Cautions and Hints About Sampling

This purpose of this discussion is to help you to understand sampling and to help you to avoid some common mistakes that people make when trying to create a random sample.

What does it mean for a sample to be random???? To create a random sample, you must select the units for the sample in such a way that

- all individual units are equally likely to be chosen,
- all pairs of units are equally likely to be chosen,
- all groups of three units are equally likely to be chosen,
- all groups of four units are equally likely to be chosen,
- all groups of five units are equally likely to be chosen, etc.

That is, to create a random sample, you much select the units for the sample in such a way that each individual unit and groups of any number of units are all equally likely to be chosen.

Can I create a random sample by listing the names of the units on slips of paper, folding the slips of paper, putting the folded slips of paper into a hat or a bowl, and then drawing slips of paper out of the hat or bowl???? No, I cannot create a random sample by selecting slips of paper out of a bowl or hat because each slip of paper does not have the same likelihood of being selected and each group ofany number of slips of paper does not have the same likelihood of being chosen at any time from the hat or bowl. Please keep in mind that when one draws the skips of paper out of the hat or the bowl, one chooses the location from which to draw the slip of paper and at the that time none of the other slips of paper can be selected. For example, if a person has a habit of choosing slips of paper from the top of the hat or bowl, the slips of paper in the middle, on the sides, or on the bottom do not have any chance of being selected: the individual slips of paper or the groups of any number of slips of paper do not have the same likelihood of being selected.

Why does the author of the textbook mention this as a way (selecting slips of paper from a hat or a bowl) in which to create a random sample???? He uses this as a simple example. Unfortunately, while easy to comprehend, this is not an acceptable way in which to create a random sample since this method does not create a random sample, and you will earn no credit for using this as your description of how to create a random sample.

What is the most basic way in which to create a random sample???? The most basic type of random sample is a simple random sample. To create a simple random sample of n units from a population,

- we need a numbered list of the units in the population (if we do not have a list of the units in the population, we create one, and if we have a list of the units in the population then we number them),
- we use a random digits table, a (pseudo) random number generator, or a chance device to select n random numbers, and
- the units in the population that correspond to the n random numbers become the units in the population.

What is a (pseudo) random number generator???? Calculators and software such as MS Excel are programmed with algorithms that generate pseudo random number. These “random numbers” are frequently referred to as pseudo random numbers since numbers generated by an algorithm are not (and never can be) truly random.

How do I use the random digits table???? We can use the random digits table to select any number or “random numbers” by selecting a row or column in which to start and then work horizontally (left or right), vertically (up or down), diagonally (up or down, left or right), one row or column at a time, skip rows or column, pair rows or columns, or, in short, use any scheme to allow us to determine/generate the n numbers using digits present in the table that correspond to the numbers for our numbered units of the population that will become the n units in the sample.

Can we make a simple random sample using the Random Digits Table???? Suppose we want to be able to select five random numbers between 1 and 100, inclusive, in order to select a sample of five rectangles on the Random Rectangles handout. To do this, we must make sure that all the numbers, 1 to 100, inclusive can be selected. Initially, one might think that 100, since it is not a two-digit number, is not possible. However, examining the random digits table, we see that 00 is a possibility. So, we can use one of two schemes to assign 100 to a two-digit number on the random digits table:

- we can assign 100 to 00 or
- we can add one to each two-digit number so that 00 corresponds to 1, 01 corresponds to 2, 02 corresponds to 3, …, 98 corresponds to 99, and 99 corresponds to 100.

