

[Home](#) »

Sampling

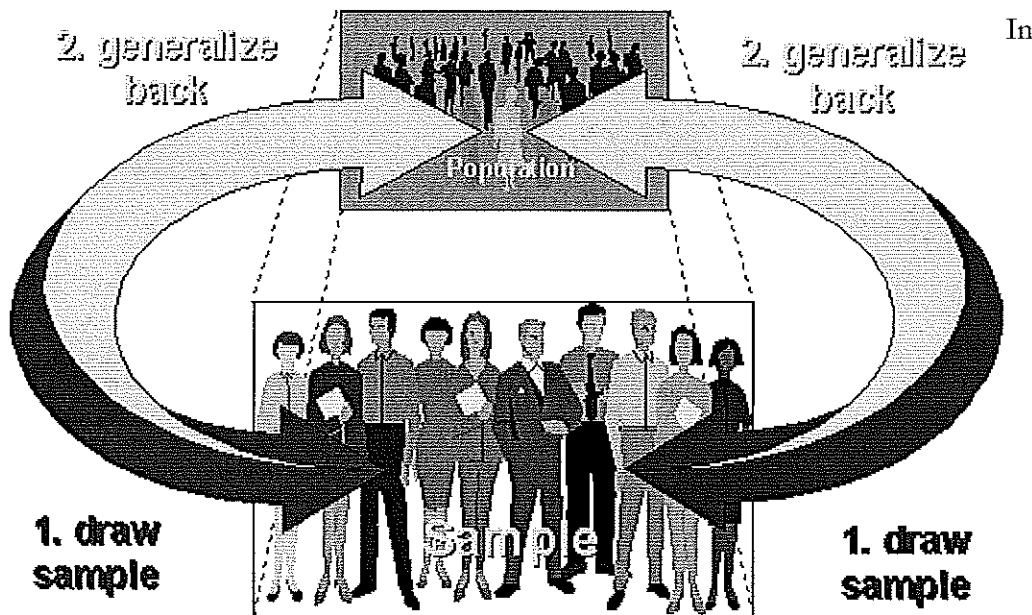
Sampling is the process of selecting units (e.g., people, organizations) from a population of interest so that by studying the sample we may fairly generalize our results back to the population from which they were chosen. Let's begin by covering some of the key terms in sampling like "population" and "sampling frame." Then, because some types of sampling rely upon quantitative models, we'll talk about some of the statistical terms used in sampling. Finally, we'll discuss the major distinction between probability and Nonprobability sampling methods and work through the major types in each.

Copyright ©2006, William M.K. Trochim, All Rights Reserved
Purchase a printed copy of the Research Methods Knowledge Base
Last Revised: 10/20/2006

[Home](#) » [Sampling](#) »

External Validity

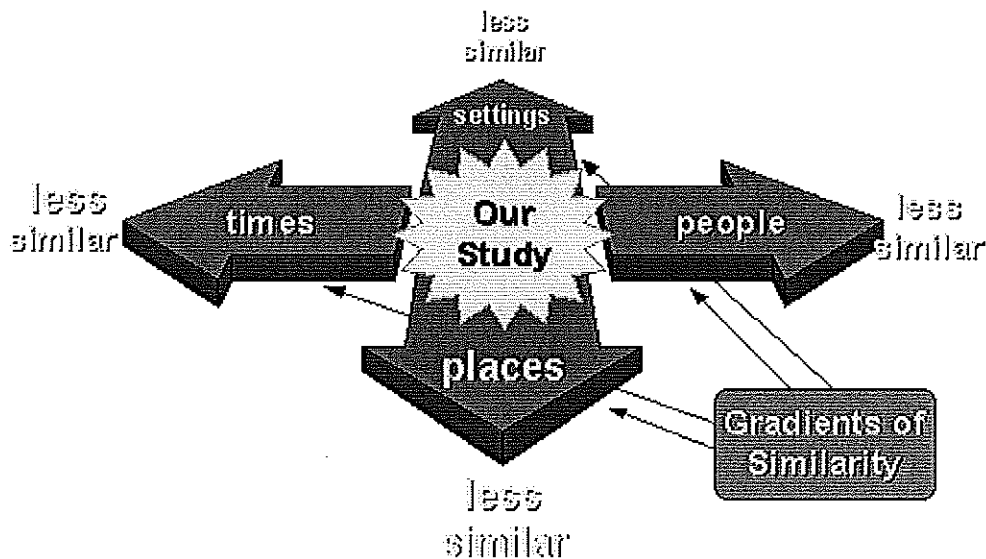
External validity is related to generalizing. That's the major thing you need to keep in mind. Recall that validity refers to the approximate truth of propositions, inferences, or conclusions. So, *external* validity refers to the approximate truth of conclusions the involve generalizations. Put in more pedestrian terms, external validity is the degree to which the conclusions in your study would hold for other persons in other places and at other times.



science there are two major approaches to how we provide evidence for a generalization. I'll call the first approach the **Sampling Model**. In the sampling model, you start by identifying the population you would like to generalize to. Then, you draw a fair sample from that population and conduct your research with the sample. Finally, because the sample is representative of the population, you can automatically generalize your results back to the population. There are several problems with this approach. First, perhaps you don't know at the time of your study who you might ultimately like to generalize to. Second, you may not be easily able to draw a fair or representative sample. Third, it's impossible to sample across all times that you might like to generalize to (like next year).

I'll call the second approach to generalizing the **Proximal Similarity Model**. 'Proximal' means 'nearby' and 'similarity' means... well, it means 'similarity'. The term *proximal similarity* was suggested by Donald T. Campbell as an appropriate relabeling of the term *external validity* (although he was the first to admit that it probably wouldn't catch on!). Under this model, we begin by thinking about different generalizability contexts and developing a theory about which contexts are more like our study and which are less so. For instance, we might imagine several settings that have people who are more similar to the people in our study or people who are less similar. This also holds for times and places. When we place different contexts in terms of their relative similarities, we can call this implicit theoretical a *gradient of*

similarity. Once we have developed this proximal similarity framework, we are able to generalize. How? We conclude that we can generalize the results of our study to other persons, places or times that are more like (that is, more proximally similar) to our study. Notice that here, we can never generalize with certainty -- it is always a question of more or less similar.



