Sampling

Concept
Sampling is the process of selecting a few (a sample) from a bigger group (the sampling population) to become the basis for estimating or predicting the prevalence of an unknown piece of information, situation or outcome regarding the bigger group. A sample is a subgroup of the population you are interested in.

This process of selecting a sample from the total population has advantages and disadvantages. The advantages are that it saves time as well as financial and human resources. However, the disadvantage is that you do not find out the information about the population’s characteristics of interest to you but only estimate or predict them. Hence, the possibility of an error in your estimation exists. Sampling is thus a trade-off between certain gains and losses. While on the one hand you save time and resources, on the other hand you may compromise the level of accuracy in your findings. Through sampling you only make an estimate about the actual situation prevalent in the total population from which the sample is drawn. If you ascertain a piece of information from the total sampling population, and if your method of inquiry is correct, your findings should be reasonably accurate. However, if you select a sample and use this as a basis from which to estimate the situation in the total population, an error is possible. Tolerance of this possibility of error is an important consideration in selecting a sample.

The concept of sampling in qualitative research
In qualitative research the issue of sampling has little significance as the main aim of most qualitative inquiries is either to explore or describe the diversity in a situation, phenomenon or issue. Qualitative research does not make an attempt to either quantify or determine the extent of this diversity. You can select even one individual as your sample and describe whatever the aim
of your inquiry is. A study based upon the information obtained from one individual, or undertaken to describe one event or situation is perfectly valid. In qualitative research, to explore the diversity, you need to reach what is known as saturation point in terms of your findings; for example, you go on interviewing or collecting information as long as you keep discovering new information. When you find that you are not obtaining any new data or the new information is negligible, you are assumed to have reached saturation point. Some researchers prefer to select a sample using non-probability designs and to collect data till they have reached saturation point. Keep in mind that saturation point is a subjective judgment which you, as a researcher, decide.

**Sampling aspects**

**Population or study population**
The group researcher intends to generalize to is called the population in the study. It is, however, pertinent here to distinguish between theoretical and accessible population.

**Theoretical population**
The population researcher intends to generalize to is called theoretical population.

**Accessible population**
The population that will be accessible to researcher is called accessible population or study population.

**Sample**
The sample is the group of people whom researcher selects to be in particular study being undertaken.

**Sub-sample**
The group that actually completes a particular study is a sub-sample of sample….it does not include new respondents or dropouts.
Sample size
The number of respondents from whom researcher obtains the required information is called the sample size. It is denoted by letter 'n'.

Sampling design or strategy
The way researcher selects the respondents is called the sampling design or strategy.

Sampling unit or sampling element
Each individual that becomes the basis for selecting a sample is called the sampling unit or sampling element.

Sampling frame
A list identifying each respondent in study population is called sampling frame.

Note: - If all the elements in the sampling frame can not be individually identified, researcher can not have a sampling frame for that study population.

Saturation point
When the researcher reaches a stage (in qualitative research*) where no new information is coming from respondents is called saturation point.

*Qualitative research is the research that is conducted to explore or describe the diversity in a situation, phenomenon or issue.

Statistical Terms in Sampling

Response, Statistics and Parameter
While conducting research, 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 we 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.

**Sampling distribution**

![Graph showing sampling distribution](https://via.placeholder.com/150)

The distribution of an infinite number of samples of the same size as the sample in particular study is known as the **sampling distribution**. To construct a sampling distribution, we would have to take an *infinite* number of samples. It would be worthwhile mentioning here that infinite is not a number we know how to reach. We need to realize here the fact that a 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.

**Standard Deviation**

A standard deviation is the spread of the scores around the average in a single sample. 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**.

**Standard Error**

The standard error is the spread of the averages around the average of averages in a sampling distribution.

**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.
For calculation of 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. It is so 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 would be the parameter.

