UNIT 6: SAMPLING AND SAMPLE DESIGN

Learning outcomes

By the end of this unit, you should be able to:

- Explain what sampling is and why we sample
- Explain the concept of statistical inference
- Highlight the basic concepts and factors to consider when determining the sample size
- Distinguish between probability and non-probability sampling
- Highlight the different sampling techniques and their limitations

6.1 Basics of Sampling Theory

Before moving on, let us define some terms we will use in the following sections.

Population: includes specified elements of interest to the researcher, which must be related to your research problem. A population is specifically defined using person, time and place. For example, your population could be all UNZA students in the School of Agriculture Sciences in the 2022 academic year.

Target population is the aggregation of the population you have identified from which you are to draw your sample. For example, your target population could be UNZA students enrolled for the research methodology course in 2022.

Sampling unit or element: is a case or a single unit available for selection from the target population, forming the basis for analysis, e.g. a person or thing. For instance, using the earlier examples, the sampling unit or element would be one student from the target population.

Sampling frame: A sample frame is a list that contains all eligible sampling units in your population of interest. From this list, the researcher selects units that will be in their study sample. For example, our sampling frame would be an electronic database of all students enrolled in the research methodology module in 2022.

Sample: A sample is a set of units selected from the sampling frame. In our case, it would be the students included in the study. What usually happens is that we study our sample, get the results and then mention generalisations about the original population. From the results we get from studying the students in our sample, we want to make a statement about all students enrolled in the research methodology module.

It is important to remember that the population does not have to mean everyone; as we said, it is specifically defined. For example, if your population of interest are female schoolteachers, it is okay to omit male schoolteachers from the sample. However, remember that the estimate you get from studying female schoolteachers cannot be used to make any generalisations about male schoolteachers.

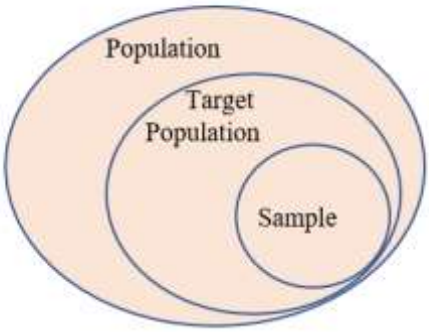

Figure 6.1: Illustration of a population, target population and sample

6.2 What is sampling and why do we sample?

Sampling is the act, process, or technique of selecting a suitable sample or a representative part of a population to determine the parameters or characteristics of the whole population.

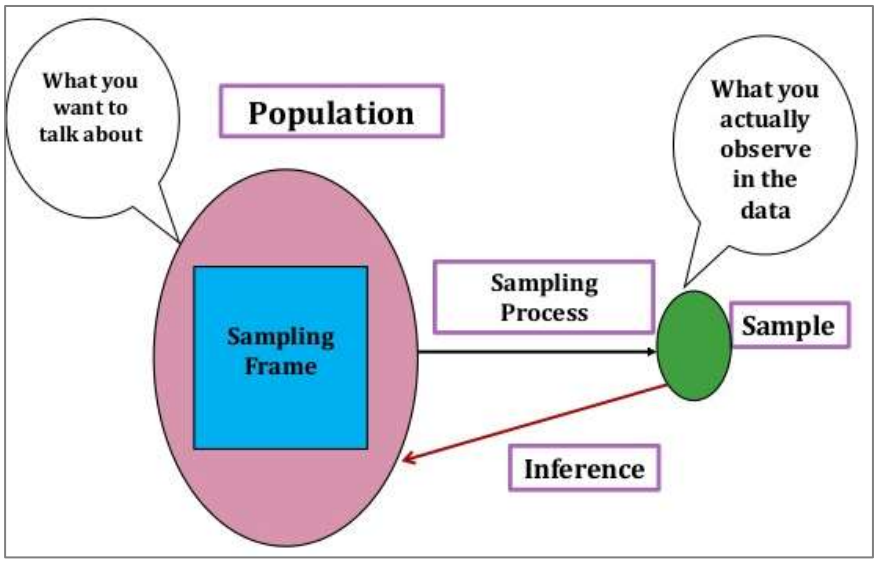

Figure 6.2: Illustration of sampling

We have probably touched on this but let us speak about it again. Sampling is a process of selecting the subset of the population, which we call a sample. Our sample is supposed to provide an adequate description of the target population. So, the figure above illustrates what sampling is. In the outer box, there is the population you want to talk about; inside, you have your sampling frame, which is the list of eligible units you will sample from. The next step is to conduct your sampling process using methods you will learn about in the following sections. So you execute those methods to get your sample which is the population subset. The sample is what you observe in the data as this is what you study. Finally, from the results you get from your sample, you then make an inference about your original population. In simple terms, this means that the results you get from your sample are true for your study population. So sampling allows us to make a statement about our original population.

