บทที่ 3
แผนสำรวจแบบแบ่งชั้นภูมิ
Stratified Sampling Plan

3.1 ความหมายและเหตุผล

ในงานสำรวจนั้นบ่อยครั้งที่เราพบปัญหาในทางปฏิบัติ เช่น หน่วยสำรวจกระจัดกระจายทั่วไปทำให้การควบคุมงานสนามเป็นไปได้โดยยาก บางครั้งเมื่อใช้วิธี SRS เลือกตัวอย่าง ปรากฏว่ากลุ่มตัวอย่างไม่ใช่ตัวแทนที่ดีพอเพราะบางส่วนของประชากรไม่ได้รับการเลือกหรือได้รับเลือกเข้ามาในปริมาณต่ำทำให้ค่าประมาณ Overestimate หรือ Underestimate และบางครั้งเราอาจต้องการทราบข้อมูลจากประชากรย่อยบางกลุ่มแต่ปรากฏว่าขาดข้อมูลอย่างเพียงพอหรือนัยหนึ่งหน่วยสำรวจจากกลุ่มประชากรย่อยที่สนใจได้รับการเลือกมาเป็นตัวอย่างน้อยเกินไปหรือไม่ปรากฏเลย เหล่านี้เป็นปัญหาพื้นฐานที่นำไปสู่การหาทางพัฒนาแผนสำรวจใหม่ขึ้นมาที่สามารถปิดกั้นปัญหาเหล่านี้พร้อมทั้งเพิ่มระดับความแม่นยำของการประมาณค่าให้สูงขึ้นอีกด้วย

ตัวอย่างเช่นการสำรวจยอดขายของห้างสรรพสินค้าถ้าใช้วิธี SRS จะเกิดปัญหาสำคัญ 2 ประการคือ

ประการแรก เนื่องจากห้างสรรพสินค้ามีขนาดแตกต่างกัน ซึ่งห้างสรรพสินค้าขนาดต่างกันย่อมมียอดขายต่างกัน และ ถ้าเราแบ่งห้างสรรพสินค้าเป็น 3 ประเภท คือ ห้างสรรพสินค้าขนาดเล็ก ขนาดกลาง และขนาดใหญ่ ซึ่งโดยปกติห้างสรรพสินค้าขนาดเล็กจะมีจำนวนมากที่สุด ห้างสรรพสินค้าขนาดใหญ่จะมีจำนวนน้อยที่สุด ถ้าใช้แผนสำรวจแบบ SRS กลุ่มตัวอย่างห้างสรรพสินค้าอาจได้มาจากเฉพาะกลุ่มห้างขนาดเล็กและขนาดกลาง

ซึ่งถ้าเป็นเช่นนี้ ค่าประมาณของยอดรวมที่ขายจะต่ำกว่าความเป็นจริง และถ้าตัวอย่างห้างสรรพสินค้าส่วนใหญ่ได้มาจากกลุ่มห้างขนาดกลางและขนาดใหญ่ จะมีผลให้ค่าประมาณยอดขายสูงกว่าความเป็นจริง ดังนี้เป็นต้น จะเห็นได้ว่า วิธี SRS ไม่อาจควบคุมให้กลุ่มตัวอย่างกระจายไปทั่วกลุ่มประชากรในทุกลักษณะย่อยได้

ประการที่ 2 ถ้านักวิจัยปรารถนาจะทราบข้อสนเทศเฉพาะห้างสรรพสินค้าขนาดใดขนาดหนึ่ง วิธี SRS อาจไม่เอื้ออำนวยให้ได้รับสิ่งที่ต้องการได้ เช่น ต้องการทราบข้อมูลบางประการที่เกี่ยวกับห้างสรรพสินค้าขนาดใหญ่ เช่น ลักษณะการบริหารงาน ยอดขายต่อวัน ฯลฯ แต่ปรากฏว่าไม่มีห้างสรรพสินค้าขนาดใหญ่ได้รับเลือกเป็นตัวอย่างเลยหรือได้รับเลือกเป็นตัวอย่างเพียงไม่กี่แห่ง เราย่อมไม่อาจกะประมาณพารามิเตอร์สำหรับในส่วนที่เกี่ยวกับห้างขนาดใหญ่ได้ หรือแม้กระทั่งจะกระทำได้ ก็ไม่น่าเชื่อถือและขาดความแม่นยำเพราะกลุ่มตัวอย่างเฉพาะส่วนของห้างขนาดใหญ่มีน้อยเกินไป

ในการปฏิบัติงานสนามนั้น ถ้าเราจำแนกประชากรออกเป็นกลุ่ม ๆ ตามเขตการปกครอง เช่น หมู่บ้าน แล้วค่อยดำเนินการสำรวจให้เสร็จสิ้นคราวละหมู่บ้าน กรณีเช่นนี้ย่อมเป็นการสะดวกต่อการควบคุมงานสนาม เพราะผู้ควบคุมงานสนามสามารถควบคุมให้การปฏิบัติงานเป็นไปตามแผนได้อย่างใกล้ชิด วิธีนี้ย่อมดีกว่า วิธี SRS ที่ปล่อยพนักงานสำรวจกระจัดกระจายไปทั่วทุกหมู่บ้านพร้อมกัน ซึ่งยากต่อการควบคุมงานสนาม

ด้วยปัญหาดังกล่าวจึงได้มีการพัฒนาแผนสำรวจขึ้นมาใหม่ เรียกว่า แผนสำรวจแบบชั้นภูมิ (Stratified Sampling Plan) แผนดังกล่าวให้เริ่มต้นด้วยการจำแนกประชากรออกเป็นกลุ่มประชากรย่อยที่ไม่ซ้ำซ้อนกัน (Nonoverlapping Supopulation) เรียกว่า ชั้นภูมิ โดยพยายามจัดให้หน่วยสำรวจที่มีธรรมชาติคล้ายคลึงกันไว้ในชั้นภูมิเดียวกันหรือนัยหนึ่งภายในชั้นภูมิเดียวกันจะต้องมีความคล้ายคลึงกันหรือเป็นเนื้อเดียวกัน (Homogeneous) ที่สุด เท่าที่จะพึงเป็นไปได้ ต่างชั้นภูมิกันมีความแตกต่างกันมากที่สุด จากนั้นจึงดำเนินการสุ่ม