Threats to External Validity

A threat to external validity is an explanation of how you might be wrong in making a generalization. For instance, you conclude that the results of your study (which was done in a specific place, with certain types of people, and at a specific time) can be generalized to another context (for instance, another place, with slightly different people, at a slightly later time). There are three major threats to external validity because there are three ways you could be wrong -- people, places or times. Your critics could come along, for example, and argue that the results of your study are due to the unusual type of people who were in the study. Or, they could argue that it might only work because of the unusual place you did the study in (perhaps you did your educational study in a college town with lots of high-achieving educationally-oriented kids). Or, they might suggest that you did your study in a peculiar time. For instance, if you did your smoking cessation study the week after the Surgeon General issues the well-publicized results of the latest smoking and cancer studies, you might get different results than if you had done it the week before.

Improving External Validity

How can we improve external validity? One way, based on the sampling model, suggests that you do a good job of drawing a sample from a population. For instance, you should use random selection, if possible, rather than a nonrandom procedure. And, once selected, you should try to assure that the respondents participate in your study and that you keep your dropout rates low. A second approach would be to use the theory of proximal similarity more effectively. How? Perhaps you could do a better job of describing the ways your contexts and others differ, providing lots of data about the degree of similarity between various groups of people, places, and even times. You might even be able to map out the degree of

proximal similarity among various contexts with a methodology like concept mapping. Perhaps the best approach to criticisms of generalizations is simply to show them that they're wrong -- do your study in a variety of places, with different people and at different times. That is, your external validity (ability to generalize) will be stronger the more you **replicate** your study.

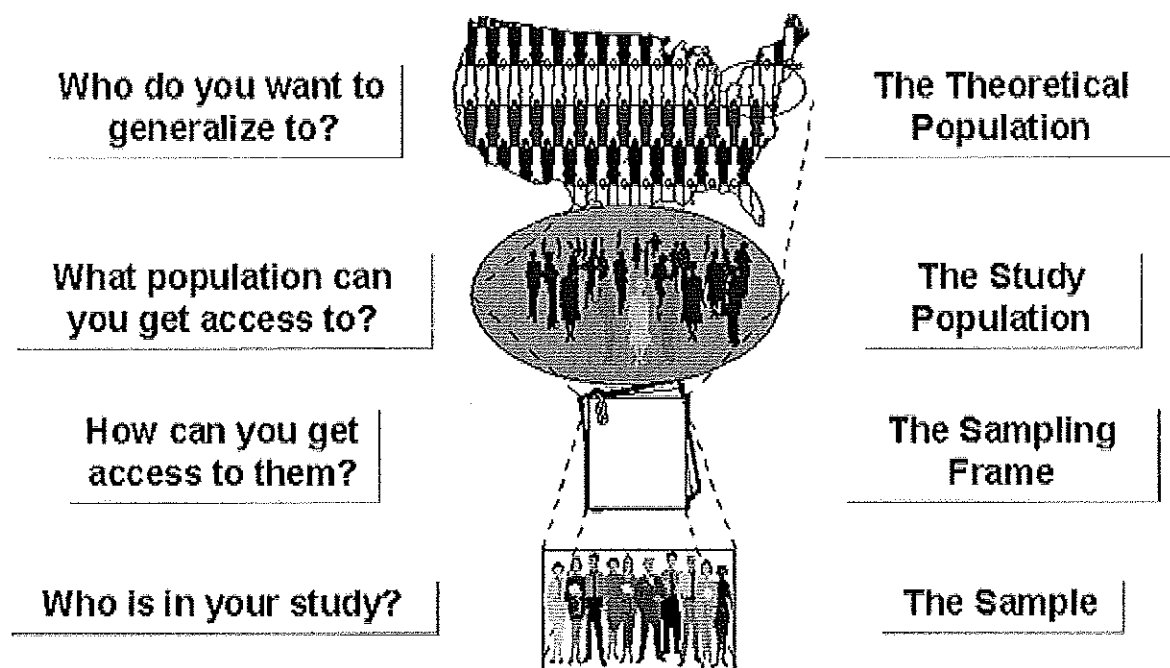
Copyright ©2006, William M.K. Trochim, All Rights Reserved
Purchase a printed copy of the Research Methods Knowledge Base
Last Revised: 10/20/2006

[Home](#) » [Sampling](#) »

Sampling Terminology

As with anything else in life you have to learn the language of an area if you're going to ever hope to use it. Here, I want to introduce several different terms for the major groups that are involved in a sampling process and the role that each group plays in the logic of sampling.

The major question that motivates sampling in the first place is: "Who do you want to generalize to?" Or should it be: "To whom do you want to generalize?" In most social research we are interested in more than just the people who directly participate in our study. We would like to be able to talk in general terms and not be confined only to the people who are in our study. Now, there are times when we aren't very concerned about generalizing. Maybe we're just evaluating a program in a local agency and we don't care whether the program would work with other people in other places and at other times. In that case, sampling and generalizing might not be of interest. In other cases, we would really like to be able to generalize almost universally. When psychologists do research, they are often interested in developing theories that would hold for all humans. But in most applied social research, we are interested in generalizing to specific groups. The group you wish to generalize to is often called the **population** in your study. This is the group you would like to sample from because this is the group you are interested in generalizing to. Let's imagine that you wish to generalize to urban homeless males between the ages of 30 and 50 in the United States. If that is the population of interest, you are likely to have a very hard time developing a reasonable sampling plan. You are probably not going to find an accurate listing of this population, and even if you did, you would almost certainly not be able to mount a national sample across hundreds of urban areas. So we probably should make a distinction between the population you would like to generalize to, and the population that will be accessible to you. We'll call the former the **theoretical population** and the latter the **accessible population**. In this example, the accessible population might be homeless males between the ages of 30 and 50 in six selected urban areas across the U.S.



Once you've identified the theoretical and accessible populations, you have to do one more thing before you can actually draw a sample -- you have to get a list of the members of the accessible population. (Or, you have to spell out in detail how you will contact them to assure representativeness). The listing of the accessible population from which you'll draw your sample is called the **sampling frame**. If you were doing a phone survey and selecting names from the telephone book, the book would be your sampling frame. That wouldn't be a great way to sample because significant subportions of the population either don't have a phone or have moved in or out of the area since the last book was printed. Notice that in this case, you might identify the area code and all three-digit prefixes within that area code and draw a sample simply by randomly dialing numbers (cleverly known as *random-digit-dialing*). In this case, the sampling frame is not a list *per se*, but is rather a procedure that you follow as the actual basis for sampling. Finally, you actually draw your sample (using one of the many sampling procedures). The **sample** is the group of people who you select to be in your study. Notice that I didn't say that the sample was the group of people who are actually *in* your study. You may not be able to contact or recruit all of the people you actually sample, or some could drop out over the course of the study. The group that actually completes your study is a subsample of the sample -- it doesn't include nonrespondents or dropouts. The problem of nonresponse and its effects on a study will be addressed elsewhere.