The 68, 95, 99 Percent Rule

There is a general rule that applies whenever we have a normal or bell-shaped distribution. Starting 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. This same rule holds for standard deviation and standard error units (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, it would be used in context of standard deviation units. If it's a sampling distribution, the same would be referred in context of 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. 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 as the -- 68, 95 and 99% confidence intervals.

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.

**Example**

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. Assume that the mean of the sampling distribution is the mean of the sample, which are 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 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 (of which real value was 3.72). So, we have correctly estimated that value with our sample.

**Principles of sampling**

Theory of sampling is guided by these principles which are crucial to keep in mind when researcher is determining the sample size needed to for a particular level of accuracy, and in selecting a sampling strategy best suited to study being undertaken.

**Principle: 1** In a majority of cases of sampling there will be a difference between the sample statistics and the true population mean, which is attributable to the selection of the units in the sample.

**Explanation:**

Suppose there are four individuals: A, B, C and D. A is 18 years of age, B is 20, C is 23 and D is 25. As you know their ages, you can find out (calculate) their average age by simply adding 18+20+23+25=86 and dividing by 4. This gives the average age of A, B, C and D as 21.5 years. Now let us assume that you want to select a sample of two individuals to make an estimate of the average age of the four individuals. If you adopt the theory of probability, we can have six
possible combinations of two: A and B; A and C; A and D; B and C; B and D; and C and D. Let us take each of these pairs to calculate the average age of the sample:

1. A + B = 18 + 20 = 38/2 = 19.0 Years;
2. A + C = 18 + 23 = 41/2 = 20.5 Years;
3. A + D = 18 + 25 = 43/2 = 21.5 Years;
4. B + C = 20 + 23 = 43/2 = 21.5 Years;
5. B + D = 20 + 25 = 45/2 = 22.5 Years;
6. C + D = 23 + 25 = 48/2 = 24.0 Years;

Notice that in most cases the average age calculated on the basis of these samples of two (sample statistics) is different. Now compare these sample statistics with the average of all four individuals—the population mean (population parameter) of 21.5 years. Out of a total of six possible sample combinations, only in the case of two is there no difference between the sample statistics and the population mean. Where there is a difference, this is attributed to the sample and is known as **sampling error**. Let us consider the difference in the sample statistics and the population mean for each of the six samples.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sample average (sample statistics)</th>
<th>Population mean (population parameter)</th>
<th>Difference between (1) and (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.0</td>
<td>21.5</td>
<td>-2.5</td>
</tr>
<tr>
<td>2</td>
<td>20.5</td>
<td>21.5</td>
<td>-1.5</td>
</tr>
<tr>
<td>3</td>
<td>21.5</td>
<td>21.5</td>
<td>0.0</td>
</tr>
<tr>
<td>4</td>
<td>21.5</td>
<td>21.5</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>22.5</td>
<td>21.5</td>
<td>+1.0</td>
</tr>
<tr>
<td>6</td>
<td>24.0</td>
<td>21.5</td>
<td>+2.5</td>
</tr>
</tbody>
</table>

**Principle: 2** The greater the sample size, the more accurate will be the estimate of the true population mean.

**Explanation:**

Let us continue with the above example, but instead of a sample of two individuals take a sample of three. There are four possible combinations of three that can be drawn.

1. A + B + C = 18 + 20 + 23 = 61/3 = 20.33 Years;
2. A + B + D = 18 + 20 + 25 = 63/3 = 21.00 Years;
3. A + C + D = 18 + 23 + 25 = 66/3 = 22.00 Years;
4. B + C + D = 20 + 23 + 25 = 68/3 = 22.67 Years;

Now, let us compare the difference between the sample statistics and the population mean.
Table #2: The difference between a sample and population average

