10.14456/dmhjournal.2007.15

บทความพิเศษ : การสุ่มตัวอย่าง

*จุฬาลักษณ์ โกมลตรี, ปร.ด.**

การวิจัยเชิงสำรวจ (survey research) ในด้านต่างๆ เช่น ทางธุรกิจ, สังคมศาสตร์, ประชากรศาสตร์, การศึกษา, การแพทย์และสาธารณสุข มักจะใช้การสุ่มตัวอย่าง (sampling) เพื่อให้ได้ตัวอย่างที่มีลักษณะเหมือนประชากรที่ต้องการศึกษา (study population) แต่อย่างไรก็ตาม ผู้วิจัยส่วนใหญ่ยังขาดความรู้ ความเข้าใจที่ถูกต้องเกี่ยวกับวิธีการสุ่มตัวอย่างและการวิเคราะห์ข้อมูลที่ได้จากการสุ่มตัวอย่าง บทความนี้จะกล่าวถึงวิธีการสุ่มตัวอย่างเท่านั้น ส่วนวิธีการประมาณค่าและการวิเคราะห์ทางสถิติที่เหมาะสมกับวิธีการสุ่มตัวอย่างต่างๆ จะได้กล่าวในบทความอื่นต่อไป

การสำรวจ (Survey)

การวิจัยเชิงสำรวจ หมายถึงการเก็บข้อมูลเกี่ยวกับมนุษย์ทั้งทางตรง (เช่น การสัมภาษณ์ทางโทรศัพท์, การสัมภาษณ์ส่วนตัว, การตอบแบบสอบถามทางไปรษณีย์, การสังเกต) และทางอ้อม (เช่น การบันทึกข้อมูลจากแฟ้มประวัติผู้ป่วยในโรงพยาบาล) โดยใช้วิธีการเก็บข้อมูลอย่างเป็นระบบ (เช่นแบบสอบถาม)

การสำรวจจัดเป็นการศึกษาแบบ observational study (nonexperimental study) และเนื่องจากมักจะเป็นการสำรวจเพียง 1 ครั้ง ณ เวลาใดเวลาหนึ่งจึงมักจะเป็น cross-sectional study ด้วย การสำรวจส่วนใหญ่มีวัตถุประสงค์หลักเพื่อประมาณค่าในประชากร เช่น สัดส่วนการได้รับวัคซีนในเด็กอายุน้อยกว่า 1 ปี จึงมักจะเป็นการศึกษาเชิงพรรณนา (descriptive study) ในบางครั้งอาจจะมีวัตถุประสงค์รองเพื่อการทดสอบสมมติฐานด้วย

ตัวอย่างการวิจัยเชิงสำรวจ เช่น การทำสำมะโนประชากร, การสำรวจรายได้ของประชากรไทย พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2542 ได้ให้ความหมายของคำว่าสำมะโนไว้ดังนี้ “**สำมะโน** (กฎ) น. การเก็บรวบรวมข้อมูลเกี่ยวกับประชากร เคหะ การเกษตร อุตสาหกรรม ธุรกิจและการอื่น เพื่อใช้ประโยชน์ในทางสถิติ โดยการแจงนับจากทุกหน่วยเกี่ยวกับเรื่องนั้นๆ... **สำมะโนประชากร** น. การเก็บรวบรวมข้อมูลเกี่ยวกับจำนวนและลักษณะต่างๆ ของราษฎรทุกคนในครัวเรือน ในระยะเวลาใดเวลาหนึ่งเพื่อใช้ประโยชน์ในทางสถิติ.” กล่าวโดยสรุปคือ สำมะโนหมายถึงการศึกษาในประชากรทั้งหมด

การทำสำมะโนมีข้อดีคือได้ข้อมูลของประชากรไม่มีอคติจากการสุ่มตัวอย่าง แต่มีข้อเสียในด้านเวลา, ค่าใช้จ่าย, ความเป็นไปได้, คุณภาพ (ความถูกต้อง, ความสมบูรณ์) ของข้อมูล

ประชากร (population, universe)

ประชากร คือ สิ่งที่ผู้วิจัยสนใจทั้งหมด ประชากรอาจจะเป็นคน, สัตว์ หรือสิ่งของก็ได้ ประชากรแบ่งเป็น 2 ประเภทคือ target population และ study population (survey population) เช่น ในการศึกษาถึงความชุกของโรคเบาหวานในผู้ใหญ่ของประเทศไทย target population คือ ประชากรเพศชายและหญิงที่มีอายุ 18 ปีขึ้นไปทั่วประเทศไทยในเดือนมกราคมปี พ.ศ. 2550 study population จะเป็นส่วนหนึ่ง (subset) ของ target population ถ้า study population มีลักษณะเหมือน target population จะทำให้สามารถสรุปผลการศึกษาจาก study populationไปสู่ target population ได้ซึ่งเรียกว่าการศึกษานั้นมี external validity (generalizability) แต่ในหลายครั้งจะมีข้อจำกัดในการดำเนินการศึกษา study population จึงมีลักษณะไม่เหมือน target population เช่น study population คือ ประชากรเพศชายและหญิงที่มีอายุ 18 ปีที่มีภูมิลำเนาอยู่ในกรุงเทพมหานครในเดือนมกราคมปี พ.ศ. 2550 เนื่องจากความชุกของโรคเบาหวานของคนที่อยู่ในกรุงเทพมหานครและจังหวัดอื่นๆ มีความแตกต่างกัน จึงไม่สามารถอ้างค่าความชุกที่ได้จากคนในกรุงเทพมหานครไปสู่คนไทยทั้งประเทศได้ จึงเรียกว่าการศึกษานี้ไม่มี external validity จาก study population จะมีการสุ่มเลือกตัวอย่างเพื่อให้ได้กลุ่มตัวอย่างที่จะศึกษา ถ้ามีการสุ่มตัวอย่างที่ถูกต้องตามหลักสถิติ, มีจำนวนคนเพียงพอและมีวิธีดำเนินการศึกษาที่ดี ไม่มีอคติ จะสามารถนำค่าความชุกของเบาหวานที่ได้จากตัวอย่างที่ศึกษาไปประมาณค่าความชุกใน study population ได้ ซึ่งจะเรียกว่าการศึกษานั้นมี internal validity internal validity เป็นสิ่งที่จำเป็นในทุกการศึกษา

ประชากรที่สนใจในแต่ละการศึกษาอาจจะมีจำนวนมากจนนับไม่ได้ (infinite) หรือนับจำนวนได้ (finite) การนับได้หรือนับไม่ได้ของประชากรจะมีความสำคัญต่อวิธีการสุ่มตัวอย่าง, การประมาณค่าและการวิเคราะห์ผลทางสถิติ

ชนิดของการสุ่มตัวอย่าง

การสุ่มตัวอย่างแบ่งเป็น 2 ชนิดคือ

1. Nonprobability sampling

Nonprobability sampling คือการเลือกตัวอย่างโดยไม่ใช้วิธีการสุ่ม (random, chance) ดังนั้นจึงไม่สามารถทราบถึงโอกาสที่แต่ละคนจะถูกเลือกเข้าไปในการศึกษา Nonprobability sampling มีหลายอย่างเช่น

1.1 Haphazard sampling เช่น การศึกษาในอาสาสมัคร (volunteer sampling) ซึ่งจะทำให้เกิดปัญหาเรื่องความเป็นตัวแทน

(representativeness) ของประชากร เช่น ในรายการโทรทัศน์ที่ให้ผู้ชมร่วมแสดงความคิดเห็น (เห็นด้วย, ไม่เห็นด้วย) ในเรื่องใดเรื่องหนึ่งผ่านโทรศัพท์ คนที่ร่วมให้ความคิดเห็นจะเป็นเฉพาะคนที่ชมรายการนั้นเท่านั้น, มีโทรศัพท์, มีความรู้สึกอยากแสดงความคิดเห็น, มีความคิดเห็นด้วยหรือไม่เห็นด้วยอย่างมาก ดังนั้นตัวอย่างจึงเป็นกลุ่มคนที่มีลักษณะต่างไปจากคนทั่วๆ ไป นอกจากนี้แต่ละคนยังสามารถแสดงความคิดเห็นได้มากกว่า 1 ครั้ง ค่าที่ได้จากตัวอย่าง (ร้อยละของการเห็นด้วย) จึงไม่สามารถนำไปอ้างอิงได้ ถึงแม้จะมาจากตัวอย่างคนเป็นจำนวนมากเช่นเกิน 10000 คน ค่าที่ได้จากตัวอย่างที่มีจำนวนน้อยกว่าเช่น 1000 คน แต่มาจากการสุ่มอย่างถูกต้องตามหลักสถิติ จะมีความน่าเชื่อถือมากกว่าและไม่มีความลำเอียง (bias)