6.2.1 Why do we sample?

If all population members were identical, the population would be homogenous, which means that the characteristics of any individual in the population would be the same as any other individual in that population. However, we know this is not the case as in the real world, each member is different from another individual, meaning our population is heterogeneous. In simpler terms, this means that there is a significant variation among individuals.

So, we use sampling because it is almost impossible to study everyone; it is expensive, unreasonable and impractical. Studying everyone is rarely done; a typical example would be a census conducted every ten years. Therefore, with a sample, we can make an inference by drawing conclusions from a sample about our original population. It also allows us to make generalisations about other populations. In other words, you can make conclusions about other populations you have not studied.

6.2.2 What do we mean by Statistical inference?

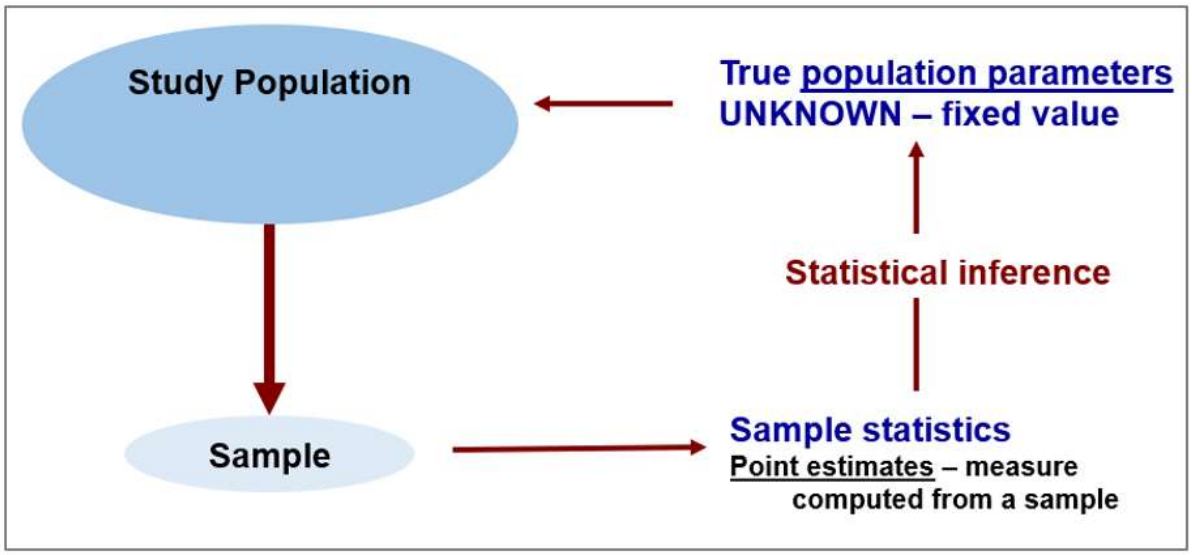

Figure 6.3: Illustration of how statistical inference helps us make inferences about the study population using sample statistics

Statistical inference uses sample statistics to make inferences about the study population (Casella & Berger, 2021), as shown in Figure 6.3. Statistical inference allows us to make conclusions about our population from the results that we get from our samples. Using Figure 6.3 as an example, suppose we want to study all UNZA students in the School of Agriculture; because there are so many, we take a sample of all Agriculture economics students registered at UNZA. Then we will study them and determine the proportion of international students among the sampled students. So our sample statistics will be the proportion of international students. We will use statistical inference to assess whether the proportion we have attained from our sample mirrors the actual population.

6.2.3 Sample statistic

If you are selecting a subset of your population which is your sample, it is possible that your sample may not be very well representative of your original population. Even if you used the correct sampling methods, it is possible that by chance, the sample that you get may not represent well your original population; this is called *a random sampling error*. This error can be minimised only if the sample highly represents the original population.