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sample average</th>
<th>Population average</th>
<th>Difference between (1) And (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.33</td>
<td>21.5</td>
<td>-1.17</td>
</tr>
<tr>
<td>2</td>
<td>21.00</td>
<td>21.5</td>
<td>-0.5</td>
</tr>
<tr>
<td>3</td>
<td>22.00</td>
<td>21.5</td>
<td>+0.5</td>
</tr>
<tr>
<td>4</td>
<td>22.67</td>
<td>21.5</td>
<td>+1.17</td>
</tr>
</tbody>
</table>

Now, compare the difference between the difference calculated in the First Table given in Principle 1 and the Second Table given in Principle 2. In the First Table the difference between the sample statistics and the population mean lies between -2.5 and +2.5 years, whereas in the second, it is between -1.17 and +1.17 years. The gap between the sample statistics and the population mean is reduced in Second Table. This reduction is attributed to the increase in the sample size.

**Principle: 3** The greater the difference in the variable under study in a population for a given sample size, the greater will be the difference between the sample statistics and the true population mean. This sampling principle is particularly important as a number of sampling strategies, such as stratified and cluster sampling, are based on it.

**Explanation:**
To understand third principle, let us continue with the same example but use slightly different data. Suppose the ages of four individuals are markedly different: A = 18, B = 26, C = 32 and D = 40. In other words, we are visualizing a population where the individuals with respect to age—the variable we are interested in—are markedly different.

Let us follow the same procedure, selecting samples of two individuals at a time and then three. If we work through the same procedures (described above) we will find that the difference in the average age in the case of samples of two ranges between -7.00 and +7.00 years and in the case of the sample of three ranges between -3.67 and +3.67. In both cases the range of the difference is greater than previously calculated. This is attributable to the greater difference in the ages of the four individuals—the sampling population. In other words, the sampling population is more heterogeneous in regard to age.

**Factors effecting the inferences drawn from sample**
Principle of sampling suggest following factors that may influence the degree of certainty about the inferences drawn from sample.
The size of the sample
Findings based upon larger samples have more certainty than those based on smaller ones. As a rule, the larger the sample size, the more accurate will be findings.

The extent of variation in sampling population—the greater the variation in the study population with respect to characteristics under study for a given sample size, the greater will be uncertainty. (In technical terms, the greater the standard deviation, the greater will be the standard error for a given sample size in your estimates.) If a population is homogeneous with respect to characteristics under study, a small sample can provide a reasonably good estimate, but if it is heterogeneous, you need to select a larger sample to obtain the same level of accuracy. Of course, if all the elements in a population are identical, then selection of even one will provide absolutely accurate estimate. As a rule, the higher the variation with respect to characteristics under study in the study population, the greater will be uncertainty for a given sample size.

Aims in selecting a sample
The aims in selecting a sample are to:

- achieve maximum precision in your estimates within a given sample size.
- avoid bias in the selection of your sample.

Bias in the selection of a sample can occur if:

- sampling is done by a non-random method—that is, if the selection is consciously or unconsciously influenced by human choice;
- the sampling frame—list, index or other population records—which serves as the basis of selection, does not cover the sampling population accurately and completely.
- a section of a sampling population is impossible to find or refuses to cooperate.
Types of sampling

The various sampling strategies can be categorized as per following figure:

Methods of drawing a random sample

Of the methods that you can adopt to select a random sample, the three most common are:

- **The fishbowl draw**—if your total population is small, an easy procedure is to number each element using separate slips of paper for each element, pull all the slips into a box, and then pick them out one by one without looking, until the number of slips selected equals the sample size you decided upon. This method is used in some lotteries.

- **Computer program**—there are number of programs that can help you to select a random sample.

- **A table of random numbers**—most books on research methodology and statistics include a table of randomly generated numbers in their appendices (see, for example, table #3). You can select you sample using these tables according to the procedure.

Procedure for using a table of random numbers

**Step 1**: Identify the total number of elements in the study population, for example 50, 100, 430, 795, or 1265. The total number of elements in a study population may
run up to four or more digits (if your total sampling populating is 9 or less, it is one digit; if it is 99 or less, it is two digits;...).