ในบางครั้งตัวอย่างได้จากคนเดินถนน, เดินตามห้างสรรพสินค้า ตัวอย่างเหล่านี้จะมีลักษณะอย่างไรขึ้นกับปัจจัยหลายอย่างเช่น การตัดสินใจของพนักงานสัมภาษณ์เกี่ยวกับคนที่จะเข้าไปขอความร่วมมือ, สถานที่ วันและเวลาที่สัมภาษณ์ ดังนั้นอาจมีอคติในการเลือกตัวอย่าง

1.2 Purposive (judgment) sampling คือการเลือกตัวอย่างโดยอาศัยการพิจารณา, การตัดสินของผู้วิจัยว่าตัวอย่างเหล่านั้นเป็นตัวแทนที่ดีของประชากร วิธีนี้มีปัญหาคือผู้วิจัยแต่ละคนอาจจะมีความเห็นที่ไม่เหมือนกันเกี่ยวกับความเป็นตัวแทนที่ดีของประชากร

1.3 Quota sampling เป็นรูปแบบหนึ่งของ purposive sampling ที่ใช้กันมากในการวิจัยตลาด เช่น เพื่อให้กลุ่มตัวอย่างมีลักษณะทางประชากรบางอย่าง (เช่นเพศ, อายุ) ที่คล้ายกับประชากร ผู้วิจัยจะกำหนดจำนวนขนาดตัวอย่างซึ่งจะทำให้ตัวอย่างมีสัดส่วนของเพศและกลุ่มอายุ ที่คล้ายกับสัดส่วนจริงในประชากร

2. Probability sampling

Probability sampling คือการเลือกตัวอย่างด้วยวิธีการสุ่ม จึงทำให้ทราบถึงโอกาสที่ตัวอย่างแต่ละตัวอย่างจะถูกเลือกเข้าในการศึกษา ตัวอย่างแต่ละราย จะมีความน่าจะเป็นที่จะถูกเลือกซึ่งจะไม่เท่ากับศูนย์ แต่จะเท่ากันหรือไม่เท่ากันก็ได้ การใช้ probability sampling ทำให้สามารถประมาณค่าที่สนใจในประชากร (เช่นความชุกของโรคเบาหวาน) จากค่าสถิติที่ได้จากตัวอย่างได้อย่างถูกต้อง

ในการเลือกตัวอย่าง แต่ละตัวอย่างอาจจะถูกเลือกเพียงครั้งเดียวหรือถูกเลือกซ้ำได้อีก การถูกเลือกเพียงครั้งเดียวเกิดจากการไม่นำตัวอย่างที่ถูกเลือกในครั้งแรกใส่คืนในประชากรเพื่อการเลือกครั้งต่อไป (sampling without replacement, WOR) ส่วนการถูกเลือกซ้ำได้อีกเกิดจากการนำตัวอย่างที่ถูกเลือกในครั้งแรกกลับคืนสู่ประชากรก่อนการเลือกครั้งต่อไป (sampling with replacement, WR) โดยส่วนมากการเลือกตัวอย่างจะเป็นแบบ without

replacement เนื่องจากต้องการทราบข้อมูลจากแต่ละตัวอย่างเพียง 1 ครั้ง การวิเคราะห์ทางสถิติเมื่อมีการสุ่มแบบ without replacement จะยุ่งยากกว่าเมื่อมีการสุ่มแบบ with replacement เนื่องจากความน่าจะเป็นที่แต่ละคนจะถูกเลือกในแต่ละครั้งจะเปลี่ยนไปคือน้อยลงเรื่อยๆ แต่อย่างไรก็ตามประชากรมักมีจำนวนมาก (อนันต์) จึงทำให้การสุ่มตัวอย่างแบบ without replacement ไม่แตกต่างจากการสุ่มแบบ with replacement คือความน่าจะเป็นที่แต่ละคนจะถูกเลือกในแต่ละครั้งจะมีค่าคงที่ การวิเคราะห์ทางสถิติจึงสามารถใช้แบบ with replacement ได้

Probability sampling

การสุ่มตัวอย่างแบบ probability sampling แบ่งเป็น 4 ชนิดใหญ่ๆคือ

1. การสุ่มตัวอย่างแบบง่าย (Simple random sampling, SRS)
2. การสุ่มตัวอย่างแบบมีระบบ (Systematic sampling)
3. การสุ่มตัวอย่างแบบชั้นภูมิ (Stratified sampling)
4. การสุ่มตัวอย่างแบบกลุ่ม (Cluster sampling)

ในการสุ่มตัวอย่าง สิ่งที่ได้รับการสุ่มตัวอย่างจะเรียกว่า sampling unit (selection unit) ซึ่ง sampling unit อาจจะเป็นคน (หรือสัตว์หรือสิ่งไม่มีชีวิตก็ได้) หรือกลุ่มคน (เช่น โรงเรียน, ห้องเรียน) ก็ได้ ในที่นี้จะสมมติว่าเป็นการสุ่มตัวอย่างในคน ส่วนคนที่ได้รับการบันทึกค่าตัวแปรที่สนใจจะเรียกว่า elementary unit (element)

1. การสุ่มตัวอย่างแบบง่าย (Simple random sampling, SRS)

การสุ่มตัวอย่างแบบง่าย คือการสุ่มตัวอย่าง (แบบ without replacement) คน n คนจากประชากรจำนวน N คน ซึ่งจะเขียนได้สั้นๆว่า SRS (n จาก N) ในการสุ่มตัวอย่างแบบง่ายแต่ละคนจะมีโอกาสที่จะถูกเลือกเท่ากันคือ n/N ในการสุ่มแบบนี้จะต้องมีรายชื่อของประชากรทั้งหมด (sampling frame, บางครั้งจึงเรียกการสุ่มตัวอย่างแบบง่ายว่า list sampling) จึงเป็นวิธีการสุ่มที่ไม่สะดวกในทางปฏิบัติ ไม่ค่อยนิยมใช้ แต่อย่างไรก็ตามวิธีการวิเคราะห์ทางสถิติทั้งหมดจะตั้งอยู่บนพื้นฐานของการสุ่มตัวอย่างแบบง่าย ตัวอย่างที่ได้จากการสุ่มแบบนี้จะเรียกว่า simple random sample ซึ่งมักจะเรียกสั้นๆว่า random sample

การสุ่มอย่างง่ายอาจทำได้โดยการใช้ตารางเลขสุ่ม (Table of random numbers) ตัวเลขต่างๆ ในตารางเลขสุ่มมีคุณสมบัติคือ

ก. ตัวเลขแต่ละตัว (0, 1, 2, ..., 9) มีโอกาสเกิดขึ้นเท่าๆ กัน

ข. ตัวเลขแต่ละคู่ (00, 01, 02, ..., 98, 99) มีโอกาสเกิดขึ้นเท่าๆ กัน

ค. ตัวเลข 3 ตัวต่างๆ (000, 001,

002, ..., 989, 999) มีโอกาสเกิดขึ้นเท่าๆ กัน

ง. ตัวเลขต่างๆ เป็นอิสระต่อกัน

ตัวอย่างที่ 1

ผู้วิจัยสนใจในการสำรวจความคิดเห็นเรื่องการเรียนต่อในชั้นมัธยม 1 (เปลี่ยน, ไม่เปลี่ยนโรงเรียน) จึงจะทำการศึกษาในนักเรียนชั้นประถมศึกษาปีที่ 6 ในโรงเรียนแห่งหนึ่งซึ่งมีถึงชั้นมัธยมปีที่ 6 จากจำนวนนักเรียนทั้งหมด 300 คน (N = 300) ผู้วิจัยจะทำการศึกษาในตัวอย่างนักเรียนจำนวน 30 คน (n = 30) เนื่องจากมีรายชื่อของนักเรียน ป.6 ทั้งหมด จึงจะใช้การสุ่มตัวอย่างแบบง่าย วิธีการสุ่มตัวอย่างแบบง่ายทำได้ดังนี้

