# A Guide to Sampling for Community Health Assessments and Other Projects

## Introduction

Healthy Carolinians defines a community health assessment as "a process by which community members gain an understanding of the health, concerns, and health care systems of the community by identifying, collecting, analyzing, and disseminating information on community assets, strengths, resources, and needs."

This guide will provide you with a crash course in some of the basic ideas that underlie the sampling methods used in community health assessments. We will start with some of the essentials of sampling(simple random samples, cluster sampling, census geography, and randomization), and then we'll move into some issues more specific to community health assessment (sample size issues and oversampling). Finally, we'll end with weighting, an important but often under-used statistical technique that might be helpful during your analysis.

For more information on conducting a community health assessment, refer to the *Community Assessment Guide Book* available at:

*http://www.healthycarolinians.org/pdfs/02Guidebook.pdf*

**What's Inside:**



## Let's Get Started...

Imagine you work in a county health department, and one morning your boss rushes into your office ranting about health insurance. He tells you to figure out how many adults in your county have health insurance.

You know it would be impossible to ask every adult in the county, and you need to find the best way to get a reliable estimate by talking to a smaller number of people.

Then, your boss tells you that no money was budgeted for this project, so you have to conduct your investigation as cheaply as possible.

And then he mumbles something about using census data and goes off to a meeting, leaving you a bit overwhelmed.

While you have a difficult task ahead of you, it can certainly be done. But first, let's review some basic concepts.


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

# Census Geography

You might be wondering what exactly a census block is! Census geography is the way the U.S. Census Bureau divides the county.

*Census blocks* are the smallest census unit and are formed by streets, roads, railroads, streams and other bodies of water, other visible physical and cultural features, and the legal boundaries shown on Census Bureau maps. Census blocks never cross county boundaries. Although most people intuitively think of census blocks as being rectangular or square, of about the same size, and occurring at regular intervals, in many areas of the United States, census block configurations actually are quite different. Patterns, sizes, and shapes of census blocks vary within and between areas. Factors that influence the overall configuration of census blocks include topography, the size and spacing of water features, the land survey system, and the extent, age, type, and density of urban and rural development.

*Block groups* are clusters of census blocks. Block groups usually include between 200 and 600 housing units (between 600 and 3,000 people, with an ideal size of about 1,500 people).

*Census tracts* are sets of contiguous census blocks and block groups; they are relatively permanent geographic entities within counties. Generally, census tracts have between 2,500 and 8,000 (the average is about 4,000) residents and boundaries that follow visible features. When first established, census tracts are to be as uniform as possible with respect to population characteristics, economic status, and living conditions. In other words, people who live within a tract should more similar to one another than to those who live in another census block.

We will now look at randomization, which is a crucial part of sampling.

*Adapted from http://www.census.gov/geo/www/GARM/Ch11GARM.pdf and http://www.census.gov/geo/www/cob/bg_metadata.html*

# Randomization

**Why randomize?**

When carried out correctly, randomization helps to ensure that the *sample population* (the people interviewed) represents the *target population* (the community as a whole). Sample sizes that are large enough are important to allow the process of randomization to work (i.e. to generate a sample population that accurately represents the target population).

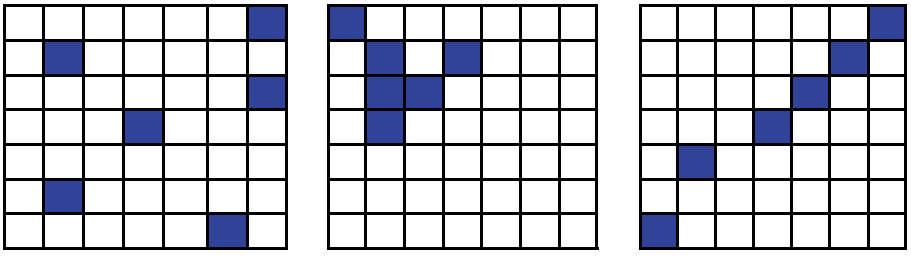
**Which is the <u>most</u> random?**

The grids to the right each contain six randomly selected squares, shown in blue. Even though the squares in the first box look more "random," each square in each grid had the same probability of being selected, so each grid reflects a random pattern; it just so happened that the first grid ended up with the squares spread out and that the second and third grids did not. Even though it may not look like it, these squares were all randomly selected; the first grid is not more random than the others!

