บทที่ 6
การสุ่มตัวอย่างด้วยความน่าจะเป็นในการเลือกไม่คงที่
(Unequal Probability Sampling)

6.1 เหตุผลและความจำเป็น

จากการศึกษาในบทที่ผ่านมาโดยเฉพาะอย่างยิ่ง 2CS 3CS และ multi stage cluster sampling ลักษณะอื่นนั้นเราถือว่าหน่วยสำรวจ (Sampling unit) มีโอกาสได้รับเลือกเท่ากันเสมอไม่ว่าจะมีขนาด มูลค่า หรือความสำคัญอย่างไร เช่น กรอบตัวอย่างประกอบด้วยโรงงานขนาดต่างๆ 500 โรงงาน เราสุ่มมา 33 โรงงาน โดยไม่สนใจว่าโรงงานจะมีขนาดแตกต่างกันหรือไม่ ผลิตสินค้าในปริมาณผิดแผกกันเพียงใด หรือกรอบตัวอย่างประกอบด้วยอำเภอต่างๆ 600 อำเภอ เราสุ่มมาเป็นตัวอย่าง 40 อำเภอ โดยไม่สนใจว่าอำเภอต่างๆ เหล่านี้จะมีขนาด (พื้นที่หรือจำนวนประชากร) แตกต่างกันหรือไม่ เพียงใด เมื่อพิจารณาเรื่องนี้โดยละเอียดแล้วจะเห็นได้ว่าการถือว่าหน่วยสำรวจมีโอกาสรับเลือก (Selection probability) เท่ากันทั้งๆ ที่ความจริงแล้วไม่น่าจะเป็นเช่นนั้น เช่น โรงงานที่ใหญ่กว่ามีคนงานประมาณ 9,000 คน ควรมีโอกาสรับเลือกสูงกว่าโรงงานขนาดเล็กที่มีคนงานประมาณ 60 คน จะส่งผลให้ค่าประมาณขาดความแม่นยำ (คือเมื่อเปรียบเทียบตัวประมาณค่าใน unbiased class เดียวกันจะพบว่า ถ้าให้ Selection probability คงที่เท่าเดิมตลอดแล้วตัวประมาณนั้นจะให้ variance of estimate สูงกว่า) ด้วยเหตุนี้ถ้าเรามีหนทางใดที่ทำให้สามารถหา Selection probability ที่เหมาะสมได้ เราก็ควรนำ Selection probability นั้นมาใช้พัฒนาตัวประมาณค่าเพื่อจะได้ตัวประมาณค่าที่มีคุณภาพดีขึ้น

ความจริงแล้ว unequal probability Sampling เป็นกรณีทั่วไปของ CS ซึ่งกล่าวถึงแล้วในบทที่ 5 แต่มาแยกศึกษาต่างหากด้วยเกรงว่าจะเป็นภาระในการศึกษา CS

คำถามก็คือ ทราบได้อย่างไรว่าถ้าใช้ unequal probability Sampling แล้วจะทำให้ $V(\theta)$ มีค่าต่ำ และ selection probability ดังกล่าวกำหนดมาได้อย่างไร เรื่องนี้

ตอบได้ดังนี้

Selection probability นั้นเราควรกำหนดให้มีค่าผันแปรไปตามความสำคัญของหน่วยสำรวจ คือหน่วยสำรวจใดมีความสำคัญมากก็ให้มีโอกาสได้รับเลือกมาก หน่วยใดมีความสำคัญน้อยก็ให้มีโอกาสได้รับเลือกน้อย คือ

$$p_i \propto \text{ ความสำคัญของหน่วยที่ } i \; ; i = 1,2,\ldots,N$$

แต่เรื่องความสำคัญนี้เป็นเรื่องเชิงอัตนิยม หมายความว่าขึ้นอยู่กับบุคลิกภาพของผู้พิจารณา เห็นได้ว่าอาจมีความขัดแย้งจึงให้กำหนดเสียใหม่ว่าให้ Selection probability ผันแปรไปตามค่าของข้อมูลของหน่วยสำรวจที่ i คือ

$$p_i \propto x_i \; ; i = 1,2,\ldots,N$$

เช่นว่าสนใจจะสำรวจเรื่องกำลังการผลิตของโรงงานต่างๆ โรงงานใดมีกำลังการผลิตสูงกว่าก็สมควรมีโอกาสรับเลือกเป็นตัวอย่างได้มากกว่าโรงงานที่มีกำลังผลิตน้อย หรือถ้าเราสนใจเรื่องปัญหาสังคม ชุมชนใดมีปัญหาสังคมมากกว่าเราสมควรให้ชุมชนนั้นมีโอกาสรับเลือกเป็นตัวแทนสูงกว่าชุมชนที่มีปัญหาสังคมน้อยกว่า