While either scheme will work fine, the former is simpler than the latter. We will use the former, assigning 00 to 100. Examining the Random Rectangles handout, we see that it can act as our frame for the population since it is a numbered list of the units in the population. Since all the numbers assigned to the rectangles are possible to obtain using the random digits table, we can now select a scheme with which to use the table. Suppose we work vertically down starting with the fifth column of five digits at the top; the digits for the first row of this column is 34673. Reading downward along the right-hand side of this column and using digits to form two-digit numbers, we take five two-digit numbers in order to determine our sample of five rectangles. If we obtain a repeated two-digit number, we simply select another two-digit number to replace it since each rectangle can be in the sample only once. Our random numbers are 35, 91, 27, 38 since we reject the repeated 35, and 95. So, our sample will contain the numbered rectangles that correspond to these random numbers. To help you to understand the difference between the rectangles in the sample and the random numbers, suppose we want to determine the average area for the rectangles in the sample. In order to do this, we must examine each of the numbered rectangles for our sample and determine their individual areas; the average area for the sample will be the sum of the areas of the rectangles divided by the number of rectangles. So, to determine this average area for the rectangles in our sample, we do not take the average of the random numbers that we obtained from the random digits table – the random numbers are only used to select the rectangles for the sample. So, examining the table, we see that

- the area for rectangle number 35 is 4 square units,
- the area for rectangle number 91 is 3 square units,
- the area for rectangle number 27 is 4 square units,
- the area for rectangle number 38 is 9 square units, and
- the area for rectangle number 95 is 6 square units.

Since the sum of the areas of the rectangles in the sample is 26 square units, the average area for the sample of rectangles is 5 1/6 square units; notice that this is definitely different from determining the average of the random numbers. It is important to remember that the sample consists of the units, in this case the rectangles, corresponding to the random numbers that you selected. The random digits table does not determine the sample: the random digits table allows you to obtain pseudo random numbers and it is *the units* (in this case the rectangles) *that correspond to these pseudo random numbers that become the units in the sample*.

What is a chance device???? An example of a chance device is a fair die. If we need to select one of the digits 1, 2, 3, 4, 5, or 6 then we could roll a fair die in order to select a digit. If we need to select one of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9, we could use a ten-sided die. If we need to select a number from 1 to 30 inclusive, we could use a 30-sided figure that bears these numbers or we could use a combination of fair dice, for example one to create the first digit of the number and then one to create the second digit – provided each die is fair (the possible digits for an individual die having the same likelihood of being selected).

What distinguishes a stratified random sample from a cluster sample???? For a stratified random sample, we have strata that we want to preserve. Strata are non-overlapping subgroups of the population, for example, gender (male and female), income level (low income, middle income, high income), size, shape, color, and occupation. For a stratified random sample, we determine the proportions for the strata in the population and then select the number of units for the sample in such a way that these proportions are maintained. So, for example, if the population is such that one third of the population is male and two-thirds of the population is female then, for a sample of nine units from this population, we must have three males and six females. To create a stratified random sample,

- we must determine the proportions of the strata in the population,
- we must determine the now many units from each stratum must be included in the sample so that the proportions of the strata in the population are maintained in the sample,
- for each stratum, we create a numbered list of the units in the strata,
- using a random digits table, a pseudo random number generator, or a chance device, we generate the number of random numbers for the number of units that we want to include from each stratum – for each stratum, the units from the stratum that correspond to the random numbers selected become the units from the stratum that are included in the sample.

The important thing to remember about a stratified random sample is that the proportions of the strata in the population are maintained in the sample and that to select the units for the sample, we must take a simple random sample from each strata. For a cluster sample, we take a simple random sample of cluster; clusters are groups of units. So, to create a cluster sample,

- we must create a numbered list of clusters,
- we use a random digits table, a pseudo random number generator, or a chance device to generate the n random numbers if we want to select n clusters, and
- then the clusters that correspond to the n random numbers selected become the n clusters in the sample – all the units in the clusters become the units in the sample.

So, for a cluster sample, it is important to note that we take a simple random sample of clusters and all the units in the cluster become the units in the sample.

What is a cluster sample???? (A little repetition never hurts.) A cluster is a group of units. A cluster is not a characteristic in the population only a group of units within a population. Examples of clusters include schools, classes, and even the books on the individual shelves of your bookcase. Suppose we want to make a cluster sample of n clusters from a population that has N clusters,

- we create a numbered list of the N clusters in the population,
- we use a random digits table, a pseudo random number generator, or a chance device to generate n random numbers, and then
- all of the units *in* the clusters corresponding to the n random numbers selected become the units in the sample.

It is important to note that for a cluster sample, we randomly select clusters and all the units in the selected clusters become the units in the sample.