หน่วยสำรวจมาจากทุกชั้นภูมิในปริมาณมากน้อยแตกต่างกันไปตามความเหมาะสม[1] และจะใช้แผนการเลือกตัวอย่างมาจากแต่ละชั้นภูมิแบบใดก็ได้[2] การกระทำดังกล่าวจะก่อให้เกิดผลดีหลายประการคือ

1. ง่ายต่อการควบคุมงานสนาม เพราะการปฏิบัติงานสนามอาจทำให้เสร็จสิ้นทีละขั้นภูมิ เรื่อยไปจนครบทุกขั้นภูมิ ผู้ควบคุมงานสนามสามารถควบคุมงานได้อย่างใกล้ชิด และสามารถตัดสินใจแก้ปัญหาเฉพาะหน้าได้ทันท่วงที

2. สามารถทราบข้อมูลและสามารถกะประมาณค่าพารามิเตอร์เป็นรายชั้นภูมิ หรือเฉพาะชั้นภูมิที่สนใจได้ ผลการประมาณค่าพารามิเตอร์ของในแต่ละชั้นภูมิมีความถูกต้องแม่นยำและน่าเชื่อถือเพียงพอ และเมื่อนำค่าประมาณของพารามิเตอร์ (Stratum Parameter) เดียวกันจากทุกชั้นภูมิมารวมกัน (pooling estimates) จะได้ค่าประมาณของพารามิเตอร์ (Population Parameter) หรือโดยนัยกลับกันการสำรวจแบบแบ่งชั้นภูมินั้น นอกจากจะสามารถกะประมาณค่าพารามิเตอร์ของกลุ่มประชากรได้แล้ว ยังสามารถกะประมาณพารามิเตอร์ของชั้นภูมิเฉพาะชั้นภูมิที่สนใจหรือทุกชั้นภูมิได้ อีกทั้งยังทำให้สามารถเปรียบเทียบความแตกต่างในระหว่างชั้นภูมิได้ นอกจากนี้ ถ้าเราต้องการเน้นความสำคัญที่ชั้นภูมิใดเราก็สามารถเพิ่มจำนวนตัวอย่างให้แก่ชั้นภูมินั้นได้

3. กลุ่มตัวอย่างรวม (Combined Sample) เป็นตัวแทนที่ดีของกลุ่มประชากร เพราะจะประกอบไปด้วยหน่วยตัวอย่างที่มาจากทุกส่วนของกลุ่มประชากร

4. ค่าประมาณของพารามิเตอร์มีความยำกว่าค่าที่ประมาณโดยวิธี SRS กล่าวคือ $V(\hat{\theta}_{str}) \leq V(\hat{\theta}_{ran})$ หรือนัยหนึ่ง วิธีสำรวจแบบแบ่งชั้นภูมิให้ช่วงของการประมาณค่าแคบกว่าวิธี SRS

---

[1] จะกล่าวถึงรายละเอียดเกี่ยวกับเรื่องนี้ในเรื่องการจัดสรรขนาดตัวอย่างให้แก่ชั้นภูมิ

[2] ถ้าใช้แผน SRS เลือกตัวอย่างมาจากแต่ละชั้นภูมิ เรียกแผนสำรวจนี้ว่า Stratified Random Sampling ถ้าใช้แผน Systematic Sampling เลือกตัวอย่างจากแต่ละชั้นภูมิเรียกแผนสำรวจนี้ว่า Stratified Systematic Sampling ถ้าใช้แผน Cluster Sampling เลือกตัวอย่างมาจากแต่ละ ชั้นภูมิเรียกแผนสำรวจนี้ว่า Stratified Cluster Sampling ในที่นี้จะกล่าวถึงเฉพาะ Stratified Random Sampling เท่านั้น แผนผสมอีก 2 แบบจะเว้นไว้ เพราะการประมาณค่าจะยึดถือแนวเดียวกัน

5. เหตุที่การสุ่มตัวอย่างจากแต่ละชั้นภูมิเป็นไปได้โดยอิสระจะมีผลให้ได้กลุ่มตัวอย่างอิสระ L กลุ่ม ตามจำนวนชั้นภูมิ ซึ่งประโยชน์ประการสำคัญอีกประการหนึ่งที่ได้รับจากแผนนี้ก็คือทำให้เราสามารถเปรียบเทียบพารามิเตอร์ระหว่างชั้นภูมิได้ เช่น สามารถทดสอบสมมุติฐานต่อไปนี้คือ

$$\text{Ho:}\quad \bar{X}_1 = \bar{X}_2 = \ldots = \bar{X}_L \qquad \text{vs} \qquad \text{H}_1\text{:}\quad \bar{X}_1 \neq \bar{X}_2 \neq \ldots \neq \bar{X}_L$$

$$\text{Ho:}\quad P_1 = P_2 = \ldots = P_L \qquad \text{vs} \qquad \text{H}_1\text{:}\quad P_1 \neq P_2 \neq \ldots \neq P_L$$

$$\text{Ho:}\quad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_L^2 \qquad \text{vs} \qquad \text{H}_1\text{:}\quad \sigma_1^2 \neq \sigma_2^2 \neq \ldots \neq \sigma_L^2$$

หรือสมมุติฐานอื่น ๆ

อย่างไรก็ตาม แผนสำรวจแบบชั้นภูมิแม้จะมีข้อได้เปรียบหลายประการแต่ก็มีจุดอ่อนที่สำคัญคือการประมาณค่ามีความยุ่งยากซับซ้อนกว่า ยิ่งจำแนกประชากรออกเป็นหลายชั้นภูมิมากเพียงใด ความยุ่งยากก็จะเพิ่มมากขึ้นเพียงนั้น ดังนั้นในทางปฏิบัติจึงไม่ควรจำแนกประชากรให้มีจำนวนชั้นภูมิมากเกินไปแม้ว่าการกระทำเช่นนั้นจะให้ค่าประมาณที่มีประสิทธิภาพสูงก็ตาม ปัญหาสำคัญจึงอยู่ที่ควรจำแนกประชากรออกเป็นกี่ชั้นภูมิจึงถือว่าเพียงพอหรือเหมาะสม