การเลือกตัวอย่างโดยกำหนดให้โอกาสรับเลือกผันแปรไปตามขนาดของข้อมูล(คือ $p_i \propto x_i$; $i = 1,2,\ldots,N$) นี้เรียกว่า Sampling with probability proportional to size (pps)

แต่อย่างไรก็ตาม ในชั้นปฏิบัติแล้ว pps ก็ยังมีปัญหากล่าวคือขนาดของข้อมูลจาก x_i ; $i = 1,2,\ldots,N$ นั้นเป็นค่าจริง (ขอให้สังเกตว่า i วิ่งจาก 1 ถึง N) และค่าจริงนั้นเป็นค่าที่จะทราบได้เฉพาะเมื่อสำรวจแล้วเท่านั้น แปลว่าก่อนสำรวจค่านี้เป็นตัวไม่ทราบค่า เมื่อยังคงเป็น ตัวไม่ทราบค่า อยู่ในระยะก่อนสำรวจการกำหนดค่า p ก่อนสำรวจจึงเป็นเรื่องที่ทำไม่ได้ ทางออกของปัญหานี้ก็คือให้พยายามหาดัชนีบางอย่างที่พอจะใช้คาดเดามูลค่าของตัวแปรได้ เรียกว่า measure of size แล้วให้นำ measure of size ดังกล่าวซึ่งโดยปกติแล้วเราทราบค่าได้มาทำหน้าที่แทน x_i เมื่อพูดถึง measure of size เราก็ต้องหันไปพิจารณาเฉพาะ multistage sampling scheme เพราะโดยปกติเรามักใช้จำนวน

ประชากรเป็น measure of size เช่น มีตำบลอยู่ 19 ตำบล แต่ละตำบลมีขนาดแตกต่างกัน ขนาดของตำบลวัดด้วยจำนวนหมู่บ้าน ตำบลที่มีขนาดใหญ่กว่าก็ควรจะได้รับเลือกเป็นตัวอย่างมากกว่า (Sampling with probability proportional to measure of size (ppms)) หรือ sampling with probability proportional to estimated size (ppes)

ให้ M_i = measure of size ของ psu ที่ i ดังนั้น

$$p_i \propto M_i \ ; i = 1,2,\ldots,L$$

6.2 pps และ ppms

จากสมการ $p_i \propto x_i$ จะพบว่า

$$p_i = kx_i$$

ดังนั้น $\sum_i^N p_i = k \sum_i^N x_i$

$$\therefore \quad k = 1/\sum_i^N x_i$$

ดังนั้น $p_i = x_i/\sum_i^N x_i$ คือ Selection probability ใน pps

สำหรับในกรณีของ ppms นั้นเราสามารถคำนวณหา Selection probability ได้หลายรูป ขึ้นอยู่กับแผนกสำรวจว่าเป็น 1CS 2CS 3CS หรือแผนสำรวจ multistage ลักษณะอื่น และขึ้นอยู่กับการนิยามหน่วยสำรวจชั้นสุดท้ายว่าเป็นหน่วยเดี่ยวหรือกลุ่มของหน่วยเดี่ยว โดยปกติถ้าหน่วยสำรวจขั้นสุดท้ายเป็นกลุ่มของหน่วยเดี่ยว (เช่น ครัวเรือน เป็นหน่วยเดี่ยว หมู่บ้านเป็นกลุ่มของหน่วยเดี่ยว) จำนวนชั้นของ multistage sampling scheme จะลดลง 1 ชั้นและหน่วยสำรวจชั้นสุดท้ายเช่นนี้จะมี Selection probability ที่แตกต่างกัน

ก. 2CS

ให้ N_i; i = 1,2,...,M คือจำนวน ssu ใน psu ที่ i เราเลือก psu มาเป็นตัวอย่าง m หน่วย และเราเข้าทำการสำรวจแก่ ssu ทุกหน่วยในแต่ละ psu ที่เป็นตัวอย่างเหล่านั้น (นี่คือ 2CS ลดลงเป็น 1CS)