**Step 2:** Number each element starting from 1.

**Step 3:** If the table for random number is on more than one page, choose the starting page by a random procedure. Again, select a column or row that will be your starting point with a random procedure and proceed from there in a predetermined directing.

**Step 4:** Corresponding to the number of digits to which the total population runs, select the same number, randomly, of columns or rows of digits from the table.

**Step 5:** Decide on your sample size.

**Step 6:** Select the required number of elements for your sample from the table. If you happen to select the same number twice, discard it and go to the next. This can happen as the table for random numbers is generated by sampling with replacement.
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48461</td>
<td>14952</td>
<td>72619</td>
<td>73689</td>
<td>52059</td>
<td>37086</td>
<td>60050</td>
<td>86192</td>
<td>67049</td>
<td>64739</td>
</tr>
<tr>
<td>2</td>
<td>76534</td>
<td>30239</td>
<td>49692</td>
<td>31366</td>
<td>52093</td>
<td>15422</td>
<td>20498</td>
<td>33901</td>
<td>10319</td>
<td>43397</td>
</tr>
<tr>
<td>3</td>
<td>70437</td>
<td>25861</td>
<td>38504</td>
<td>14752</td>
<td>23757</td>
<td>29660</td>
<td>67844</td>
<td>78815</td>
<td>23758</td>
<td>86814</td>
</tr>
<tr>
<td>4</td>
<td>59584</td>
<td>03370</td>
<td>42806</td>
<td>11393</td>
<td>71722</td>
<td>93804</td>
<td>09095</td>
<td>07856</td>
<td>55589</td>
<td>46820</td>
</tr>
<tr>
<td>5</td>
<td>04285</td>
<td>58554</td>
<td>16085</td>
<td>51555</td>
<td>27501</td>
<td>73883</td>
<td>33427</td>
<td>33343</td>
<td>45307</td>
<td>50633</td>
</tr>
<tr>
<td>6</td>
<td>77340</td>
<td>10412</td>
<td>69189</td>
<td>85171</td>
<td>29802</td>
<td>44785</td>
<td>86368</td>
<td>02583</td>
<td>96483</td>
<td>76553</td>
</tr>
<tr>
<td>7</td>
<td>59183</td>
<td>62687</td>
<td>91778</td>
<td>80354</td>
<td>23512</td>
<td>97219</td>
<td>65921</td>
<td>02035</td>
<td>95847</td>
<td>91403</td>
</tr>
<tr>
<td>8</td>
<td>91800</td>
<td>04281</td>
<td>39979</td>
<td>03927</td>
<td>82564</td>
<td>28777</td>
<td>59049</td>
<td>97532</td>
<td>54540</td>
<td>79472</td>
</tr>
<tr>
<td>9</td>
<td>12066</td>
<td>24817</td>
<td>81099</td>
<td>48940</td>
<td>69554</td>
<td>55925</td>
<td>48379</td>
<td>12866</td>
<td>42820</td>
<td>21580</td>
</tr>
<tr>
<td>10</td>
<td>69907</td>
<td>91751</td>
<td>53512</td>
<td>23748</td>
<td>65906</td>
<td>91385</td>
<td>84983</td>
<td>27915</td>
<td>48491</td>
<td>91688</td>
</tr>
</tbody>
</table>

Table#3: Selecting a sample using a table for random numbers.
### Table#4: Selecting elements using the table of random numbers

<table>
<thead>
<tr>
<th>Column Table</th>
<th>Elements selected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>232</td>
</tr>
<tr>
<td></td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>246</td>
</tr>
<tr>
<td></td>
<td>161</td>
</tr>
<tr>
<td></td>
<td>243</td>
</tr>
<tr>
<td></td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>213</td>
</tr>
<tr>
<td></td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>234</td>
</tr>
<tr>
<td></td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>211</td>
</tr>
<tr>
<td></td>
<td>156</td>
</tr>
<tr>
<td></td>
<td>220</td>
</tr>
<tr>
<td></td>
<td>231</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>246</td>
</tr>
</tbody>
</table>

### Different systems of drawing a random sample

There are two ways of selecting a random sample:

- **Sampling without replacement**
  
  In this system of selecting a random sample, each element has an equal and independent chance of selection.