1) เรียงชื่อนักเรียน ป.6 ทั้งหมดตามตัวอักษร แล้วให้หมายเลขจาก 1 ถึง 300

2) สุ่มตัวเลขจากตารางเลขสุ่ม โดยมีขั้นตอนคือ

– สุ่มแถวและสดมภ์

– จากแถวและสดมภ์ที่สุ่มได้กำหนดลักษณะของการอ่านตัวเลข เช่น อ่านจากซ้ายไปขวาและจากบนลงล่าง

– อ่านตัวเลข 3 หลัก ถ้าตัวเลขสุ่มนั้นมีค่าไม่เกิน 300 (จำนวนประชากร) ก็เลือกตัวอย่างหมายเลขนั้นได้เลย แต่ถ้าตัวเลขสุ่มนั้นมีค่าเกิน 300 ให้เอาตัวเลขสุ่มนั้นลบด้วย 1*300 หรือ 2*300 หรือ 3*300 เช่น ถ้าสุ่มได้เลข 374 ให้เลือกนักเรียนหมายเลข 74 (= 374-300) ถ้าสุ่มได้เลข 826 ให้เลือกนักเรียนหมายเลข 226 (= 826-600) ถ้าสุ่มได้เลข 912 ให้เลือกนักเรียนหมายเลข 12 (= 912-900)

ในการเลือกตัวอย่าง ถ้าเลือกได้หมายเลขเดิมให้เลือกใหม่ แต่ละคนจะได้รับเลือกเพียง 1 ครั้งเท่านั้น

ตัวเลขข้างล่างเป็นส่วนหนึ่งของตารางเลขสุ่ม ถ้า 546 คือตัวเลขแรกที่สุ่มได้จะสุ่มได้คนหมายเลขต่างๆ ดังนี้ 246 (= 546-300), 29 (= 329-300), 037, 194 (= 494-300), 143, 192 (= 492-300), 58 (= 958-900), 272 (=872-600), 043, 035, 70 (= 970-900) เรื่อยไปจนครบ 30 คน

32924	22324	18125	09077
54632	90374	94143	49295
88720	43035	97081	83373
21727	11904	41513	31653
80985	70799	57975	69282

ในปัจจุบันการสุ่มสามารถทำงานได้ด้วยการใช้โปรแกรมคอมพิวเตอร์ ซึ่งจะสะดวกและรวดเร็วกว่าการสุ่มด้วยมือจากตารางเลขสุ่ม

2. การสุ่มตัวอย่างแบบมีระบบ (Systematic sampling)

จากตัวอย่างข้างต้นจะเห็นได้ว่าในการสุ่มแบบง่าย ตัวอย่างที่ได้จะกระจายกันไปในประชากร จึงอาจจะทำให้เกิดความไม่สะดวกในการเก็บข้อมูลโดยเฉพาะเมื่อประชากรมีขนาดใหญ่ ผู้วิจัยจึงอาจจะใช้การสุ่มตัวอย่างแบบมีระบบซึ่งทำได้ง่ายกว่าและค่าประมาณที่ได้มีความถูก

ต้องเช่นเดียวกับการสุ่มอย่างง่าย

ถ้า N = จำนวนประชากร

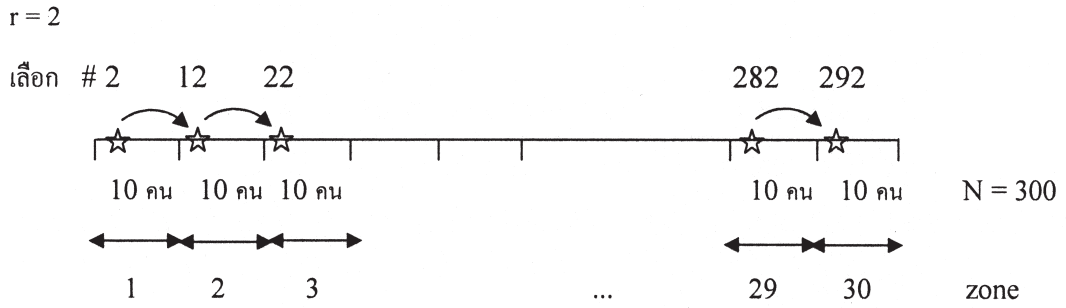
n = จำนวนขนาดตัวอย่าง

การสุ่มแบบมีระบบมีขั้นตอนคือ

1) คำนวณ k โดยที่

k = N/n = sampling interval (selection interval, skip number)

จากตัวอย่างแรก N = 300, n = 30, k = 300/30 = 10 หมายถึงประชากรมีจำนวนเป็น 10 เท่าของตัวอย่าง ดังนั้นทุกๆ ประชากร 10 คนจะเลือกตัวอย่างมา 1 คน โอกาสที่แต่ละคนจะถูกเลือกจะเท่ากันคือเท่ากับ n/N หรือ 0.1 หรือจะกล่าวได้ว่า ในการสุ่มแบบมีระบบจะแบ่งประชากรเป็น 30 zone (n) ในแต่ละ zone มี 10 คน (k) และในแต่ละzone จะสุ่มตัวอย่างมา 1 คน ดังแสดงในภาพข้างล่าง



2) สุ่มเลือกตัวเลข 1 ตัวระหว่าง 1 ถึง k เป็นจุดเริ่มต้น (random start number) ซึ่งจะเรียกว่า r

3) ตัวอย่างที่เลือกคนแรกคือประชากรหมายเลข r หลังจากนั้นเว้นไป k คนเรื่อยไป กล่าวคือเลือกคนที่ r, r+k, r+2k, ..., r+(n-1)k จะได้ตัวอย่าง n คน เช่น ในตัวอย่างแรก ถ้าสุ่มได้ r = 2 จะเลือกตัวอย่างจากประชากรหมายเลข 2, 12, 22, 32, ..., 292 จะเห็นได้ว่าตัวอย่างคน 30 คนนี้จะกระจายไปเท่าๆ กันในประชากร

ในการสุ่มแบบมีระบบจะมีการสุ่มตัวเลขเพียง 1 ครั้งเท่านั้นในครั้งแรก (r) และหมายเลขที่ถูกเลือกในครั้งต่อๆ ไปจะถูกกำหนดโดยค่า r ที่สุ่มได้ในครั้งแรก การสุ่มแบบนี้จึงง่ายและสะดวกกว่าการสุ่มแบบง่าย

เนื่องจากในการสุ่มตัวอย่างแบบมีระบบนี้ ตัวอย่างคนที่สุ่มได้จะกระจายไปเท่าๆ กันในประชากร ลำดับที่ในประชากรจึงมีความสำคัญต่อค่าประมาณที่ได้จากตัวอย่าง เช่น ถ้าเรียงชื่อประชากรตามตัวอักษร ก. ถึง ฮ. การใช้การสุ่มตัวอย่างแบบมีระบบจะทำให้ตัวอย่างมีสัดส่วนของตัวอักษรต่างๆ เหมือนกับในประชากร ดังนั้นถ้าสนใจในการประมาณค่าใช้จ่ายเฉลี่ยของผู้ป่วยในและเรียงชื่อผู้ป่วยตามรหัสของโรค การใช้การสุ่มตัวอย่างแบบมีระบบ