การกำหนดจำนวนชั้นภูมิและวิธีแบ่งชั้นภูมินั้นโดยปกติจะขึ้นอยู่กับตัวแปรที่เราสนใจจะศึกษาและสามัญวินิจฉัยของนักวิจัยเป็นสำคัญ เช่น การประมาณยอดขายของห้างสรรพสินค้าตัวแปรคือยอดขาย แต่เนื่องจากห้างสรรพสินค้าขนาดต่างกันจะมียอดขายต่างกัน ดังนั้น ถ้าต้องการกะประมาณยอดขายก็ควรแบ่งชั้นภูมิโดยยึดถือขนาดของห้างเป็นเกณฑ์จำแนก ควรจำแนกเป็นกี่ชั้นภูมิย่อมขึ้นอยู่กับสามัญวินิจฉัยของนักวิจัยแต่ถ้าปรารถนาจะประมาณรายได้ของห้างควรใช้ทำเลที่ตั้งเป็นเกณฑ์จำแนกชั้นภูมิดังนี้เป็นต้น หลักเกณฑ์ทั่วไปสำหรับเรื่องนี้มีลักษณะเป็นเชิงอัตนิยม (Subjective Judgement) อยู่มาก แต่ปัจจัยที่สำคัญยิ่งที่พึงยึดถือไว้ให้เป็นปกติก็คือควรจำแนกประชากรออกเป็นกลุ่มย่อย ๆ ให้ภายในกลุ่มเดียวกันมีความละม้ายคล้ายคลึงกันมากที่สุดและต่างกลุ่มมีความแตกต่าง

กันอย่างเห็นได้ชัด เมื่อยึดถือตามแนวนี้แล้วแม้จะได้ชั้นภูมิกี่ชั้นภูมิก็ไม่ใช่สิ่งที่ต้องวิตกกังวล เพราะถ้าสามารถทำได้เช่นนี้ ค่าประมาณจะมีประสิทธิภาพสูง

อย่างไรก็ตามเทคนิคดังกล่าวเป็นวิธีที่ใช้กันในทางปฏิบัติและมีลักษณะค่อนข้างไปในทางอัตนิยม ซึ่งขึ้นอยู่กับประสบการณ์ของนักวิจัยเป็นสำคัญ และอาจนำไปสู่ความขัดแย้งกันในทางทฤษฎีได้ ในทางทฤษฎีเรามีเทคนิคสำหรับการกำหนดจำนวนชั้นภูมิและการจัดชั้นภูมิหลายประการ เช่น เทคนิคที่เสนอโดย Daleneous and Hodge เทคนิคที่เสนอโดย Aoyama เทคนิคที่เสนอโดย Ekman และอื่น ๆ ซึ่งจะได้กล่าวถึงในลำดับต่อไป

อนึ่ง การศึกษาแผนสำรวจแบบแบ่งชั้นภูมิในที่นี้จะกล่าวถึงเฉพาะการแบ่งชั้นภูมิและเลือกตัวอย่างจากแต่ละชั้นภูมิโดยวิธี SRS ที่เรียกว่า Stratified Random Sampling เท่านั้น จะไม่กล่าวถึง Stratified Systematic Sampling,Stratified Cluster Sampling และแบบผสมอื่น ๆ เพราะถ้ามีความเข้าใจในวิธีการของ Stratified Random Sampling ได้ดีแล้ว แบบผสมแบบอื่น ๆ ก็มิใช่เป็นเรื่องยากอีกต่อไป

3.2 นิยามและสัญลักษณ์

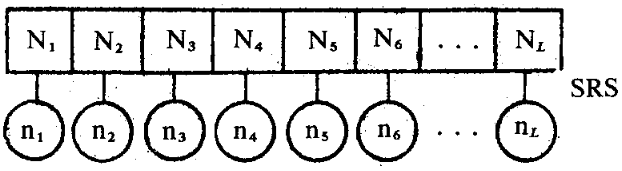
สมมุติเราจำแนกประชากรออกเป็น L ชั้นภูมิ แต่ละชั้นภูมิมีขนาด (Stratum Size หรือ Subpopulation Size) N_h; $h = 1, 2, \ldots, L$ [1] โดยที่ขนาดของชั้นภูมิทุกชั้นภูมิรวมกันจะต้องเท่ากับขนาดของประชากร (Population Size)

คือ $\sum_{h}^{L} N_h = N$ แล้วสุ่มตัวอย่างขนาด n_h ; $h = 1, 2, \ldots, L$ มาจากทุกชั้นภูมิ โดยที่ $\sum_{h}^{L} n_h = n$ ซึ่งการสุ่มตัวอย่างจากแต่ละชั้นภูมินั้นเราอาจใช้แผนสำรวจแบบใดที่เห็นว่าเหมาะสมก็ได้ ในที่นี้จะเสนอไว้เฉพาะกรณีที่มีการสุ่มตัวอย่างจากแต่ละชั้นภูมิโดยใช้แผน SRS [2] ดังไดอะแกรมต่อไปนี้

---

[1] ใช้ h เป็น running index แทนชั้นภูมิ

[2] ถ้าเข้าใจวิธีการสุ่มตัวอย่างจากแต่ละชั้นภูมิมาโดยวิธี SRS ดีแล้ว นักศึกษาสามารถพัฒนาทฤษฎีการประมาณค่าสำหรับกรณีที่สุ่มตัวอย่างจากแต่ละชั้นภูมิโดยวิธีอื่นได้โดยไม่ยากนัก



ดังนั้น เราจึงสามารถนิยามสัญลักษณ์และตัวสถิติตลอดจนพารามิเตอร์ที่เกี่ยวข้องได้ดังนี้

N_h; $h = 1,2,\ldots,L$ คือจำนวนหน่วยสำรวจทั้งหมดในชั้นภูมิที่ h หรือนัยหนึ่ง N_h คือขนาดของชั้นภูมิที่ h (Stratum Size)

n_h; $h = 1,2,\ldots,L$ คือจำนวนหน่วยตัวอย่างในกลุ่มตัวอย่างที่สุ่มมาจากชั้นภูมิที่ h [1]

x_{hi}; $i = 1,2,\ldots,n_h$; $h = 1,2,\ldots,L$ คือค่าของตัวแปรสุ่มที่ i ที่สุ่มมาจากชั้นภูมิที่ h

$W_h = \dfrac{N_h}{N}$; $h = 1,2,\ldots,L$ คือน้ำหนัก (weight) ของชั้นภูมิที่ h

$\bar{X}_h = \dfrac{1}{N_h}\sum\limits_{i}^{N_h} x_{hi}$; $h = 1,2,\ldots,L$ คือค่าเฉลี่ยจริงของชั้นภูมิที่ h (True Stratum Mean)