Therefore, we can assess how close our sample estimate is to the actual population using statistical inferences. This type of statistical inference is called precision, it shows us how close the sample estimate is to the actual population. *It is worth keeping in mind that we do not know the actual value of the population.*

a. The first way we can check how close or far off our sample estimate is from the true unknown population estimate or value is called estimation. With estimation, we use confidence intervals to check how far away or close the result we obtained from our sample is from the true population value. The Central limit theorem explains the background theory as to how we can use confidence intervals to verify whether the results we have from our sample are close or far away from the unknown population parameter. *The central limit theorem states that when an infinite number of successive random samples are taken from a population, the sampling distribution of the means of those samples will become approximately normally distributed as the sample size (n) becomes larger, irrespective of the shape of the population distribution.*

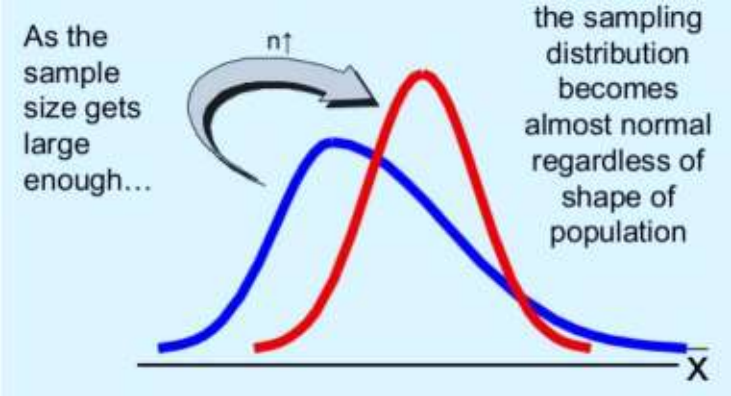

Figure 6.4: The Central limit theorem

- **Standard error of the mean:** the standard error of the mean (SEM=S/√n) measures the degree of variation between means we would expect from repeated samples of the population.

- **SEM=S/√n where S= sample standard deviation and n = sample size :** the standard error of the mean allows us to calculate the 95% confidence interval. The 95% confidence interval is an interval within which the population parameter will likely lie within that specific degree of confidence. Remember, 95% CI tells us that we are 95% confident that the true population value would lie between that range's minimum value and maximum value.

- **The interval $\bar{x} \pm 1.96 * S/\sqrt{n}$ is called the 95% CI for the population mean:** So as you can see from this formula, if you calculate your 95% confidence interval, you must include your standard deviation divided by the square root of your sample size, "which is your standard error of means". Remember that when we use this formula to calculate the 95% CI, we use the standard value of 1.96 and then the mean you get from your sample.

Class Activity: Standard error of the mean

1. If we take repeated samples from a population, do you think the sample estimate will always be exactly equal to the population parameter?

2. Will the sample estimates always be equal to each other?

b. Hypothesis testing (p-values): the second method used to test how close or far our sample estimate is from the true population value is called hypothesis testing. With this, we use p-values to estimate how far or close our estimate is to the unknown population parameter.

- A *statistical hypothesis* is an assumption about a population parameter. This assumption may or may not be true.

- *Hypothesis testing* is used to determine the probability – how likely it is that observed results in our sample are entirely due to sampling error rather than true underlying population parameter.

Decision Errors

Table 6.1: Decision errors

		Truth (for population studied)	
		Null Hypothesis True	**Null Hypothesis False**
Decision (based on sample)	Reject Null Hypothesis	Type I Error	Correct decision
	Fail to reject Null Hypothesis	Correct Decision	Type II Error

To make correct decisions about the result you get from your sample or to infer the population parameter correctly, you have to set your significance level to be as low as possible. Also, you have to set the power of your statistical test to be as high as possible.

6.3 Sample size estimation

Estimating the sample size is an integral part of all studies. We need an appropriate sample size for our study to minimise the chance of having a sample not representative of the source population (Kadam & Bhalerao, 2010). A sample size that is too large will waste resources, while a sample that is too small may lead to inconclusive or imprecise results. A result is inconclusive if the sample size is too small to declare an observed effect statistically significant. A result is imprecise if the small sample size results in wide confidence limits for the estimated population characteristic.