จะทำให้กลุ่มตัวอย่างประกอบด้วยผู้ป่วยด้วยโรคต่างๆ ในสัดส่วนที่เหมือนกับในประชากร เนื่องจากค่าใช้จ่ายของผู้ป่วยโรคเดียวกันจะใกล้เคียงกัน การสุ่มตัวอย่างแบบมีระบบจึงคล้ายกับการสุ่มตัวอย่างแบบชั้นภูมิ (ชั้นภูมิคือรหัสโรค) ซึ่งจะได้กล่าวต่อไป แต่ถ้าเรียงชื่อผู้ป่วยตามตัวอักษร หรือวันที่ที่ออกจากโรงพยาบาล การสุ่มตัวอย่างแบบมีระบบจะคล้ายกับการสุ่มตัวอย่างแบบง่าย เนื่องจากชื่อผู้ป่วยหรือ วันที่ออกจากโรงพยาบาลไม่มีความสัมพันธ์กับค่าใช้จ่ายของผู้ป่วย

การใช้การสุ่มแบบมีระบบจะไม่เหมาะสมเมื่อมี monotonic trend ในตัวแปรที่ศึกษา เช่นในการประมาณค่าเงินเดือนเฉลี่ยในปัจจุบัน ถ้าหมายเลขในประชากรเรียงตามเงินเดือนในปีที่แล้ว (จากน้อยไปมาก) และใช้การสุ่มแบบมีระบบโดยที่ N = 3000, n = 30, k = 100 ค่าเฉลี่ยของเงินเดือนที่ได้ตัวอย่างหมายเลข (1, 101, 201, ..., 2901; random start r = 1) จะน้อยกว่าค่าเฉลี่ยที่ได้ตัวอย่างหมายเลข (100, 200, 300, ..., 3000; r = 100)

ในกรณีที่ k ไม่ใช่เลขจำนวนเต็ม (integer) เช่น N = 920, n = 100 จะได้ k = 9.2 วิธีการแก้ปัญหามีหลายวิธี แต่วิธีที่ง่ายคือการพิจารณาหมายเลข 1-920 เหมือนเป็นวงกลม การสุ่มตัวอย่างทำโดยสุ่ม random start (r) ซึ่งเป็นตัวเลขจาก 1 ถึง 920 หมายเลขแรกที่เลือกคือ r และหมายเลขต่อไปคือ r+k, r+2k เรื่อยไปจนครบ n คน ถ้าหมายเลขที่สุ่มได้เกินจำนวนคนในประชากรก็ให้ต่อด้วยหมายเลข 1 เช่น ถ้า r = 120 คนแรกที่เลือกคือหมายเลข 120 คนต่อไปคือหมายเลข 120+k หรือ 120+9 = 129, 120+2k = 138, 120+3k = 147, ..., 120+88k = 912, 120+89k = 921 = 1 (เนื่องจากตัวเลขที่สุ่มได้มีค่าเกินจำนวนประชากร N = 920 จึงนำตัวเลขนั้นลบด้วย 920 กลายเป็นตัวเลขสุ่มตัวใหม่ที่มีค่าน้อยกว่า 920), 1+k = 10, 1+2k = 19, ..., 1+9k = 82, 1+10k = 91

3. การสุ่มตัวอย่างแบบชั้นภูมิ (Stratified sampling)

ในการสุ่มตัวอย่างแบบชั้นภูมิ จะมีการแบ่งประชากรออกเป็นกลุ่มย่อย (subgroup, stratum) ตามตัวแปรที่มีความสัมพันธ์กับสิ่งที่สนใจจะศึกษา แต่ละกลุ่มจะไม่ซ้ำซ้อนกัน (เช่น กลุ่มอายุ ≤30, 31-50, >50 ปี) และแต่ละคนจะถูกจัดอยู่ในกลุ่มย่อยกลุ่มใดกลุ่มหนึ่งเท่านั้น

ตัวอย่างที่ 2

ผู้วิจัยสนใจที่จะประเมินรายได้เฉลี่ยของพนักงานทั้งหมดในธนาคาร จากประสบการณ์ผู้วิจัยคิดว่ารายได้ขึ้นกับสาขาที่จบการศึกษา, อายุ แต่ไม่ขึ้นกับเพศ วิธีการสุ่มตัวอย่างที่เหมาะสมควรเป็นวิธีใด

การเลือก Stratification variable

เนื่องจากรายได้ของพนักงานขึ้นกับสาขาที่จบการศึกษาจึงแบ่งพนักงานเป็นกลุ่มตามสาขาที่จบ สาขาที่จบเป็นตัวแปรที่ใช้ในการแบ่งกลุ่มหรือแบ่งชั้นภูมิซึ่งเรียกว่า stratification variable ในที่นี้มี stratification variable 1 ตัว ซึ่งประกอบไปด้วย 3 stratum คือเศรษฐศาสตร์, บัญชี และอื่นๆ (ตารางที่ 1) การแบ่งพนักงานออกเป็น 2 กลุ่มตามเพศอาจจะไม่เหมาะสมถ้ารายได้ไม่ขึ้นกับเพศ ส่วนการแบ่งพนักงานเป็นกลุ่มตามสาขาที่จบ (เศรษฐศาสตร์, บัญชี และอื่นๆ) และกลุ่มอายุ (<30, 30-39, 40-49, ≥50 ปี) ดังตารางที่ 2 อาจจะดีกว่าการแบ่งกลุ่มตามสาขาที่จบแต่เพียงอย่างเดียวเนื่องจากรายได้ขึ้นกับอายุด้วย การทำเช่นนี้ทำให้มี stratification variable 2 ตัวคือสาขาที่จบ (3 สาขา) และกลุ่มอายุ (4 กลุ่ม) จำนวนกลุ่ม (ชั้นภูมิ) ทั้งหมดจึงเท่ากับ 12 กลุ่ม พนักงานแต่ละคนจะถูกจัดอยู่ในกลุ่มใดกลุ่มหนึ่งใน 12 กลุ่ม การมี stratification variable 2 ตัวเช่นนี้มีข้อดีทางสถิติคือค่าเฉลี่ยของรายได้ของพนักงานที่ได้จากตัวอย่างจะมีความถูกต้อง (ใกล้เคียงกับค่าเฉลี่ยที่แท้จริงของพนักงานทั้งหมดในประชากร) มากกว่าการจัดกลุ่มตามสาขาที่จบเพียงอย่างเดียว

การที่คนที่จบสาขาต่างกันได้เงินเดือนต่างกัน หมายถึงมีความแตกต่างในตัวแปรที่ศึกษา (เงินเดือน) ระหว่างกลุ่ม, ชั้นภูมิ (สาขาที่จบ) ซึ่งในทางสถิติเรียกว่า heterogeneous among strata และผู้วิจัยยังคาดว่าคนที่จบสาขาเดียวกันน่าจะได้เงินเดือนใกล้เคียงกัน หรือไม่มีความแตกต่างในตัวแปรที่ศึกษา (เงินเดือน) ระหว่างคนในกลุ่ม, ชั้นภูมิเดียวกัน ซึ่งเรียกว่า homogeneous within stratum (internally homogeneous) ดังนั้นการสุ่มแบบชั้นภูมิจะเหมาะสมเมื่อมี heterogeneous (ของตัวแปรที่ศึกษา) among strata และ homogeneous within stratum หรือจะกล่าวได้สั้นๆ ว่า stratification variable ที่ดีต้องมีความสัมพันธ์กับตัวแปรที่จะศึกษา

การที่ stratification variable มีความสัมพันธ์กับตัวแปรที่จะศึกษาจะทำให้ค่าประมาณที่ได้จากตัวอย่างมีความคลาดเคลื่อน (standard error, SE) ต่ำกว่าค่าประมาณที่ได้จากการใช้การสุ่มตัวอย่างแบบง่าย ในทางตรงข้ามถ้า stratification variable ไม่มีความสัมพันธ์กับตัวแปรที่จะศึกษา การสุ่มแบบชั้นภูมิจะไม่มีประโยชน์และไม่ดีกว่า (ในทางสถิติ) การสุ่มแบบง่าย ดังนั้นจึงควรใช้การสุ่มแบบง่ายซึ่งจะสะดวกและง่ายกว่าในทางปฏิบัติเนื่องจากไม่ต้องมีการจัดคนเป็นกลุ่มๆ แต่ยังคงให้ค่าประมาณที่ถูกต้อง