กรณีนี้ measure of size คือ N_i ดังนั้น

$p_i \propto N_i \ ; i = 1,2,\ldots, M$

นั่นคือ $p_i = N_i / \sum_{j}^{M} N_i$

ข. 3CS

ให้ $M_i ; i = 1,2,\ldots, L$ คือจำนวน ssu ใน psu ที่ i เราเลือก psu มาเป็นตัวอย่าง 1 หน่วย

ให้ $N_{ij} ; j = 1,2,\ldots, M_i \ ; i = 1,2,\ldots, L$ คือจำนวน tsu ใน ssu ที่ j จาก psu ที่ i และเราเข้าทำการสำรวจแก่ tsu **ทุกหน่วยในแต่ละ** ssu ที่เป็นตัวอย่างเหล่านั้น (นี่คือ 3CS ลดลงเป็น 2CS)

ดังนั้น $p_i \propto M_i$ และ $p_{ij} \propto N_{ij}$ **และพบว่า**

$$p_i = M_i / \sum_{i}^{L} M_i$$

$$p_{ij} = N_{ij} / \sum_{i}^{L} \sum_{j}^{M_i} N_{ij}$$

6.3 การประมาณค่าพารามิเตอร์ใน 1 CS

การประมาณค่าพารามิเตอร์ในตอนนี้หรือในตอนอื่นต่อจากนี้ เราจะศึกษาเฉพาะการประมาณค่ายอดรวม T เท่านั้น ที่ศึกษาเพียงเท่านี้เพราะค่าประมาณของ $\overline{X}$, P และ R ล้วนเป็นกรณีเฉพาะของค่าประมาณของยอดรวมทั้งสิ้น

พิจารณา $\hat{T}$ ในกรณี equal probability จะพบว่า

$$\begin{aligned}\hat{T} &= N\bar{x} \\ &= N\frac{1}{n}\sum_{j}^{n} x_i \\ &= \frac{1}{n}\sum_{i}^{n} \frac{x_i}{1/N}\end{aligned}$$

ถ้าเราเปลี่ยน 1/N เป็น p_i เราจะได้สูตรทั่วไปสำหรับประมาณยอดรวมใน 1 CS ดังนี้

$$\hat{T} = \frac{1}{n}\sum_{i}^{n} \frac{x_i}{p_i}$$

สูตรทั่วไปนี้สามารถใช้ได้ทั้งในกรณีที่ p_i (Selection probability) มีค่าเท่ากันเท่ากับ 1/N หรือมีค่าแตกต่างกันได้ เราให้ชื่อกรณีแรกว่า $\hat{T}$equal และให้ชื่อ

กรณีหลังว่า $\hat{T}$unequal กล่าวคือ

$$\hat{T}\text{equal} = N\bar{x} = \frac{1}{n}\sum_{i}^{n}\frac{X_i}{1/N}$$

$$\hat{T}\text{unequal} = \frac{1}{n}\sum_{i}^{n}\frac{X_i}{P_i}$$

และเราสามารถพิสูจน์ได้โดยง่ายว่า $\hat{T}$unequal เป็นตัวประมาณค่าที่ปราศจากอคติของ T

ทฤษฎี 6.1 ในการดำเนินการสำรวจแบบ 1 CS กำหนดให้ p_i คือค่าความน่าจะเป็นที่หน่วยสำรวจหน่วยที่ i จะได้รับเลือกมาเป็นตัวอย่าง ดังนั้น

1. $E(\hat{T}\text{unequal}) = T$
2. $V(\hat{T}\text{unequal}) = \frac{1}{n}\sum_{j}^{n} p_i \left(\frac{x_i}{p_i} - T\right)^2$
3. $E[\hat{V}(\hat{T}\text{unequal})] = V(\hat{T}\text{unequal})$ โดยที่

$$V(\hat{T}\text{unequal}) = \frac{1}{n}\sum_{j}^{n}\frac{1}{n-1}\left(\frac{x_i}{p_i} - \hat{T}\text{unequal}\right)^2$$

4. ถ้าใช้ pps คือกำหนดให้ $p_i = x_i/\sum_{i}^{n} x_i$ จะมีผลให้ $V(\hat{T}\text{unequal}) = 0$