$\bar{x}_h = \dfrac{1}{n_h}\sum\limits_{i}^{n_h} x_{hi}$; $h = 1,2,\ldots,L$ คือค่าเฉลี่ยในกลุ่มตัวอย่างจากชั้นภูมิที่ h

$P_h = \dfrac{1}{N_h}\sum\limits_{i}^{N_h} x_{hi}$; $h = 1,2,\ldots,L$ โดยที่ $x_{hi} = 1$ ถ้า x_{hi} C_h และ $x_{hi} = 0$ ถ้า x_{hi} C_h' คือสัดส่วนจริงของชั้นภูมิที่ h (True Stratum Proportion)

---

[1] $N_1 + N_2 + \ldots + N_L = N$ และ $n_1 + n_2 + \ldots + n_L = n$

$\hat{P}_h = p_h = \frac{1}{n_h} \sum_{i}^{n_h} x_{hi}\,; h=1,2,\ldots,L$ โดยที่ $x_{hi}=1$ ถ้า $x_{hi} \in C_h$ และ $x_{hi} = 0$ ถ้า $x_{hi} \in C'_h$

คือสัดส่วนในกลุ่มตัวอย่างจากชั้นภูมิที่ h

$R_h = \sum_{i}^{N_h} x_{hi} / \sum_{i}^{N_h} y_{hi} = \bar{X}_h/\bar{Y}_h\,; h=1,2,..,L$ คืออัตราส่วนจริงของชั้นภูมิที่ h (True Ratio)

$\hat{R}_h = \sum_{i}^{n_h} x_{hi} / \sum_{i}^{n_h} y_{hi} = \bar{x}_h/\bar{y}_h\,; h=1,2,..,L$ คืออัตราส่วนในกลุ่มตัวอย่างจากชั้นภูมิที่ h

$\bar{X} = \frac{1}{N} \sum_{h}^{L} \sum_{i}^{N_h} x_{hi}$ คือค่าเฉลี่ยจริงของประชากร[1]

$S_h^2 = \frac{1}{N_h-1} \sum_{i}^{N_h} (x_{hi}-\bar{X}_h)^2\,; h=1,2,\ldots,L$ คือความแปรปรวนจริงของชั้นภูมิที่ h (True Stratum Variance)

$s_h^2 = \frac{1}{n_h-1} \sum_{i}^{n_h} (x_{hi}-\bar{x}_h)^2\,; h=1,2,..,L$ คือความแปรปรวนในกลุ่มตัวอย่างชั้นภูมิที่ h

ข้อสังเกต

สัญลักษณ์และนิยามเหล่านี้มิได้แตกต่างไปจากที่เคยกล่าวถึงมาแล้วในบทที่ 2 ต่างกันในเพียงในเรื่องนี้มี subscript h เพิ่มเข้ามาเพื่อชี้ให้เห็นว่าเรากำลังพูดถึงชั้นภูมิใดเท่านั้น วิธีศึกษาเรื่องนี้ให้เข้าใจง่ายก็คือให้ถือว่าชั้นภูมิหนึ่ง ๆ ทำหน้าที่เสมือนกลุ่มประชากรที่เคยกล่าวถึงในบทที่ 2 การศึกษาก็ศึกษาเป็นรายชั้นภูมิ ศึกษาถึงชั้นภูมิใดก็ให้หมายเลข

---

1 $\sum_{h=1}^{L} \sum_{i}^{N_h} x_{hi}$ คือผลรวมของค่าของตัวแปรสุ่มทุกหน่วยในทุกชั้นภูมิ

$\sum_{h}^{L} \sum_{i}^{N_h} x_{hi} = (x_{11}+x_{12}+\ldots+x_{1N_1}) + (x_{21}+x_{22}+\ldots+x_{2N_2}) + \ldots + (x_{L1}+x_{L2}+\ldots+x_{LN_L})$

(subscript) ระบุไว้แก่ชั้นภูมินั้น เช่น ศึกษาถึงชั้นภูมิที่ 6 (h = 6) ก็ให้ระบุหมายเลข 6 ไว้ เช่น N_6, n_6, x_{6i}, W_6, $\bar{X}_6$, $\bar{x}_6$, P_6, p_6, R_6, $\hat{R}_6$, S_6^2, s_6^2 เป็นต้น ส่วนโครงสร้างภายในของสัญลักษณ์เหล่านี้จะมีลักษณะเช่นเดียวกับที่เคยศึกษาผ่านมาแล้ว

3.3 การประมาณค่าเฉลี่ยและยอดรวม

(Estimation of Population Mean, $\bar{X}$, and Populaion Total)

3.3.1 คำแนะนำทั่วไป

ดังที่ได้กล่าวมาแล้วในตอน 3.2 ว่า เทคนิคที่ใช้ในส่วนที่เกี่ยวกับแผนสำรวจแบบแบ่งชั้นภูมินั้นมิได้แตกต่างไปจากวิธีการของ SRS เลย เพียงแต่แผนสำรวจแบบแบ่งชั้นภูมิเกี่ยวข้องอยู่กับกลุ่มประชากรย่อยหลายกลุ่ม ซึ่งเราใช้อักษร h เป็น running index ที่ช่วยชี้ให้เห็นว่าเรากำลังเกี่ยวข้องอยู่กับกลุ่มย่อยที่เท่าไรเท่านั้น