ดังนั้นการเลือก stratification variable จึงมีความสำคัญในการสุ่มแบบชั้นภูมิ

ตารางที่ 1 การแบ่งกลุ่มพนักงานตามสาขาที่จบ

ชั้นภูมิ	สาขาที่จบ	ประชากร (พนักงานทั้งหมด) จำนวน (N_h)	ตัวอย่าง (พนักงานที่สุ่มได้) จำนวน (n_h)
1	เศรษฐศาสตร์		
2	บัญชี		
3	อื่นๆ		
รวม		N	n

ตารางที่ 2 การแบ่งกลุ่มพนักงานตามสาขาที่จบและกลุ่มอายุ

ชั้นภูมิ	สาขาที่จบ	กลุ่มอายุ (ปี)	ประชากร (พนักงานทั้งหมด) จำนวน (N_h)	ตัวอย่าง (พนักงานที่สุ่มได้) จำนวน (n_h)
1	เศรษฐศาสตร์	<30		
2	เศรษฐศาสตร์	30-39		
3	เศรษฐศาสตร์	40-49		
4	เศรษฐศาสตร์	≥50		
5	บัญชี	<30		
6	บัญชี	30-39		
7	บัญชี	40-49		
8	บัญชี	≥50		
9	อื่นๆ	<30		
10	อื่นๆ	30-39		
11	อื่นๆ	40-49		
12	อื่นๆ	≥50		
รวม			N	n

จำนวน Stratification variable(s) และจำนวน stratum

การมี stratification variable มากเกินไปจะทำให้เกิดความยุ่งยากในการสุ่มตัวอย่าง แม้ว่าค่าประมาณที่ได้จะถูกต้องมากขึ้น ดังนั้นการออกแบบวิธีการสุ่มตัวอย่างจึงต้องคำนึงถึงความถูกต้องทางสถิติและความยากง่ายในทางปฏิบัติไปพร้อมกัน

ผู้วิจัยควรคำนึงถึงจำนวน stratum ในแต่ละ stratification variable ด้วย จำนวน stratum ขึ้นกับจำนวนประชากรที่มีในแต่ละ stratum และความแตกต่างในตัวแปรที่ศึกษาระหว่างแต่ละ stratum เช่น ถ้าจำนวนพนักงานธนาคารที่มีอายุมากกว่า 50 ปีมีน้อย ก็ควรรวมเป็นกลุ่มอายุ ≥40 ปี

การสุ่มตัวอย่างในแต่ละชั้นภูมิ

หลังจากที่ได้จัดคนในประชากรเป็นชั้นภูมิต่างๆ แล้ว เพื่อให้ได้มาซึ่งกลุ่มตัวอย่างจำนวน n คน จะต้องทำการสุ่มตัวอย่างจากประชากร ซึ่งจะเป็นการสุ่มตัวอย่างในแต่ละชั้นภูมิ (สุ่ม n_h คนจาก N_h คน) การสุ่มตัวอย่างในแต่ละชั้นภูมิอาจจะใช้วิธีการสุ่มที่เหมือนกันหรือต่างกันก็ได้ เช่น ใช้การสุ่มตัวอย่างแบบง่ายในทุกชั้นภูมิ (จึงเรียกวิธีการสุ่มแบบนี้ว่า Stratified simple random sampling) หรือใช้การสุ่มตัวอย่างแบบง่ายในบางชั้นภูมิและใช้การสุ่มตัวอย่างแบบมีระบบในบางชั้นภูมิ

การแบ่งจำนวนคนทั้งหมดในตัวอย่างออกตามชั้นภูมิต่าง ๆ

วิธีการแบ่งจำนวนคนในตัวอย่าง n คนออกตามชั้นภูมิต่างๆ (stratum allocation) มี 4 วิธีคือ

1) Proportionate allocation
2) Balanced allocation
3) Optimum allocation
4) Disproportionate allocation

ตารางที่ 3 แสดงจำนวนประชากรทั้งหมดและจำนวนประชากรในแต่ละชั้นภูมิ โดยที่

N = จำนวนประชากรทั้งหมด = 1000

N_h = จำนวนประชากรในชั้นภูมิ h, h = 1, 2, 3, 4
(N_1 = 100, N_2 = 300, N_3 = 400, N_4 = 200 และ $N_1+N_2+N_3+N_4$ = 1000)

P_h = สัดส่วนของประชากรในชั้นภูมิ h = N_h/N
(P_1 = 0.1, P_2 = 0.3, P_3 = 0.4, P_4 = 0.2 และ $P_1+P_2+P_3+P_4$ = 1.0)

ตารางที่ 4 แสดงการแบ่งจำนวนตัวอย่างทั้งหมดออกตามชั้นภูมิต่างๆ โดยวิธีต่างๆ โดยที่

n = จำนวนตัวอย่างทั้งหมด
n_h = จำนวนตัวอย่างในชั้นภูมิ h, h = 1, 2, 3, 4 (และ $n_1+n_2+n_3+n_4 = n$)
p_h = สัดส่วนของตัวอย่างในชั้นภูมิ h = n_h/n (และ $p_1+p_2+p_3+p_4 = 1.0$)
f_h = sampling rate ในชั้นภูมิ h = n_h/N_h

Proportionate allocation

การแบ่งจำนวนตัวอย่างในแต่ละชั้นภูมิด้วยวิธี proportionate allocation (ตารางที่ 4 ก.) มีลักษณะคือ

1) sampling rate ในแต่ละชั้นภูมิจะเท่ากัน ($f_1 = f_2 = f_3 = f_4$) และเท่ากับ sampling rate ของตัวอย่างนั้น ($f_1 = f_2 = f_3 = f_4 = f$) เนื่องจาก n = 200, N = 1000 ดังนั้นจึงสุ่มมา 20% หรือ sampling rate ในการศึกษานี้เท่ากับ 0.2 (f = n/N) sampling rate ในแต่ละชั้นภูมิจึงควรเท่ากับ 20% ด้วย กล่าวคือในชั้นภูมิที่ 1 มีประชากรจำนวน 100 คนจึงสุ่มมา 20% หรือ 20 คน ในชั้นภูมิที่ 2 มีประชากร 300 คนและสุ่มมา 20% เช่นกัน ดังนั้นจึงสุ่มมา 60 คน ในชั้นภูมิที่ 3 มี 400 คนและชั้นภูมิที่ 4 มี 200 คน เมื่อสุ่มมา 20% จะได้จำนวนตัวอย่าง 80, 40 คนตามลำดับ

2) สัดส่วนของจำนวนตัวอย่างในแต่ละชั้นภูมิเท่ากับสัดส่วนของจำนวนประชากรในแต่ละชั้นภูมิ ($p_1 = P_1$, $p_2 = P_2$, $p_3 = P_3$, $p_4 = P_4$) กล่าวคือ ตัวอย่างมีการกระจายของ stratification variable เหมือนกับในประชากร จากตารางที่ 3 ประชากรประกอบด้วยจำนวนคนในชั้นภูมิที่ 1, 2, 3 และ 4 ร้อยละ 10, 30, 40 และ 20 ตามลำดับ และตัวอย่างก็มีจำนวนคนในชั้นภูมิที่ 1, 2, 3 และ 4 ร้อยละ 10, 30, 40 และ 20 ด้วยเช่นกัน ตัวอย่างจึงมีลักษณะเหมือนกับประชากร

Balanced allocation

การแบ่งจำนวนตัวอย่างในแต่ละชั้นภูมิด้วยวิธี balanced allocation หมายถึง ในตัวอย่าง แต่ละชั้นภูมิจะมีจำนวนคนเท่ากัน จากจำนวนตัวอย่างทั้งหมด 200 คน ถ้าแบ่งเป็น 4 ชั้นภูมิ จะได้จำนวนคนในแต่ละชั้นภูมิเท่ากับ 50 (ตารางที่ 4 ข.) วิธีการแบ่งแบบนี้เป็นวิธีที่ง่าย แต่ตัวอย่างจะมีลักษณะ (ของ stratification variable) แตกต่างจากประชากร

ในด้าน sampling rate ในแต่ละชั้นภูมิจะมี sampling rate (f_h) แตกต่างกันไป คือเท่ากับ 50%, 17%, 13%, และ 25% ในชั้นภูมิที่ 1, 2, 3 และ 4 ตามลำดับ เมื่อเทียบกับ sampling rate ของตัวอย่างทั้งหมด (f) ซึ่งเท่ากับ 20% จะเห็นว่ามีการสุ่มตัวอย่างในชั้นภูมิที่