พิสูจน์ 1. $E(\hat{T}\text{unequal}) = T$

$$E(\hat{T}\text{unequal}) = E\left[\frac{1}{n}\sum_{i}^{n}\frac{x_i}{p_i}\right]$$

$$= \frac{1}{n}\sum_{i}^{n} E\left(\frac{x_i}{p_i}\right)$$

$$= \frac{1}{n}\sum_{i}^{n}\left[\sum_{i}^{N}\left(\frac{x_i}{p_i}\right)p_i\right]$$

$$= \frac{1}{n}\, n\sum_{i}^{N} x_i$$

$$= \quad T$$

2. $V(\hat{T}unequal) = V\left[\frac{1}{n}\sum_{i}^{n}\frac{x_i}{p_i}\right]$

$$= \frac{1}{n^2}\sum_{i}^{n} V\left(\frac{x_i}{p_i}\right)$$

$$= \frac{1}{n^2}\, nV\left(\frac{x_i}{p_i}\right)$$

$$= \frac{1}{n} E\left[\frac{x_i}{p_i} - E\left(\frac{x_i}{p_i}\right)\right]^2$$

$$= \frac{1}{n}\sum_{i}^{N}\left[\frac{x_i}{p_i} - E\left(\frac{x_i}{p_i}\right)\right]^2 p_i$$

ดังนั้น $V(\hat{T}unequal) = \frac{1}{n}\sum_{i}^{N} p_i\left(\frac{x_i}{p_i} - T\right)^2$

หมายเหตุ $E\left(\frac{x_i}{p_i}\right) = \sum_{i}^{N}\left(\frac{x_i}{p_i}\right)p_i = \sum_{i}^{N} x_i = T$

3. $E[\hat{V}(\hat{T}unequal)] = E\left[\frac{1}{n}\sum_{i}^{n}\frac{1}{n-1}\left(\frac{x_i}{p_i} - \hat{T}unequal\right)^2\right]$

$$= \frac{1}{n(n-1)} E\sum_{i}^{n}\frac{1}{n}\left(\frac{x_i}{p_i} - \hat{T}unequal\right)^2$$

$$= \frac{1}{n(n-1)} E\sum_{i}^{n}\left[\left(\frac{x_i}{p_i}\right)^2 + \hat{T}^2unequal - 2\hat{T}unequal\left(\frac{x_i}{p_i}\right)\right]$$

$$= \frac{1}{n(n-1)} E\left[\sum_{i}^{n}\left(\frac{x_i}{p_i}\right)^2 + n\hat{T}^2unequal - 2\hat{T}unequal\sum_{i}^{n}\frac{x_i}{p_i}\right]$$

$$= \frac{1}{n(n-1)} E\left[\sum_i^n \left(\frac{x_i}{p_i} \right)^2 - n\hat{T}^2\text{unequal} \right]$$

$$\because \quad \sum_i^n \frac{x_i}{p_i} = n\hat{T} \text{ unequal}$$

$$= \frac{1}{n(n-1)} E\left[\sum_i^n \left(\frac{x_i}{p_i} \right)^2 - n\left(\frac{1}{n} \sum_i^n \frac{x_i}{p_i} \right)^2 \right]$$

$$= \frac{1}{n(n-1)} E\left[\sum_i^n \left(\frac{x_i}{p_i} \right)^2 - \frac{1}{n} \sum_i^n \left(\frac{x_i}{p_i} \right)^2 - \frac{1}{n} \sum_{i \neq j}^n \left(\frac{x_i}{p_i} \right)\left(\frac{x_j}{p_j} \right) \right]$$

$$= \frac{1}{n(n-1)} E\left[\frac{n-1}{n} \sum_i^n \left(\frac{x_i}{p_i} \right)^2 - \frac{1}{n} \sum_{i \neq j}^n \left(\frac{x_i}{p_i} \right)\left(\frac{x_j}{p_j} \right) \right]$$

$$= \frac{1}{n^2} \sum_i^n E\left(\frac{x_i}{p_i} \right)^2 - \frac{1}{n^2(n-1)} \sum_{i \neq j}^n E\left(\frac{x_i}{p_i} \right) E\left(\frac{x_j}{p_j} \right)$$

$$= \frac{1}{n} E\left(\frac{x_i}{p_i} \right)^2 - \frac{n(n-1)}{n^2(n-1)} T^2$$

เนื่องจากเป็นการเลือกแบบใส่คืนดังนั้น

$$E\left(\frac{x_i}{p_i} \right)\left(\frac{x_j}{p_j} \right) = E\left(\frac{x_i}{p_i} \right) E\left(\frac{x_j}{p_j} \right) = T^2$$

