บทที่ 5
แผนสำรวจแบบ Cluster
Cluster Sampling Plan (CS)

5.1 เหตุผล ความจำเป็นและแนวคิดพื้นฐาน

ในกรณีของการสำรวจโครงการใหญ่ ๆ ที่เกี่ยวข้องกับประชากรขนาดใหญ่ เราจะพบปัญหาในทางปฏิบัติหลาย ๆ ประการที่แผนสำรวจทั้งแบบ SRS แบบแบ่งชั้นภูมิ และแบบ Systematic ไม่อาจแก้ไขได้ ปัญหาดังกล่าวคือ

1. ปัญหาเรื่องการเตรียมกรอบตัวอย่าง

ในกรณีที่กลุ่มประชากรที่เกี่ยวข้องเป็นกลุ่มประชากรขนาดใหญ่ปัญหาที่สำคัญยิ่งก็คือ การสร้างกรอบตัวอย่าง เช่น เมื่อกลุ่มประชากรคือครัวเรือน เราจำเป็นจะต้องจัดทำบัญชีรายชื่อหัวหน้าครัวเรือนพร้อมทั้งที่ตั้งของครัวเรือน อาจใช้เลขบ้านหรือแผนที่ตั้งบ้านอย่างใดอย่างหนึ่ง หรือทั้งสองอย่าง การกระทำดังกล่าวต้องสิ้นเปลืองค่าใช้จ่ายและเวลามาก ยกตัวอย่างเช่นการสำรวจลักษณะรายได้-รายจ่ายของประชากรในเขตกรุงเทพมหานคร ก่อนสำรวจนักวิจัยจะต้องสร้างกรอบตัวอย่างขึ้นมาก่อน โดยสำรวจบัญชีรายชื่อหัวหน้าครัวเรือนจากที่ทำการเขตต่าง ๆ พร้อมทั้งที่ตั้งของครัวเรือน พร้อมกันนี้นักวิจัยจะต้องปรับปรุงกรอบตัวอย่างให้ทันสมัยอยู่เสมอ ด้วยการออกสำรวจที่ตั้งบ้านเรือนตามที่ปรากฏในทะเบียน ทั้งนี้เพราะบางครั้งอาจมีการโยกย้าย ทั้งย้ายเข้าและย้ายออก ทำให้กรอบตัวอย่างเปลี่ยนแปลงไปโดยที่สำนักงานทะเบียนท้องถิ่นยังไม่ได้รับแจ้ง ตลอดจนครัวเรือนที่ยังไม่มีบ้านเลขที่ซึ่งปรากฏขึ้นเสมอ จะเห็นได้ว่าการสำรวจกรอบตัวอย่างและปรับปรุง

กรอบตัวอย่างเป็นกระบวนการที่ทำให้เสียทั้งเวลาและค่าใช้จ่ายไปเป็นจำนวนมากมิใช่น้อย ซึ่งในแผนสำรวจทั้ง 3 แบบที่ผ่านมาแล้วนั้นนักวิจัยจำเป็นต้องสร้างกรอบตัวอย่างเช่นนี้ขึ้นมาก่อนเสมอและจะเป็นสิ่งที่แทบจะเป็นไปได้เลยสำหรับโครงการใหญ่ ๆ เพราะเวลาที่ใช้ในการวิจัยตามโครงการอาจหมดไปกว่าครึ่งเพราะการสร้างและปรับปรุงกรอบตัวอย่าง นอกจากนี้โครงการวิจัยอาจล้มเหลวโดยสิ้นเชิงเพราะงบประมาณในการดำเนินการส่วนหนึ่งซึ่งเป็นจำนวนมิใช่น้อยต้องสูญเสียไปเพราะกระบวนการดังกล่าว

2. ปัญหาเรื่องค่าใช้จ่ายในการสำรวจ

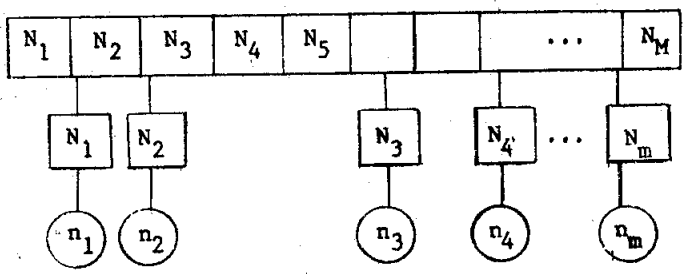
ในกรณีของโครงการใหญ่แต่ใช้แผนสำรวจแบบ SRS หรือแบบ Systematic ในการเลือกตัวอย่าง จะปรากฏปัญหาที่สำคัญคือหน่วยสำรวจกระจัดกระจายไปทั่วประชากร ซึ่งอาจกระจายไปในส่วนต่าง ๆ ของประชากรในลักษณะที่เราไม่อาจควบคุมทิศทางไว้ได้ ซึ่งในกรณีเช่นนี้การเข้าถึงหน่วยสำรวจย่อมสิ้นเปลืองมากกว่าในกรณีที่สามารถควบคุมทิศทางได้ ซึ่งส่งผลสะท้อนให้ค่าใช้จ่ายในการสำรวจสูงมากกว่าปกติ แม้แต่ในแผนสำรวจแบบแบ่งชั้นภูมิก็เช่นกัน ถ้างานสำรวจเป็นโครงการใหญ่จะมีผลให้ชั้นภูมิมีขนาดใหญ่และหน่วยสำรวจจะกระจายไปทั่วชั้นภูมิซึ่งส่งผลสะท้อนให้ค่าใช้จ่ายต่อหน่วยในชั้นภูมิผิดปกติได้เช่นกัน

3. ปัญหาเรื่องการควบคุมงานสนาม

ในกรณีที่กลุ่มตัวอย่างมีอยู่กระจัดกระจายไปทั่วกลุ่มประชากร นอกจากจะเพิ่มภาระด้านค่าใช้จ่ายในการสำรวจแล้วยังเป็นอุปสรรคในการวางแผนควบคุมงานสนามอีกด้วย เพราะถ้าหน่วยสำรวจอยู่กระจายไปทั่วห่าง ๆ กันย่อมสร้างความยุ่งยากในการติดต่อสื่อสารและการควบคุมงานสนามของผู้คุมงานหรือเจ้าของโครงการ

จากประเด็นของปัญหาทั้ง 3 ประการนี้ทำให้เกิดมีความจำเป็นต้องพัฒนาแผนสำรวจที่เหมาะกับโครงการขนาดใหญ่ขึ้นมาเรียกแผนสำรวจนี้ว่า Cluster Sampling คำว่า Cluster มีความหมายเช่นเดียวกันกับชั้นภูมิหรือกลุ่มประชากรย่อย การกำหนด Cluster กำหนดได้เช่นเดียวกับการกำหนดชั้นภูมิ หรือกล่าวโดยสรุปก็คือแผนสำรวจแบบนี้มีลักษณะทั่วไป

คล้ายแผนสำรวจแบบแบ่งชั้นภูมิ แตกต่างกันเฉพาะแผนสำรวจแบบแบ่งชั้นภูมิจะเลือกตัวอย่างจากทุกชั้นภูมิ ส่วนการสำรวจแบบ Cluster จะเลือกชั้นภูมิขึ้นมาเป็นตัวอย่างเพียงบางส่วนแล้วเลือกตัวอย่างหน่วยสำรวจจากแต่ละ Cluster ที่เป็นตัวอย่างเหล่านั้นขึ้นมาเป็นตัวอย่างอีกต่อหนึ่ง ลองพิจารณาไดอะแกรมต่อไปนี้จะเห็นว่า



ไดอะแกรมแสดง 2CS

กลุ่มประชากรขนาด N ถูกจำแนกออกเป็นส่วน ๆ ตามตัวแปรที่สนใจ M ส่วน (Cluster) แต่ละ Cluster มีขนาดแตกต่างกันไปคือ $N_1,N_2,...,N_M$ ตามลำดับ และเนื่องจากแต่ละ Cluster ย่อมมีธรรมชาติภายในต่าง ๆ แตกต่างกันไป เราจึงสุ่มตัวอย่าง Cluster เหล่านี้ไปเป็นตัวอย่างต่อไป เรียก Cluster ในชั้นนี้ว่า หน่วยสำรวจชั้นที่ 1 (Primary Sampling Unit, psu)

1. สุ่มตัวอย่าง psu มาเป็นตัวอย่าง m หน่วย

2. จากแต่ละ psu ที่สุ่มมานั้นให้สุ่มตัวอย่างหน่วยสำรวจไปเป็นตัวอย่าง $n_1,n_2,...,n_m$ หน่วยตามลำดับ หน่วยสำรวจในชั้นนี้เรียกว่า หน่วยสำรวจชั้นที่ 2 (Secondary Sampling Unit, ssu)