- **Sampling with replacement**
  
  In this system of selecting a sample, the selected element is replaced in the same population and if it is selected again, it is discarded and the next one is selected.

### Explanation:

Suppose you want to select a sample of 20 farmers out of a total of 80. The first farmer is selected out of the total group, and so the probability of selection for the first farmer is 1/80. When you select the second farmer there are only 79 left in the group and the probability of selection for the second farmer is not 1/80 but 1/79. The probability of selecting the next farmer is 1/78. By the time you select the 20th farmer, the probability of his selection is 1/61. This type of sampling is called *sampling without replacement*. But this is contrary to our basic definition of randomization; that is, each element has an equal and independent chance of selection. In the second system, called *sampling with replacement*, the selected element is replaced in the sampling population and if it is selected again, it is discarded and the next one is selected. If the sampling population is fairly large, the probability of selecting the same element twice is fairly remote.

### Basic Terms in sampling

Before explaining the various probability sampling methods, it is worthwhile 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

With those terms defined we can begin to define the different probability sampling methods.

**Specific random/probability sampling designs**

There are three commonly used types of random sampling design.

**Simple random sampling (SRS):**

In this sampling type, each element in the population is given an equal & independent chance of selection. This is the most commonly used method of selecting a probability sample. In line with the definition of randomization, whereby each element in the population is given an equal and independent chance of selection. The procedure for selecting simple random sample is as follows:-

- **Step 1** Identify by a number all elements or sampling units in the population.
- **Step 2** Decide on the sample size (\( n \)).
- **Step 3** Select (\( n \)) using either the fishbowl draw, the table of random numbers or a computer program.

**Explanation 1:**

Let's assume that we are doing some research with a small insecticide company that wishes to assess farmers’ views of quality of insecticide over past year. First, we have to get the sampling frame organized. To accomplish this, we'll go through company’s records to identify every dealer over the past 12 months. Then, we will pursue dealers to get the record of their regular customers…the farmers. After that, we will have to actually draw the sample. Decide on the number of buyers we would like to have in the final sample. For example sake, let's say we intend 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. Among those options, a relatively better procedure should be opted.
Here's a simple procedure that's especially useful if you have the names of the buyers 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 buyers’ 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. You could probably accomplish the whole thing in under a minute.

**Explanation 2:-**

To illustrate SRS in another way, let us take example of farming group. Suppose there are 80 farmers in the group. The first step is to identify each farmer by a number from 1 to 80. Suppose you decide to select a sample of 20 using the simple random sampling technique. Use the fishbowl draw, the table of random numbers or a computer program either to select the 20 farmers. In this way, these 20 farmers become the basis of your inquiry.

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:**

As already discussed, the accuracy of your estimate largely depends on the extent of variability or heterogeneity of the study population with respect to the characteristics that have a strong correlating with what you are trying to ascertain (Principle three). It follows; therefore, that if the heterogeneity in the population can be reduced by some means for a given sample size you can achieve greater accuracy in your estimate. Stratifies random sampling is based upon this logic. In stratified random sampling the researcher attempts to stratify the population in such a way that the population within a stratum is homogeneous with respect to the characteristic on the basis of which it is being stratified. It is important that the characteristics chosen as the basis of stratification are clearly identifiable in the study population. For example, it is much easier to stratify a population on the basis of gender than on the basis of age, income or attitude. It is also important for the characteristic that becomes the basis of stratification to be related to the main
variable that you are exploring. Once the sampling population has been separated into non-overlapping groups you select the required number of elements from each stratum, using the simple random sampling technique. There are two types of stratified sampling: *proportionate* and *disproportionate stratified sampling*. With proportionate stratified sampling, the number of elements from each stratum in relation to its proportion in the total population is selected, whereas in disproportionate stratified sampling, consideration is not given to the size of the stratum.