What is a two-stage sample???? A two-stage sample, is a cluster sample followed by a simple random sample from each cluster. Suppose we want to make a two-stage sample from a population that has N clusters,

- we create a numbered list of the N clusters in the population,
- we use a random digits table, a pseudo random number generator, or a chance device to generate n random numbers,
- the n clusters that correspond to the n random numbers selected are then sampled to select the units for the sample – we take a simple random sample of the units in each cluster and the selected units from each cluster are combined to form the sample.

It is important to note that, once the clusters have been selected, we must create a numbered list for each of these clusters and use this numbered list together with a random digits table, a pseudo random number generator, or a chance device to determine the units from each cluster that will be included in the sample. So, we need a list of the population and a list for each of the clusters: for a two-stage sample, we need n + 1 lists where n is the number of clusters that we initially select and from which we select units for the sample. If we select n clusters in the first stage and m units from each cluster in the second stage then there are mn units in the sample; that is, the sample size is mn.

What is the difference between a cluster sample and a two-stage sample???? For a cluster sample, we take a simple random sample of clusters and all the units in the selected clusters become the units in the sample. However, for a two-stage sample, since we take a simple random sample of clusters followed by a simple random sample of the units in each of the clusters, not all the units in the selected clusters become units in the sample. Since we include all the units in the selected clusters when we create a cluster sample and since we include only a subset of the units in the selected clusters when we create a two-stage sample, the sample size for a cluster sample is greater than the sample size for a two-stage sample.

When would I choose one sample type over another???? A stratified random sample is desirable since a stratified random sample preserves the proportions of the characteristics of the population in the sample; this provides greater precision for statistical analysis (the sample statistics are closer to the population statistics). However, if the population frame does not provide information about the desired characteristics of the population then we cannot create strata and we cannot create a stratified random sample. If the frame is a list of groups within the population then we have clusters: if we have a list of clusters in the population then we can create a cluster sample. If the population frame is a list of clusters in the population and if we do not want to include all the units in each cluster but rather we want to take a sample of clusters and a sample of the units in the clusters then we would use a two-stage sample. If we have a list of units for the population (a frame) that does not provide characteristics that can be used to create strata and if the frame does not group the units into clusters then we could use a simple random sample: we need distinct characteristics in order to create strata (non-overlapping subgroups) and to take a stratified random sample, and we need groups within the population in order to have clusters if we want to take a cluster sample or for a two-stage sample. If we want to take a simple random sample, we need a list of the units in the population.

What do I need if I want to create a systematic sample???? In order to create a systematic sample,

- we must have order (for example, people in a line or a list of units rather than an unordered group),
- we need a "count off" number, and
- we need a random start number.

We determine the "count off" number, k, by taking the quotient of the population size, N, and the sample size, n where we round down to the nearest integer: k = N/n, rounding down to the nearest integer. The random start number is a counting number between 1 and k, inclusive; the counting numbers, also known as the natural numbers, are the numbers with which it is natural to count, that is,1, 2, 3, 4, 5, …. So, suppose we have 100 people waiting in line at the airport and we want to create a systematic sample of twenty people. We determine the count off number, k = 100/20 = 5, we take a randomly selected counting number between 1 and 5, inclusive, as our random start (here, let us take three), and, we create our sample by taking the third person in line and every fifth person after that; if we were to number the people waiting in line, we would take person number 3, 8, 13, 18, 23, 28, 33, 38, 43, 48, 53, 58, 63, 68, 73, 78, 83, 88, 93, and 98 to be in our sample. Notice that the units selected for the sample are *systematically* and equally positioned throughout the line.

Are there other kinds of sample???? If we combine various sample types we can create three-stage, four-stage, …, that is, multi-stage samples.

Please see my PowerPoint slides on random sampling for discussion, examples, and exercises on creating

- simple random samples,
- stratified random samples,
- cluster samples,
- two-stage samples, and
- multi-stage samples.

Please note that the sample size and the population size are numbers. In addition, it is important to remember that the term *unit* has different meanings depending on its context: the units for a study are not the same as the units for a variable.