Two-Stage Cluster Sampling: General Guidance for Use in Public Heath Assessments

Introduction to Cluster Sampling

Cluster sampling involves dividing the specific population of interest into geographically distinct groups or clusters, such as neighborhoods or families. Because the information is readily available, many people use census blocks or block groups for their clusters.

A random sample of clusters is obtained, and then members of the selected clusters are then surveyed (either randomly or as a census). Contrast this with stratified sampling, in which the population is divided into distinct groups (e.g., states or ethnicities) and then random samples are obtained from each group.

A commonly used two-stage cluster sampling scheme, the "30 x 7" sample was developed by the World Health Organization with the aim of calculating the prevalence of immunized children within +/- 10 percentage points. That is, if the true prevalence was 40%, one would expect an estimate between 30% and 50% when using the 30x7 method.

This design has been adopted for other purposes such as rapid needs assessments with no (or only slight) modification. This sampling scheme is thought to be sufficient for most sampling of community health factors.

30 x 7 means that you randomly select 30 census blocks from a list from all the census blocks in your county and then randomly select seven interview sites per block. The 30x7 method is an example of what is known as a *two-stage cluster sample*. In the first stage, census blocks are randomly selected, while in the second stage, interview locations are randomly selected within each census block. Census blocks are the primary sampling units, while the random interview locations are your secondary sampling units

Census blocks may be selected in stage one through a method known as "probability proportionate to population size," which means that a census block with more households is more likely to be included than one with fewer households.

What's Inside:

Want a basic introduction to sampling?

See "A Guide to Sampling for Community Health Assessments and Other Projects" available at:

http://nccphp.sph.unc.edu/PHRST5

From the North Carolina Center for Public Health Preparedness in cooperation with North Carolina Public Health Regional Surveillance Team 5

How the Numbers of Clusters and Interviews Affect the Data

Two locations drawn randomly from within the same census block are likely to be more similar than two locations from different census blocks (remember, the goal of census blocks is to be as uniform as possible with respect to population characteristics, economic status, and living conditions). So, two locations within the same census block do not each contribute completely independent information; this is known as the "intra-cluster correlation" or ICC.

In statistical terms, this correlation always increases the variance of your estimate, which reduces the precision. As a result, it's always better to randomly select more clusters than to randomly select more points within any particular cluster. In other words: selecting an additional cluster provides more information than selecting additional points within a cluster.

The variance of an estimate is a measure of its dispersion over all possible samples. For any sampling situation, there are an extremely large number of possible samples, each one of which would produce an estimate of the value of interest. The variance gives an indication of the likely "spread" of the estimates, or the range that values that might result from all these different estimates. A lower variance or larger sample size results in greater precision.

Precision also gives you an idea of the width of the confidence interval around the estimated prevalence. With an estimated prevalence of 35% and a 95% confidence interval of 25% to 45%, a correct interpretation would be that you are 95% confident that the true (unknown) prevalence lies somewhere between 25% and 45%, but it's just as likely to be near 25% or 45% as to be near 35%. Of course, you never know the true prevalence; otherwise, there would be no reason to do the study!

Choosing the Right Number of Clusters and Interviews

Since selecting more clusters rather than more points within any cluster improves precision, using a "40 x 5" method likely yield estimates with more precision than the 30 x 7 method, even though it involves fewer total interviews (200 compared to 210). So shouldn't you always choose more clusters and fewer interviews?

If survey costs or time are important, such as during a rapid needs assessment, additional clusters may be more costly or time-consuming than additional locations within a cluster, so the reduction in sample size from 210 to 200 might not necessarily lead to improved efficiency or timeliness. For example, interviewer travel time to 40 census blocks may be much greater than travel time to 30 census blocks, which may cost more and delay reporting important results to authorities. Additionally, you may not have more than 30 census blocks in a county or an area affected by a natural disaster or an outbreak.

On the other hand, a 20 x 10 or 15 x 14 method may save time or money, but would result in a substantially less precise estimate. If it's not possible to include more than 15 or 20 census blocks in the first stage of your sample, you may need to increase the number of interview locations in the second stage to as many as 18 or more in order to achieve the same statistical precision as with a 30 x 7 design. The actual number of interview locations depends on the ICC. The higher this correlation, the larger total sample size you need. In fact, if the ICC is too high, it's not always possible to achieve the same level of precision by sampling 15 or 20 blocks as compared to 30 blocks, no matter how many locations are sampled within each block. If the ICC were as high as 0.20, you would need to sample 96 locations in 20 blocks (for a sample size of 1,920) to achieve the same precision as a 30 x 7 design!