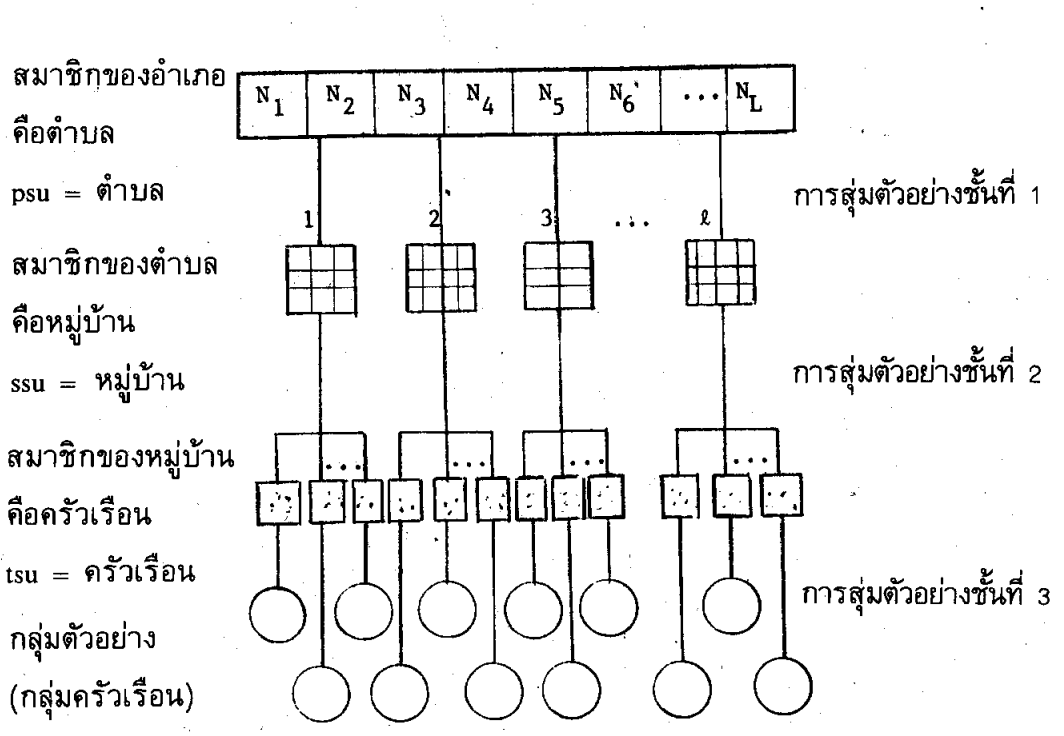
3. พิจารณาแต่ละ ssu ว่ามีขนาดใหญ่เกินกว่าจะสำรวจได้หรือไม่ ถ้าเป็นหน่วยที่อยู่ในวิสัยที่สามารถสำรวจได้ให้ดำเนินการสำรวจได้เลย แต่ถ้าเป็นหน่วยขนาดใหญ่และไม่อยู่ในวิสัยที่จะสำรวจได้ให้ดำเนินการสุ่มตัวอย่างขั้นต่อไป และให้สุ่มตัวอย่างต่อไปเรื่อย ๆ จนกว่าหน่วยสำรวจในชั้นสุดท้ายจะมีขนาดเหมาะสมและอยู่ในวิสัยที่สามารถสำรวจได้

ตัวอย่างเช่นในการสำรวจทัศนคติของประชาชนต่อโครงการวางแผนครอบครัวในเขตอำเภอโพธาราม จังหวัดราชบุรี จะพบว่าเราสามารถจำแนกประชากรในเขตสำรวจโพธารามออกเป็นส่วน ๆ ตามเขตการปกครองคือ ตำบล ดังนั้นตำบลจึงเป็น psu เลือกตำบลขึ้นมาเป็นตัวอย่าง m ตำบล แล้วเลือกครัวเรือนตัวอย่างจากแต่ละตำบลมา $n_1,n_2,\ldots,n_m$ ครัวเรือนตามลำดับ ซึ่งถ้ากระทำดังนี้เราเรียกครัวเรือนว่าหน่วยสำรวจชั้นที่ 2 แต่ขอให้พิจารณาดูว่า การสุ่มตัวอย่างครัวเรือนมาจากตำบลเป็นสิ่งที่อยู่ในวิสัยที่กระทำได้หรือไม่ เมื่อพิจารณาแล้วจะเห็นว่าตำบลเป็นกลุ่มของ ssu ที่ใหญ่เกินไป ถ้าสุ่มครัวเรือนมาจากตำบลโดยตรงจะเกิดปัญหา 3 ประการข้างต้น คือ ก. ต้องสร้างกรอบตัวอย่างของตำบลตัวอย่างทุกตำบล ข. หน่วยสำรวจกระจายทั่วพื้นที่ในตำบลทำให้เกิดผลเสียในด้านค่าใช้จ่ายในการสำรวจสูงเกินไป ค. หน่วยสำรวจกระจายทั่วไปมากทำให้ไม่อาจควบคุมงานสนามได้อย่างมีประสิทธิภาพ ด้วยเหตุนี้จึงเห็นว่าเราจำเป็นต้องสุ่มตัวอย่างมากกว่า 2 ชั้นคือแทนที่จะสุ่มครัวเรือนจากตำบลให้สุ่มหมู่บ้านจากตำบลตัวอย่างในกรณีนี้หมู่บ้านคือ ssu แล้วสุ่มครัวเรือนจากหมู่บ้านที่ตกเป็นตัวอย่าง ในที่นี้ครัวเรือนจะเป็นหน่วยสำรวจชั้นที่ 3 (third sampling unit, tsu) ซึ่งในกรณีนี้ จะเห็นได้ว่าหมู่บ้านซึ่งเป็น Cluster ของครัวเรือนมีขนาดที่เหมาะสมและอยู่ในวิสัยที่เราจะสุ่มตัวอย่างครัวเรือนขึ้นมาเป็นตัวอย่างได้โดยคาดว่าปัญหาทั้ง 3 ประการจะไม่ปรากฏขึ้น

แผนสำรวจที่ต้องสุ่มตัวอย่าง 2 ชั้นเรียกว่า two - stage cluster sampling ซึ่งจะเขียนย่อ ๆ เป็น 2 CS แผนสำรวจที่ต้องสุ่มตัวอย่าง 3 ชั้นเรียกว่า three-stage cluster sampling ซึ่งจะเขียนย่อ ๆ เป็น 3CS แผนสำรวจที่ต้องสุ่มตัวอย่าง 4 ชั้นเรียกว่า four-stage cluster sampling, 4CS แผนสำรวจที่ต้องสุ่มตัวอย่าง 5 ชั้นเรียกว่า five stage cluster sampling, 5CS ดังนี้เรื่อย ๆ ไปอาจเรียกชื่อรวม ๆ ว่า Multi-Stage Cluster Sampling โดยปกติเราจะไม่นิยมใช้เกิน 3 ชั้น ถ้าไม่จำเป็นเพราะ $V(\hat{\theta})$ จะมีค่าสูงเพิ่มมากขึ้นตามจำนวนชั้นที่สุ่มตัวอย่าง เพราะการสุ่มตัวอย่างครั้งหนึ่งย่อมก่อให้เกิดความผันแปรหรือความแปรปรวนขึ้นในระหว่างหน่วยที่เกี่ยวข้องเสมอ เช่น เมื่อสุ่มครัวเรือน ย่อมเกิดความแปรปรวนในระหว่างครัวเรือน เมื่อสุ่มหมู่บ้านย่อมเกิดความแปรปรวนระหว่างหมู่บ้าน สุ่มตำบลย่อมเกิดความแปรปรวนระหว่างตำบล เป็นต้น และความแปรปรวนทั้งหลายเหล่านี้จะสะสมเข้าเป็น

$V(\hat{\theta})$ ดังนั้นถ้าเรายิ่งวางแผนให้มีการสุ่มตัวอย่างมากชั้นขึ้นเพียงใด $V(\hat{\theta})$ ก็ย่อมมีค่าสูงมากขึ้นเพียงนั้น และเท่าที่ปรากฏในทางปฏิบัติในกรณีโครงการใหญ่จริง ๆ จะใช้แผนสำรวจเพียง 5 ชั้นคือแผนการสำรวจหลังสำมะโน โดยใช้ psu = จังหวัด ssu = อำเภอ, tsu = ตำบล, fsu = หมู่บ้าน, ffsu = ครัวเรือน แต่โดยปกติถ้าไม่จำเป็นจริง ๆ เราจะหลีกเลี่ยง 4CS และ 5CS และนิยมใช้เพียง 2CS และ 3CS เท่านั้น ในที่นี้ขอสรุปลักษณะของ 3CS ตามตัวอย่างข้างต้นดังไดอะแกรมต่อไปนี้