The appropriate sample size will depend on the following:

- The Research question(s) - particularly variation within the population concerning the characteristic of interest that you are investigating
- The Study design
- Types of outcome measures (categorical or numerical data)
- The expected value of the population parameter (like the expected prevalence of a stunted growth condition)
- Required level of precision- How much sampling error can be tolerated (margin of error around the estimate)?
- The formulae differ depending on whether one is interested in estimation or hypothesis testing.

6.3.1 Sample Size Calculation for a Prevalence Study (Single Proportion)

You will need to specify the following:

- The anticipated population proportion (p): e.g., the prevalence of mycotoxins in the population's animal feed is usually derived from the literature or a pilot study.
- The confidence interval: states that the population proportion is within certain specified limits; usually, researchers use a 95% confidence level;
- The desired precision (d): The acceptable margin of error on either side of the proportion, as this will determine the width of the confidence interval.

Assuming random sampling from a large population, we use the following formula to calculate the minimum sample size for estimating a single proportion (prevalence study).

$$\mathbf{n= (}p\ (1-p)\ z^2)/d^2$$

Where:

n is the minimum sample size
p is the anticipated population proportion
d is the precision required on either side of the proportion
z = 1.96 (the cut-off value of the Normal distribution at the 95% confidence level).

Example: Sample size calculation for a single propotion

In this exmaple we will use a reseach paper: **Mokubedi, S.M., Phoku, J.Z., Changwa, R.N., Gbashi, S. and Njobeh, P.B., 2019. Analysis of mycotoxins contamination in poultry feeds manufactured in selected provinces of South Africa using UHPLC-MS/MS. *Toxins*, *11*(8), p.452.**

A researcher wants to estimate the sample size for a prevalence study of mycotoxins contamination in poultry feed manufactured in Lusaka Zambia. How many retailers should be included in the sample if:

They need a precision (d) of 5%

They would like to use 95% confidence intervals (z)

Anticipated population proportion (p): 67%

$n= (p\ (100-p)\ z^2)/d^2$

$n= (0.67 * (1-0.67) * 1.96\ ^2)/0.05^2$

n= 339.6

6.3.2 Sample size formula for testing a difference between two proportions

$$n = \frac{n'}{4}\left[1+\sqrt{1+\frac{4}{n'|p_2-p_1|}}\right]^2$$

$$\text{where } n' = \frac{\left[Z_{1-\alpha}\sqrt{2p(1-p)}+Z_{1-\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}\right]^2}{(p_1-p_2)^2}$$

$$\text{and } \quad p = \frac{p_1+p_2}{2}$$

For α = 0.05 (two-sided test) $Z_{1-\alpha} = 1.96$
For 80% power $Z_{1-\beta} = 0.84$
p1 and *p2* are the anticipated proportions

Stata immediate commands:

To compute sample size for two-sample proportions with 80% power

sampsi p1 p2, p(0.8)

How to compute power given n1 and n2?

sampsi p1 p2, n1(n1) n2(n2)

6.3.3 Sample size formula for testing a difference between two means

$$n = \frac{\left(\sigma_1^2+\sigma_2^2\right)(Z_{1-\alpha}+Z_{1-\beta})^2}{(\mu_1-\mu_2)^2}$$

For α = 0.05 (two-sided test) $Z_{1-\alpha} = 1.96$
For 80% power $Z_{1-\beta} = 0.84$
The anticipated mean and variance of the two samples are μ_1, σ_1^2 and μ_2, σ_2^2

Stata immediate commands:

To compute sample size for two-sample means with 80% power

sampsi mean1 mean2, sd1(sd1) sd2(sd2) p(0.8)

How to compute power given n1 and n2?

sampsi mean1 mean2, n1(n1) n2 (n2) sd1(sd1) sd2(sd2)

Class Activity: Testing differences between proportions

An existing survey a few years ago found that 70% of new farmers in district A and 60% of new farmers in district B had adequate knowledge about the on-going Enhanced Smallholder Agribusiness Promotion Program (E-SAPP).

A student is interested in doing a new survey given that guidelines related to market-oriented agriculture have changed.