1 และ 4 มากเกินไป (oversample) และสุ่มตัวอย่างในชั้นภูมิที่ 2 และ 3 น้อยเกินไป (undersample)

Optimum allocation

การแบ่งแบบ optimum allocation มีหลักการคือ จำนวนตัวอย่างในแต่ละชั้นภูมิจะแปรผกผันกับค่าใช้จ่ายในการเก็บข้อมูลของชั้นภูมินั้น (J_h) และแปรผันตาม variation (standard deviation, SD) ของตัวแปรที่ศึกษาในแต่ละชั้นภูมิ (S_h, within-stratum variation) กล่าวคือจำนวนตัวอย่างในชั้นภูมิจะน้อยถ้าต้องเสียค่าใช้จ่ายในการเก็บข้อมูลของชั้นภูมินั้นมาก (n_h จะน้อยถ้า J_h มีค่ามาก) และมีความแตกต่างน้อยในตัวแปรที่ศึกษาระหว่างคนที่อยู่ในชั้นภูมิเดียวกัน (n_h จะน้อยถ้า S_h มีค่าน้อย)

$$n_h \alpha S_h/\sqrt{J_h}$$

โดยส่วนมากค่าใช้จ่ายในการเก็บข้อมูลในแต่ละชั้นภูมิมักจะไม่แตกต่างกันมาก จำนวนตัวอย่างในแต่ละชั้นภูมิจึงขึ้นอยู่กับ SD ของตัวแปรที่ศึกษาในแต่ละชั้นภูมิเพียงอย่างเดียว ซึ่งการแบ่งเช่นนี้เรียกว่า Neyman allocation

$$n_h \alpha S_h$$

ในแง่ sampling rate ในแต่ละชั้นภูมิจะมีความแตกต่างกันไปเมื่อใช้ optimum allocation

Disproportionate allocation

การแบ่งแบบ disproportionate allocation เป็นการแบ่งที่ผู้วิจัยกำหนดจำนวนตัวอย่างในแต่ละชั้นภูมิเอง เช่น ถ้าชั้นภูมิใดมีจำนวนประชากรน้อยมาก ก็จะเก็บตัวอย่างมามากหรืออาจจะทั้งหมด (เช่น ชั้นภูมิที่ 1 ในตารางที่ 4 ค.) เพื่อให้ชั้นภูมินั้นมีจำนวนคนเพียงพอที่จะสามารถประมาณค่าตัวแปรที่สนใจในชั้นภูมินั้นได้อย่างน่าเชื่อถือ ชั้นภูมิใดที่ผู้วิจัยมีความสนใจเป็นพิเศษก็จะเก็บตัวอย่างมามาก การแบ่งแบบนี้จึงทำให้ตัวอย่างมีลักษณะไม่เหมือนกับประชากร

เช่นเดียวกับ balanced และ optimum allocation disproportionate allocation จะทำให้ sampling rate ในแต่ละชั้นภูมิแตกต่างกัน บางชั้นภูมิก็จะมี oversample หรือ undersample เช่น สุ่มคนในชั้นภูมิที่ 1 มามากถึง 100% ทำให้ตัวอย่างมีคนในชั้นภูมิที่ 1 มากถึง 50% เทียบกับเพียง 10% ในประชากร คนในชั้นภูมิที่ 2, 3 และ 4 มีการสุ่มเพียง 10%, 10% และ 15% ตามลำดับ ทำให้ตัวอย่างมีคนในชั้นภูมิเหล่านี้น้อยคือ 15%, 20% และ 15% ตามลำดับเทียบกับ 30%, 40% และ 20% ในประชากร

ตารางที่ 3 จำนวนประชากรในแต่ละชั้นภูมิ

ชั้นภูมิ	ประชากร	
	จำนวน (N_h)	สัดส่วน ($P_h = N_h/N$)
1	$N_1 = 100$	$P_1 = 0.1$
2	$N_2 = 300$	$P_2 = 0.3$
3	$N_3 = 400$	$P_3 = 0.4$
4	$N_4 = 200$	$P_4 = 0.2$
รวม	$N = 1000$	

ตารางที่ 4 การแบ่งจำนวนตัวอย่างในแต่ละชั้นภูมิ

ก. Proportionate allocation

ชั้นภูมิ	ตัวอย่าง : Proportionate allocation		
	จำนวน (n_h)	สัดส่วน ($p_h = n_h/n$)	Sampling rate ($f_h = n_h/N_h$)
1	$n_1 = 20$	$p_1 = 0.1$	$f_1 = 0.2$
2	$n_2 = 60$	$p_2 = 0.3$	$f_2 = 0.2$
3	$n_3 = 80$	$p_3 = 0.4$	$f_3 = 0.2$
4	$n_4 = 40$	$p_4 = 0.2$	$f_4 = 0.2$
รวม	$n = 200$		$f = 0.2$

ข. Equal allocation

ชั้นภูมิ	ตัวอย่าง : Equal allocation		
	จำนวน (n_h)	สัดส่วน ($p_h = n_h/n$)	Sampling rate ($f_h = n_h/N_h$)
1	$n_1 = 50$	$p_1 = 0.25$	$f_1 = 0.5$
2	$n_2 = 50$	$p_2 = 0.25$	$f_2 = 0.167$
3	$n_3 = 50$	$p_3 = 0.25$	$f_3 = 0.125$
4	$n_4 = 50$	$p_4 = 0.25$	$f_4 = 0.25$
รวม	$n = 200$		$f = 0.2$

ค. Disproportionate allocation

ชั้นภูมิ	ตัวอย่าง : Disproportionate allocation		
	จำนวน (n_h)	สัดส่วน ($p_h = n_h/n$)	Sampling rate ($f_h = n_h/N_h$)
1	$n_1 = 100$	$p_1 = 0.5$	$f_1 = 1$
2	$n_2 = 30$	$p_2 = 0.15$	$f_2 = 0.1$
3	$n_3 = 40$	$p_3 = 0.2$	$f_3 = 0.1$
4	$n_4 = 30$	$p_4 = 0.15$	$f_4 = 0.15$
รวม	$n = 200$		$f = 0.2$

ในการสุ่มแบบชั้นภูมิ ผู้วิจัยต้องทราบก่อนทำการสุ่มว่าแต่ละคนอยู่ในชั้นภูมิใด ซึ่งเรียกว่า pre-stratification ในบางครั้งจะไม่ทราบว่าแต่ละคนอยู่ในชั้นภูมิใดจนกว่าจะได้เก็บข้อมูลเสร็จแล้ว จึงไม่สามารถทำ pre-stratification ได้ แต่อย่างไรก็ตามในการวิเคราะห์ทางสถิติผู้วิจัยสามารถทำการวิเคราะห์แยกในแต่ละชั้นภูมิได้ ซึ่งเรียกว่า post-stratification

การสุ่มแบบชั้นภูมิ มีข้อดีคือ

1) ช่วยลดความคลาดเคลื่อน (SE) ในค่าประมาณที่ได้จากตัวอย่างที่ศึกษา ถ้า stratification variable(s) มีความสัมพันธ์กับตัวแปรที่ศึกษา

2) การสุ่มตัวอย่างในแต่ละชั้นภูมิเป็นอิสระต่อกัน ซึ่งอาจจะใช้วิธีการสุ่มตัวอย่างที่เหมือนกันหรือต่างกันก็ได้

3) สามารถกำหนดขนาดตัวอย่างในแต่ละชั้นภูมิได้เอง เช่น บางชั้นภูมิที่มีจำนวนประชากรน้อย (เช่น จบเศรษฐศาสตร์, อายุ ≥50 ปี) อาจจะทำการศึกษาในประชากรทุกคน ไม่ต้องใช้การสุ่มตัวอย่าง