**Definition of Random:**
being or relating to a set or to an element of a set each of whose elements has equal probability of occurrence

*From the Merriam-Webster Dictionary*

Now that you understand the basics of census geography and randomization, you need to think about what type of sampling you are going to conduct.



# Simple Random Samples

You may have heard the term, "simple random sample." Like the name suggests, it is a relatively simple concept!

The first step in selecting a simple random sample is to list and then assign a number to each element (in the case of the health insurance sample, each adult in your county). The next step is to choose adults in a way that ensures that each adult has exactly the same chance of being selected. Choosing adults could be done with a computer, a random number table, or certain types of calculators.

Simple random samples are not always feasible because they can be expensive and not terribly practical – you would need a list of every single adult in the whole county before you could even get started. For this and a few other reasons, other study designs can be used, although these other study designs usually incorporate a random element—such as the way census blocks are selected or how households are selected within the census block.

Next let's look carefully at another type of sampling: cluster sampling.

# Cluster Sampling

Cluster sampling involves dividing the specific population into geographically distinct groups or clusters, such as neighborhoods or villages. Because the information is widely available, many people use census blocks for their clusters. A sample of clusters is then chosen, and everyone (or in the case of a two stage cluster sample, a selection of people within those clusters) is included in the survey.

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. 30 x 7 means that you randomly select 30 census blocks from all of the census blocks in your county and then randomly select seven interview sites per block.

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.

The 30 x 7 method is an example of what is known as a *two-stage cluster sample*. At the first stage, census blocks are randomly selected, while at the second stage, interview locations are randomly selected within each census block. Census blocks in stage one may be selected through a method known as "probability proportionate to size," which means that a census block with more households is more likely to be included than one with fewer households.

Two interview 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 areas 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; some of the information is redundant.

In statistical terms, this type of 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.

## Sample Size Concerns

One of the first questions you'll need to answer when performing a CHA is: *how many people do you need to include in your sample?*

Deciding how large a sample size you need boils down to two factors: reliability and cost. As a rule of thumb, a larger sample size will increase the reliability of your estimates. If you only ask two people how many vegetables they eat and they both say 5 servings a day, can you assume that everyone in your county eats 5 servings of vegetables a day?

Of course not! A sample size that small will not come up with reliable estimates. (Be sure not to confuse *reliable* with *valid*; everyone may not be honest about how many vegetables they eat, but that is a problem with the way in which vegetable consumption is being measured, not the estimate's *reliability*.)

The prevalence of a characteristic in your population will affect reliability of your estimate. If only a small number of people in the population have the characteristic that you are interested in, you will have to interview more people to have reliable estimates.

For example, HIV prevalence will probably be low in your community. Therefore, a study of HIV will require more interviews to be reliable than a study of a more common characteristic. Prior knowledge of the prevalence of a characteristic can be a helpful tool when calculating how many people you need to interview (sample size calculations are often done using computer software such as Epi Info.)

This is where it gets tricky! If you're collecting information on a lot of different characteristics (how many people are uninsured, how many people are worried about drunk driving, how many people know daycare is available, etc.), how can you tell how many people you need to interview?

One strategy is to select a few of the characteristics you are most interested in, calculate how many people you would need for each, and choose the largest number calculated. Unfortunately, as the sample size increases, so does the cost of conducting the study.

## Oversampling

Oversampling is done to make sure you have enough information from a particular population subgroup to generate statistically reliable estimates of their characteristics.

Oversampling means that you interview more people from a particular subgroup than you normally would. Unlike with a simple random sample, certain groups are sampled with higher selection probabilities than others. The share of the oversampled group in the sample is greater than its share in the population from which it was drawn. Why? If you don't collect information from enough members of the group, you may not be able to generate reliable estimates for that subpopulation.