We need to know:

Desired power

Expected proportions in each group (p1 and p2)

STATA immediate commands to compute sample size for two-sample proportions with 80% power

sampsi p1 p2, p(0.8)

Note that the software we will use is STATA.

```
. sampsi 0.7 0.6, p(0.8)

Estimated sample size for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
                    and p2 is the proportion in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.8000
            p1 =   0.7000
            p2 =   0.6000
         n2/n1 =   1.00

Estimated required sample sizes:

            n1 =      376
            n2 =      376
```

Interpretation: The student will need to recruit 376 farmers at each locality or geographical area to detect a 10% difference in the proportion with adequate knowledge.

Class Activity: Testing differences between means

But what if they are interested in describing knowledge as a continuous score? The previous survey used a score out of 20 and the mean score in clinic A was 18 (SD 8) and in clinic B was 15 (SD 10)

Therefore, we need to know the:

Expected means (mean1, mean2)

Expected standard deviations (sd1, sd2)

Desired power

STATA immediate command:

To compute sample size for two-sample means with 80% power

sampsi mean1 mean2, sd1(sd1) sd2(sd2) p(0.8)

How to compute power given n1 and n2?

sampsi mean1 mean2, n1(n1) n2(n2) sd1(sd1) sd2(sd2)

```
. sampsi 18 15, sd1(8) sd2(10) p(0.8)

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2
Assumptions:

         alpha =    0.0500   (two-sided)
         power =    0.8000
            m1 =        18
            m2 =        15
           sd1 =         8
           sd2 =        10
         n2/n1 =      1.00

Estimated required sample sizes:

            n1 =       144
            n2 =       144
```

Interpretation: The student needs to recruit 144 women from each clinic to detect a 3-unit difference in mean knowledge score.

KEY POINTS TO NOTE

- If you are estimating *more than one outcome*, you should calculate the sample size for the **rarest outcome**, or the **outcome with the smallest mean**.
- A larger sample size is needed to detect small differences or a low prevalence.
- The calculated sample size should be large enough to allow for **non-response** or **loss to follow-up**.
- What if required sample is too large to be feasible?
  i. Consider alternative study designs
  ii. Consider if larger minimum difference would be possible
  iii. Lower your precision
  iv. Reduce the desired power

6.4 Ways of Sampling

There are two broad classes of sampling techniques: random and non-random (see Table 6.2). Random sampling techniques are methods where each member of the population has a non-zero equal probability of being selected, which means that the likelihood of being selected is known in advance because we can calculate it, and it always exists in each member. We also refer to these methods as probability sampling methods as they ensure the independent selection of members. Each member has an equal chance of being selected for the sample. The second category of sampling techniques is the non-random sampling method, where members of the population have unknown different chances or probabilities of being selected into the sample. In other words, the probability of being selected is not equal for all members.

Table 6.2: Distinction between probability and non-probability sampling techniques

Probability Sampling	Non-Probability Sampling
Each sampling unit has a known and equal chance (probability) of being selected for the sample	Chance of each element being included in the sample cannot be calculated before hand
The **sample will be representative of the target population** regarding the variables of interest.	The selection of sampling units is based on the judgement of the researcher and **may or may not be representative of the target population.**
Allows us to make inferences i.e. the estimation of population parameters	Used mostly in Qualitative research
Removes conscious and unconscious sampling bias, hence more accurate than non-probability samples	Does not remove conscious and unconscious sampling bias
Examples: *Simple random sampling, systematic random sampling, stratified random sampling and cluster sampling.*	Examples: *Convenience sampling, snowball sampling and quota sampling.*

6.4.1 Random sampling methods

a. **A simple random sample**

The first random sampling method is called simple random sampling. Remember, for random sampling; we said that each member of the population has a known non-zero equal chance of being selected into the sample. Hence, we can calculate that using the formula below:

Probability = sample size/ population size

Supposing we have a population of 10,000 people and need a sample size of 400. So, the likelihood of being selected is 400/10,000, which gives 0.04, which means that each member of the population of 10,000 has an equal 4% chance of being selected for the sample of 400 people.

Method

i. Assign each member of the population a number
ii. Decide on the n(sample size) that you need
iii. Use a hat or random number table, or random number generator to select the "n" members

The pros of this method are that it accurately represents the population and is easy to use. The cons of this method are that if the sampling frame is too large, it would be impractical because we would have to assign each member of the population a number. Another con is that the minority subgroup of interest in a population may not be present in the sample.

b. A systematic random sampling

The second random sampling method is systematic random sampling. With this method, you use a sample frame to select members at fixed intervals.

