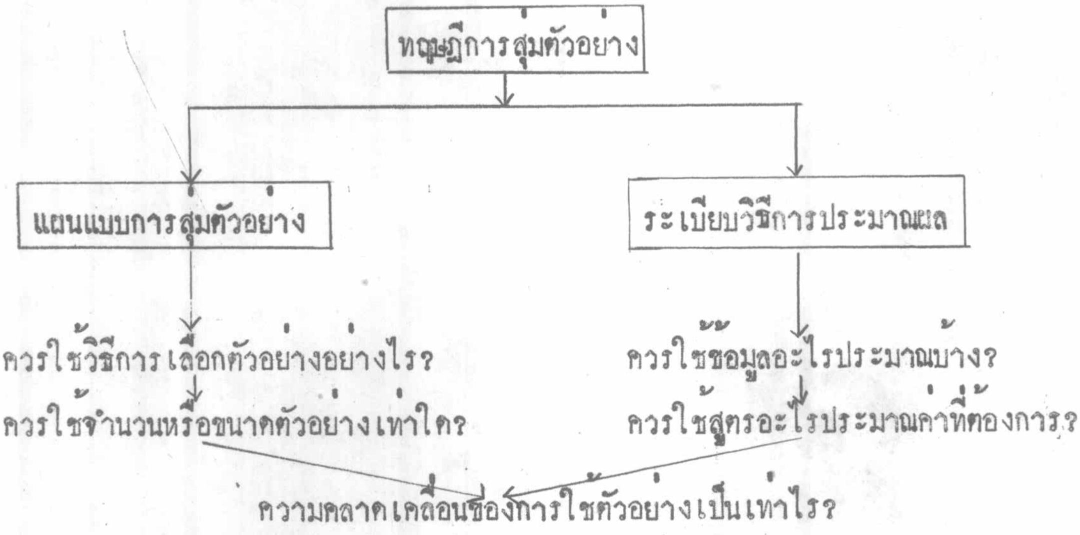
๒.๑ องค์ประกอบที่สำคัญในทฤษฎีการสุ่มตัวอย่าง

ตามทฤษฎีการสุ่มตัวอย่างมีองค์ประกอบที่สำคัญ ๆ ๒ ส่วน คือ

๑. แผนแบบการสุ่มตัวอย่าง (Sample Design)

๒. ระเบียบวิธีการประมาณผล (Estimation Procedure)

และอาจเขียนแผนผังแสดงความสัมพันธ์ได้ดังข้างล่างนี้



องค์ประกอบทั้งสองส่วนย่อมมีความสัมพันธ์กันมาก หากขาดสิ่งหนึ่งสิ่งใดแล้วจะไม่สามารถทำให้การนำเอาวิธีการสำรวจด้วยตัวอย่างมาใช้ได้ตามจุดประสงค์ที่กำหนดไว้ ในการวางแผนเลือกตัวอย่างโดยมากต้องอาศัยหลักของ Probability Sampling[๑/] เพื่อช่วยในการประมาณค่าต่างๆ ที่ต้องการ ในการวางแผนเลือกตัวอย่าง จะต้องทำตามกฎเกณฑ์ต่าง ๆ

๑/ Probability Sampling คือแผนแบบตัวอย่างที่สามารถคำนวณความน่าจะเป็นที่หน่วยหนึ่งหน่วยใดจะถูกเลือก

ของการเลือกตัวอย่าง เพื่อให้ได้แผนแบบในการเลือกตัวอย่างที่ดีทำให้การประมาณผลมีประสิทธิภาพ แผนแบบที่ดีนั้น ควรจะมีการประมาณค่าความคลาดเคลื่อนทางสถิติได้และสามารถทำให้ควบคุมคุณภาพของงานแต่ละขั้นได้

การประมาณค่าของสิ่งที่สนใจโดยอาศัยข้อมูลที่รวบรวมได้จากตัวอย่าง ตามแผนแบบการเลือกตัวอย่างนั้นมิใช่เป็นการคาดคะเนจากประสพการณ์หรือเดาเอง ทั้งนี้เพราะถือว่าตัวอย่างที่ได้มานั้นเป็นตัวแทนของประชากรทั้งหมด การประมาณค่าของสิ่งที่สนใจ เช่น ค่าเฉลี่ยจากตัวอย่าง ($\bar{x}$) จะมีประสิทธิภาพเพียงไรนั้นขึ้นอยู่กับสูตรที่ใช้ประมาณและแผนแบบในการเลือกตัวอย่าง ประสิทธิภาพของค่าประมาณนี้วัดได้ด้วยความคลาดเคลื่อนของตัวประมาณค่า [illegible] เช่น การประมาณค่าความแปรปรวนจากประชากร (population variance $= \sigma^2$) ด้วยค่าความแปรปรวนจากตัวอย่าง (sample variance $= s^2$) ซึ่ง