People often confuse what is meant by random selection with the idea of random assignment. You should make sure that you understand the distinction between random selection and random assignment.

At this point, you should appreciate that sampling is a difficult multi-step process and that there are lots of places you can go wrong. In fact, as we move from each step to the next in identifying a sample, there is the possibility of introducing systematic error or **bias**. For instance, even if you are able to identify perfectly the population of interest, you may not have access to all of them. And even if you do, you may not have a complete and accurate enumeration or sampling frame from which to select. And, even if you do, you may not draw the sample correctly or

accurately. And, even if you do, they may not all come and they may not all stay. Depressed yet? This is a very difficult business indeed. At times like this I'm reminded of what Donald Campbell used to say (I'll paraphrase here): "Cousins to the amoeba, it's amazing that we know anything at all!"

Copyright ©2006, William M.K. Trochim, All Rights Reserved

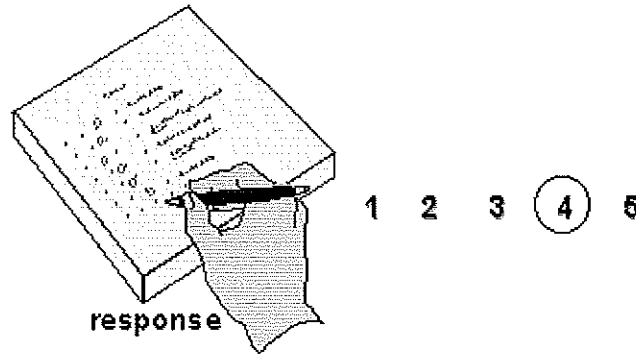
Purchase a printed copy of the Research Methods Knowledge Base

Last Revised: 10/20/2006

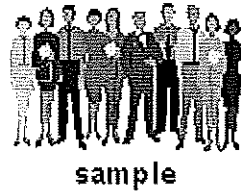
[Home](#) » [Sampling](#) »

Statistical Terms in Sampling

Variable

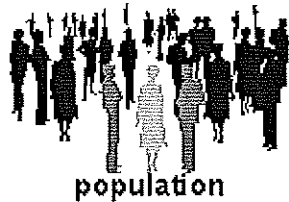


Statistic



Average = 3.75

Parameter



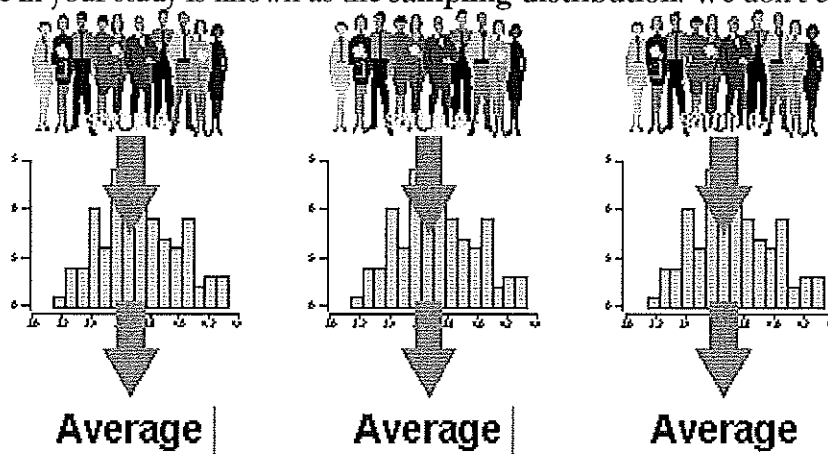
Average = 3.72

Let's begin by defining some very simple terms that are relevant here. First, let's look at the results of our sampling efforts. When we sample, the units that we sample -- usually people -- supply us with one or more responses. In this sense, a **response** is a specific measurement value that a sampling unit supplies. In the figure, the person is responding to a survey instrument and gives a response of '4'. When we look across the responses that we get for our entire sample, we use a **statistic**. There are a wide variety of statistics we can use -- mean, median, mode, and so on. In this example, we see that the mean or average for the sample is 3.75. But the reason we sample is so that we might get an estimate for the population we sampled from. If we could, we would much prefer to measure the entire population. If you measure the entire population and calculate a value like a mean or average, we don't refer to this as a statistic, we call it a **parameter** of the population.

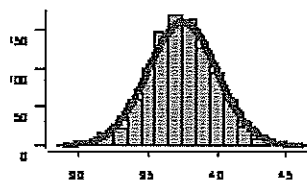
The Sampling Distribution

So how do we get from our sample statistic to an estimate of the population parameter? A crucial midway concept you need to understand is the **sampling distribution**. In order to understand it, you have to be able and willing to do a thought experiment. Imagine that instead of just taking a single sample like we do in

a typical study, you took three independent samples of the same population. And furthermore, imagine that for each of your three samples, you collected a single response and computed a single statistic, say, the mean of the response. Even though all three samples came from the same population, you wouldn't expect to get the exact same statistic from each. They would differ slightly just due to the random "luck of the draw" or to the natural fluctuations or vagaries of drawing a sample. But you would expect that all three samples would yield a similar statistical estimate because they were drawn from the same population. Now, for the leap of imagination! Imagine that you did an *infinite* number of samples from the same population and computed the average for each one. If you plotted them on a histogram or bar graph you should find that most of them converge on the same central value and that you get fewer and fewer samples that have averages farther away up or down from that central value. In other words, the bar graph would be well described by the *bell curve* shape that is an indication of a "normal" distribution in statistics. The distribution of an infinite number of samples of the same size as the sample in your study is known as the **sampling distribution**. We don't ever actually