**Procedure for selecting a stratified sample**

**Steps to be followed:**

- **Step 1** Identify all elements or sampling units in the sampling population.
- **Step 2** Decide upon the different strata (k) into which you want to stratify the population.
- **Step 3** Place each element into the appropriate stratum.
- **Step 4** Number every element in each stratum separately.
- **Step 5** Decide the total sample size (n).
- **Step 6** Decide whether you want to select proportionate or disproportionate stratified sampling and follow accordingly the steps given below:

<table>
<thead>
<tr>
<th>Disproportionate stratified sampling</th>
<th>Proportionate stratified sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 7</strong> Determine the number of elements to be selected from each stratum</td>
<td><strong>Step 7</strong> Determine the proportion of each stratum in the study populating (p)</td>
</tr>
<tr>
<td>$\frac{\text{sample size (n)}}{\text{no.,of strata (k)}}$</td>
<td>$\frac{\text{elements (#) in each stratum}}{\text{total population size}}$</td>
</tr>
<tr>
<td><strong>Step 8</strong> Select the required number of elements from each stratum with SRS technique</td>
<td><strong>Step 8</strong> Determine the number of elements to be selected from each stratum</td>
</tr>
<tr>
<td>$= (\text{sample size} \times \text{p})$</td>
<td>$= (\text{sample size}) \times (\text{p})$</td>
</tr>
<tr>
<td><strong>Step 9</strong> Select the required number of elements from each stratum with SRS technique</td>
<td><strong>Step 9</strong></td>
</tr>
</tbody>
</table>

*As this method does not take the size of the stratum into consideration in the selection of the sample, it is called disproportionate stratified sampling.*

*As the sample selected is in proportion to the size of each stratum in the population, this method is called proportionate stratified sampling.*

**Cluster sampling**—simple random and stratifies sampling techniques are based on a researcher’s ability to identify each element in a population. It is easy to do this if the total sampling population is small, but if the populating is large, as in the case of a city, state or
country, it becomes difficult and expensive to identify each sampling unit. In such cases the use of cluster sampling is more appropriate.

Cluster sampling is based on the ability of the researcher to divide the sampling population into groups, called clusters, and then select elements within each cluster, using the SRS technique. Clusters can be formed on the basis of geographical proximity or a common characteristic that has a correlation with the main variable of the study (as in stratified sampling). Depending on the level of clustering, sometimes sampling may be done at different levels. The first level of cluster sampling could be at the province level, second could be at the district level & third could be the tehsil level. These levels constitute the different stages (single, double or multi) of clustering.

In cluster sampling, we follow these steps:

- Divide population into clusters (usually along geographic boundaries). Clusters may be grouped according to similar characteristics that ensure their compatibility in terms of focused population.
- randomly sample clusters
- measure all units within sampled clusters

**Multi-Stage Sampling**

The three methods we've covered so far -- simple, stratified and cluster -- are the simplest random sampling strategies. In most real applied social research, we 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**.

**Example:**

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.

**Non-random/non-probability sampling designs**

Non-probability sampling designs do not follow the theory of probability in the choice of elements from the sampling population. These sampling designs are used when the number of
elements in a population is either unknown or can not be individually identified. In these circumstances, it is not feasible, practical or theoretically sensible to do random sampling. Here, we consider a wide range of non-probabilistic alternatives.

We can divide non-probability 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**

Accidental sampling is also based upon convenience in accessing the sampling population. Whereas quota sampling attempts to include people possessing a obvious/visible characteristic, accidental sampling makes no such attempt.

This method of sampling is common among market research and newspaper reporters. It has more or less the same advantages and disadvantages as quota sampling. As you are guided by obvious characteristics, some people contacted may not have the required information.

