ONLINE MATERIAL THAT ACCOMPANIES CHAPTER 4

Box 1. Sampling the right way and the wrong way: why we need a simple random sample

Question: Suppose we wish to estimate the proportion of individuals with obesity in New York City. We have funds to collect data on 500 individuals. We decide to take a snapshot of New York City, sampling individuals at a given moment in time without following them forward. Thus, our aim is to estimate the proportion of people with obesity by taking a single sample of people without regard to their obesity status and measuring the proportion in that sample. Given that there are more than 8 million individuals residing in New York City, each individual has a very low probability of being selected for the sample. We decide to take our sample by sending field representatives to the neighborhoods surrounding the hospital where we work. They are able to recruit 500 people into the study within a four-block radius of the hospital. Is this a problematic way to take a sample?

Answer: Yes, this is a problematic method to take a sample. The individuals within a four-block radius of the hospital may not be representative of the general population of New York City. If the intention is to infer from the sample the proportion of individuals with obesity in New York City, then the sample needs to reflect the full range of the distribution of obesity status in New York City. We will not achieve an effective random sample of the population of New Yorkers by collecting our sample from a specific geographic area.

We illustrate this below with three basic figures. The first is a population (Box 1, Figure 1). There are two characteristics that define this population, denoted by X (color) and Y (dots). We can imagine that these characteristics may be factors such as age, sex, or geographic location. In this population, the proportion of individuals with black color and dots is shown in Box 1, Figure 1. as follows:

Box 1, Figure 1. A Sample Population

Proportion with black: 30/60 = 50%
Proportion with dots: 15/60 = 25%

We also see that individuals with dots are clustered in the top left-hand corner of the figure.

Taking a sample: the wrong way

Suppose we have funds to take a sample of 15 individuals from this population (Box 1, Figure 2). If we choose from the upper left hand corner (say, the area around the hospital), we would find the following estimates for the proportion with dots and the proportion black:

Box 1, Figure 2. Incorrect Sampling Method

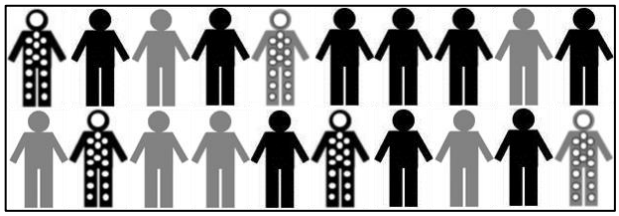
Proportion with black: 5/15 = 33.33%
Proportion with dots: 11/15 = 73.33%

Thus, we would underestimate the proportion of individuals who are black and overestimate the proportion with dots.

Taking a sample: the right way

We randomly select 20 individuals using a random number generator (Box 1, Figure 3). Beginning with the top left individual and proceeding right we assign each individual a random number. If the random number ends in 1, 2, or 3, we will select them for the sample. Each individual selected for the sample is shown with a box around it. Now, when we examine the estimates for the proportion with dot and the proportion black, we find:

Box 1, Figure 3. Correct Sampling Method



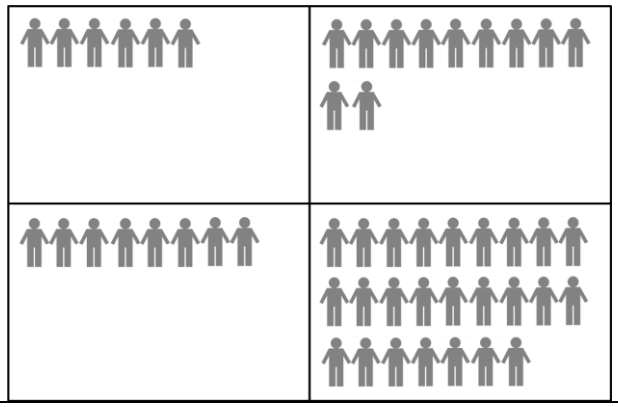
Proportion with black: 10/20 = 50%
Proportion with dots: 5/20 = 25%

We see that the estimates in the sample are now in line with the true population proportion.

Box 2. Beyond the simple random sample: other ways to take a sample

A simple random sample is conceptually the most straightforward way to take a sample from an underlying population. But often, a simple random sample is neither efficient nor effective. Why would this be? Consider, for example, conducting a study where the population of interest is the population of the United States. We have sufficient funds to interview 20,000 people. The population size of the US based on the 2010 census is 308,745,538. Thus, our sample is only going to capture 20,000/308,745,538 =0.0000648 or 0.00648% of the population. It is unlikely that based on only a simple random sample we will ensure that there would be sufficient numbers of people in every US state and from remote rural areas as well as large urban areas to truly have a representative sample of people in the US. Because of this, many large studies use stratified random sampling in order to achieve representativeness of the sample. We illustrate a simple example of stratified random sampling in Box 2, Figure 1.