**The Sampling
Distribution...**



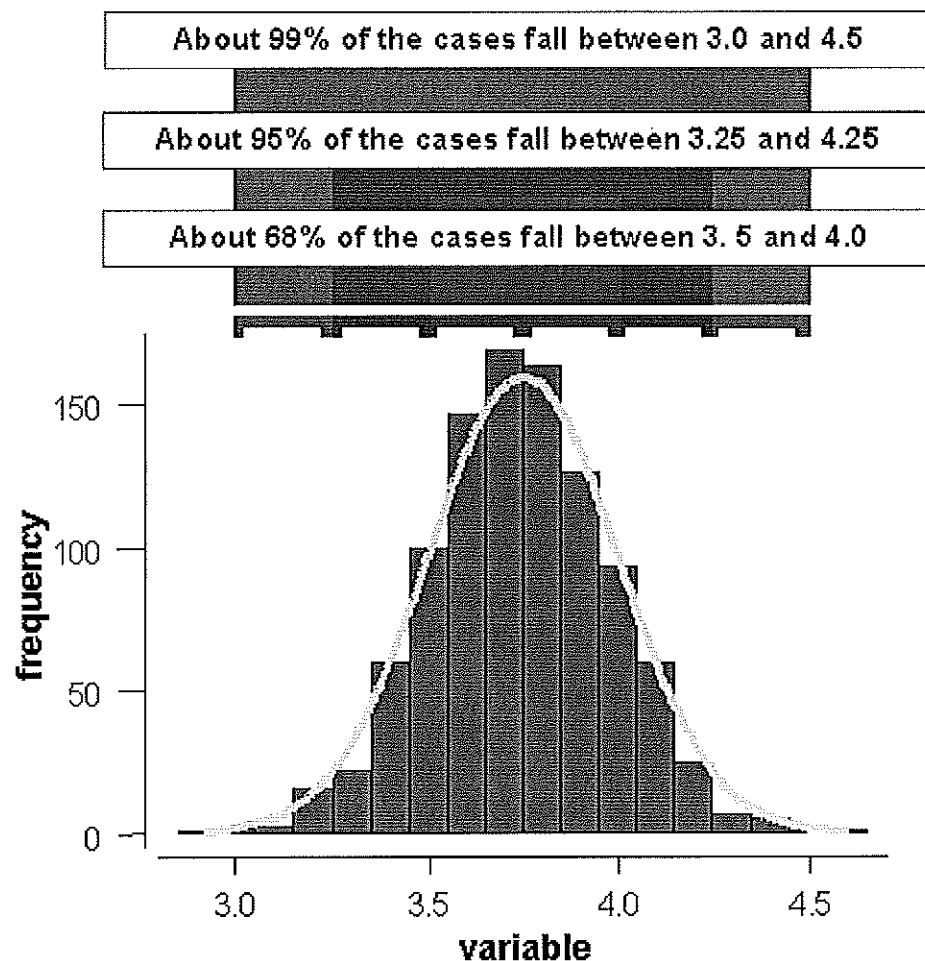
**...is the distribution
of a statistic
across
an infinite number
of samples**

construct a sampling distribution. Why not? You're not paying attention! Because to construct it we would have to take an *infinite* number of samples and at least the last time I checked, on this planet infinite is not a number we know how to reach. So why do we even talk about a sampling distribution? Now that's a good question! Because we need to realize that our sample is just one of a potentially infinite number of samples that we could have taken. When we keep the sampling distribution in mind, we realize that while the statistic we got from our sample is probably near the center of the sampling distribution (because most of the samples would be there) we could have gotten one of the extreme samples just by the luck of the draw. If we take the average of the sampling distribution -- the average of the averages of an infinite number of samples -- we would be much closer to the true population average -- the parameter of interest. So the average of the sampling distribution is essentially equivalent to the parameter. But what is the standard deviation of the sampling distribution (OK, never had statistics? There are any number of places on the web where you can learn about them or even just brush up if you've gotten rusty. This isn't one of them. I'm going to assume that you at least

know what a standard deviation is, or that you're capable of finding out relatively quickly). The standard deviation of the sampling distribution tells us something about how different samples would be distributed. In statistics it is referred to as the **standard error** (so we can keep it separate in our minds from standard deviations. Getting confused? Go get a cup of coffee and come back in ten minutes...OK, let's try once more... A standard deviation is the spread of the scores around the average in a single sample. The standard error is the spread of the averages around the average of averages in a sampling distribution. Got it?)

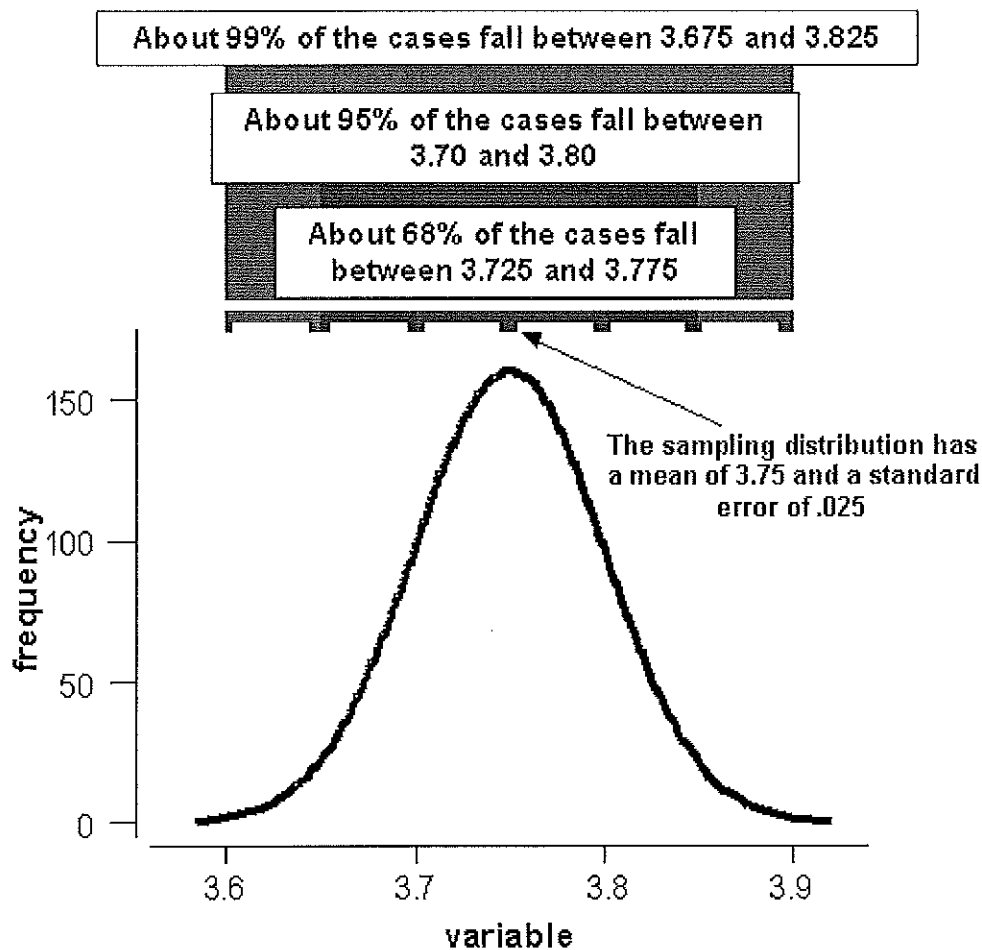
Sampling Error

In sampling contexts, the standard error is called **sampling error**. Sampling error gives us some idea of the precision of our statistical estimate. A low sampling error means that we had relatively less variability or range in the sampling distribution. But here we go again -- we never actually see the sampling distribution! So how do we calculate sampling error? We base our calculation *on the standard deviation of our sample*. The greater the sample standard deviation, the greater the standard error (and the sampling error). The standard error is also related to the sample size. The greater your sample size, the smaller the standard error. Why? Because the greater the sample size, the closer your sample is to the actual population itself. If you take a sample that consists of the entire population you actually have no sampling error because you don't have a sample, you have the entire population. In that case, the mean you estimate is the parameter.