**Judgmental or purposive sampling**

The primary consideration in purposive sampling is the judgment of the researcher as to who can provide the best information to achieve the objectives of the study.

In purposive sampling, we sample with a *purpose* in mind. We usually would have one or more specific predefined groups we are seeking. It is most likely that persons running into people in a mall or in the street carrying a clipboard, stopping various people and asking for interviews are conducting a purposive sampling (and most likely they are engaged in market research). 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 the concern of verifying 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 overweigh subgroups in your population that are more readily accessible.

This type of sampling is extremely useful when you want to construct a historical reality, describe a phenomenon or develop something about which only a little is known.

All of the methods mentioned hereunder 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 reason being that 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 sub case 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*.

**Proportional quota sampling**

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 that whether it will be by age, education, race or religion etc.

Nonproportional quota sampling
This category of 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 being adequately represented in your sample.

Heterogeneity Sampling
We Heterogeneity sampling is used when we sample for including 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.

*Concept mapping is a structured process, focused on a topic or construct of interest, involving input from one or more participants, that produces an interpretable pictorial view (concept map) of their ideas and concepts and how these are interrelated. Concept mapping helps people to think more effectively as a group without losing their individuality. It helps groups to manage the complexity of their ideas without trivializing them or losing detail.

Snowball sampling
Snowball sampling is the process of selecting a sample using networks. To start with, a few individuals in a group or organization are selected and the required information is collected from
them. They are then asked to identify other people in the group or organization, and the people
selected by them become the part of the sample. Information is collected from them, and then
these people are asked to identify other members of the group and, in turn, those identified
become the basis of further data collection. This process is continued until the required number
or a saturation point has been reached, in terms of the information being sought.

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.

This method of selecting a sample is useful for studying communication patterns, decision
making or diffusion of knowledge within a group. There are disadvantages to this technique,
however. The choice of the entire sample rests upon the choice of individuals at the first stage. If
they belong to a particular faction or have strong biases, the study may be biased. Also, it is
difficult to use this technique when the sample becomes fairly large.

‘Mixed’ sampling Design

Systematic sampling design

Systematic sampling has been classified under the ‘mixed’ sampling category because it has the
characteristics of both random and non-random sampling designs.

In systematic sampling the sampling frame is first divided into a number of segments called
intervals. Then, from the first interval, using the SRS technique, one element is selected. The
selection of subsequent elements from other intervals is dependent upon the order of the element
selected in the first interval. If in the first interval it is the fifth element, the fifth element of each
subsequent interval will be chosen. Notice that from the first interval the choice of an element is
on a random basis, but the choice of the elements from subsequent intervals is dependent upon
the choice from the first, and hence cannot be classified as a random sample for this reason it has
been classified here as ‘mixed’ 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
- \( k = \frac{N}{n} \) = the interval size
- randomly select an integer between 1 to \( k \)
• then take every kth unit

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.

The calculation of sample size

‘How big a sample should I select?’, ‘What should be my sample size?’ and ‘How many cases do I need?’ These are the most common questions asked. Basically, it depends on what you want to do with the findings and what type of relationships you want to establish. Your purpose in undertaking research is the main determinant of the level of accuracy required in the results, and this level of accuracy is an important determinant of sample size. However, in qualitative research, as the main focus is to explore or describe a situation, issue, process or phenomenon, the question of sample size is less important. You usually collect data till you think you have reached saturation point in terms of discovering new information. Once you think you are not getting much new data from you respondents, you stop collecting further information. Of course, the diversity or heterogeneity in what you are trying to find out about plays an important role in
how fast you will reach saturation point. And remember: the greater the heterogeneity or diversity in what you are trying to find out about, the greater the number of respondents you need to contact to reach saturation point. In determining the size of your sample for quantitative studies and in particular for cause-and-effect studies, you need to consider the following:

- At what **level of confidence** do you want to test your results, findings or hypotheses?
- With what **degree of accuracy** do you wish to estimate the population parameters?
- What is the estimated **level of variation** (standard deviation), with respect to the main variable you are studying, in the study population?