ตัวอย่างที่ 3

ในการศึกษาภาคตัดขวางเพื่อประมาณค่าสัดส่วนของผู้ต้องขังที่มีสุขภาพจิตต่ำกว่าเกณฑ์เฉลี่ย ประชากรคือผู้ต้องขังในเรือนจำและทัณฑสถานทั่วประเทศจำนวน 200,000 คน กลุ่มตัวอย่างมีจำนวน 5,000 คนคือเป็น 2.5% ของประชากร เนื่องจากคาดว่าสัดส่วนการมีสุขภาพจิตต่ำกว่าเกณฑ์เฉลี่ยจะแตกต่างกันในแต่ละเรือนจำ จึงแบ่งเรือนจำเป็น 4 กลุ่ม (ชั้นภูมิ) คือ เรือนจำกลาง, เรือนจำจังหวัด, เรือนจำพิเศษ และทัณฑสถานบำบัดพิเศษ ในแต่ละชั้นภูมิจะสุ่มจำนวนผู้ต้องขังมา 2.5% (proportionate allocation) โดยใช้การสุ่มแบบเป็นระบบจากรายชื่อผู้ต้องขังทั้งหมดในชั้นภูมินั้นๆ ดังนั้นการสุ่มตัวอย่างจึงเป็นการสุ่มแบบชั้นภูมิอย่างมีระบบ (stratified systematic random sampling)

ตารางที่ 5 การสุ่มผู้ต้องขังแบบชั้นภูมิ

ชั้นภูมิ	ประชากร จำนวนผู้ต้องขัง (%)	ตัวอย่าง จำนวนผู้ต้องขัง
เรือนจำกลาง	N_1 (a_1%)	$n_1 = 0.025\ N_1$
เรือนจำจังหวัด	N_2 (a_2%)	$n_2 = 0.025\ N_2$
เรือนจำพิเศษ	N_3 (a_3%)	$n_3 = 0.025\ N_3$
ทัณฑสถานบำบัดพิเศษ	N_4 (a_4%)	$n_4 = 0.025\ N_4$
รวม	200,000 (N)	5,000 (n)

4. การสุ่มตัวอย่างแบบกลุ่ม (Cluster sampling)

ในการสุ่มแบบง่าย, แบบมีระบบต้องมีรายชื่อของทุกคนในประชากร ซึ่งบางครั้งจะเป็นไปไม่ได้เนื่องจากไม่มีรายชื่อในประชากร การหารายชื่อในประชากรสามารถทำได้แต่เสียค่าใช้จ่ายมาก หรือบางครั้งมีรายชื่อแต่การสุ่มจากรายชื่อจะทำให้ได้ตัวอย่างที่กระจายกันไปในประชากร ทำให้เสียค่าใช้จ่ายในการเก็บข้อมูลมาก วิธีการสุ่มที่เหมาะสมกว่าคือทำการสุ่มเลือกกลุ่ม (cluster) ก่อน ซึ่งอาจจะต้องเลือกหลายชั้นแล้วจึงเลือกคนเป็นอันดับสุดท้าย ซึ่งเรียกว่า

multistage cluster sampling

Cluster sampling เป็นวิธีการสุ่มตัวอย่างที่ sampling unit ประกอบไปด้วยคนมากกว่า 1 คน sampling unit จึงเป็นกลุ่มของคน ตัวอย่างของ cluster เช่นโรงเรียน, ห้องเรียน, จังหวัด, อำเภอ, ตำบลและหมู่บ้าน cluster จะมีลักษณะที่ตรงข้ามกับ stratum คือ แต่ละ cluster มีความเหมือนกันในตัวแปรที่จะศึกษา (homogeneity among clusters) คนที่อยู่ภายใน cluster เดียวกันมีความแตกต่างกันในตัวแปรที่จะศึกษา (heterogeneity within cluster) ส่วน stratum มีลักษณะคือ homogeneity within stratum และ heterogeneity among strata

ในการศึกษาถึงสภาวะสุขภาพจิต (ปกติ+ดีกว่าปกติ, ไม่ปกติ) ในบุคลากรสาธารณสุขในภาคใต้ผู้วิจัยสุ่มเลือกจังหวัด 5 จังหวัด ในแต่ละจังหวัดสุ่มเลือก 3 โรงพยาบาล ในแต่ละโรงพยาบาลสุ่มเลือกบุคลากรสาธารณสุข 20 คน จังหวัด, โรงพยาบาล คือ cluster ซึ่งผู้วิจัยคิดว่าสุขภาพจิตของบุคลากรสาธารณสุขในแต่ละจังหวัด, โรงพยาบาลจะคล้ายกัน ในตัวอย่างนี้มีการสุ่ม 3 ครั้งจึงเรียกว่า 3-stage cluster sampling ในการสุ่มแต่ละครั้งมี sampling unit ต่างกันคือ จังหวัดเป็น primary sampling unit (PSU), โรงพยาบาลคือ secondary sampling unit (SSU) และบุคลากรสาธารณสุขคือ tertiary sampling unit (TSU)

โดยปกติแต่ละ cluster จะมีขนาดไม่เท่ากัน (unequal-sized cluster) เช่น แต่ละจังหวัดจะมีจำนวนโรงพยาบาลไม่เท่ากัน แต่ละโรงพยาบาลจะมีจำนวนบุคลากรสาธารณสุขไม่เท่ากัน ในบางครั้ง cluster อาจมีขนาดเท่ากัน (equal-sized cluster) เช่น บุหรี่แต่ละห่อจะมี 20 ซอง แต่ละซองมี 20 มวน การสุ่มอาจเริ่มจากการสุ่มบุหรี่ 5 ห่อ ในแต่ละห่อสุ่มบุหรี่ 2 มวน

Cluster sampling มีข้อดีคือ

1) ในทางปฏิบัติ cluster sampling จะง่ายกว่า, เสียค่าใช้จ่ายน้อยกว่าการสุ่มแบบง่าย, แบบมีระบบ

2) ใน cluster sampling ต้องการรายชื่อเฉพาะ cluster ที่ถูกสุ่มเลือก เช่น ต้องการรายชื่อบุคลากรสาธารณสุขเฉพาะโรงพยาบาลที่สุ่มได้เท่านั้น

Cluster sampling มีข้อเสียคือ

1) ค่าประมาณที่ได้จาก cluster sampling มีความคลาดเคลื่อนสูงกว่าค่าที่ได้จากการสุ่มแบบง่าย เนื่องจากคนที่อยู่ใน cluster เดียวกันมักจะมีความคล้ายกันในตัวแปรที่ศึกษา เช่น ในการสำรวจความเห็นเกี่ยวกับการใส่เครื่องแบบนักเรียนที่รัดรูป นักเรียนในโรงเรียนเดียวกันอาจมีความคิดเห็นคล้ายๆ กัน ดังนั้นในแต่ละโรงเรียนจึงไม่ควรใช้นักเรียนจำนวนมาก ซึ่งจะไม่ได้ประโยชน์เนื่องจากมีความคิดเห็นคล้ายๆ กัน แต่ควรให้มีหลายๆ โรงเรียน เช่น

ถ้าต้องการทำศึกษาในตัวอย่างนักเรียนทั้งหมด 1000 คน การสุ่มมา 10 โรงเรียนๆ ละ 100 คน จะดีกว่าการสุ่มมาเพียง 5 โรงเรียนๆ ละ 200 คน

2) การวิเคราะห์ผลทางสถิติจะยากกว่าการสุ่มแบบง่าย, แบบมีระบบ

ตัวอย่างที่ 4

ผู้วิจัยสนใจที่จะสำรวจความคิดเห็น (เห็นด้วย, ไม่เห็นด้วย) ต่อการสอบ A-net ของนักเรียนชั้นมัธยม 6 ในกรุงเทพมหานคร ถ้าใช้การสุ่มแบบง่าย หรือการสุ่มแบบมีระบบ จะต้องมีรายชื่อของนักเรียน ม. 6 ทั้งหมดใน กทม. แล้วสุ่มตัวอย่างจากรายชื่อเหล่านั้น ซึ่งไม่สะดวกในทางปฏิบัติ การสุ่มแบบ 2-stage cluster sampling จะสะดวกกว่า ซึ่งมีขั้นตอนคือ