Method

i. You first assign each member of the population a number
ii. Decide on the sample size that you want
iii. Calculate a selection interval, that is: the interval size(k) = N/n
= 10000/400
= 25
iv. Randomly select a starting point using a random number table or random number generator
v. Then finally, take every "kth " unit

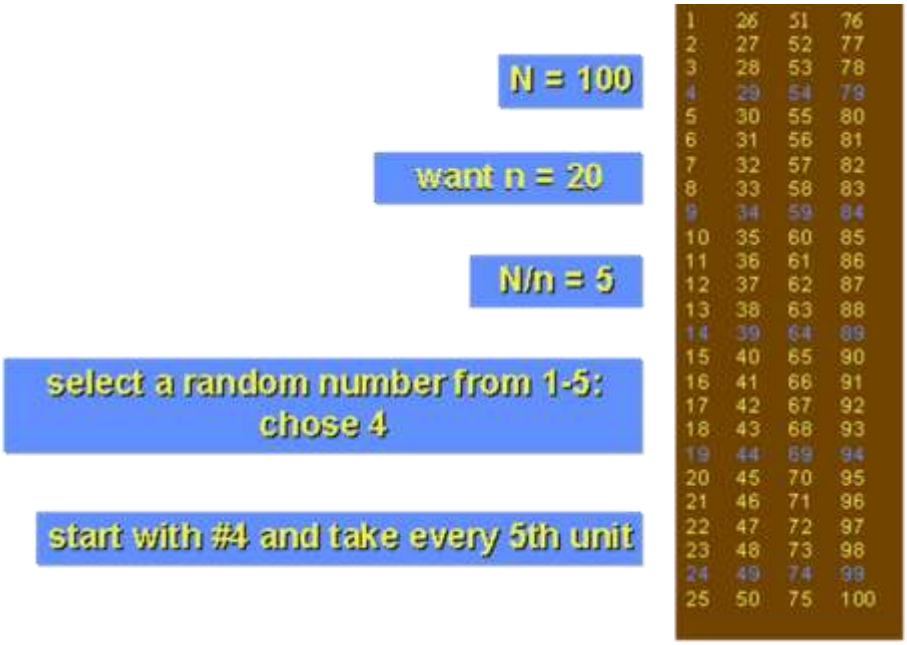

Figure 6.4: Illustration of how to select a systematic random sample

For instance, if you want a sample of 20 from 100 and you have their names listed on a piece of paper may be in alphabetical order. Using systematic random sampling, divide 100 by 20 to get 5. Randomly select any number between 1 and five. Suppose the number you pick is 4; that will be your starting number. So participant number 4 has been selected. From there, you will select every fifth name until you reach the last one, number one hundred. You will end up with 20 selected participants.

The advantage of this method is that it is simple, especially when a sample frame is not available, and it distributes the sample evenly over the entire population.

c. A stratified sample

The third random sampling method is called **stratified random sampling.** A stratum is a segment/subgroup of the population that shares at least one common characteristic, e.g. birth years, education, socio-economic status or gender. A stratified sample is obtained by independently selecting a simple random sample from each population stratum.

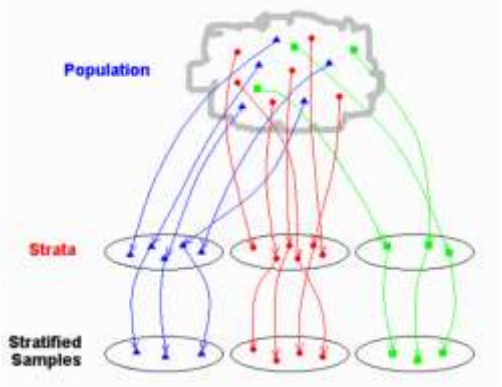

Figure 6.5: Illustration of a stratified random sample

Method

i. First, you distinguish all members in the population (that is, in the sampling frame) according to their value on some relevant characteristic. These are the sampling strata.
ii. Next, you randomly sample members from within these strata.
iii. Every unit in a stratum has a known non-zero equal probability of being selected.