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \bar{\bar{x}})^2}{N} \quad \text{และ} \quad \bar{\bar{x}} = \text{ค่าเฉลี่ยของประชากร}$$

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \quad \text{และ} \quad \bar{x} = \text{ค่าเฉลี่ยของตัวอย่าง}$$

<u>ตัวประมาณค่าที่ดีควรมีคุณสมบัติต่อไปนี้คือ</u>

๑. <u>Unbiasedness</u> ถ้า $\hat{\theta}$ เป็น unbiased estimator ของ θ นั่นก็คือ $E(\hat{\theta}) = \theta$

๒. <u>Minimum Variance หรือ Least Variance</u> ในกรณีที่ตัวประมาณค่าหลายตัวต่างก็เป็น unbiased estimator จะเลือกตัวประมาณค่าที่มีประสิทธิภาพสูงสุด หรือมีค่าความแปรปรวนน้อยสุด ประสิทธิภาพของตัวประมาณค่าจึงมีความสัมพันธ์กับความแปรปรวน

คือ ประสิทธิภาพจะเป็นส่วนกลับกับความแปรปรวน เช่น ในการเปรียบเทียบประสิทธิภาพของตัวประมาณค่า ๒ ตัว ก็อาศัยพิจารณาจากค่าประสิทธิภาพสัมพัทธ์โดยใช้สูตร

$$\text{Efficiency of A relative to B} = \frac{V(B)}{V(A)} \times 100 \ \%$$

ถ้าค่าที่ได้ออกมาเกิน ๑๐๐ % ก็หมายความว่าตัวประมาณค่า A มีประสิทธิภาพมากกว่าของ B ในทางตรงกันข้ามถ้าผลออกมาน้อยกว่า ๑๐๐ % หมายความว่าตัวประมาณค่า B มีประสิทธิภาพมากกว่า A นั่นคือเราจะเลือกตัวประมาณค่าที่ให้ค่าความแปรปรวนน้อยสุด

(๓) Ease of Computation ในกรณีที่ตัวประมาณค่า ๒ ตัวมีลักษณะเหมือนกันทุกประการ เราจะเลือกตัวที่ง่ายไม่สลับซับซ้อนในการคำนวณเป็นการทุ่นค่าใช้จ่ายและประหยัดเวลา

(๔) Consistency คือ ตัวประมาณค่าที่จะมีค่าเข้าไปใกล้ค่าพารามิเตอร์ที่ต้องการประมาณเมื่อใช้ขนาดตัวอย่างมากขึ้น นั่นคือ

$$\lim_{n \to \infty} \text{Prob.} \left\{ |\hat{\theta}_n - \theta| < \epsilon \right\} = 1$$

$\hat{\theta}$ เป็นตัวประมาณค่าของ θ

๒.๒ ลักษณะสำคัญ ๆ ของข้อมูลที่เป็นเป้าหมายของการประมาณผลในการสำรวจโดยทั่ว ๆ ไป

พารามิเตอร์ต่าง ๆ ที่สำคัญที่ใช้ในการประมาณผลของโครงการต่าง ๆ มี ๔ ตัว คือ ยอดรวม (total) ค่าเฉลี่ย (mean) สัดส่วน (proportion) และอัตราส่วน (ratio) รวมทั้งค่าความแปรปรวนของพารามิเตอร์ทั้งสี่นี้ด้วย

ให้ $x_1, x_2, \ldots\ldots\ldots, x_n$ เป็นลักษณะที่สนใจจะศึกษาของหน่วยที่ i ซึ่งเลือกโดยวิธีสุ่มอย่างง่ายจากประชากรทั้งหมด N หน่วย

- ยอดรวม T(x)

$$T(x) = \sum_{i=1}^{N} x_i$$

ซึ่งค่าประมาณของยอดรวม คือ

$$\hat{T}(X) = \frac{N}{n} \sum_{i=1}^{n} x_i$$

โดยค่าความแปรปรวนของค่าประมาณของผลรวมคือ

$$V(\hat{T}(x)) = \frac{N^2}{n} \sigma^2 \quad ; \quad \sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \bar{\bar{X}})^2}{N}$$