The 68, 95, 99 Percent Rule

You've probably heard this one before, but it's so important that it's always worth repeating... There is a general rule that applies whenever we have a normal or bell-shaped distribution. Start with the average -- the center of the distribution. If you go up and down (i.e., left and right) one standard unit, you will include approximately 68% of the cases in the distribution (i.e., 68% of the area under the curve). If you go up and down two standard units, you will include approximately 95% of the cases. And if you go plus-and-minus three standard units, you will include about 99% of the cases. Notice that I didn't specify in the previous few sentences whether I was talking about standard deviation units or standard error units. That's because the same rule holds for both types of distributions (i.e., the raw data and sampling distributions). For instance, in the figure, the mean of the distribution is 3.75 and the standard unit is .25 (If this was a distribution of raw data, we would be talking in standard deviation units. If it's a sampling distribution, we'd be talking in standard error units). If we go up and down one standard unit from the mean, we would be going up and down .25 from the mean of 3.75. Within this range -- 3.5 to 4.0 -- we would expect to see approximately 68% of the cases. This section is marked in red on the figure. I leave to you to figure out the other ranges. But what does this all mean you ask? If we are dealing with raw data and we know the mean and standard deviation of a sample, we can *predict* the intervals within which 68, 95 and 99% of our cases would be expected to fall. We call these intervals the -- guess what -- 68, 95 and 99% confidence intervals.



Now, here's where everything should come together in one great aha! experience if you've been following along. If we had a *sampling distribution*, we would be able to predict the 68, 95 and 99% confidence intervals for where the population parameter should be! And isn't that why we sampled in the first place? So that we could predict where the population is on that variable? There's only one hitch. We don't actually have the sampling distribution (now this is the third time I've said this in this essay)! But we do have the distribution for the sample itself. And we can from that distribution estimate the standard error (the sampling error) because it is based on the standard deviation and we have that. And, of course, we don't actually know the population parameter value -- we're trying to find that out -- but we can use our best estimate for that -- the sample statistic. Now, if we have the mean of the sampling distribution (or set it to the mean from our sample) and we have an estimate of the standard error (we calculate that from our sample) then we have the two key ingredients that we need for our sampling distribution in order to estimate confidence intervals for the population parameter.

Perhaps an example will help. Let's assume we did a study and drew a single sample from the population. Furthermore, let's assume that the average for the sample was 3.75 and the standard deviation was .25. This is the raw data distribution depicted above. now, what would the sampling distribution be in this case? Well, we don't actually construct it (because we would need to take an infinite number of samples) but we *can* estimate it. For starters, we assume that the mean of the sampling distribution is the mean of the sample, which is 3.75. Then, we calculate the standard error. To do this, we use the standard deviation for our sample and the sample size (in this case $N=100$) and we come up with a standard error of .025 (just trust me on this). Now we have everything we need to estimate a confidence interval for the population parameter. We would estimate that the probability is 68% that the true parameter value falls between 3.725 and 3.775 (i.e., 3.75 plus and minus .025); that the 95% confidence interval is 3.700 to 3.800; and that we can say with 99% confidence that the population value is between 3.675 and 3.825. The real value (in this fictitious example) was 3.72 and so we have correctly estimated that value with our sample.

Copyright ©2006, William M.K. Trochim, All Rights Reserved
[Purchase a printed copy of the Research Methods Knowledge Base](#)
Last Revised: 10/20/2006

[Home](#) » [Sampling](#) »

Probability Sampling

A **probability sampling** method is any method of sampling that utilizes some form of *random selection*. In order to have a random selection method, you must set up some process or procedure that assures that the different units in your population have equal probabilities of being chosen. Humans have long practiced various forms of random selection, such as picking a name out of a hat, or choosing the short straw. These days, we tend to use computers as the mechanism for generating random numbers as the basis for random selection.

Some Definitions

Before I can explain the various probability methods we have to define some basic terms. These are:

- N = the number of cases in the sampling frame
- n = the number of cases in the sample
- NC_n = the number of combinations (subsets) of n from N
- $f = n/N$ = the sampling fraction

That's it. With those terms defined we can begin to define the different probability sampling methods.

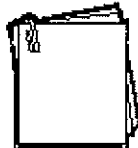
Simple Random Sampling

The simplest form of random sampling is called **simple random sampling**. Pretty tricky, huh? Here's the quick description of simple random sampling:

- **Objective:** To select n units out of N such that each NC_n has an equal chance of being selected.
- **Procedure:** Use a table of random numbers, a computer random number generator, or a mechanical device to select the sample.

A somewhat stilted, if accurate, definition. Let's see if we can make it a little more real. How do we select a simple random sample? Let's assume that we are doing

List of Clients



some research with a small service agency that wishes to assess client's views of quality of

Random Subsample



service over the past year. First, we have to get the sampling frame organized. To

accomplish this, we'll go through agency records to identify every client over the past 12 months. If we're lucky, the agency has good accurate computerized records and

can quickly produce such a list. Then, we have to actually draw the sample. Decide on the number of clients you would like to have in the final sample. For the sake of the example, let's say you want to select 100 clients to survey and that there were 1000 clients over the past 12 months. Then, the sampling fraction is $f = n/N = 100/1000 = .10$ or 10%. Now, to actually draw the sample, you have several options. You could print off the list of 1000 clients, tear them into separate strips, put the strips in a hat, mix them up real good, close your eyes and pull out the first 100. But this mechanical procedure would be tedious and the quality of the sample would depend on how thoroughly you mixed them up and how randomly you reached in. Perhaps a better procedure would be to use the kind of ball machine that is popular with many of the state lotteries. You would need three sets of balls numbered 0 to 9, one set for each of the digits from 000 to 999 (if we select 000 we'll call that 1000). Number the list of names from 1 to 1000 and then use the ball machine to select the three digits that selects each person. The obvious disadvantage here is that you need to get the ball machines. (Where do they make those things, anyway? Is there a ball machine industry?).