For example, say 5% of your county is Latino and you also want to be able to analyze the Latino population independently. So instead of a sample that is 5% Latino, you might need one that is 15% Latino.

When analyzing data for the entire county, the unequal selection probabilities will require using weights to remedy the imbalance, which we'll discuss on the next page.

For example, the Census Bureau oversampled low-income households in 1996 by determining which households were likely to be low-income, then sampling these low-income households at 1.66 times the rate of high-income households.

This oversampling produced an 18 percent increase in the number of households in and near poverty, with increases up to 24 percent in some subgroups, such as Black and Hispanic households in poverty. These increases strengthened the ability of the analysis to detect important factors in these subgroups. However, sample sizes for higher income and age groups were reduced.

*Adapted from http://www.sipp.census.gov/sipp/oversample.html*

# Weighting

Weighting addresses the different probabilities that certain households are selected as part of the sample. So why would there be any difference in the chance a household has of being selected?

One reason was mentioned above: deliberate oversampling. In the example from the Census Bureau, there is a greater proportion of low-income households in the sample than in the population. You must weight the sample to correct the proportions before doing any population-level calculations.

Another reason for different selection probabilities is the cluster sampling scheme often used. If you choose 10 households from each census block selected, but Block 1 has 3,000 households and Block 2 has 7,000 households, a household in Block 1 has a higher probability of being selected than Block 2.

Let's look at the math:

10 households from Block 1: 10/3000; chance of being selected = 1 out of 300

10 households from Block 2: 10/7000; chance of being selected = 1 out of 700

When households are selected with non-equal selection probabilities, the data need to be **weighted** during the analysis. How? Weights are calculated in ratio to the inverse of the probability of selection. If household X has half the chance of selection of household Y has, then household X will be given a weight twice as large as that of Y.

We'll look at an example using the census blocks in Table 1 below.

The three census blocks cover about 14,500 households, but only 30 will be selected for the survey. To be representative of the 14,500, the 30 households must be weighted using the numbers in the far right column.

To calculate the mean income of your county, you multiply the mean income for each tract by its weight, then divide by the weight for the total population:

[(T1 mean income x T1 weight) + (T2 mean income x T2 weight) + (T3 mean income x T3 weight)] / Total weight =

[(43,193 x 300) + (41,616 x 700) + (40,252 x 450)] / 1450 = 41,519

**Table 1. Calculation of Mean Income with Selection Probability Weights**

| Tract # | Mean income | Households per tract | # Households selected | Selection probability | Weight |
|---|---|---|---|---|---|
| T1 | 43,193 | 3,000 | 10 | 10/3,000 = 0.0033 | 300 |
| T2 | 41,616 | 7,000 | 10 | 10/7,000 = 0.0014 | 700 |
| T3 | 40,252 | 4,500 | 10 | 10/4,500 = 0.0022 | 450 |
| TOTAL | | 14,500 | 30 | | 1450 |

*Adapted from http://www.napier.ac.uk/depts/fhls/peas/theoryweighting.asp*

# Generalizability

When sampling is carried out correctly, the information collected from the sample population can be generalized to the target population, meaning what is true for the people you interviewed is also (roughly) true for the rest of the population you are studying.

There is, however, an important limitation to generalization – you can't generalize your findings to units other than those from which you sampled.

When a sample has been properly drawn from the county, you can draw conclusions such as, "Approximately 30% of the people in our county do not have health insurance."

However, you cannot use this data to claim that 30% of the people in a certain neighborhood or in the entire state don't have health insurance.

## 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.

# Conclusion

Now that you've reviewed these concepts, you should be well on your way to assessing the health insurance situation in your county.

Sampling is difficult work, but if done correctly, it can give you very valuable information about your population of interest.

For more information on two-stage sampling,

view "Two-Stage Cluster Sampling: General Guidance for Use in Public Heath Assessments" available at

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