๒. ค่าเฉลี่ย ให้ค่าเฉลี่ยของลักษณะที่สนใจจะศึกษาของประชากรเป็น

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

$$= T(X) / N$$

และค่าประมาณของค่าเฉลี่ยคือ

$$\bar{X} = \sum_{i=1}^{n} x_i / n = \hat{T}(X) / N$$

โดยค่าความแปรปรวนของค่าเฉลี่ย $V(\bar{X}) = \frac{\sigma^2}{n}$

๓. อัตราส่วน จากหน่วยตัวอย่าง n หน่วย ถ้าจะศึกษาลักษณะสองลักษณะคู่กัน คือ x, y ดังนี้จะได้ข้อมูล $(X_1, Y_1), (X_2, Y_2) \ldots (X_n, Y_n)$ ที่ให้ค่าประมาณอัตราส่วนของประชากร

$$R = \sum_{i=1}^{N} Y_i / \sum_{i=1}^{N} X_i$$

เป็น

$$\hat{R} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i = \bar{y} / \bar{x}$$

โดยมีค่าความแปรปรวนของอัตราส่วนเป็น

$$V(\hat{R}) \doteq R^2 \left[\frac{\sigma_{\bar{x}}^2}{\bar{\bar{X}}^2} + \frac{\sigma_{\bar{y}}^2}{\bar{\bar{Y}}^2} - \frac{2\sigma_{\bar{x}\bar{y}}}{\bar{\bar{X}}\bar{\bar{Y}}} \right]$$

๔. สัดส่วน

ในการเก็บรวบรวมข้อมูลสถิติ บางครั้งจะไม่บันทึกข้อมูลเชิงปริมาณเป็น $x_1, x_2 \ldots\ldots\ldots\ldots x_n$ อาจต้องการข้อมูลซึ่งได้จากการนับหลังจากมีการจัดจำแนกหน่วยแจงนับในเชิงคุณภาพ เช่น ในหน่วยแจงนับซึ่งมี N หน่วย เราอาจต้องการประมาณจำนวนบ้านในกรุงเทพฯ ซึ่งใช้ผงซักฟอกชนิดหนึ่ง สถิติจำนวนบ้านที่ใช้ผงซักฟอกเป็นสถิติที่ได้จากการนับหลังจากจำแนกคุณภาพ

ถ้าให้ A = จำนวนบ้านที่ใช้ผงซักฟอก

$P = \frac{A}{N}$

ในการสุ่มตัวอย่างเพื่อประมาณค่า A ให้ n เป็นขนาดของตัวอย่างและจะจัดจำแนกตัวอย่าง n หน่วย ที่สุ่มได้จากตัวอย่างมีจำนวนเท่าไรที่มีลักษณะที่ต้องการ

สมมุติให้มีลักษณะที่ต้องการ = a

ค่าประมาณของสัดส่วน คือ $\hat{P} = \frac{a}{n}$

ค่าความแปรปรวนของค่าประมาณของสัดส่วน $V(\hat{P}) = \frac{P(1-P)}{n}$

คุณสมบัติทางคณิตศาสตร์ของความแปรปรวน

๑. ความแปรปรวนจะมีค่าเป็น + เสมอ

๒. $V\left(\sum_{i=1}^{N} x_i\right) = \sum_{i=1}^{N} V(x_i) + \sum_{i \neq j} cov.(x_i, x_j)$

๓. $V(X_1 \pm a) = V(X_1)$

๔. $V(ax) = a^2 V(x)$

๕. $V(a) = 0$ ในเมื่อ a = ค่าคงที่ (constant)

๖. $V\left(\sum_{i=1}^{N} a_i x_i\right) = \sum_{i=1}^{N} a_i^2 V(x_i) + \sum_{i \neq j}^{N} a_i a_j\, cov.(x_i, x_j)$

Covariance เป็นค่าที่ใช้วัดการกระจายร่วมหรือแปรปรวนร่วม (covariate) ของข้อมูล ๒ ชุด เช่น X_i, Y_j ซึ่งเราให้คำจำกัดความของ covariance ระหว่างตัวแปร x_i, y_j คือ

$$\mathrm{Cov}(X_i, Y_j) = E(X_i - E(X_i))(Y_j - E(Y_j))$$

หรือ $$\mathrm{Cov}(X_i, Y_j) = E((X_i - \bar{\bar{X}})(Y_j - \bar{\bar{Y}}))$$

$$= \sum_i \sum_j (x_i - \bar{\bar{x}})(y_j - \bar{\bar{y}})\ \mathrm{prob}(x_i, y_j)$$

ค่าของ covariance อาจพิจารณาวัดได้จากค่าการกระจายร่วมของตัวแปร X_i, Y_j ค่าของ covariance จะมีค่ามากและเป็น + เมื่อ X_i, Y_j แปรไปในทางเดียวกัน และมีค่าเป็น – เมื่อ X_i, Y_j แปรไปในทางตรงกันข้าม