ไดอะแกรมของ 3CS



สำหรับรายละเอียดของแผนสำรวจแบบ 2CS 3CS หรือ MCS (Multi-Stage Cluster Sampling) จะได้กล่าวถึงโดยละเอียดในลำดับต่อ ๆ ไป ในที่นี้ผู้เขียนต้องการชี้ให้เห็นเค้าโครงหรือลักษณะทั่ว ๆ ไปของแผนสำรวจแบบ Cluster ว่าหมายถึงอะไร และสามารถจำแนกเป็นแผนย่อย ๆ ได้อย่างไรเท่านั้น

ขอให้สังเกตว่าเมื่อเราใช้แผนสำรวจแบบ Cluster แล้ว เราสามารถแก้ปัญหาสำคัญ 3 ประการข้างต้นได้ดังนี้

ก. เราจะสร้างกรอบตัวอย่างหรือปรับปรุงกรอบตัวอย่างเฉพาะใน Cluster ที่ถูกเลือกมาเป็นตัวอย่างเท่านั้น ไม่ต้องสร้างหรือปรับปรุงกรอบตัวอย่างของประชากรทั้งกลุ่ม ดังตัวอย่างเรื่อง "การสำรวจทัศนคติของประชากรในเขตอำเภอโพธารามต่อการวางแผนครอบครัว" ถ้าใช้แผน SRS หรือแผนสำรวจแบบแบ่งชั้นภูมิหรือแผนสำรวจแบบ Systematic นักวิจัยต้องสำรวจรายชื่อหัวหน้าครัวเรือนและที่ตั้งทุกครัวเรือนในท้องที่อำเภอโพธาราม แต่ถ้าใช้แผนสำรวจแบบ 2 CS นักวิจัยจะสร้างและปรับปรุงกรอบตัวอย่างเฉพาะตำบลตัวอย่างเท่านั้น และเมื่อใช้แผนสำรวจแบบ 3 CS นักวิจัยจะสร้างและปรับปรุงกรอบตัวอย่างเฉพาะหมู่บ้านตัวอย่างเท่านั้น จึงเห็นได้ว่า CS ช่วยประหยัดแรงงานและงบประมาณในส่วนที่เกี่ยวข้องกับการสร้างกรอบตัวอย่างไปได้มาก

ข. การใช้แผนสำรวจแบบ CS ทำให้เราสามารถควบคุมหน่วยสำรวจได้ว่าจะตกอยู่ในส่วนใดของประชากร ในตำบลใด ในหมู่บ้านใด เมื่อเป็นเช่นนี้หน่วยสำรวจจึงกระจายไปเฉพาะใน Cluster ที่เป็นตัวอย่างเท่านั้นไม่กระจายแผ่กว้างไปทั่วกลุ่มประชากรซึ่งทำให้การควบคุมงานสนามเป็นไปได้อย่างมีประสิทธิภาพ เพราะผู้ควบคุมงานสนามสามารถกำหนดการปฏิบัติงานสนามและควบคุมให้ทำงานเสร็จไปคราวละ Cluster ก่อนที่จะย้ายไปสู่ Cluster ใหม่ ตามตัวอย่าง ผู้ควบคุมงานสนามสามารถควบคุมงานโดยกำหนดให้พนักงานสำรวจเข้าทำการสำรวจคราวละหมู่บ้าน หรือเสร็จเป็นหมู่บ้านไปแล้วจึงย้ายไปสำรวจหมู่บ้านอื่น ๆ วิธีนี้นอกจากง่ายต่อการควบคุมงานสนามแล้วยังประหยัดค่าใช้จ่ายในการสำรวจเพราะหน่วยสำรวจกระจายอยู่ในกลุ่มที่เราสามารถเข้าถึงได้พร้อมกัน ผู้คุมงานสนามไม่ต้อง "วิ่งรอก" รับส่งพนักงานสำรวจคราวเดียวกันหลายหมู่บ้าน

สำหรับแผนสำรวจแบบ CS นี้การสุ่มตัวอย่างในแต่ละขั้น เราอาจใช้แผนสำรวจแบบใดก็ได้ อาจใช้ SRS Systematic หรือแบบแบ่งชั้นภูมิก็ได้[1] สูตรที่ใช้วิเคราะห์จะแตกต่างกันไปซึ่งเราสามารถพัฒนาสูตรได้เองโดยไม่ยากนัก ในที่นี้จะกล่าวถึงเฉพาะกรณีที่สุ่มตัวอย่างโดยใช้แผนสำรวจแบบ SRS แก่แต่ละชั้นภูมิเท่านั้น ถ้านักศึกษามีความรู้ความเข้าใจได้ดี การพัฒนาสูตรสำหรับกรณีที่ใช้แผนสำรวจแบบอื่น ๆ ก็มิใช่เรื่องยากอีกต่อไป

การศึกษาในลำดับต่อไปนี้จะเริ่มศึกษาเป็นลำดับ ๆ ไปเริ่มตั้งแต่ 2CS 3CS และ pps (Probability Proportional to Size) อาจเพิ่มนแผนการสำรวจแบบ Stratified Cluster Sampling ในลักษณะการแนะนำไว้ท้ายบท

5.2.2 CS^2

5.2.1 นิยามและสัญลักษณ์

M = จำนวน psu หรือจำนวน cluster

$N_i; i = 1,2,...,M$ คือจำนวนหน่วยสำรวจ (หรือ ssu) ใน Cluster ที่ i (หรือ psu ที่ i)

---

1. ถ้าใช้แผนสำรวจแบบ SRS สุ่มตัวอย่างในแต่ละชั้น เรียกแผนสำรวจนั้นว่า Simple Cluster Sampling ถ้าใช้แผนสำรวจแบบแบ่งชั้นภูมิสุ่มตัวอย่างในแต่ละชั้นเรียกแผนสำรวจนั้นว่า Stratified Cluster Sampling และถ้าใช้แผนสำรวจแบบ Systematic สุ่มตัวอย่างในแต่ละชั้นเรียกแผนสำรวจนั้นว่า Systematic Cluster Sampling เทคนิคของการพัฒนาสูตรก็คือ วิเคราะห์ข้อมูลตามวิธีเฉพาะแผนแล้วนำมาประกอบกันตามวิธีของ CS นอกจากวิธีดังกล่าวแล้ว เราอาจผสมผสานแผนสำรวจพื้นฐานต่าง ๆ เหล่านี้เข้าด้วยกันได้อีก ซึ่งเทคนิคการพัฒนาสูตรก็คงเป็นไปในลักษณะเดิมคือใช้สูตรเฉพาะแผนแล้วรวมกันตามวิธี CS เช่น ชั้นที่ 1 สุ่มตำบลมาโดยวิธี SRS ชั้นที่ 2 สุ่มหมู่บ้านมาโดยวิธี SRS ชั้นที่ 3 สุ่มครัวเรือนมาโดยวิธี Systematic หรือ ชั้นที่ 1 สุ่มตำบลโดยวิธี SRS ชั้นที่ 2 สุ่มครัวเรือนจากทุกหมู่บ้านในตำบลตัวอย่างโดยวิธีแบ่งชั้นภูมิ เป็นต้น

2. โดยทั่วไปนิยมเรียกว่า Single Stage Cluster Sampling หรือ Subsampling แต่ผู้เขียนใช้ 2CS เพราะถือว่ามีการสุ่มตัวอย่างคราวหนึ่งก็ถือว่าเป็น stage หนึ่ง กล่าวคือ เมื่อสุ่ม psu ถือว่าเป็น 1^{st} stage สุ่ม ssu ถือว่าเป็น 2^{nd} stage

n = จำนวน Cluster ที่ได้รับเลือกเป็นตัวอย่างในการสุ่มชั้นที่ 1

n_i; $i = 1,2,\ldots,m$ คือจำนวนหน่วยสำรวจที่ได้รับเลือกเป็นตัวอย่างมาจาก Cluster ตัวอย่าง (Sampled Cluster) ที่ i

N = จำนวนหน่วยสำรวจทั้งสิ้นในกลุ่มประชากรที่เราสนใจโดยที่ $\sum_{i}^{M} N_i = N$

n = จำนวนหน่วยสำรวจในกลุ่มตัวอย่าง โดยที่ $\sum_{i}^{m} n_i = n$

T_i; $i = 1,2,\ldots,m$ = คือยอดรวมของ Characteristic ที่เราสนใจใน Cluster ที่ i

$\hat{T}_i$; $i = 1,2,\ldots,m$ คือค่าประมาณของยอดรวมของ Characteristic ที่สนใจใน Cluster ที่ i

T = $\sum_{i}^{M} T_i$ = ยอดรวมของ Characteristic ที่สนใจในกลุ่มประชากร

$\hat{T}$ = ค่าประมาณของ T

x_{ij} = ค่าของตัวแปร x หน่วยที่ j ที่สุ่มมาจาก Cluster ที่ i ; $j = 1,2,\ldots,n_i$, $i = 1,2,\ldots,m$