สำหรับเทคนิคการประมวลผลเพื่อประมาณค่าเฉลี่ยนั้น วิธีที่ง่ายที่สุดซึ่งจะได้กล่าวถึงและใช้เป็นหลักต่อไป ซึ่งรวมตลอดถึงการประมาณสัดส่วนและอัตราส่วนด้วยนั้นก็คือ ให้ประมาณยอดรวมของแต่ละชั้นภูมิก่อนโดยใช้สูตร $\hat{T}_h = N_h\bar{x}_h$ เมื่อเกี่ยวข้องอยู่กับชั้นภูมิที่ h จากนั้นให้นำยอดรวมเหล่านี้ของทุกชั้นภูมิมารวมกันคือ $\hat{T}_1 + \hat{T}_2 + \ldots + \hat{T}_L$ หรือนัยหนึ่ง $N_1\bar{x}_1 + N_2\bar{x}_2 + \ldots + N_L\bar{x}_L$ หรือ $\sum_h^L N_h\bar{x}_h$ ผลลัพธ์ก็คือค่าประมาณของยอดรวมของประชากร (Population Total) ภายหลังจากนั้น ถ้าปรารถนาจะประมาณค่าเฉลี่ยของกลุ่มประชากร (Population) ก็ทำได้โดยง่าย เพียงแต่นำขนาดประชากรคือ N ไปหารยอดรวมดังกล่าว จะได้ผลลัพธ์ตามต้องการ คือ $\hat{\bar{X}} = \frac{\hat{T}}{N} = \frac{1}{N}\sum_h^L N_h\bar{x}_h$ ส่วนการวิเคราะห์เพื่อพัฒนาสูตรความแปรปรวนก็ทำได้โดยง่ายโดยอาศัยสูตรหรือหลักเกณฑ์ที่เกี่ยวกับความแปรปรวน เช่น $V(C) = 0$ ถ้า C เป็นค่าคงที่ หรือ $V(aX) = a^2V(X)$ ถ้า a เป็นค่าคงที่และ X เป็นตัวแปรสุ่ม เป็นต้น

อนึ่ง ขอให้นักศึกษาสังเกตว่าขนาดของชั้นภูมิคือ $N_1, N_2, \ldots, N_L$ หรือเขียนย่อ ๆ ว่า $N_h; h=1,2,\ldots,L$ นั้น เป็นค่าคงที่ ดังนั้นน้ำหนักของชั้นภูมิคือ $W_1 = \frac{N_1}{N}$, $W_2 = \frac{N_2}{N}$, ..., $W_L = \frac{N_L}{N}$ หรือเขียนย่อ ๆ ว่า $W_h = \frac{N_h}{N}$; $h=1,2,\ldots,L$ ย่อมเป็นค่าคงที่ด้วย ความเข้าใจประการนี้ จะช่วยให้นักศึกษาเข้าใจการพัฒนาสูตรความแปรปรวนของตัวประมาณค่าได้เป็นอย่างมาก ผู้เขียนขอย้ำคำว่า "พัฒนา" เพราะเมื่อเผชิญปัญหาในทางปฏิบัตินั้นมิใช่ว่าเราจะสามารถพบสถานการณ์ที่ตรงกับกรณีที่กำลังศึกษาเสมอไป อาจต้องพลิกแพลงแผนสำรวจไปตามสถานการณ์ซึ่งถ้านักศึกษาเข้าใจแนวคิดเบื้องต้นต่าง ๆ ดีพอนักศึกษาย่อมสามารถพัฒนาตัวประมาณค่าต่าง ๆ ได้เองทั้งสามารถพิสูจน์ให้เห็นคุณสมบัติที่น่าพึงปรารถนาของตัวประมาณค่าที่ดี (Good Estimator) ได้อีกด้วย

สำหรับแผนสำรวจแบบแบ่งชั้นภูมินี้ เราจะใช้อักษร 'st'' เป็น subscript ของตัวประมาณค่า เช่น $\bar{x}_{st}$, p_{st}, R_{st} เพื่อชี้ให้เห็นว่าตัวประมาณค่าเหล่านี้เป็นตัวประมาณค่าที่ได้มาจากกลุ่มตัวอย่างที่สุ่มโดยอาศัยแผนสำรวจแบบแบ่งชั้นภูมิ (Stratified Sampling Plan) อักษร st ย่อมาจากคำว่า Stratification

3.3.2 คุณสมบัติของตัวประมาณค่า

ทฤษฎี 3.1 เมื่อดำเนินการสำรวจโดยใช้แผนสำรวจแบบแบ่งขั้นภูมิ โดยที่การสุ่มตัวอย่างจากแต่ละชั้นภูมิใช้แผนสำรวจแบบ SRS แล้ว

ก. $\hat{\bar{X}} = \bar{x}_{st} = \frac{1}{N}\sum_{h}^{L} N_h\bar{x}_h$ โดยที่ $\bar{x}_h = \frac{1}{n_h}\sum_{n_h}^{n_h} x_{hi}$ และ $N = \sum_{h}^{L} N_h$ จะเป็นตัวประมาณค่าที่ปราศจากอคติของ $\bar{X}$ หรือนัยหนึ่งเราสามารถใช้ค่าของ $\bar{x}_{st}$ เป็นตัวแทนของ $\bar{X}$ ได้

ข. ความแปรปรวนของตัวประมาณค่า $\bar{x}_{st}$ คือ

$$V(\bar{x}_{st}) = \frac{1}{N^2}\sum_{h}^{L} \frac{N_h-n_h}{N_h}\frac{N_h^2S_h^2}{n_h} \qquad \text{โดยที่ } S_h^2 = \frac{1}{N_h-1}\sum_{h}^{L}(x_{hi}-\bar{X}_h)^2$$

ค. ตัวประมาณค่าที่ปราศจากอคติของ $V(\bar{x}_{st})$ คือ $\hat{V}(\bar{x}_{st})$

$$\hat{V}(\bar{x}_{st}) = \frac{1}{N^2}\sum_{h}^{L}\frac{N_h-n_h}{N_h}\frac{N_h^2s_h^2}{n_h} \quad \text{โดยที่ } s_h^2 = \frac{1}{n_h-1}\sum_{i}^{n_h}(x_{hi}-\bar{x}_h)^2$$

พิสูจน์ ก. $E(\bar{x}_{st}) \stackrel{?}{=} \bar{X}$

$$E(\bar{x}_{st}) = E\left(\frac{1}{N}\sum_{h}^{L}N_h\bar{x}_h\right)$$

$$= \frac{1}{N}E\left(\sum_{h}^{L}N_h\bar{x}_h\right)$$

$$= \frac{1}{N}\sum_{h}^{L}N_hE(\bar{x}_h)$$

เมื่อพิจารณาเฉพาะในชั้นภูมิที่ h ใด ๆ เราสามารถพิสูจน์ได้ว่า $E(\bar{x}_h) = \bar{X}_h$ (ทฤษฎี 2.1)