Box 2, Figure 1. Stratified Random Sampling



In the population depicted in Box 2, Figure 1, we have 50 individuals, but they are not evenly distributed throughout the space of the figure. If we took a simple random sample of this population, say, a sample of size 16, we would collect more individuals from the lower left hand corner than the upper right hand corner, because there are more individuals in that corner. If each individual has an equal probability of selection, more people from the lower left will be selected than from the upper right. One solution would be to stratify the population into four quadrants (see Box 2, Figure 1) and then take a simple random sample of equal size within each quadrant. For example, we could collect a sample of four individuals from each quadrant, thereby still obtaining our sample size of 16 but ensuring that they are representative of all four corners of the underlying population.

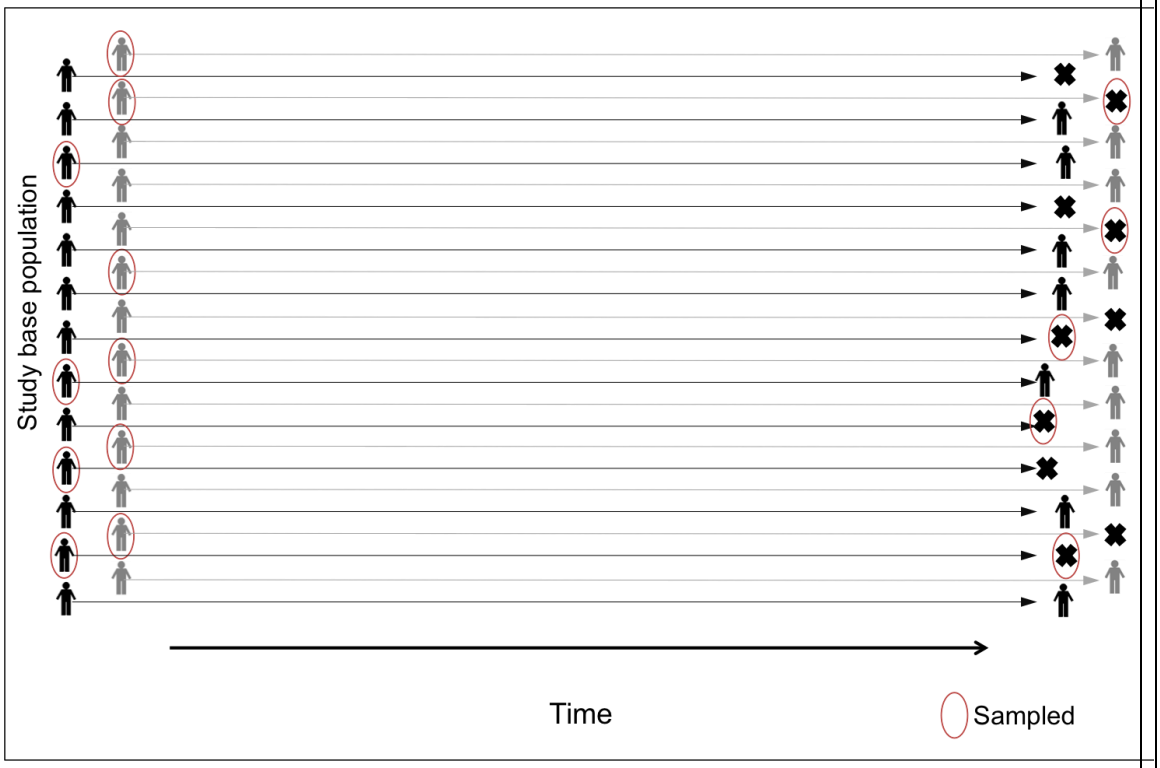
Box 3. Sampling based on health indicator status: approaches

All studies in which individuals are sampled based on health indicator status (presence or absence of the health indicator in question, or cases and controls, respectively) are founded on the idea that individuals with and without the health indicator are sampled from an underlying cohort or base of individuals moving through time. In this text we focus on the most common approach to sampling on health indicator status, which is sampling individuals with the health indicator of interest, and simultaneously sample a group of individuals who have survived without developing the health indicator of interest to the time that the cases are sampled. This type of sampling is often described as cumulative incidence case control or survivor sampling. There are two other common approaches, however, to sampling based on health indicator status in a case control study. Understanding these other approaches to sampling based on health indicator status is helpful in approaching the concept of sampling based on health indicator status.

Sample controls at the beginning of a follow-up period

One way in which to sample controls is to select them at the beginning of a follow-up period. Remember that in prospective cohort studies, all individuals are free of the health indicator at the start of the follow-up period, and all are at-risk of developing the health indicator of interest. Therefore, all individuals in the population base for which follow-up will occur can be considered for the control group. Box 1, Figure 1 below shows a schematic of the approach.

Box 3, Figure 1. Example of control sampling at the beginning of a follow-up period



When controls are selected at the beginning of the follow-up period, they may later develop the health indicator of interest and be eligible for the case group. Thus, an individual in the study can serve as both a case and a control in this approach. The third person in Box 1, Figure 1 has this quality. This person was disease free at the beginning of the follow-up period, and was therefore eligible, and selected, for the control group. This individual later developed the health indicator of interest, and was therefore eligible, and selected, for the case group.