$\bar{x}_i$ = ค่าเฉลี่ยต่อหน่วยสำรวจชั้นที่ 2 ของ Cluster ที่ i โดยที่

$$\bar{\bar{x}}_i = \frac{1}{n_i} \sum_{j}^{n_i} x_{ij} \; ; \; i = 1,2,\ldots,m$$

$\bar{\bar{x}}$ = ค่าเฉลี่ยต่อหน่วยชั้นที่ 2 หรือค่าเฉลี่ยต่อหน่วยสำรวจ โดยที่

$$\bar{\bar{x}} = \frac{1}{n} \sum_{i}^{m} \sum_{j}^{n_i} x_{ij}$$

$\bar{\bar{X}}_i$ = $\frac{1}{N_i} \sum_{j}^{N_i} x_{ij}$ = ค่าเฉลี่ยจริงต่อหน่วยสำรวจขั้นที่ 2 ของ Cluster ที่ i

$\bar{\bar{X}}$ = $\frac{1}{N} \sum_{i}^{M} \sum_{j}^{N_i} x_{ij}$ = ค่าเฉลี่ยจริงต่อหน่วยสำรวจที่ 2 (ssu)

S_i^2 = ความแปรปรวนภายใน Cluster ที่ i โดยที่ $S_i^2 = \frac{1}{N_i - 1} \sum_{j}^{N_i} (x_{ij} - \bar{\bar{X}}_i)^2$

s_i^2 = ความแปรปรวนภายในกลุ่มตัวอย่าง (หรือระหว่าง ssu) ที่ได้มาจาก Cluster ที่ i โดยที่ $s_i^2 = \frac{1}{n_i - 1} \sum_{j}^{n_i} (x_{ij} - \bar{x}_i)^2$

S_b^2 = ความแปรปรวนระหว่าง psu โดยที่ $S_b^2 = \frac{1}{M-1} \sum_{i}^{M} (T_i - \bar{T})^2$

เมื่อ $\bar{T}$ คือค่ายอดรวมเฉลี่ยต่อ Cluster โดยที่ $\bar{T} = \frac{1}{M} \sum_{i}^{M} T_i$

s_b^2 = ความแปรปรวนระหว่าง Cluster (psu) ตัวอย่าง โดยที่

$s_b^2 = \frac{1}{m-1} \sum_{i}^{m} (\hat{T}_i - \hat{\bar{T}})^2$ เมื่อ $\hat{\bar{T}} = \frac{1}{m} \sum_{i}^{m} \hat{T}_i$

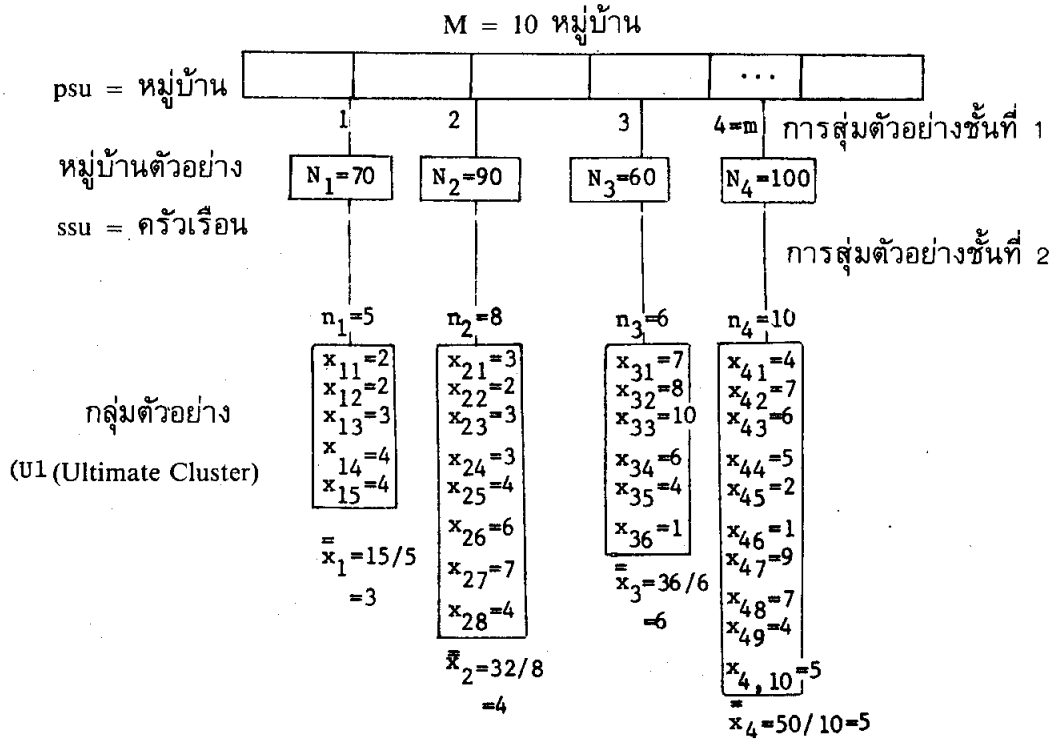
เพื่อให้เข้าใจนิยามและสัญลักษณ์ข้างต้น ขอให้นักศึกษาพิจารณาเปรียบเทียบจากตัวอย่างต่อไปนี้ (ข้อมูลสมมุติ)

ในการสำรวจภาวะเศรษฐกิจสังคมของท้องถิ่นตำบลหนึ่งซึ่งประกอบด้วยราษฎรทั้งสิ้น 500 ครัวเรือน ซึ่งประกอบด้วยหมู่บ้านต่าง ๆ 10 หมู่บ้าน

ในการสำรวจชั้นที่ 1 ใช้ psu = หมู่บ้าน เลือกหมู่บ้านมาเป็นตัวอย่าง 4 หมู่บ้าน ซึ่งแต่ละหมู่บ้านประกอบด้วยครัวเรือนราษฎรทั้งสิ้น 70,90,60 และ 100 ครัวเรือนตามลำดับ

ในการสำรวจชั้นที่ 2 ใช้ ssu = ครัวเรือน สุ่มครัวเรือนจากแต่ละหมู่บ้านตัวอย่างมา 5,8,6 และ 10 ครัวเรือนตามลำดับ และปรากฏจำนวนสมาชิกในครัวเรือนที่มีอายุต่ำกว่า 20 ปี จากหมู่บ้านตัวอย่างเท่ากับ (2,2,3,4,4), (3,2,3,3,4,6,7,4), (7,8,10,6,4,1) และ (4,7,6,5,2,1,9,7,4,5) ตามลำดับ

จากข้อมูลข้อสนเทศข้างต้น เราสามารถสรุปแผนสำรวจ 2CS นี้ได้ดังไดอะแกรมต่อไปนี้



จากไดอะแกรมจะพบว่า

M = จำนวน psu = 10 หมู่บ้าน

m = จำนวน psu ตัวอย่าง = 4 หมู่บ้าน

N_1 = 70 = จำนวน ssu ในหมู่บ้านตัวอย่างที่ 1, N_2 = 90 = จำนวน ssu ในหมู่บ้านตัวอย่างที่ 2, N_3 = 60 = จำนวน ssu ในหมู่บ้านตัวอย่างที่ 3 และ $N_4 = N_m$ = 100 = จำนวน ssu ในหมู่บ้านตัวอย่างที่ m = 4

n_i = จำนวนครัวเรือนในกลุ่มตัวอย่างที่สุ่มมาจากหมู่บ้านที่ i; i = 1,2,...,m ในที่นี้ $n_1 = 5, n_2 = 8, n_3 = 6, n_4 = 10$

x_{ij} = ค่าของตัวแปรสุ่มหน่วยที่ j ที่สุ่มจาก psu ที่ i เช่น $x_{35} = 4$ หมายความว่า จำนวนบุคคลที่มีอายุต่ำกว่า 20 ปีในครัวเรือนที่ 5 ของหมู่บ้านตัวอย่างที่ 3 มีอยู่ทั้งสิ้น 4 คน

$\bar{x}_i = \frac{1}{n_i} \sum_{j}^{n_i} x_{ij}$ จำนวนบุคคลที่มีอายุต่ำกว่า 20 ปีเฉลี่ยต่อ 1 ครัวเรือนของหมู่บ้านที่ i ซึ่งเป็นค่าประมาณที่ปราศจากอคติของ $\bar{\bar{X}}_i = \frac{1}{N_i} \sum_{j}^{N_i} x_{ij}$

เช่น $\bar{x}_1 = \frac{1}{n_1} \sum_{j}^{n_1} x_{1j} = \frac{1}{5}(2+2+3+4+4) = \frac{15}{5} = 3$ แสดงว่าครัวเรือนหมู่บ้านตัวอย่างที่ 1 มีจำนวนบุคคลที่มีอายุต่ำกว่า 20 ปีถัวเฉลี่ยครัวเรือนละ 3 คน ค่า $\bar{x}_1 = 3$ เป็นค่าประมาณของ $\bar{\bar{X}}_1$ ซึ่งเป็นค่าจริงของจำนวนบุคคลที่มีอายุต่ำกว่า 20 ปีต่อครัวเรือนของหมู่บ้านตัวอย่างที่ 1