Neither of these mechanical procedures is very feasible and, with the development of inexpensive computers there is a much easier way. Here's a simple procedure that's especially useful if you have the names of the clients already on the computer. Many computer programs can generate a series of random numbers. Let's assume you can copy and paste the list of client names into a column in an EXCEL spreadsheet. Then, in the column right next to it paste the function `=RAND()` which is EXCEL's way of putting a random number between 0 and 1 in the cells. Then, sort both columns -- the list of names and the random number -- by the random numbers. This rearranges the list in random order from the lowest to the highest random number. Then, all you have to do is take the first hundred names in this sorted list. pretty simple. You could probably accomplish the whole thing in under a minute.

Simple random sampling is simple to accomplish and is easy to explain to others. Because simple random sampling is a fair way to select a sample, it is reasonable to generalize the results from the sample back to the population. Simple random sampling is not the most statistically efficient method of sampling and you may, just because of the luck of the draw, not get good representation of subgroups in a population. To deal with these issues, we have to turn to other sampling methods.

Stratified Random Sampling

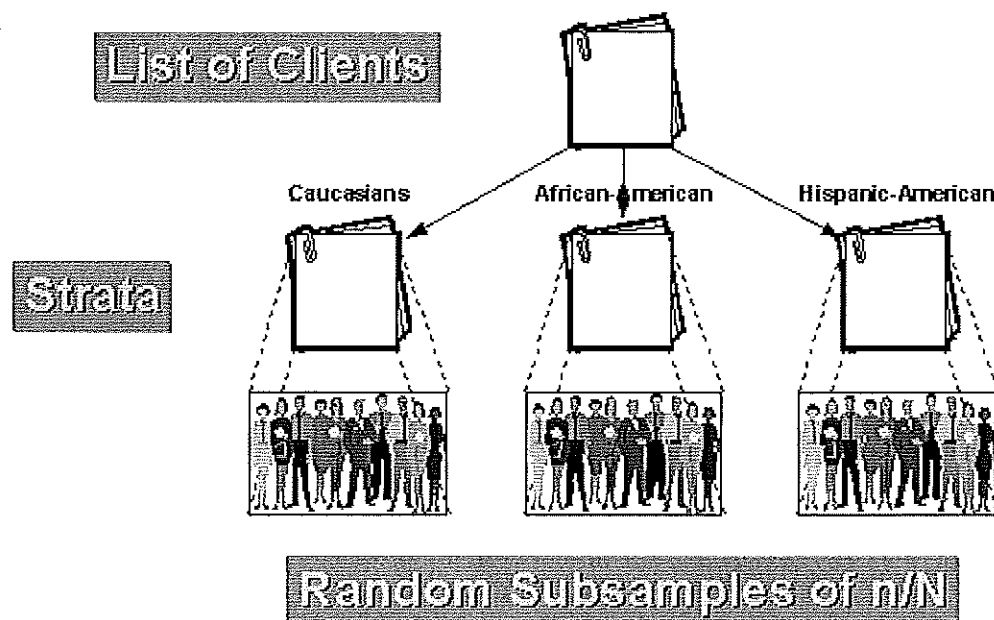
Stratified Random Sampling, also sometimes called *proportional* or *quota* random sampling, involves dividing your population into homogeneous subgroups and then taking a simple random sample in each subgroup. In more formal terms:

Objective: Divide the population into non-overlapping groups (i.e., *strata*) $N_1, N_2, N_3, \dots, N_i$, such that $N_1 + N_2 + N_3 + \dots + N_i = N$. Then do a simple random sample of $f = n/N$ in each strata.

There are several major reasons why you might prefer stratified sampling over simple random sampling. First, it assures that you will be able to represent not only the overall population, but also key subgroups of the population, especially small minority groups. If you want to be able to talk about subgroups, this may be the only way to effectively assure you'll be able to. If the subgroup is extremely small, you can

use different sampling fractions (f) within the different strata to randomly over-sample the small group (although you'll then have to weight the within-group estimates using the sampling fraction whenever you want overall population estimates). When we use the same sampling fraction within strata we are conducting *proportionate* stratified random sampling. When we use different sampling fractions in the strata, we call this *disproportionate* stratified random sampling. Second, stratified random sampling will generally have more statistical precision than simple random sampling. This will only be true if the strata or groups are homogeneous. If they are, we expect that the variability within-groups is lower than the variability for the population as a whole. Stratified sampling capitalizes on that fact.

For



example, let's say that the population of clients for our agency can be divided into three groups: Caucasian, African-American and Hispanic-American. Furthermore, let's assume that both the African-Americans and Hispanic-Americans are relatively small minorities of the clientele (10% and 5% respectively). If we just did a simple random sample of $n=100$ with a sampling fraction of 10%, we would expect by chance alone that we would only get 10 and 5 persons from each of our two smaller groups. And, by chance, we could get fewer than that! If we stratify, we can do better. First, let's determine how many people we want to have in each group. Let's say we still want to take a sample of 100 from the population of 1000 clients over the past year. But we think that in order to say anything about subgroups we will need at least 25 cases in each group. So, let's sample 50 Caucasians, 25 African-Americans, and 25 Hispanic-Americans. We know that 10% of the population, or 100 clients, are African-American. If we randomly sample 25 of these, we have a within-stratum sampling fraction of $25/100 = 25\%$. Similarly, we know that 5% or 50 clients are Hispanic-American. So our within-stratum sampling fraction will be $25/50 = 50\%$. Finally, by subtraction we know that there are 850 Caucasian clients. Our within-stratum sampling fraction for them is $50/850 = \text{about } 5.88\%$. Because the groups are more homogeneous within-group than across the population as a whole, we can expect greater statistical precision (less variance). And, because we stratified, we know we will have enough cases from each group to make meaningful subgroup inferences.

Systematic Random Sampling

Here are the steps you need to follow in order to achieve a **systematic random sample**:

- number the units in the population from 1 to N
- decide on the n (sample size) that you want or need
- $k = N/n =$ the interval size
- randomly select an integer between 1 to k
- then take every kth unit

N = 100

want n = 20

$N/n = 5$

select a random number from 1-5:
chose 4

start with #4 and take every 5th unit

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

All
of
this
will
be

much clearer with an example. Let's assume that we have a population that only has N=100 people in it and that you want to take a sample of n=20. To use systematic sampling, the population must be listed in a random order. The sampling fraction would be $f = 20/100 = 20\%$. in this case, the interval size, k, is equal to $N/n = 100/20 = 5$. Now, select a random integer from 1 to 5. In our example, imagine that you chose 4. Now, to select the sample, start with the 4th unit in the list and take every k-th unit (every 5th, because k=5). You would be sampling units 4, 9, 14, 19, and so on to 100 and you would wind up with 20 units in your sample.