ดังนั้น $$E\left[\hat{V}\left(\hat{T}\text{unequal} \right) \right] = \frac{1}{n} \left[E\left(\frac{x_i}{p_i} \right)^2 - T^2 \right]$$

แต่

$$V\left(\frac{x_i}{p_i} \right) = E\left(\frac{x_i}{p_i} - T \right)^2 = E\left[\left(\frac{x_i}{p_i} \right)^2 + T^2 - 2T\left(\frac{x_i}{p_i} \right) \right] = E\left(\frac{x_i}{p_i} \right)^2 - T^2$$

ดังนั้น $E\left[\hat{V}\left(\hat{T}unequal\right)\right] = \frac{1}{n}V\left(\frac{x_i}{p_i}\right)$

$$= \frac{1}{n}E\left(\frac{x_i}{p_i} - \hat{T}\right)^2$$

$$= \frac{1}{n}\sum_{i}^{n} p_i \left(\frac{x_i}{p_i} - T\right)^2$$

$$= V\left(\hat{T}unequal\right)$$

แสดงว่า $V\left(\hat{T}unequal\right) = \frac{1}{n}\sum_{i}^{n}\frac{1}{n-1}\left(\frac{x_i}{p_i} - \hat{T}unequal\right)^2$ เป็นตัวประมาณ ค่าที่ปราศจากอคติของ $V\left(\hat{T}unequal\right)$

4. จาก $V\left(\hat{T}unequal\right) = \frac{1}{n}\sum_{i}^{N} p_i \left(\frac{x_i}{p_i} - \hat{T}\right)^2$

เมื่อแทนที่ p_i ด้วย $\frac{x_i}{\sum_{i}^{N} x_i}$ ซึ่งเป็นค่าความน่าจะเป็นตามวิธี pps

จะพบว่า

$$V\left(\hat{T}unequal\right) = \frac{1}{n}\sum_{i}^{N}\left(\frac{x_i}{\sum_{i}^{N} x_i}\right)\left(\frac{x_i}{\frac{x_i}{\sum_{i}^{N} x_i}} - T\right)^2 = 0$$

แสดงว่าถ้าเราเลือก Selection probability ในลักษณะ pps แล้วจะมีผลให้ค่าประมาณยอดรวมคือ $\hat{T}unequal$ มีความแม่นยำสูงสุดคือ $V(\hat{T}unequal) = 0$

แผนสำรวจแบบ unequal probability sampling ไม่ว่าจะเป็น pps หรือ ppms เป็นแผนสำรวจที่เราเลือกตัวอย่างแบบใส่คืน (sampling with replacement) แปลว่าเรายอมให้หน่วยสำรวจคือ psu หรือ ssu (หรืออื่นๆ แล้วแต่ความเหมาะสมว่าพึงใช้ multistage sampling ไกลถึงขั้นไหน) มีโอกาสถูกเลือกซ้ำ กล่าวคือหน่วยสำรวจคือ psu ที่มี measure of size ใหญ่มากย่อมมีโอกาสได้รับเลือกมาก ซึ่งจาก psu นี้เมื่อเราเลือกได้เราจะเลือก ssu ออกมา เมื่อเลือกได้แล้ว ทราบแล้วและเลือก ssu แล้วให้คืน psu นี้ลงใน Collection เคล้าให้ทั่วแล้วเลือก psu หน่วยถัดไป หากหน่วยเดิมถูกเลือกมาอีก (เพราะมี measure of size สูงโอกาสถูกเลือกซ้ำย่อมมีได้มากเป็นธรรมดา) เราก็ไม่ได้

รังเกียจ ให้ทำการเลือก ssu จาก psu นี้ต่อไป และเพราะเหตุที่ psu นี้มีขนาดใหญ่เรา จึงเชื่อว่า ssu ที่ได้ในคราวหลัง จะไม่ซ้ำกับ ssu เมื่อคราวก่อน (คือชุดก่อนหน้านี้) เมื่อเลือก psu และ ssu ครบจำนวนแล้วให้เลือก tsu หรือ หน่วยสำรวจ ในลำดับขั้นสูงอื่นๆ ต่อไปตามวิธีเดียวกัน

โดยปกติแผนสำรวจแบบ unequal probability sampling นิยมใช้วิธี pps หรือ ppms กับเฉพาะ psu เท่านั้น การเลือกหน่วยสำรวจลำดับถัดไปเช่น ssu tsu หรืออื่นๆ จะใช้วิธี SRS แต่ถ้าจะใช้ pps หรือ ppms กับหน่วยสำรวจลำดับถัดไปด้วยก็ไม่เป็นปัญหา เพียงแต่โครงสร้างของสูตรจะรุงรังกว่าเดิมเท่านั้น