$\hat{T}_i = N_i \bar{x}_i; i = 1,2,...,m$ คือค่าประมาณของยอดรวมจำนวนบุคคลที่มีอายุต่ำกว่า 20 ปีของหมู่บ้านที่ i เช่น $\hat{T}_4 = N_4 \bar{x}_4 = 100 \times 5 = 500$ แสดงว่าในหมู่บ้านที่ 4 จะมีบุคคลที่มีอายุต่ำกว่า 20 ปี ประมาณ 500 คน

$\hat{\bar{T}} = \sum_{i}^{m} \frac{\hat{T}_i}{m}$ = จำนวนบุคคลที่มีอายุต่ำกว่า 20 ปี เฉลี่ยต่อหมู่บ้าน ในที่นี้

$$\hat{\bar{T}} = \frac{1}{4}(\hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \hat{T}_4) = \frac{1}{4}\{(70 \times 3) + (90 \times 4) + (60 \times 6) + (100 \times 5)\}$$

$$= \frac{1}{4}(210 + 360 + 360 + 500) = \frac{1,430}{4} = 357.5 \simeq 358$$

$\hat{\bar{T}} \simeq 358$ แสดงว่าโดยถัวเฉลี่ยแล้ว หมู่บ้านหนึ่ง ๆ จะมีบุคคลที่มีอายุต่ำกว่า 20 ปี ประมาณ 358 คน

$\hat{\bar{T}}$ เป็นค่าประมาณของ $\bar{T} = \frac{1}{M} \sum_{i}^{M} T_i$

$s_i^2 = \frac{1}{n_i - 1} \sum_{j}^{n_i} (x_{ij} - \bar{x}_i)^2$ = ความแปรปรวนภายในกลุ่มตัวอย่างที่สุ่มจากหมู่บ้านที่ i; i = 1,2,...,m เช่น

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{j}^{n_1} (x_{1j} - \bar{x}_1)^2$$

$$= \frac{1}{5-1} \left\{ \sum_{j}^{5} x_{1j}^2 - \left(\sum_{j}^{5} x_{1j} \right)^2 / 5 \right\}$$

$$= \frac{1}{4} \left\{ (2^2 + 2^2 + 3^2 + 4^2 + 4^2) - 15^2/5 \right\}$$

$$= \frac{1}{4} (49 - 45)$$

$$= 1$$

s_i^2 เป็นค่าประมาณของ $S_i^2 = \frac{1}{N_i - 1} \sum_{j}^{N_i} (x_{ij} - \bar{X}_i)^2$ โดยที่ $\bar{X}_i = \frac{1}{N_i} \sum_{j}^{N_i} x_{ij}$

$s_b^2 = \frac{1}{m-1} \sum_{i}^{m} (\hat{T}_i - \bar{\hat{T}})^2$ = ความแปรปรวนในระหว่าง psu ตัวอย่าง ในที่นี้คือความแปรปรวนในจำนวนบุคคลที่มีอายุต่ำกว่า 20 ปี ในระหว่างหมู่บ้านตัวอย่าง

$$s_b^2 = \frac{1}{4-1} \left\{ \sum_{i}^{4} \hat{T}_i^2 - \left(\sum_{i}^{4} \hat{T}_i \right)^2 / 4 \right\}$$

$$= \frac{1}{3} \left\{ (210^2 + 360^2 + 360^2 + 500^2) - \frac{1{,}430^2}{4} \right\}$$

$$= \frac{1}{3} (553{,}300 - 511{,}225)$$

$$= 14{,}025$$

s_b^2 เป็นค่าประมาณของ $S_b^2 = \frac{1}{M-1} \sum_{i}^{M} (T_i - \bar{T})^2$ โดยที่ $T_i = \sum_{j}^{N_i} x_{ij}$

และ $\bar{T} = \frac{1}{M} \sum_{i}^{M} T_i$

5.2.2 การประมาณค่ายอดรวมประชากร T

เพื่อความสะดวกและง่ายต่อการศึกษาต่อเนื่องเราควรจะเริ่มศึกษาวิธีประมาณค่าพารามิเตอร์ด้วยการประมาณค่ายอดรวมประชากร (T) เสียก่อน เพราะเมื่อเราประมาณค่ายอดรวมประชากรได้แล้วการประมาณค่าเฉลี่ยประชากร ($\bar{\bar{X}}$) ย่อมกระทำได้โดยง่ายเพียงแต่นำขนาดประชากร (N) ไปหารค่าประมาณของยอดรวมก็จะได้ค่าประมาณของค่าเฉลี่ยประชากรตามต้องการ และด้วยเหตุที่สัดส่วนประชากร (P) คือกรณีเฉพาะของ $\bar{\bar{X}}$ [1] เมื่อ $x_{ij} = 0,1$ ดังนั้น เมื่อเราสามารถประมาณค่าเฉลี่ยประชากรได้ การประมาณค่าสัดส่วน-P ก็จะเป็นเพียงกรณีเฉพาะที่สามารถพัฒนาขึ้นมาได้โดยง่าย

เทคนิคการ ประมวลผลให้ดำเนินการวิเคราะห์ในลักษณะ "ล่างขึ้นบน" กล่าวคือ

(1) ให้เริ่มต้นด้วยการคำนวณหา $\bar{x}_i;\ i=1,2,\ldots,m$

(2) นำผลลัพธ์จากขั้นที่ (1) มาคำนวณหา $\hat{T}_i = N_i\bar{x}_i;\ i=1,2,\ldots,m$

(3) นำผลลัพธ์ในขั้นที่ (2) มาคำนวณหาค่าเฉลี่ยของยอดรวมต่อ psu คือ

$$\hat{\bar{T}} = \frac{1}{m}\sum_{i}^{m}\hat{T}_i$$

(4) คำนวณหาค่าประมาณของยอดรวมประชากร $\hat{T}$ โดยที่ $\hat{T} = M\hat{\bar{T}}$

(5) ถ้าต้องการประมาณค่าเฉลี่ยของประชากรต่อ ssu คือ $\bar{\bar{X}}$ ก็กระทำได้โดยง่ายเพียงแต่หาร $\hat{T}$ ด้วย N คือ $\hat{\bar{\bar{X}}} = \frac{\hat{T}}{N}$

ทฤษฎี 5.1 เมื่อดำเนินการสำรวจแบบ 2CS โดยที่การสุ่มตัวอย่าง psu และ ssu ใช้แผนสำรวจแบบ SRS จะพบว่าค่าประมาณของยอดรวม (T) สามารถคำนวณได้ดังนี้

1. $\hat{T} = \frac{M}{m}\sum_{i}^{m}\frac{N_i}{n_i}\sum_{j}^{n_i}x_{ij}$ หรือ $\hat{T} = M\hat{\bar{T}}$

---

[1] เครื่องหมาย " = " ที่เขียนเหนือ X แสดงว่า (1) $\bar{\bar{X}}$ = ค่าเฉลี่ยต่อ ssu (2) แผนสำรวจนี้เป็นแผนสำรวจแบบ 2 ลำดับขั้น (Two Stage)

2. ช่วงเชื่อมั่น $(1-\alpha)100\%$ ที่คาดว่าค่าจริงของ T จะปรากฏอยู่คือ

$$\hat{T} + Z_{\alpha/2}\sqrt{V(\hat{T})} \leq T \leq \hat{T} + Z_{1-\alpha/2}\sqrt{V(\hat{T})}$$

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

เมื่อ $S_b^2 = \frac{1}{M-1}\sum_{i}^{M}(T_i-\bar{T})^2$ และ $S_i^2 = \frac{1}{N_i-1}\sum_{j}^{N_i}(x_{ij}-\bar{\bar{X}}_i)^2$

และค่าประมาณของ $V(\hat{T})$ คือ $\hat{V}(\hat{T})$ โดยที่[1]

$$\hat{V}(\hat{T}) = M^2 \frac{M-m}{M}\frac{s_b^2}{m} + \frac{M}{m}\sum_{i}^{m} N_i^2 \frac{N_i-n_i}{N_i}\frac{s_i^2}{n_i}$$

เมื่อ $s_b^2 = \frac{1}{m-1}\sum_{i}^{m}(\hat{T}_i-\bar{\hat{T}})^2$, $s_i^2 = \frac{1}{n_i-1}\sum_{j}^{n_i}(x_{ij}-\bar{\bar{x}}_i)^2$

ข้อแนะนำ การพิสูจน์ทฤษฎี 5.1 นี้ค่อนข้างยุ่งยากโดยเฉพาะในส่วนของ $V(\hat{T})$ ดังนั้น นักศึกษาอาจผ่านการพิสูจน์ $V(\hat{T})$ ไปก่อนแล้วค่อยย้อนมาดูการพิสูจน์ในภายหลังก็ได้แต่ไม่ควรผ่านวิธีพิสูจน์สำหรับ $\hat{T}$ ไปเพราะเป็นสิ่งที่จำเป็นต่อการเข้าใจ 3CS, 4CS และอื่น ๆ