ดังนั้น
$$E(\bar{x}_{st}) = \frac{1}{N}\sum_{h}^{L}N_h\bar{X}_h = \frac{1}{N}\sum_{h}^{L}N_h\left(\frac{1}{N_h}\sum_{i}^{N_h}x_{hi}\right)$$

$$= \frac{1}{N}\sum_{h}^{L}\sum_{i}^{N_h}x_{hi}$$

$$= \bar{X}$$

ข. $$V(\bar{x}_{st}) = V\left(\frac{1}{N}\sum_{h}^{L}N_h\bar{x}_h\right) = V\left(\sum_{h}^{L}W_h\bar{x}_h\right) \text{ เมื่อ } W_h = N_h/N$$

$$= \sum_{h}^{L}V(W_h\bar{x}_h) = \sum_{h}^{L}W_h^2V(\bar{x}_h) = \frac{1}{N^2}\sum_{h}^{L}N_h^2V(\bar{x}_h)$$

เมื่อพิจารณาเฉพาะชั้นภูมิที่ h ใด ๆ เราสามารถพิสูจน์ได้ว่า

$$V(\bar{x}_h) = \frac{N_h-n_h}{N_h}\frac{S_h^2}{n_h} \quad \text{(ทฤษฎี 2.2)}$$

ดังนั้น
$$V(\bar{x}_{st}) = \frac{1}{N^2}\sum_{h}^{L}N_h^2\cdot\frac{N_h-n_h}{N_h}\frac{S_h^2}{n_h}$$

$$= \frac{1}{N^2}\sum_{h}^{L}\frac{N_h-n_h}{N_h}\frac{N_h^2S_h^2}{n_h}$$

ค. เมื่อพิจารณาเฉพาะชั้นภูมิที่ h เราสามารถพิสูจน์ได้ว่า $E(s_h^2) = S_h^2$ (ทฤษฎี 2.3) หรือนัยหนึ่ง s_h^2 เป็นตัวประมาณค่าที่ปราศจากอคติของ S_h^2

ดังนั้นตัวประมาณค่าที่ปราศจากอคติของ $V(\bar{x})$ คือ

$$\hat{V}(\bar{x}_{st}) = \frac{1}{N^2} \sum_{h}^{L} \frac{N_h - n_h}{N_h} \frac{N_h^2 s_h^2}{n_h}$$

ทั้งนี้เพียงแต่แทนที่ S_h^2 ด้วย s_h^2 เท่านั้น

ข้อสังเกต

1. จากสูตร $V(\bar{x}_{st})$ ขอให้สังเกตว่า $V(\bar{x}_{st})$ เกิดขึ้นจากผลรวมของ $V(\bar{x}_h)$; $h = 1,2,\ldots,L$ แล้วหาค่าโดยหารผลรวมนั้นด้วย N^2

2. S_h^2 เรียกว่า Stratum Variance หรือ Within Stratum Variance ถ้า S_h^2; $h = 1,2,\ldots,L$ มีค่าต่ำจะมีผลให้ $V(\bar{x}_{st})$ มีค่าต่ำด้วย S_h^2 จะมีค่าต่ำได้ก็ต่อเมื่อตัวแปรสุ่ม X ในชั้นภูมิเดียวกันมีความละม้ายคล้ายคลึงกัน (Honogeneous) หรือโดยนัยตรงข้ามกัน ถ้าสามารถจัดให้สมาชิกภายในชั้นภูมิเดียวกันมีความคล้ายคลึงกันได้จะมีผลให้ Within Stratum Variance มีค่าต่ำและมีผลโดยตรงให้ $V(\bar{x}_{st})$ มีค่าต่ำหรือการประมาณค่าเฉลี่ยด้วยแผนสำรวจแบบแบ่งชั้นภูมิมีความแม่นยำสูง ความจริงประการนี้ชี้ให้เห็นว่าเพราะเหตุใดเราจึงควรจำแนกประชากรออกเป็นชั้นภูมิ

3. ถ้า $N_h >> n_h$; $h = 1,2,\ldots,L$ จะมีผลให้

$$\frac{N_h - n_h}{N_h} \cong 1 \text{ หรือ } \frac{n_h}{N_h} \cong 0 \text{ ดังนั้น } V(\bar{x}_{st}) \cong \frac{1}{N^2} \sum_{h}^{L} \frac{N_h^2 S_h^2}{n_h}$$

บทแทรก 3.1 เมื่อดำเนินการสำรวจโดยใช้แผนสำรวจแบบแบ่งชั้นภูมิโดยที่การสุ่มตัวอย่างจากแต่ละชั้นภูมิใช้แผนสำรวจแบบ SRS แล้ว

ก. $\hat{T}_{st} = N\bar{x}_{st} = \sum_{h}^{L} N_h\bar{x}_h$ จะเป็นตัวประมาณค่าที่ปราศจากอคติของ T โดยที่

$$T = \sum_{h}^{L} \sum_{i}^{N_h} x_{hi}$$

ข. ความแปรปรวนของตัวประมาณค่า $\hat{T}_{st}$ คือ

$$V(\hat{T}_{st}) = \sum_{h}^{L} \frac{N_h - n_h}{N_h} \frac{N_h^2 S_h^2}{n_h} \quad \text{โดยที่} \quad S_h^2 = \frac{1}{N_h - 1} \sum_{i}^{N_h} (x_{hi} - \bar{X}_h)^2$$

ค. ตัวประมาณค่าที่ปราศจากอคติของ $V(\hat{T}_{st})$ คือ $\hat{V}(\hat{T}_{st})$ โดยที่

$$\hat{V}(\hat{T}_{st}) = \sum_{h}^{L} \frac{N_h - n_h}{N_h} \frac{N_h^2 s_h^2}{n_h} \quad \text{เมื่อ} \quad s_h^2 = \frac{1}{n_h - 1} \sum_{i}^{n_h} (x_{hi} - \bar{x}_h)^2$$

พิสูจน์ การพิสูจน์ยึดถือแนวเดียวกับทฤษฎี 3.1 ซึ่งนักศึกษาสามารถพิสูจน์ได้เองโดยง่าย และขอเว้นไว้เป็นแบบฝึกหัด