So consider the balance of timeliness and precision in choosing your study design. If you have the time and money, choosing a two-stage cluster sampling design with more census blocks and fewer interviews per block will give you the most precise estimates of the variables you are trying to measure in your community.

Stratification and Subgroups

Stratification is the process of sorting individuals into homogeneous groups prior to sampling. Groups, or strata, should be mutually exclusive (*no* member should belong to *more than one* group) and exhaustive (*all* members should belong to *some* group). After individuals are divided into groups, sampling can be done within each group. An example of mutually exclusive and exhaustive groups would be males and females.

Typically, proportionate allocation should be used when conducting stratified sampling. If 60% of your community members are female and 40% are male, then your sample should be 60% female and 40% male. In this way, stratification ensures that all groups are sufficiently represented.

Since stratification almost always increases the precision of your estimates and narrows the confidence intervals around your estimates, it may be desirable to stratify.

This is particularly true if the stratification factor is at least moderately related to your outcome variable. An example might be the stratification of properties by designated flood zone in a rapid needs assessment following a hurricane. The statistical precision gained from stratification such as this may result in needing fewer census block clusters in your study than you would with an unstratified design

While it is statistically valid and maybe even statistically desirable to stratify, analysis of stratified data collected using a two-stage cluster sampling method can be complex. Your analysis must account for both stratification and clustering when computing your estimates and standard errors. While this is not too difficult to do in sophisticated statistical computing software (usually SAS or SUDAAN), it would require someone with special expertise in this subject area and the software.

The 30 x 7 cluster sampling method was not designed to collect stratified data, but rather to provide overall estimates for a designated assessment area. If a stratified design is used, it is always possible to obtain estimates for each individual stratum (e.g., if the design were stratified by flood plain, you would obtain an overall estimate, but could also obtain estimates within each flood plain). Of course, the se stratum-specific estimates will always be less precise than the overall estimate.

If you wish to calculate separate estimates of equal precisions for specific population subgroups using the 30 x 7 or a similar two-stage cluster design, you must obtain sufficient samples from each subgroup to achieve that level of precision; taking separate 30 x 7 samples from each subgroup may even be necessary .

Of course, it is always possible to obtain estimates within specific subgroups of your population (as long as members of that subgroup were actually sampled), even if the original sample design was not stratified according to these subgroups. For example, it will be possible to estimate the outcome proportion within census blocks that are predominantly Hispanic even if the design was not stratified by ethnicity, as long as some of the census blocks sampled are predominantly Hispanic. Once again, the estimates within any subgroup will be less precise than the overall estimate.

It is also possible to increase the precision of your overall estimate using "post-stratification" in which a stratified analysis is conducted even though the sampling design used to obtain the data was not necessarily stratified. As with stratified sampling, post-stratification will be especially useful if the stratification factor is at least moderately associated with the outcome of interest. Post-stratification can be easily accomplished by including the stratification factor as a covariate in a regression model using a statistical software package such as SAS.

The North Carolina Center for Public Health Preparedness

North Carolina Institute for Public Health
Gillings School of Global Public Health
University of North Carolina at Chapel Hill
400 Roberson Street
Campus Box 8165
Chapel Hill, NC 27599-8165

Phone: 919-843-5561
Fax: 919-843-5563
Email: nccphp@unc.edu

The North Carolina Center for Public Health Preparedness (NCCPHP) offers a variety of training activities and technical support to local and state public health agencies. NCCPHP is funded by the Centers for Disease Control and Prevention under grant/cooperative agreement number U90/424255 to improve the capacity of the public health workforce to prepare for and respond to terrorism and other emerging public health threats. The contents of this newsletter are the responsibility of the authors and do not necessarily represent the official views of CDC.

To learn more about NCCPHP's training and technical assistance, visit http://nccphp.sph.unc.edu.

Public Health Regional Surveillance Team 5

Team Leader: Steven Ramsey
Phone: 336-641-8192
E-mail: sramsey@co.guilford.nc.us

Public Health Regional Surveillance Team 5 is one of seven "PHRSTs" created by the Office of Public Health Preparedness and Response in the North Carolina Division of Public Health to provide support to local health agencies serving all 100 counties. The host counties for these regional offices are Buncombe, Mecklenburg, Guilford, Durham, Cumberland, Pitt, and New Hanover. Each team includes an epidemiologist, an industrial hygienist, a nurse consultant, and administrative specialist.

Reference

1. Henderson RH and Sundaresan T. Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. *Bulletin of the World Health Organization*, 60(2):253-260. Available at: http://whqlibdoc.who.int/bulletin/1982/Vol60-No2/bulletin_1982_60(2)_253-260.pdf