อย่างไรก็ตาม ขอให้ข้อสังเกตเกี่ยวกับ $V(\hat{T})$ ไว้หลายประการดังนี้

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

จะเห็นว่าโครงสร้างของ $V(\hat{T})$ สามารถจำแนกออกได้เป็น 2 ส่วนคือ $M^2\frac{M-m}{M}\frac{S_b^2}{m}$ และ $\frac{M}{m}\sum_{i}^{M}\frac{N_i-n_i}{N_i}N_i^2\frac{S_i^2}{n_i}$ ในส่วนแรก $M^2\frac{M-m}{M}\frac{S_b^2}{m}$ คือความแปรปรวนที่เกิดจากการสุ่มตัวอย่างชั้นที่ 1 หรือความแปรปรวนอันเนื่องมาจากการสุ่มตัวอย่างของ psu ขอให้สังเกตว่า ความแปรปรวนในส่วนนี้มีโครงสร้างคล้ายโครงสร้างสูตรความแปรปรวน

[1] $\hat{V}(\hat{T})$ เป็นค่าประมาณที่ปราศจากอคติของ $V(\hat{T})$ แม้ว่า $E(s_b^2) \neq S_b^2$ ซึ่งความจริงข้อนี้จะได้พิสูจน์ให้เห็นในลำดับต่อไป

$V(\bar{x}_{ran})$ ในแผนสำรวจแบบ SRS ทั้งนี้เพราะเราเลือก psu มาโดยวิธี SRS เพียง m หน่วยจาก psu ทั้งสิ้น M หน่วย ส่วนที่แตกต่างกันก็คือในที่นี้มี M^2 คูณอยู่กับ $\frac{M-m}{M}\ \frac{S_b^2}{m}$ สำหรับในส่วนที่ 2 คือ $\frac{M}{m}\ \sum_{i}^{M}\ \frac{N_i-n_i}{N_i}\ N_i^2\ \frac{S_i^2}{n_i}$ คือความแปรปรวนระหว่าง ssu ที่สุ่มมาจาก psu ต่าง ๆ หรือความแปรปรวนอันเนื่องมาจากการสุ่มตัวอย่างขั้นที่ 2 ขอให้สังเกตว่าโครงสร้างของความแปรปรวนในส่วนนี้คล้ายคลึงกับโครงสร้างของความแปรปรวน $V(\bar{x}_{st})$ ในแผนสำรวจแบบแบ่งชั้นภูมิ ที่เป็นเช่นนี้เพราะในทุก ๆ psu ที่เป็นตัวอย่างเราจะสุ่ม ssu มาเป็นตัวอย่างโดยไม่เว้น psu ใด ส่วนที่แตกต่างไปจากโครงสร้างของ $V(\bar{x}_{st})$ คือความแปรปรวนส่วนนี้มี $\frac{M}{m}$ คูณอยู่

2. ถ้าเป็นแผนสำรวจแบบ 3CS สูตร $V(\hat{T})$ จะมีความแปรปรวนอื่นสมทบเข้าไปอีก 1 ส่วน รวมเป็น 3 ส่วน ถ้าเป็นแผนสำรวจแบบ 4CS สูตร $V(\hat{T})$ จะมีความแปรปรวนอื่นสมทบเข้าไปอีก 2 ส่วน รวมเป็น 4 ส่วน ดังนี้เรื่อย ๆ ไป ดังนั้น ถ้าเราใช้แผนสำรวจที่มีการสำรวจมากขั้นขึ้นเพียงใด $V(\hat{\theta})$ ก็จะมีค่าสูงมากขึ้นเพียงนั้นและผลสะท้อนที่สำคัญก็คือการประมาณค่าพารามิเตอร์ขาดความแม่นยำ ดังนั้น ถ้าไม่มีความจำเป็นจริง ๆ แล้วนักวิจัยไม่ควรวางแผนสำรวจที่ต้องทำการสำรวจมากเกินกว่า 2 หรือ 3 ขั้น

3. ถ้าสุ่มตัวอย่าง ssu มาจากทุก psu แผนสำรวจแบบ 2CS จะกลายเป็นแผนสำรวจแบบแบ่งชั้นภูมิ กล่าวคือ ถ้าสุ่มตัวอย่าง ssu มาจากทุก psu ก็ย่อมหมายความว่า $m=M$ ดังนั้น $M-m=0$

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

$$= \sum_{i}^{M}\ \frac{N_i-n_i}{N_i}\ \frac{N_i^2 S_i^2}{n_i}$$ [1]

กล่าวได้อีกนัยหนึ่งก็คือแผนสำรวจแบบแบ่งชั้นภูมิก็คือแผนสำรวจที่เป็นกรณีเฉพาะของแผนสำรวจแบบ 2CS เมื่อ $M=m$

---

[1] ในแผนสำรวจแบบแบ่งชั้นภูมิ $V(\hat{T}) = \sum_{i}^{L}\ \frac{N_h-n_h}{N_h}\ \frac{N_h^2 S_h^2}{n_h}$ L = จำนวนชั้นภูมิ

4. ถ้าสำรวจทุก ssu ใน psu ตัวอย่างแผนสำรวจแบบ 2CS จะกลายเป็นแผนสำรวจแบบ SRS กล่าวคือ ถ้าสำรวจทุก ๆ ssu ใน psu ตัวอย่าง แสดงว่า $n_i = N_i$ หรือ $N_i - n_i = N_i - N_i = 0$

ดังนั้น $V(\hat{T}) = M^2 \frac{M-m}{M} \frac{S_b^2}{m}$

5. ถ้าดำเนินการสำรวจทุก ssu จากทุก psu กล่าวคือ $n_i = N_i$; $i = 1,2,...,m$ และ $M = m$ แล้วจะพบว่า $V(\hat{T}) = 0$ หรือไม่มี Sampling error เพราะการกระทำเช่นนี้คือการสำมะโน ความแปรปรวนในการเลือกตัวอย่างย่อมไม่ปรากฏ แต่ทั้งนี้มิได้หมายความว่าถ้าใช้วิธีสำมะโนแล้วงานจะมีความถูกต้อง 100% ทั้งนี้เพราะความผิดพลาดอาจเกิดจากสาเหตุอื่นเช่น การสำรวจซ้ำซ้อน การประมวลผล (Data Processing) มีข้อบกพร่องในบางขั้นตอน เป็นต้น

6. ถ้า $M >> m$ และ $N_i >> n_i$ ย่อมมีผลให้ $\frac{m}{M}$ และ $\frac{n_i}{N_i}$ มีค่าน้อยมากซึ่งเพื่อความสะดวกในการวิเคราะห์เราจะถือว่า $\frac{m}{M} \simeq 0$ และ $\frac{n_i}{N_i} \simeq 0$

ดังนั้น $\frac{M-m}{M} = 1 - \frac{m}{M} \simeq 1$ และ $\frac{N_i - n_i}{N_i} = 1 - \frac{n_i}{N_i} \simeq 1$

$$=> V(\hat{T}) \simeq M^2 \frac{S_b^2}{m} + \frac{M}{m} \sum_i^M \frac{N_i^2 S_i^2}{n_i}$$

และ $$\hat{V}(\hat{T}) \simeq M^2 \frac{s_b^2}{m} + \frac{M}{m} \sum_i^m \frac{N_i^2 s_i^2}{n_i}$$

7. ถ้า $M >> m$ เราสามารถจัดรูป $\hat{V}(\hat{T})$ ในข้อ 6 ให้เป็นรูปที่ง่ายยิ่งขึ้นได้ดังนี้

$$V(\hat{T}) = M^2 \frac{S_b^2}{m} + \frac{M}{m} \sum_i^m N_i^2 \frac{S_i^2}{n_i}$$

$$= \left(\frac{M}{m}\right)^2 \left(mS_b^2 + \frac{m}{M} \sum_i^m \frac{N_i^2 S_i^2}{n_i}\right)$$

$\because M >> m = \frac{m}{M} \to 0$ ดังนั้น

$$V(\hat{T}) \simeq (\frac{M}{m})^2 m S_b^2 = M^2 \frac{S_b^2}{m}$$

และในทำนองเดียวกัน

$$\hat{V}(\hat{T}) \simeq (\frac{M}{m})^2 m s_b^2 = M^2 \frac{s_b^2}{m}$$

กรณีเช่นนี้ควรนำมาใช้เมื่อ $\frac{m}{M} < .01$ หรือ $\frac{m}{M} < 1\%$

8. ในหลายกรณีเรามักจะพบว่า Cluster ต่าง ๆ มีขนาดใกล้เคียงกัน หรือนัยหนึ่ง psu ต่าง ๆ ประกอบไปด้วย ssu ในจำนวนพอ ๆ กัน เช่น ชั้นเรียนหนึ่ง ๆ (psu) ประกอบด้วยนักเรียน (ssu) ประมาณ 45 คน หมู่บ้านหนึ่ง ๆ (psu) ประกอบไปด้วยครัวเรือน (ssu) ใกล้เคียงกันเป็นต้น ซึ่งในกรณีเช่นนี้เราจะถือได้ว่า $N_i = \frac{1}{M} \sum_i^M N_i = \frac{N}{M} = \bar{N}$