ตัวอย่าง 3.1 ต้องการกะประมาณการใช้ประโยชน์ของพื้นที่ทำกินของเกษตรกรในท้องที่แห่งหนึ่ง โดยจำแนกเกษตรกรออกเป็น 4 กลุ่มตามขนาดพื้นที่ถือครอง คือกลุ่มที่มีพื้นที่ถือครองไม่เกิน 100 ไร่ 100-200 ไร่ 200-300 ไร่ และ 300 ไร่ขึ้นไปซึ่งแต่ละกลุ่มเหล่านี้มีจำนวนเกษตรกร 100,80,60 และ 40 ครอบครัวตามลำดับ สุ่มตัวอย่าง (โดยวิธี SRS) เกษตรกรจากแต่ละชั้นภูมิมา 6,5,5 และ 4 ครอบครัวตามลำดับ แล้วเข้าดำเนินการสำรวจการใช้ประโยชน์ในพื้นที่ถือครอง ปรากฏข้อมูลจำนวนพื้นที่ที่ใช้ประโยชน์ดังนี้

ชั้นภูมิ	ขนาดของชั้นภูมิ (N_h)	ขนาดตัวอย่าง (n_h)	พื้นที่ที่ทำประโยชน์(x_{hi})
0-100 ไร่	100	6	40,50,90,70,20,60
100-200 ไร่	80	5	140,150,140,130,180
200-300 ไร่	60	5	240,280,260,250,220
300 ไร่ขึ้นไป	40	4	350,330,310,380

ก. จงกะประมาณจำนวนพื้นที่ที่ทำกินถัวเฉลี่ยต่อครอบครัวที่เกษตรในท้องที่นี้ได้ใช้ประโยชน์จริง พร้อมทั้งช่วงเชื่อมั่น 95% ของการประมาณค่า

ข. จงกะประมาณพื้นที่ทำกินรวม (ยอดรวม) ที่เกษตรกรในท้องที่นี้ได้ใช้ประโยชน์จริง พร้อมทั้งช่วงเชื่อมั่น 99% ของการประมาณค่า

วิธีทำ การวิเคราะห์ข้อมูลสรุปได้ดังตารางต่อไปนี้

ขั้นภูมิที่	N_h	n_h	x_{hi}	$\sum_{i}^{n_h} x_{hi}$	$\bar{x}_h$	$N_h\bar{x}_h$	s_h^2	$N_h^2 s_h^2$
1	100	6	40,50,90,70,20,60	330	55	5500	590	5900000
2	80	5	140,150,140,130,180	740	148	11840	370	2368000
3	60	5	240,280,260,250,220	1250	250	15000	500	1800000
4	40	4	350,330,310,380	1370	342.5	13700	891.6	1426560
รวม	280	20		3690		46040		

หมายเหตุ

$$s_1^2 = \frac{1}{n_1-1}\left\{\sum_{i}^{6} x_{1i}^2 - (\sum_{i}^{6} x_{1i})^2/6\right\} = \frac{1}{5}\{21100 - (108900)/6\} = 590$$

s_2^2, s_3^2 และ s_4^2 ก็คำนวณได้ในทำนองเดียวกัน

ก. $\hat{\bar{X}} = \bar{x}_{st} = \dfrac{1}{N}\sum_{h}^{L} N_h\bar{x}_h = \dfrac{46040}{280} = 164.43$ ไร่

นั่นคือเกษตรกรในท้องที่ดังกล่าวใช้ที่ดินให้เป็นประโยชน์จริงโดยถัวเฉลี่ยครอบครัวละ 164.43 ไร่

$$\because \quad \hat{V}(\bar{x}_{st}) = \frac{1}{N^2} \sum_{h}^{L} \frac{N_h - n_h}{N_h} \frac{N_h^2 s_h^2}{n_h}$$

$$\text{ดังนั้น} \quad \hat{V}(\bar{x}_{st}) = \frac{1}{(280)^2} \left\{ \frac{100-6}{100} \cdot \frac{5900000}{6} + \frac{80-5}{80} \cdot \frac{2368000}{5} \right.$$

$$\left. + \frac{60-5}{60} \cdot \frac{1800000}{5} + \frac{40-4}{40} \cdot \frac{1426560}{4} \right\}$$

$$= \frac{1}{78400} (924333.33 + 444000 + 329999.98 + 320976)$$

$$= \frac{2019309.2}{78400} = 25.757$$

$$=> \quad s_{\bar{x}_{st}} = \sqrt{25.757} = 5.075$$

ดังนั้น ช่วงเชื่อมั่น 95% ที่คาดว่าค่าจริง $\bar{X}$ จะปรากฏอยู่คือช่วง

$$\{\bar{x}_{st} - (1.96)(5.075),\ \bar{x}_{st} + (1.96)(5.075)\} = (154.48,\ 174.38)$$

หรือนัยหนึ่ง เราสามารถเชื่อมั่นได้ถึง 95% ว่าเกษตรกรในท้องที่นี้จะใช้ที่ดินเพื่อทำประโยชน์จริงประมาณระหว่าง 154.48 ถึง 174.38 ไร่

ข. $\hat{T}_{st} = \sum_{h}^{L} N_h \bar{x}_h = 46040$

นั่นคือเกษตรกรในท้องที่ดังกล่าวใช้ที่ดินให้เป็นประโยชน์จริง ๆ รวมทั้งสิ้น 46,040 ไร่

$$\because \quad \hat{V}(\hat{T}_{st}) = \sum_{h}^{L} \frac{N_h - n_h}{N_h} \frac{N_h^2 s_h^2}{n_h}$$

$$= 2{,}019{,}309.2$$

$$=> \quad s_{\hat{T}_{st}} = 1{,}421.02$$

ช่วงเชื่อมั่น 95% ของ T คือ $\{\hat{T}_{st} - (1.96)(1{,}421.02),\ \hat{T}_{st} + (1.96)(1{,}421.02)\}$

$$= (43254.80,\ 48825.21)$$

นั่นคือ เราสามารถเชื่อมั่นได้ถึง 95% ได้ว่าเกษตรกรในท้องที่นี้ใช้ที่ดินให้เป็นประโยชน์จริงรวมทั้งสิ้นประมาณระหว่าง 43,254.8 ถึง 48,825.21 ไร่

3.3.3 การจัดสรรจำนวนตัวอย่างให้แก่ชั้นภูมิ (Allocation of Sample Size to Strata)