ทฤษฎี 6.2 ในแผนสำรวจแบบแบ่งชั้นภูมิโดยที่หน่วยสำรวจต่างๆ ในชั้นภูมิเดียวกันมีโอกาสได้รับเลือกไม่เท่ากัน เราเรียกแผนสำรวจนี้ว่า Strafified Unequal Probability Sampling จะพบว่า

$$E(\hat{T}_{str\text{-}unequal}) = T \quad \text{โดยที่} \quad \hat{T}_{str\text{-}unequal} = \sum_{h}^{n} \left[\frac{1}{n_h} \sum_{i}^{n_h} x_{hi}/p_{hi}\right]$$

$$V(\hat{T}_{str\text{-}unequal}) = \sum_{h}^{n} \frac{1}{n_h} \sum_{i}^{n} p_{hi} \left(\frac{X_{hi}}{P_{hi}} - T_h\right)^2$$

พิสูจน์ 1. ในกรณี Str ตามปกติ (คือ str-equal) เราทราบว่า

$$\begin{aligned}
\hat{T}_{str\text{-}equal} &= \sum_{h}^{L} \hat{T}_h \\
&= \sum_{h}^{L} N_h \bar{x}_h \\
&= \sum_{h}^{L} N_h \left(\frac{1}{n_h} \sum_{i}^{n_h} x_{hi}\right) \\
&= \sum_{h}^{L} \frac{1}{n_h} \left[\sum_{i}^{n_h} \frac{X_{hi}}{1/N_h}\right]
\end{aligned}$$

ถ้าเราใช้ p_{hi} แทน $^1/_{N_h}$ สูตรข้างบนนี้จะเปลี่ยนเป็นกรณีทั่วไปคือเป็นกรณีของ str-equal ถ้า p_{hi} = $^1/_{N_h}$ และจะเป็นกรณีของ str-unequal ถ้ากำหนดให้ p_{hi} มีค่าเป็นอย่างอื่น ดังนี้

$$\hat{T}\text{str-unequal} = \sum_{h}^{L} \frac{1}{n_h} \left[\sum_{i}^{n_h} \frac{x_{hi}}{P_{hi}} \right]$$

โดยที่ p_{hi} เรียกว่า Selection probability ของหน่วยสำรวจที่ i จากชั้นภูมิที่ h และพบว่า

$$E(\hat{T}\text{str-unequal}) = \sum_{h}^{L} \frac{1}{n_h} \left[\sum_{i}^{n_h} \frac{1}{P_{hi}} E(x_{hi}) \right]$$

$$= \sum_{h}^{L} \frac{1}{n_h} \left[\sum_{i}^{n_h} \frac{1}{P_{hi}} \{ \sum_{i}^{n_h} x_{hi} \cdot P_{hi} \} \right]$$

$$= \sum_{h}^{L} T_h$$

$$= T$$

แสดงว่า $\hat{T}$str-unequal เป็นตัวประมาณค่าที่ปราศจากอคติของ T

2. $$V(\hat{T}\text{str-unequal}) = V\left[\sum_{h}^{L} \frac{1}{n_h} \sum_{i}^{n_h} \frac{x_{hi}}{P_{hi}} \right]$$

$$= \sum_{h}^{L} \frac{1}{n_h} \left[\sum_{i}^{n_h} V\left(\frac{x_{hi}}{P_{hi}}\right) \right]$$

$$= \sum_{h}^{L} \frac{1}{n_h} \left[\sum_{i}^{n_h} \left\{ \sum_{h}^{N_h} \left(\frac{x_{hi}}{P_{hi}} - E\left(\frac{x_{hi}}{P_{hi}}\right) \right)^2 \cdot P_{hi} \right.\right.$$

แต่ $E\left(\frac{x_{hi}}{P_{hi}}\right) = \sum_{i}^{N_h} \frac{x_{hi}}{P_{hi}} \cdot P_{hi} = \sum_{i}^{n_h} x_{hi} = T_h$ ดังนั้น

$$V(\hat{T}\text{str-unequal}) = \sum_{h}^{L} \frac{1}{n_h} \cdot n_h \left\{ \sum_{i}^{N_h} \left(\frac{x_{hi}}{P_{hi}} - T_h \right)^2 P_{hi} \right.$$

$$= \sum_{h}^{L} \frac{1}{n_h} \sum_{i}^{N_h} P_{hi} \left(\frac{x_{hi}}{P_{hi}} - T_h \right)^2$$