ดังนั้น ถ้า $M >> m$ $N_i >> n_i$ และ $N_i = \bar{N}$

$$V(\hat{T}) = M^2 \frac{S_b^2}{m}$$

$$= M^2 \cdot \frac{1}{m} \cdot \frac{1}{m-1} \sum_i^m (\hat{T}_i - \hat{\bar{T}})^2$$

$$= M^2 \cdot \frac{1}{m} \cdot \frac{1}{m-1} \sum_i^m (\hat{T}_i - \frac{1}{m} \sum_i^m \hat{T}_i)^2$$

$$= M^2 \cdot \frac{1}{m} \cdot \frac{1}{m-1} (\frac{\bar{N}}{\bar{N}})^2 \sum_i^m (\hat{T}_i - \frac{1}{m} \sum_i^m \hat{T}_i)^2$$

$$= (M\bar{N})^2 \cdot \frac{1}{m} \frac{1}{m-1} \sum_i^m (\frac{\hat{T}_i}{\bar{N}} - \frac{1}{m} \sum_i^m \frac{1}{\bar{N}} \hat{T}_i)^2$$

$$= (M\bar{N})^2 \cdot \frac{1}{m} \frac{1}{m-1} \sum_i^m (\hat{\bar{X}}_i - \frac{1}{m} \sum_i^m \hat{\bar{X}}_i)^2$$

$$V(\hat{T}) = (M\bar{N})^2 \cdot \frac{1}{m} \frac{1}{m-1} \sum_i^m (\hat{\bar{X}}_i - \hat{\bar{X}})^2$$

นอกจากนี้ยังพบว่า

$$V(\hat{\bar{T}}/M\bar{N}) = \frac{1}{m}\frac{1}{m-1}\sum_{i}^{m}(\hat{\bar{X}}_i - \hat{\bar{\bar{X}}})^2$$

$$\Rightarrow V(\hat{\bar{X}}) = \frac{1}{m}\frac{1}{m-1}\sum_{i}^{m}(\hat{\bar{X}}_i - \hat{\bar{\bar{X}}})^2$$

พิสูจน์

1. ที่มาของสูตรสำหรับประมาณค่ายอดรวมพัฒนาขึ้นมาได้ได้ดังนี้

$\because \hat{\bar{x}}_i = \frac{1}{n_i}\sum_{j}^{n_i} x_{ij}$; i = 1,2,...,m คือค่าเฉลี่ยต่อ ssu ใน psu ตัวอย่างที่ i

ดังนั้น $\hat{T}_i = N_i\bar{x}_i$; i = 1,2,...,m คือค่าประมาณยอดรวมของ characteristic ใน psu ตัวอย่างที่ i

$\Rightarrow$ ค่าเฉลี่ยของ characteristic ต่อ psu คือ

$$\hat{\bar{T}} = \frac{1}{m}\sum_{i}^{m}\hat{T}_i = \frac{1}{m}\sum_{i}^{m}N_i\bar{x}_i = \frac{1}{m}\sum_{i}^{m}N_i\cdot\frac{1}{n_i}\sum_{j}^{n_i}x_{ij}$$

ดังนั้นยอดรวมของ characteristic ในกลุ่มประชากรที่ประกอบไปด้วย M psu คือ

$$\hat{T} = M\hat{\bar{T}} = \frac{M}{m}\sum_{i}^{m}\hat{T}_i = \frac{M}{m}\sum_{i}^{m}\frac{N_i}{n_i}\sum_{j}^{n}x_{ij}$$

เราสามารถพิสูจน์ได้ว่า $E(\hat{T}) = T$ ดังนี้

$$E(\hat{T}) = E_iE_j\left(\frac{M}{m}\sum_{i}^{m}\frac{N_i}{n_i}\sum_{j}^{n_i}x_{ij}\right)$$ [1]

$$= E_i\left(\frac{M}{m}\sum_{i}^{m}\frac{N_i}{n_i}\sum_{j}^{n_i}E_j(x_{ij})\right)$$

$$= E_i\left(\frac{M}{m}\sum_{i}^{m}\frac{N_i}{n_i}\sum_{j}^{n_i}\bar{X}_i\right) \because E(x) = \bar{X} \text{ หรือ } \mu \text{ ในที่นี้ } E_j(x_{ij})$$

จึงเท่ากับ $\bar{X}_i$ [2]

---

1. เรามีการสุ่มตัวอย่างเป็น 2 ชั้น E_i แสดงค่าคาดหมายของการสุ่มตัวอย่างชั้นที่ 1 (psu) E_j แสดงค่าคาดหมายของการสุ่มตัวอย่างชั้นที่ 2 (ssu) พูดง่าย ๆ subscript i คือ psu subscript j คือ ssu

2. หรือ $E_j(x_{ij}) = \sum_{j}^{N_i}x_{ij}Pr(x_{ij}) = \sum_{j}^{N_i}x_{ij}\frac{1}{N_i} = \bar{x}_i$

$$= E_i\left(\frac{M}{m}\sum_{i}^{m} N_i\bar{X}_i\right) \because \sum_{j}^{n_i}\bar{X}_i = n_i\bar{X}_i$$

$$= E_i\left(\frac{M}{m}\sum_{i}^{m}\sum_{j}^{N_i} x_{ij}\right) \because N_i\bar{X}_i = N_i\sum_{j}^{N_i}\frac{1}{N_i}x_{ij} = \sum_{j}^{N_i} x_{ij}$$

$$= E_i\left(\frac{M}{m}\sum_{i}^{m} T_i\right) \because T_i = \sum_{j}^{N_i} x_{ij}$$

$$= \frac{M}{m}\sum_{i}^{m} E_i(T_i)$$

$$= \frac{M}{m}\sum_{i}^{m}\frac{1}{M}\sum_{i}^{M} T_i \quad \because E(x) = \sum_{i}^{N} x_i Pr(x_i)$$ ในที่นี้

$E_i(T_i)$ จึงเท่ากับ $\sum_{i}^{M} T_i Pr(T_i)$ และ

$Pr(T_i) = 1/M$

$$= \frac{M}{m}\frac{1}{M}\sum_{i}^{m} T_i \because \sum_{i}^{M} T_i = T = \sum_{i}^{M}\sum_{j}^{N_i} x_{ij}$$

$$= \frac{M}{m}\frac{1}{M}\cdot mT$$

$$= T$$

นั่นคือ $\hat{T} = \frac{M}{m}\sum_{i}^{m}\frac{N_i}{n_i}\sum_{j}^{n_i} x_{ij}$ เป็นตัวประมาณค่าที่ปราศจากอคติของ

$T = \sum_{i}^{M}\sum_{j}^{N_i} x_{ij}$

พิสูจน์ $V(\hat{T}) = E(\hat{T} - E(\hat{T}))^2$

พิจารณา $(\hat{T} - T)^2$ จะพบว่า $(\hat{T} - T)^2 = \left(\frac{M}{m}\sum_{i}^{m}\hat{T}_i - T\right)^2$

$$= \left(\frac{M}{m}\sum_{i}^{m}\hat{T}_i - \frac{M}{m}\sum_{i}^{m} T_i + \frac{M}{m}\sum_{i}^{m} T_i - T\right)^2$$

$$= \left\{\left(\frac{M}{m}\sum_{i}^{m}\hat{T}_i - \frac{M}{m}\sum_{i}^{m} T_i\right) + \left(\frac{M}{m}\sum_{i}^{m} T_i - T\right)\right\}^2$$

$$= \left(\frac{M}{m}\sum_{i}^{m}\hat{T}_i - \frac{M}{m}\sum_{i}^{m}T_i\right)^2 + 2\left(\frac{M}{m}\sum_{i}^{m}\hat{T}_i - \frac{M}{m}\sum_{i}^{m}T_i\right)$$

$$\left(\frac{M}{m}\sum_{i}^{m}T_i - T\right) + \left(\frac{M}{m}\sum_{i}^{m}T_i - T\right)^2$$

$$\Rightarrow (\hat{T} - T)^2 = \left(\frac{M}{m}\right)^2\left\{\sum_{i}^{m}(\hat{T}_i - T_i)\right\}^2 + 2\left(\frac{M}{m}\right)\left\{\sum_{i}^{m}(\hat{T}_i - T_i)\right\}\left(\frac{M}{m}\sum_{i}^{m}T_i - T\right) +$$

$$\left(\frac{M}{m}\sum_{i}^{m}T_i - T\right)^2$$

$$= \left(\frac{M}{m}\right)^2\sum_{i}^{m}(\hat{T}_i - T_i)^2 + \left(\frac{M}{m}\right)^2\sum_{i\neq i'}^{m}(\hat{T}_i - T_i)(\hat{T}_{i'} - T_{i'}) +$$

$$2\left(\frac{M}{m}\right)\left(\frac{M}{m}\sum_{i}^{m}T_i - T\right)\sum_{i}^{m}(\hat{T}_i - T_i) + \left(\frac{M}{m}\sum_{i}^{m}T_i - T\right)^2$$

$$= A + B + C + D$$

$$\therefore V(\hat{T}) = E(\hat{T} - T)^2 = E_iE_j(\hat{T} - T)^2 = E_iE_j(A) + E_iE_j(B) + E_iE_j(C) + E_iE_j(D)$$