สิ่งที่ควรคำนึงถึงในการใช้แผนสำรวจแบบแบ่งชั้นภูมิก็คือจะจัดสรรจำนวนตัวอย่างให้แก่แต่ละชั้นภูมิอย่างไรจึงจะเหมาะสม หมายความว่า เมื่อได้กำหนดขนาดตัวอย่างขนาด n สำหรับงานสำรวจขึ้นแล้ว ควรจะแบ่งขนาดตัวอย่าง n ออกเป็นส่วน ๆ อย่างไรสำหรับใช้เป็นขนาดตัวอย่างที่จะใช้สำหรับเลือกหน่วยจากแต่ละชั้นภูมิ แบ่งขนาดตัวอย่างออกเป็น L ส่วนเท่า ๆ กันตามจำนวนชั้นภูมิที่มีอยู่ L ชั้น หรือว่าชั้นภูมิใดมีขนาดใหญ่ก็แบ่งจำนวนตัวอย่างให้มาก ชั้นภูมิใดมีขนาดเล็กก็แบ่งจำนวนตัวอย่างให้น้อย? หรือว่าต้องคำนึงถึงค่าใช้จ่ายในการสำรวจต่อหน่วยของแต่ละชั้นภูมิประกอบด้วย กล่าวคือ ชั้นภูมิใดที่เสียค่าใช้จ่ายในการสำรวจต่อหน่วยแพงกว่าก็สุ่มตัวอย่างมาน้อย ชั้นภูมิใดที่เสียค่าใช้จ่ายในการสำรวจต่อหน่วยน้อยกว่าก็สุ่มตัวอย่างมามาก เหล่านี้ล้วนเป็นปัญหาที่ต้องระลึกถึงเสมอก่อนที่จะมีการสำรวจ เกี่ยวกับเรื่องนี้ขอให้ทำความเข้าใจไว้ว่าในทางปฏิบัติเราไม่นิยมที่จะกำหนดขนาดตัวอย่างให้แก่แต่ละชั้นภูมิแล้วนำขนาดตัวอย่างนั้นมารวมกันในภายหลังเป็นตัวอย่างรวม (Combined Sample) เพราะยุ่งยากและไม่สะดวกต่อการปฏิบัติงาน แต่เรานิยมกำหนดขนาดตัวอย่างรวมขึ้นมาก่อนแล้วค่อยจัดสรรออกเป็นส่วน ๆ ตามจำนวนชั้นภูมิ ชั้นภูมิหนึ่ง ๆ จะได้รับการจัดสรรจำนวนตัวอย่างให้มากน้อยเพียงใดขึ้นอยู่กับความเหมาะสมที่นักวิจัยจะต้องพิจารณาว่าเหมาะสมกับสถานการณ์หรือไม่เพียงใดเป็นเรื่อง ๆ ไป

อนึ่ง ในการศึกษาในลำดับต่อไปนี้จะเป็นการศึกษาถึงเทคนิคการจัดสรรจำนวนตัวอย่าง โดยถือว่าทราบขนาดตัวอย่างรวมคือ n แล้ว ปัญหาเรื่องการกำหนดขนาดตัวอย่างรวมได้อย่างไรจะได้ศึกษาถึงรายละเอียดในตอนต่อไป และขอเกริ่นไว้ก่อนว่าการกำหนดขนาดตัวอย่างรวมนั้นจะต้องอาศัยพื้นฐานความรู้ความเข้าใจที่จะศึกษาต่อไปในตอนนี้เป็นหลัก

เทคนิคการจัดสรรจำนวนตัวอย่างให้แก่แต่ละชั้นภูมินั้นโดยปกติเรานิยมใช้ 4 แบบ ดังนี้คือ

1. การจัดสรรอย่างเท่าเทียมกัน (Equal Allocation)
2. การจัดสรรตามสัดส่วนหรือร้อยละของขนาดชั้นภูมิ (Proportional Allocation)
3. การจัดสรรแบบอุตมะ (Optimum Allocation)
4. การจัดสรรแบบเนย์แมน (Neyman Allocation)

1. **การจัดสรรอย่างเท่าเทียมกัน** (Equal Allocation)

การจัดสรรอย่างเท่าเทียมกันคือวิธีจำแนกขนาดตัวอย่าง n ออกเป็นส่วน ๆ ส่วนละเท่ากันให้แก่แต่ละชั้นภูมิ หมายความว่า เมื่อกำหนดขนาดตัวอย่างไว้เท่ากับ n และมีจำนวนประชากรกลุ่มย่อย หรือชั้นภูมิทั้งสิ้น L ชั้น ขนาดตัวอย่างที่พึงสุ่มจากแต่ละชั้นภูมิจะมีขนาดเท่าเทียมกันเท่ากับ n/L หรือประมาณ $\frac{n}{L}$ ในกรณีที่หารไม่ลงตัว

นั่นคือ $n_h = n/L\,;\ h = 1, 2, \ldots, L$ ทั้งนี้ $\sum\limits_{h}^{L} n_h = n$

การจัดสรรจำนวนตัวอย่างวิธีนี้เหมาะสำหรับกรณีที่ชั้นภูมิต่าง ๆ มีขนาดเท่ากันหรือใกล้เคียงกัน ไม่เหมาะที่จะใช้กับกรณีที่ชั้นภูมิมีขนาดต่างกันมาก

การวิเคราะห์ข้อมูล เช่น ค่าเฉลี่ยและยอดรวมตลอดจนถึงความแปรปรวนของค่าประมาณที่ใช้แผนสำรวจแบบนี้กระทำได้โดยง่าย เพียงแทนที่ n_h ด้วย n/L เท่านั้น ในทางปฏิบัติเราไม่มีความจำเป็นต้องจดจำสูตรเฉพาะเรื่องเพราะถ้าทราบสูตรทั่วไปของ $\bar{x}_{st}$, $\hat{V}(\bar{x}_{st})$, $\hat{T}_{st}$, $\hat{V}(\hat{T}_{st})$ เราสามารถจะทราบสูตรเหล่านี้เฉพาะกรณีจัดสรรเท่ากันได้โดยเพียงแต่แทนที่ n_h ด้วย n/L เท่านั้น นั่นคือ

1. $\hat{\bar{X}} = \bar{x}_{eq} = \frac{1}{N}\sum\limits_{h}^{L} N\bar{x}_h$ โดยที่ $\bar{x}_h = \frac{1}{n_h}\sum\limits_{i}^{n_h} x_{hi}$ เมื่อ $n_h = n/L$

หรือแทนที่ n_h ด้วย n/L จะได้ $\bar{x}_h = \frac{1}{n/L}\sum\limits_{i}^{n/L} x_{hi} = \frac{L}{n}\sum\limits_{i}^{n/L} x_{hi}$