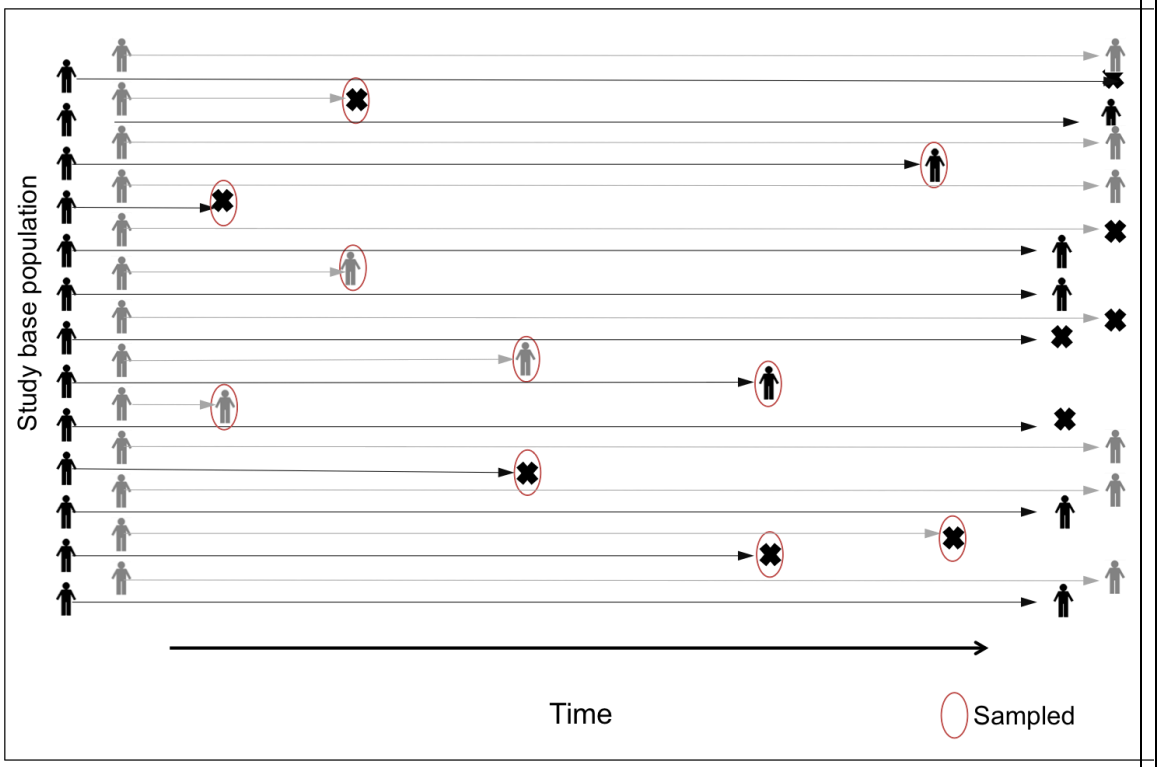
This type of sampling goes by several names, include case-cohort, case-base, and base sampling. It is most commonly implemented within ongoing cohort studies, and is therefore considered a nested type of case-control study, since it is a study nested within a larger study. Why would we want to conduct this type of study? Simply, it is an efficient way to obtain denominator information about the risk of the health indicator in the underlying population base. In this type of sampling strategy, the outcome need not be rare in the underlying population base in order to obtain a valid estimate of the association between exposure and health indicator. Therefore, within one ongoing cohort study, many different nested studies can be efficiently conducted without using resources to obtain the necessary data on the full cohort.

Sample controls at the time the case arises

A third way to obtain controls is to sample one or more individuals without the health indicator of interest at the time that the case arises. This approach is shown in Box 3, Figure 2. We begin with a cohort of individuals who are being followed prospectively. When a case arises, one or more controls are selected at that time among those without the health indicator in question.

This type of sampling is typically referred to as risk set or incidence density sampling.

Box 3, Figure 2. Example of control sampling at the time the case arises



This type of sampling shared several similarities with the approach of selecting controls at the beginning of a follow-up period. First, both require a sample that is being followed prospectively followed, therefore this design is considered a nested approach. The difference is that in the case-based approach, controls are selected at the beginning of the follow-up period; in the risk set approach, controls are selected among those without the health indicator of interest at the time that the case arises. Second, a single individual can be both a control and a case. Remember that we select controls among those who have not developed the outcome of interest at the time that the case comes to clinical attention. An individual who is selected to be a control at one time point may develop the health indicator of interest at a later time and then be eligible for the case group. The advantage of risk set sampling is that the denominator of our measures of association become a sample of the person time contribution of the underlying population base, and therefore our measure of association becomes are more rigorous estimate of the rate of disease. Therefore, again similar to selecting controls at the beginning of the follow-up period, the health indicator need not be rare in the population base in order to obtain a valid estimate of the association between exposure and the health indicator.

In summary, we can conceptualize control selection as obtaining a sample of either: a) the percentage of non-cases in the underlying population base at the end of a hypothetical follow-up period from which the cases are sampled (traditional cumulative incidence sampling); b) the overall risk of the health indicator in the underlying population base (case-cohort sampling); c) the distribution of person-time among exposed and unexposed in the underlying population base (incidence density sampling).