ตัวประมาณค่าที่ปราศจากอคติของ V (Tstr-unequal) คือ V ($\hat{T}$str-unequal) โดยที่

$$\hat{V}(\hat{T}\text{str-unequal}) = \sum_{h}^{L} \frac{1}{n_h} \sum_{i}^{N_h} \frac{1}{n_h - 1} \left(\frac{x_{hi}}{P_{hi}} - \hat{T}_h \right)^2$$

การพิสูจน์ใช้วิธีเดียวกันกับข้อ 3 ในทฤษฎี 6.1

สำหรับการประมาณ P, $\bar{X}$ และ R ให้อาศัยผลจากการประมาณยอดรวมโดยใช้หลักเกณฑ์ที่เคยศึกษาผ่านมาแล้วคือ

$$\bar{x} = \frac{\hat{T}}{N}$$

$p = \frac{\hat{T}}{N}$ โดยที่ x_i หรือ x_{hi} มีค่าเท่ากับ 0 หรือ 1 อย่างใดอย่างหนึ่ง

$$\hat{T} = \hat{T}x / \hat{T}y = \bar{x}/\bar{y}$$

6.4 การประมาณค่าพารามิเตอร์ใน 2 CS

การเลือกตัวอย่างตามวิธี two stage unequal probability sampling เป็นการเลือกตัวอย่าง 2 ระยะ ระยะแรกเลือก psu มาเป็นตัวอย่าง m หน่วยโดยที่แต่ละหน่วยมี

โอกาสถูกเลือกเท่ากับ p_i แล้วเราจึงเลือก ssu จากแต่ละ psu ที่เลือกได้ในขั้นที่ 1 มาเป็นตัวอย่าง n_i หน่วย ตามปกติเราจะเลือก ssu โดยวิธี SRS ตัวอย่างเช่น ตำบลหนึ่งประกอบไปด้วยหมู่บ้านต่างๆ ทั้งสิ้น 10 หมู่บ้าน แต่ละหมู่บ้านมีขนาดแตกต่างกันไป คือมีจำนวนครัวเรือนมากน้อยแตกต่างกัน ให้เลือกหมู่บ้านมาเป็นตัวอย่าง 4 หมู่บ้าน จากแต่ละหมู่บ้านที่เลือกได้ให้เลือกครัวเรือนมาเป็นตัวอย่าง 5,8,6 และ 10 ครัวเรือน ตามลำดับ (ดูตอน 5,2,1) จากกรณีนี้จะเห็นว่า psu คือหมู่บ้าน ssu คือครัวเรือน เรามี psu ทั้งสิ้น 10 หน่วย ถ้าพิจารณาการเลือกตัวอย่างแบบ 2 CS ตามปกติแต่ละ psu จะมีโอกาสถูกเลือกเท่ากันเท่ากับ 1/10 แต่ถ้าพิจารณาการเลือกเป็นแบบ 2CS - unequal จะพบว่า psu หนึ่งๆ จะมีโอกาสถูกเลือกเท่ากับ p_i ซึ่งเราไม่ทราบว่าควรใช้ pi เท่าไรจึงจะถือได้ว่าเหมาะสม แต่เราสามารถพิสูจน์ได้ว่า โดยวิธี pps ถ้าเราเลือก x

$$p_i = \frac{T_i}{\sum_j^{M_i} T} = \frac{T_i}{T} \quad ; \quad i = 1,2,\ldots, M$$

แล้วจะมีผลให้ V ($\hat{T}$2cs-unequal) = 0 และเราสามารถใช้วิธี ppms แทนวิธี pps คือใช้

$$p_i = \frac{N_i}{N} \quad ; \quad i = 1,2,\ldots, M$$

ได้โดยถือว่า N_i คือ measure of size เพราะเราไม่อาจทราบค่าจริงคือ T_i และ T ได้ก่อนสำรวจ (แม้หลังสำรวจก็ไม่อาจทราบได้)