For this to work, it is essential that the units in the population are randomly ordered, at least with respect to the characteristics you are measuring. Why would you ever want to use systematic random sampling? For one thing, it is fairly easy to do. You only have to select a single random number to start things off. It may also be more precise than simple random sampling. Finally, in some situations there is simply no easier way to do random sampling. For instance, I once had to do a study that involved sampling from all the books in a library. Once selected, I would have to go to the shelf, locate the book, and record when it last circulated. I knew that I had a fairly good sampling frame in the form of the shelf list (which is a card catalog where the entries are arranged in the order they occur on the shelf). To do a simple random sample, I could have estimated the total number of books and generated random numbers to draw the sample; but how would I find book #74,329 easily if that is the number I selected? I couldn't very well count the cards until I came to 74,329! Stratifying wouldn't solve that problem either. For instance, I could have stratified by

card catalog drawer and drawn a simple random sample within each drawer. But I'd still be stuck counting cards. Instead, I did a systematic random sample. I estimated the number of books in the entire collection. Let's imagine it was 100,000. I decided that I wanted to take a sample of 1000 for a sampling fraction of $1000/100,000 = 1\%$. To get the sampling interval k , I divided $N/n = 100,000/1000 = 100$. Then I selected a random integer between 1 and 100. Let's say I got 57. Next I did a little side study to determine how thick a thousand cards are in the card catalog (taking into account the varying ages of the cards). Let's say that on average I found that two cards that were separated by 100 cards were about .75 inches apart in the catalog drawer. That information gave me everything I needed to draw the sample. I counted to the 57th by hand and recorded the book information. Then, I took a compass. (Remember those from your high-school math class? They're the funny little metal instruments with a sharp pin on one end and a pencil on the other that you used to draw circles in geometry class.) Then I set the compass at .75", stuck the pin end in at the 57th card and pointed with the pencil end to the next card (approximately 100 books away). In this way, I approximated selecting the 157th, 257th, 357th, and so on. I was able to accomplish the entire selection procedure in very little time using this systematic random sampling approach. I'd probably still be there counting cards if I'd tried another random sampling method. (Okay, so I have no life. I got compensated nicely, I don't mind saying, for coming up with this scheme.)

Cluster (Area) Random Sampling

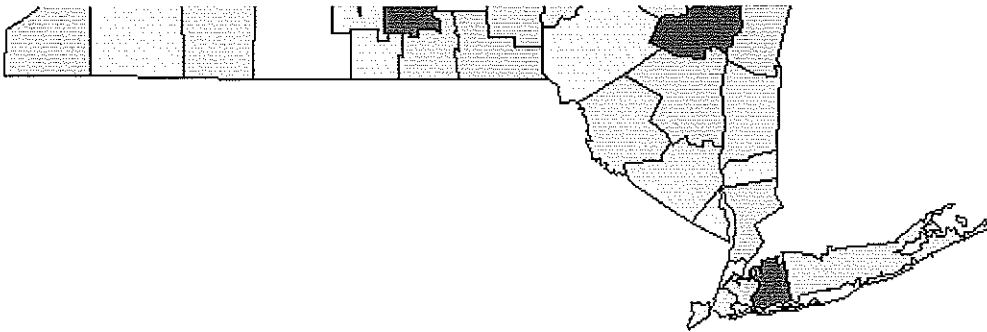
The problem with random sampling methods when we have to sample a population that's disbursed across a wide geographic region is that you will have to cover a lot of ground geographically in order to get to each of the units you sampled. Imagine taking a simple random sample of all the residents of New York State in order to conduct personal interviews. By the luck of the draw you will wind up with respondents who come from all over the state. Your interviewers are going to have a lot of traveling to do. It is for precisely this problem that **cluster or area random sampling** was invented.

In cluster sampling, we follow these steps:

- divide population into clusters (usually along geographic boundaries)
- randomly sample clusters
- measure all units within sampled clusters

For





instance, in the figure we see a map of the counties in New York State. Let's say that we have to do a survey of town governments that will require us going to the towns personally. If we do a simple random sample state-wide we'll have to cover the entire state geographically. Instead, we decide to do a cluster sampling of five counties (marked in red in the figure). Once these are selected, we go to *every* town government in the five areas. Clearly this strategy will help us to economize on our mileage. Cluster or area sampling, then, is useful in situations like this, and is done primarily for efficiency of administration. Note also, that we probably don't have to worry about using this approach if we are conducting a mail or telephone survey because it doesn't matter as much (or cost more or raise inefficiency) where we call or send letters to.

Multi-Stage Sampling

The four methods we've covered so far -- simple, stratified, systematic and cluster -- are the simplest random sampling strategies. In most real applied social research, we would use sampling methods that are considerably more complex than these simple variations. The most important principle here is that we can combine the simple methods described earlier in a variety of useful ways that help us address our sampling needs in the most efficient and effective manner possible. When we combine sampling methods, we call this **multi-stage sampling**.

For example, consider the idea of sampling New York State residents for face-to-face interviews. Clearly we would want to do some type of cluster sampling as the first stage of the process. We might sample townships or census tracts throughout the state. But in cluster sampling we would then go on to measure everyone in the clusters we select. Even if we are sampling census tracts we may not be able to measure *everyone* who is in the census tract. So, we might set up a stratified sampling process within the clusters. In this case, we would have a two-stage sampling process with stratified samples within cluster samples. Or, consider the problem of sampling students in grade schools. We might begin with a national sample of school districts stratified by economics and educational level. Within selected districts, we might do a simple random sample of schools. Within schools, we might do a simple random sample of classes or grades. And, within classes, we might even do a simple random sample of students. In this case, we have three or four stages in the sampling process and we use both stratified and simple random sampling. By combining different sampling methods we are able to achieve a rich variety of probabilistic sampling methods that can be used in a wide range of social research contexts.

Copyright ©2006, William M.K. Trochim, All Rights Reserved
 Purchase a printed copy of the Research Methods Knowledge Base

Last Revised: 10/20/2006

[Home](#) » [Sampling](#) »

Nonprobability Sampling

The difference between nonprobability and probability sampling is that nonprobability sampling does not involve *random* selection and probability sampling does. Does that mean that nonprobability samples aren't representative of the population? Not necessarily. But it does mean that nonprobability samples cannot depend upon the rationale of probability theory. At least with a probabilistic sample, we know the odds or probability that we have represented the population well. We are able to estimate confidence intervals for the statistic. With nonprobability samples, we may or may not represent the population well, and it will often be hard for us to know how well we've done so. In general, researchers prefer probabilistic or random sampling methods over nonprobabilistic ones, and consider them to be more accurate and rigorous. However, in applied social research there may be circumstances where it is not feasible, practical or theoretically sensible to do random sampling. Here, we consider a wide range of nonprobabilistic alternatives.