เราจะเริ่มหาค่าคาดหมายของ E_j (คือค่าคาดหมายของ ssu) ก่อน แล้วจึงค่อยคิดค่าคาดหมายของ psu หรือ E_i ดังนี้

1. $E_j(A) = \left(\frac{M}{m}\right)^2 E_j\sum_{i}^{m}(\hat{T}_i - T_i)^2 = \left(\frac{M}{m}\right)^2\sum_{i}^{m}E_j(\hat{T}_i - T_i)^2 = \left(\frac{M}{m}\right)^2\sum_{i}^{m}V(\hat{T}_i)$

แต่ ∵ เราสุ่ม ssu มาจากแต่ละ ssu โดยวิธี SRS ดังนั้น

$$V(\hat{T}_i) = N_i^2\frac{N_i - n_i}{N_i}\frac{S_i^2}{n_i} \text{ โดยที่ } S_i^2 = \frac{1}{N_i - 1}\sum_{i}^{N_i}(x_{ij} - \bar{X}_i)^2$$

$$\Rightarrow E_j(A) = \left(\frac{M}{m}\right)^2\sum_{i}^{m}\frac{N_i - n_i}{N_i}\frac{N_i^2S_i^2}{n_i}$$

ดังนั้น $E_iE_j(A) = \left(\frac{M}{m}\right)^2 E_i\sum_{i}^{m}\frac{N_i - n_i}{N_i}\frac{N_i^2S_i^2}{n_i}$

ให้ $\frac{N_i - n_i}{N_i}\frac{N_i^2S_i^2}{n_i} = V_i$ ดังนั้น $E_i\sum_{i}^{m}V_i = \sum_{i}^{m}E(V_i) = \sum_{i}^{m}\left(\sum_{i}^{m}V_iPr(V_i)\right)$

$$= \sum_{i}^{m} \left(\sum_{i}^{M} \frac{1}{M} V_i \right) = m \frac{1}{M} \sum_{i}^{M} V_i = \frac{m}{M} \sum_{i}^{M} \frac{N_i - n_i}{N_i} \frac{N_i^2 S_i^2}{N_i}$$

$$=> E_i E_I(A) \qquad = \left(\frac{M}{m} \right)^2 E_i \sum_{i}^{m} V_i = \frac{M}{m} \sum_{i}^{M} \frac{N_i - n_i}{N_i} \frac{N_i^2 S_i^2}{n_i}$$

2. $E_I(B) = E_I \left(\frac{M}{m} \right)^2 \sum_{i \neq i'}^{m} (\hat{T}_i - T_i)(\hat{T}_{i'} - T_{i'})$

$$= 0 = \left(\frac{M}{m} \right)^2 \sum_{i \neq i'}^{m} E_I(\hat{T}_i - T_i)(\hat{T}_{i'} - T_{i'})$$

$= 0$ ทั้งนี้เพราะ $\hat{T}_i$ เป็นอิสระกับ $\hat{T}_{i'}$

เนื่องจาก $\hat{T}_i = N_i \bar{x}_i$ และ $\hat{T}_{i'} = N_{i'} \bar{x}_{i'}$ ที่ $\bar{x}_i$ และ $\bar{x}_{i'}$ ได้มาจากกลุ่มตัวอย่างที่เป็นอิสระต่อกัน

ดังนั้น $E_I(\hat{T}_i - T_i)(\hat{T}_{i'} - T_{i'}) = Cov(\hat{T}_i, \hat{T}_{i'}) = 0$

$$E_i E_I(B) = 0 \qquad \text{.....(2)}$$

3. $E_I(C) = 2 \left(\frac{M}{m} \right) E_I \left(\frac{M}{m} \sum_{i}^{m} T_i - T \right) \sum_{i}^{m} (\hat{T}_i - T_i)$

$$= 2 \left(\frac{M}{m} \right) \left(\frac{M}{m} \sum_{i}^{m} T_i - T \right) \sum_{i}^{m} E_I(\hat{T}_i - T_i)$$

$\because E_I(\hat{T}_i - T_i) = E_I(\hat{T}_i) - E_I(T_i) = T_i - T_i = 0$

ดังนั้น $E_I(C) = 0$

$$=> E_i E_I(C) = 0 \qquad \text{.....(3)}$$

4. $E(D) = E_I \left(\frac{M}{m} \sum_{i}^{m} T_i - T \right)^2$

แต่ $\left(\frac{M}{m}\sum_{i}^{m} T_i - T\right)^2$ มิได้เกี่ยวเนื่องอยู่กับการสุ่มขั้นที่ 2 (ssu) [1] แต่เป็นเรื่องของ psu ล้วน ๆ

ดังนั้น $E_j(D) = E_j\left(\frac{M}{m}\sum_{i}^{m} T_i - T\right)^2 = \left(\frac{M}{m}\sum_{i}^{m} T_i - T\right)^2$

$$\Rightarrow E_iE_j(D) = E_i\left(\frac{M}{m}\sum_{i}^{m} T_i - T\right)^2 = E_i\left\{M^2\left(\frac{1}{m}\sum_{i}^{m} T_i - \frac{1}{M}T\right)^2\right\}$$

$$= E_i\left\{M^2\left(\frac{1}{m}\sum_{i}^{m} T_i - \bar{T}\right)^2\right\} = E_i\left\{M^2(\hat{\bar{T}} - \bar{T})^2\right\}^{2}$$

$$= M^2E_i\left(\hat{\bar{T}} - \bar{T}\right)^2$$

$$= M^2V(\hat{\bar{T}})$$

$\because$ เราสุ่ม psu มาโดยวิธี SRS ดังนั้น $V(\hat{\bar{T}}) = \frac{M-m}{M}\frac{S_b^2}{m}$ เมื่อ S_b^2 คือความแปรปรวนระหว่าง psu นั่นคือ $S_b^2 = \frac{1}{M-1}\sum_{i}^{M}(T_i - \bar{T})^2$

$$\Rightarrow E_iE_j(D) = M^2V(\hat{\bar{T}}) = M^2 \cdot \frac{M-m}{M}\frac{S_b^2}{m} \qquad \text{.....(4)}$$

ดังนั้น $V(\hat{T}) = E_iE_j(A) + E_iE_j(B) + E_iE_j(C) + E_iE_j(D)$

---

1. $\hat{T}_i$ เกี่ยวข้องกับการสุ่ม ssu คือ $\hat{T}_i = N\bar{x}_i = \frac{N_i}{n_i}\sum_{j}^{n_i} x_{ij}$ และ T_i ไม่เกี่ยวข้องกับการสุ่ม ssu กล่าวคือ $T_i = \sum_{j}^{N_i} x_{ij}$ ขอให้สังเกตเลยว่าค่า $\hat{T}_i$ จะ sum ถึงขนาดตัวอย่าง n_i เป็นการยืนยันว่า $\hat{T}$ เกี่ยวข้องอยู่กับการสุ่ม ssu มา n_i หน่วย ส่วนค่า T_i จะ sum ถึง N_i ซึ่งไม่เกี่ยวกับการสุ่มตัวอย่างแต่ประการใด คำอธิบายข้อนี้อาจช่วยให้นักศึกษาเข้าใจ $E_i(A)$ ได้ดีขึ้น

2 $\frac{1}{m}\sum_{i}^{m} T_i$ มีความหมายว่าเราสุ่ม psu มา m หน่วย และทุก ๆ ssu ในแต่ละ psu ตัวอย่างจะได้รับการสำรวจ เช่น สุ่มชั้นเรียนในโรงเรียนแห่งหนึ่งมา 10 ชั้นเรียนและนักเรียนทุกคนในแต่ละชั้นเรียนเหล่านี้ได้รับแบบสอบถาม ดังนั้น $\frac{1}{m}\sum_{i}^{m} T_i$ จึงเท่ากับ $\hat{\bar{T}}$ มีความหมายเช่นเดียวกับ sample mean ใน SRS ของ $\bar{x} = \hat{\bar{X}} = \frac{1}{n}\sum_{i}^{n} x_i$