พิจารณาโครงสร้าง $\hat{T}_{2cs}$ ที่ศึกษามาแล้วในตอน 5.2.1 จะพบว่า

$$\hat{T}_{2cs} = \frac{M}{m} \sum_i^m \frac{N_i}{n_i} \sum_i^{n_i} x_{ij}$$

$$= \frac{1}{m} / \frac{1}{1/M} \sum_i^m \frac{N_i}{n_i} \sum_j^{N_i} x_{ij}$$

ถ้าเราใช้ p_i แทน 1/M จะได้สูตรทั่วไปของ $\hat{T}$2cs-unequal ดังนี้

$$\hat{T}2cs\text{-}unequal = \frac{1}{m}\sum_{i}^{m}\frac{1}{p_i}\frac{N_i}{n_i}\sum_{j}^{n_i}x_{ij}$$

ถ้าเป็น pps เราจะให้ $p_i = \frac{N_i}{\sum N_i} = \frac{N_i}{N}$

และเนื่องจาก $\hat{V}(\hat{T}equal) = M^2\frac{M-m}{m}\frac{S_b^2}{m} + \frac{M}{m}\sum_{i}^{m}\frac{N_i-n_i}{n_i}N_i^2\frac{S_i^2}{n_i}$

โดยที่ $\hat{V}(\hat{T}equal) = M^2\frac{M-m}{m}\frac{S_b^2}{m} + \frac{M}{m}\sum_{i}^{m}\frac{N_i-n_i}{n_i}N_i^2\frac{S_i^2}{n_i}$

ดังนั้น $V(\hat{T}unequal) = \frac{1}{M}\sum_{i}^{M}\frac{1}{n-1}\left\{\frac{\hat{T}_i}{P_i} - \hat{T}unequal\right\}^2$

โดยที่ $\hat{T}_i = M_i\bar{x}_i$

กรณี Separate ratio estimate เราสามารถพิสูจน์หา $V(\hat{T}x)$ ซึ่งนำไปสู่ $V(\hat{R}s)$ ได้ดังนี้

เพราะว่า $\hat{T}_x = \sum_{h}^{L}\hat{T}_{xh}$

ดังนั้น $\hat{T}_x - T_x = \sum_{h}^{L}\hat{T}_{xh} - T_x = \sum_{h}^{L}(\hat{T}_{xh}-T_{xh})$

$V(\hat{T}_x) = \sum_{h}^{L}E(\hat{T}_x-T_x)^2 = E\left[\sum_{h}^{L}(\hat{T}_{xh}-T_{xh})\right]^2$

$$= \sum_{h}^{L} E\ (\hat{T}_{xh}-T_{xh})^2 + 2\sum_{k<l}\sum\ (\hat{T}_{xk}-T_{xk})\ (\hat{T}_{xl}-T_{xl})$$

$$= \sum_{h}^{L} E\ (\hat{T}_{xh}-T_{xh})^2\ ;\ \text{cov}\ (\hat{T}_{xk}, \hat{T}_{xl}) = 0$$ เพราะชั้นภูมิ

เป็นอิสระต่อกัน

$$= \sum_{h}^{L} T_{yh}^{2} E\ (\hat{R}_h-R_h)^2\ ;\quad T_{xh} = \hat{R}_h\ T_{yh}$$

$$= \sum_{h}^{L} Nh^2\, yh^2\, V\ (\hat{R}h)$$

แต่ $V\ (\hat{R}_h) = \frac{1}{\bar{y}_h^{\,2}}\ \frac{N_h-n_h}{N_h}\ \frac{1}{n_h}\ \{ S_{xh}^{2} + R_h^{2} S_{yh}^{2} - 2R_h\rho_h S_{xh}S_{yh} \}$

$$V(\hat{T}_x) = \sum_{h}^{L} N_h^{2}\ \frac{N_h-n_h}{N_h}\ \frac{1}{n_h} \{S_{xh}^{2} + R_h^{2}\ S_{yh}^{2} - 2R_h\rho_h\ S_{xh}S_{yh}\}$$

แต่ $\hat{R}_s = \hat{T}_x/T_y$ ดังนั้น $V\ (\hat{R}_s) = \frac{1}{T_y^{\,2}} V\ (\hat{T}_x)$

ดังนั้น $V\ (\hat{R}_s) = \frac{1}{N^2\bar{y}^{2}} \sum_{h}^{L}\ \frac{N_h-n_h}{N_h}\ \frac{N_h^{2}}{N_h} \{S_{xh}^{2}+R_h^{2}\ S_{yh}^{2} - 2R_h\rho_h\ S_{xh}S_{yh}$