1) จากจำนวนและรายชื่อโรงเรียนทั้งหมดใน กทม. A โรงเรียน จะสุ่มมา a โรงเรียน โดยใช้วิธีการสุ่มแบบง่าย ซึ่งจะเขียนสั้นๆ ว่า SRS (a จาก A โรงเรียน) โรงเรียนเป็น cluster โดยคาดว่านักเรียนในโรงเรียนต่างๆมีความคิดเห็นต่อการสอบ A-net คล้ายๆ กัน ส่วนนักเรียนที่อยู่ในโรงเรียนเดียวกันมีความคิดเห็นต่อการสอบต่างกัน เมื่อทุกโรงเรียนเหมือนกัน (ในตัวแปรที่จะศึกษา) จึงสุ่มเลือกมา a โรงเรียนได้ ในที่นี้โรงเรียนเป็น primary sampling unit (PSU)

2) ใน a โรงเรียนที่สุ่มมาได้ จะสุ่มเลือกมา b ห้องด้วยวิธีการสุ่มแบบง่าย ห้องเรียนเป็น cluster โดยถือว่า นักเรียนที่อยู่ในห้องเรียนเดียวกันมีความคิดเห็นต่อการสอบ A-net ต่างกัน นักเรียนแต่ละห้องมีความความคิดเห็นคล้ายๆ กัน เมื่อทุกห้องมีความคล้ายกัน (ในตัวแปรที่จะศึกษา) จึงสุ่มเลือกมา b ห้องได้ ในที่นี้ห้องเรียนเป็น secondary sampling unit (SSU) เนื่องจากแต่ละโรงเรียนที่สุ่มมาได้มีจำนวนห้อง ม.6 แตกต่างกันจึงเขียนว่าสุ่มแบบ SRS (b จาก B_α ห้อง) เมื่อ B_α คือ จำนวนห้อง ม.6 ในโรงเรียน α ที่สุ่มได้

3) ในแต่ละห้องเรียนที่สุ่มได้ จะสอบถามความคิดเห็นของนักเรียนทุกคนในห้อง

ถ้าเปลี่ยนวิธีการสุ่มตัวอย่างเป็น ในแต่ละโรงเรียนที่สุ่มได้ จะทำการศึกษาในทุกห้องของโรงเรียนนั้นๆ และในแต่ละห้องจะทำการศึกษาในนักเรียนทุกคน จะมีการสุ่มตัวอย่างเพียง 1 ครั้งเท่านั้น วิธีการสุ่มจึงเป็น 1-stage cluster sampling

ถ้าเปลี่ยนวิธีการสุ่มตัวอย่างเป็น ในแต่ละห้องที่สุ่มได้ จะสุ่มเลือกนักเรียน c คน ด้วยวิธีการสุ่มแบบง่าย วิธีการสุ่มจะกลายเป็น 3-stage cluster sampling สมมติว่าแต่ละห้องมีจำนวนนักเรียนเท่ากัน (C คน) จะเขียนได้ว่าสุ่มนักเรียนแบบ SRS (c จาก C คน) ถ้าแต่ละห้องมีจำนวนนักเรียนไม่เท่ากัน (C_α คน) จะเขียนได้ว่าสุ่มนักเรียนแบบ SRS (c จาก C_α คน)

ตัวอย่างที่ 5

ผู้วิจัยสนใจที่จะสำรวจระดับความเครียด (น้อย+ปานกลาง, มาก) ต่อการสอบ A-net ของนักเรียนชั้นมัธยม 6 ในกรุงเทพมหานคร ผู้วิจัยคิดว่าความเครียดขึ้นกับชนิดของโรงเรียน (รัฐบาล+สาธิต, เอกชน) และขนาดของโรงเรียน (เล็ก+กลาง, ใหญ่) จึงแบ่งโรงเรียนมัธยมใน กทม. ออกเป็นกลุ่ม 4 กลุ่ม (ชั้นภูมิ) ตามชนิดและขนาดของโรงเรียน คือ

1. โรงเรียนรัฐบาล+สาธิต ขนาดเล็ก+กลาง
2. โรงเรียนรัฐบาล+สาธิต ขนาดใหญ่
3. โรงเรียนเอกชน ขนาดเล็ก+กลาง
4. โรงเรียนเอกชน ขนาดใหญ่

ชนิดและขนาดของโรงเรียนจึงเป็น stratification variable ในที่นี้จึงมี stratification variable 2 ตัว หลังจากนั้นในแต่ละชั้นภูมิจะสุ่มเลือกโรงเรียนโดยใช้การสุ่มแบบง่าย เนื่องจากแต่ละชั้นภูมิมีจำนวนโรงเรียนไม่เท่ากัน เพื่อความง่าย ในแต่ละชั้นภูมิจึงสุ่มมา a โรงเรียน ซึ่งก็คือ balanced allocation หรืออาจจะสุ่มจำนวนโรงเรียนในแต่ละชั้นภูมิมาไม่เท่ากันโดยใช้ proportionate allocation เมื่อสุ่มเลือกโรงเรียนแล้วก็จะสุ่มห้องมา b ห้องและทำการศึกษาในนักเรียนทุกคนของห้องที่สุ่มได้ วิธีการสุ่มตัวอย่างจึงเป็น stratified 2-stage cluster sampling (โรงเรียน = primary sampling unit, ห้องเรียน = secondary sampling unit)

สมมติว่า กทม. มีโรงเรียนมัธยมทั้งหมด 200 โรงเรียน เมื่อแบ่งตามชนิดและขนาดของโรงเรียนจะได้ดังแสดงในสดมภ์ที่ 3 (N_h) ของตารางที่ 6 ถ้าจะทำการศึกษาในโรงเรียน 40 โรงเรียน (sampling rate = 20%) อาจให้แต่ละชั้นภูมิมีจำนวนโรงเรียนเท่ากันคือ 10 โรงเรียน (balanced allocation) หรือแบ่งจำนวนโรงเรียนแบบ proportionate allocation ก็ได้ โดยในแต่ละชั้นภูมิจะสุ่มโรงเรียนมา 20% เช่น ในชั้นภูมิที่ 1 มี 100 โรงเรียนจึงสุ่มมา 20 โรงเรียน การสุ่มแบบ proportionate allocation จะทำให้ตัวอย่างและประชากรมีสัดส่วนของโรงเรียนในแต่ละชั้นภูมิเหมือนกัน คือในชั้นภูมิที่ 1, 2, 3 และ 4 มีโรงเรียนร้อยละ 50, 20, 20 และ 10 ตามลำดับ

ตารางที่ 6 ตัวอย่างของ stratified 2-stage cluster sampling

ชั้นภูมิ	ชนิด, ขนาดของโรงเรียน	ประชากร	ตัวอย่าง Balanced allocation	ตัวอย่าง Proportionate allocation
		จำนวนโรงเรียนมัธยมใน กทม. (N_h)	จำนวนโรงเรียนมัธยมใน กทม. (n_h)	จำนวนโรงเรียนมัธยมใน กทม. (n_h)
1	(รัฐบาล+สาธิต) ขนาด(เล็ก+กลาง)	100	10	20
2	(รัฐบาล+สาธิต) ขนาดใหญ่	40	10	8
3	เอกชน ขนาด (เล็ก+กลาง)	40	10	8
4	เอกชน ขนาดใหญ่	20	10	4
รวม		N = 200	n = 40	n = 40

สรุป

ในการสุ่มตัวอย่าง ตัวอย่างที่ดีต้องเป็นตัวแทนของประชากร การเลือกวิธีการสุ่มตัวอย่างต้องคำนึงถึงความถูกต้องในทางสถิติและความเหมาะสมในทางปฏิบัติควบคู่ไปด้วย

เอกสารอ้างอิง

1. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology, the essentials. 2nd ed. Baltimore: Williams & Wilkins; 1988.
2. Kish L. Survey sampling. 1st ed. New York: John Wiley & Sons, Inc.; 1965.
3. Lemeshow S, Hosmer DW, Klar J, Lwanga SK. Adequacy of sample size in health studies. 1st ed. Chichester: John Wiley & Sons, Inc.; 1990.
4. Levy PS, Lemeshow S. Sampling of populations: Methods and applications. 3rd ed. New York: John Wiley & Sons, Inc.; 1999.