We can divide nonprobability sampling methods into two broad types: *accidental* or *purposive*. Most sampling methods are purposive in nature because we usually approach the sampling problem with a specific plan in mind. The most important distinctions among these types of sampling methods are the ones between the different types of purposive sampling approaches.

Accidental, Haphazard or Convenience Sampling

One of the most common methods of sampling goes under the various titles listed here. I would include in this category the traditional "man on the street" (of course, now it's probably the "person on the street") interviews conducted frequently by television news programs to get a quick (although nonrepresentative) reading of public opinion. I would also argue that the typical use of college students in much psychological research is primarily a matter of convenience. (You don't really believe that psychologists use college students because they believe they're representative of the population at large, do you?). In clinical practice, we might use clients who are available to us as our sample. In many research contexts, we sample simply by asking for volunteers. Clearly, the problem with all of these types of samples is that we have no evidence that they are representative of the populations we're interested in generalizing to -- and in many cases we would clearly suspect that they are not.

Purposive Sampling

In purposive sampling, we sample with a *purpose* in mind. We usually would have one or more specific predefined groups we are seeking. For instance, have you ever run into people in a mall or on the street who are carrying a clipboard and who are stopping various people and asking

if they could interview them? Most likely they are conducting a purposive sample (and most likely they are engaged in market research). They might be looking for Caucasian females between 30-40 years old. They size up the people passing by and anyone who looks to be in that category they stop to ask if they will participate. One of the first things they're likely to do is verify that the respondent does in fact meet the criteria for being in the sample. Purposive sampling can be very useful for situations where you need to reach a targeted sample quickly and where sampling for proportionality is not the primary concern. With a purposive sample, you are likely to get the opinions of your target population, but you are also likely to overweight subgroups in your population that are more readily accessible.

All of the methods that follow can be considered subcategories of purposive sampling methods. We might sample for specific groups or types of people as in modal instance, expert, or quota sampling. We might sample for diversity as in heterogeneity sampling. Or, we might capitalize on informal social networks to identify specific respondents who are hard to locate otherwise, as in snowball sampling. In all of these methods we know what we want -- we are sampling with a purpose.

- Modal Instance Sampling

In statistics, the *mode* is the most frequently occurring value in a distribution. In sampling, when we do a modal instance sample, we are sampling the most frequent case, or the "typical" case. In a lot of informal public opinion polls, for instance, they interview a "typical" voter. There are a number of problems with this sampling approach. First, how do we know what the "typical" or "modal" case is? We could say that the modal voter is a person who is of average age, educational level, and income in the population. But, it's not clear that using the averages of these is the fairest (consider the skewed distribution of income, for instance). And, how do you know that those three variables -- age, education, income -- are the only or even the most relevant for classifying the typical voter? What if religion or ethnicity is an important discriminator? Clearly, modal instance sampling is only sensible for informal sampling contexts.

- Expert Sampling

Expert sampling involves the assembling of a sample of persons with known or demonstrable experience and expertise in some area. Often, we convene such a sample under the auspices of a "panel of experts." There are actually two reasons you might do expert sampling. First, because it would be the best way to elicit the views of persons who have specific expertise. In this case, expert sampling is essentially just a specific subcase of purposive sampling. But the other reason you might use expert sampling is to provide evidence for the validity of another sampling approach you've chosen. For instance, let's say you do modal instance sampling and are concerned that the criteria you used for defining the modal instance are subject to criticism. You might convene an expert panel consisting of persons with acknowledged experience

and insight into that field or topic and ask them to examine your modal definitions and comment on their appropriateness and validity. The advantage of doing this is that you aren't out on your own trying to defend your decisions -- you have some acknowledged experts to back you. The disadvantage is that even the experts can be, and often are, wrong.

- Quota Sampling

In quota sampling, you select people nonrandomly according to some fixed quota. There are two types of quota sampling: *proportional* and *non proportional*. In **proportional quota sampling** you want to represent the major characteristics of the population by sampling a proportional amount of each. For instance, if you know the population has 40% women and 60% men, and that you want a total sample size of 100, you will continue sampling until you get those percentages and then you will stop. So, if you've already got the 40 women for your sample, but not the sixty men, you will continue to sample men but even if legitimate women respondents come along, you will not sample them because you have already "met your quota." The problem here (as in much purposive sampling) is that you have to decide the specific characteristics on which you will base the quota. Will it be by gender, age, education race, religion, etc.?

Nonproportional quota sampling is a bit less restrictive. In this method, you specify the minimum number of sampled units you want in each category. here, you're not concerned with having numbers that match the proportions in the population. Instead, you simply want to have enough to assure that you will be able to talk about even small groups in the population. This method is the nonprobabilistic analogue of stratified random sampling in that it is typically used to assure that smaller groups are adequately represented in your sample.

- Heterogeneity Sampling

We sample for heterogeneity when we want to include all opinions or views, and we aren't concerned about representing these views proportionately. Another term for this is sampling for *diversity*. In many brainstorming or nominal group processes (including concept mapping), we would use some form of heterogeneity sampling because our primary interest is in getting broad spectrum of ideas, not identifying the "average" or "modal instance" ones. In effect, what we would like to be sampling is not people, but ideas. We imagine that there is a universe of all possible ideas relevant to some topic and that we want to sample this population, not the population of people who have the ideas. Clearly, in order to get all of the ideas, and especially the "outlier" or unusual ones, we have to include a broad and diverse range of participants. Heterogeneity sampling is, in this sense, almost the opposite of modal instance sampling.

- Snowball Sampling

In snowball sampling, you begin by identifying someone who meets the criteria for inclusion in your study. You then ask them to recommend others who they may know who also meet the criteria. Although this method would hardly lead to representative samples, there are times when it may be the best method available. Snowball sampling is especially useful when you are trying to reach populations that are inaccessible or hard to find. For instance, if you are studying the homeless, you are not likely to be able to find good lists of homeless people within a specific geographical area. However, if you go to that area and identify one or two, you may find that they know very well who the other homeless people in their vicinity are and how you can find them.

Copyright ©2006, William M.K. Trochim, All Rights Reserved
Purchase a printed copy of the Research Methods Knowledge Base
Last Revised: 10/20/2006