Answering these questions is necessary regardless of whether you intend to determine the sample size yourself or have an expert do it for you. The size of the sample is important for testing a hypothesis or establishing an association, but for other studies the general rule is *the larger the sample size, the more accurate will be your estimate*. In practice, your budget determines the size of your sample. Your skills in selecting a sample, within the constraints of your budget, lie in the way you select your elements so that they effectively and adequately represent your sampling population.

To illustrate this procedure let us take the example of a class. Suppose you want to find out the average age of the students within an accuracy of 0.5 of a year; that is, you can tolerate an error of half a year on either side of the true average age. Let us also assume that you want to find the average age within half a year of accuracy at 95 per cent confidence level; that is, you want to be 95 per cent confident about your findings.

The formula (form statistics) for determining the confidence limits is:

\[
\hat{x} = \bar{x} \pm (t_{0.05}) \frac{\sigma}{\sqrt{n}}
\]

Where

\(\hat{x}\) = estimated value of the population mean
\(\bar{x}\) = average age calculated from the sample
\(t_{0.05}\) = value of t at 95 per cent confidence level

\(\frac{\sigma}{\sqrt{n}}\) = standard error
\(\sigma\) = standard deviation
\(n\) = sample size
\(\sqrt{\cdot}\) = square root

If we decide to tolerate an error of \(\frac{1}{2}\) years, that means:

\[
\bar{x} \pm (t_{0.05}) \frac{\sigma}{\sqrt{n}} = \bar{x} \pm 0.5
\]
In other words we would like \[(t_{0.05}) \frac{\sigma}{\sqrt{\eta}} = 0.5\]  
Or \[(1.96^*) \frac{\sigma}{\sqrt{\eta}} = 0.5\] (value of \(t_{0.05} = 1.96\))

\[
\therefore \sqrt{\eta} = \frac{1.96 \times \sigma}{0.5}
\]

*\(t\)-value from the following table

<table>
<thead>
<tr>
<th>Level</th>
<th>0.02</th>
<th>0.10</th>
<th>0.05</th>
<th>0.02</th>
<th>0.01</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>(t)-value</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
<td>3.291</td>
</tr>
</tbody>
</table>

There is only one unknown quantity in the above equation, that is \(\sigma\). Now the main problem is to find the value of \(\sigma\) without having to collect data. This is the biggest problem in estimating the sample size. Because of this it is important to know as much as possible about the study population.

The value of \(\sigma\) can be found by one of the following:

1. Guessing.
2. Consulting an expert.
3. Obtaining the value of \(\sigma\) from previous comparable studies; or
4. Carrying out a pilot study to calculate the value.

Let us assume that \(\sigma\) is 1 year.

\[
\therefore \sqrt{\eta} = \frac{1.96 \times 1}{0.5} = 3.92
\]

\[
\therefore \sqrt{\eta} = 15.37, \text{ say, 16}
\]

Hence, to determine the average age of the class at a level of 95 per cent accuracy (assuming \(\sigma = 1\) year) with \(\frac{1}{2}\) year, of error, a sample of at least 16 students in necessary.

Now assume that instead of 95 per cent, you want to be 99 per cent confident about the estimated age, tolerating an error of \(\frac{1}{2}\) year.

\[
\sqrt{\eta} = \frac{2.576 \times 1}{0.5}
\]

\[
= 5.15
\]

\[
\therefore \eta = 26.54, \text{ say, 27}
\]

Hence, if you want to be 99 per cent confident and are willing to tolerate an error of \(\frac{1}{2}\) year, you need to select a sample of 27 students. Similarly, you can calculate the sample size with varying values of \(\sigma\). Remember the golden rule: \textit{the greater the sample size, the more accurately your findings will reflect the ‘true’ picture.}