If one stratum is not included, the sample is not representative.

- *Proportional Stratified Sampling*: The sample size in each stratum is proportional to the stratum size in the population.
- *Disproportional Stratified Sampling*: The sample size in each stratum is NOT proportional to the stratum size in the population.

A population can be divided into different groups based on some characteristic or variable like education or income. We refer to these groups as strata. For instance, anybody with ten years of education will be in Group A, between 10 and 20 in Group B and between 20 and 30 in Group C. You can then randomly select from each stratum a given number of units which may be based on proportion like if group A has 100 persons while group B has 50, and C has 30, you may take 10% of each. So, you end up with ten from Group A, five from Group B and three from Group C.

d. Cluster sampling

A cluster sample is obtained by selecting clusters from the population based on simple random sampling. The sample comprises a census of each random cluster selected—for example, a cluster, maybe something like a village, school, or state. So you decide all the elementary schools in Lusaka are clusters. You want 20 schools selected. You can use simple or systematic random sampling to select the schools, and then every school selected becomes a cluster. If you are interested in interviewing teachers on their opinion of some new program that has been introduced, then all the teachers in a cluster must be interviewed. However, cluster sampling is very susceptible to sampling bias. As for the above case, you are likely to get similar responses from teachers in one school because they interact with one another.

Difference between strata and clusters

Although strata and clusters are subsets of the population, they differ in several ways:

Table 6.3 Differences between strata and clusters

Characteristic	Stratified sampling	Cluster Sampling
Desired **internal** relationship	Subjects in the same stratum are similar to one another regarding the stratifying factor (homogenous)	Subjects in the same cluster are different from one another regarding the factor of interest (heterogeneous)
Desired **external** relationship	Each stratum is different from the other strata	Each cluster is similar to other clusters
Inclusion in the sample	All strata are represented in the sample	Only a subset of clusters are in the sample.

6.4.2 Non random sampling methods

a. Convenience sampling

The first one is convenience sampling, sometimes known as grab or opportunity sampling or accidental or haphazard sampling. It is a type of non-probability sampling involving the sample being drawn from that part of the population close at hand. That is, readily available and convenient.

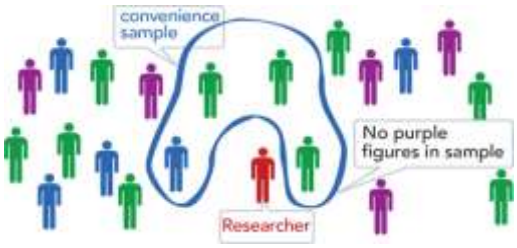

Figure 6.6: Illustration of Convenience sampling

The researcher using such a sample cannot scientifically make generalisations about the population from this sample because it needs to be more representative.

b. Purposive sampling

Purposive sampling is sometimes known as judgmental or subjective sampling. The researcher chooses the sample based on who they think would be appropriate for the study because they have specific predetermined characteristics. The technique is often used in qualitative research.

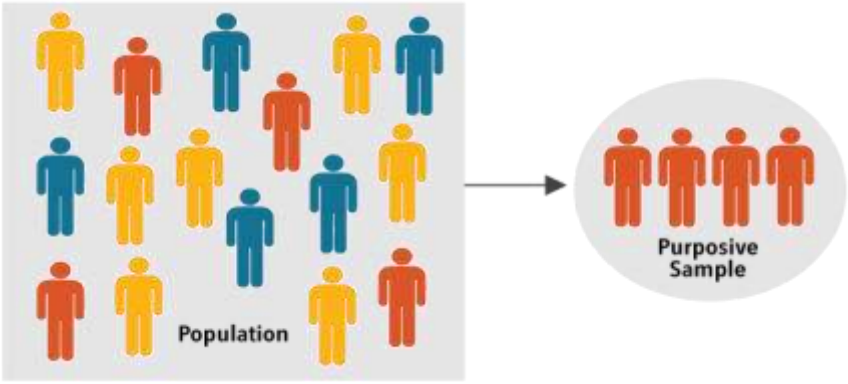

Figure 6.7: Illustration of Purposive Sampling

c. Snowball sampling

Selecting participants by finding one or two eligible people and then asking them to refer you to others. Used for populations that are difficult to access (e.g. users of a particular technology, drug users, sex workers, and homeless people).

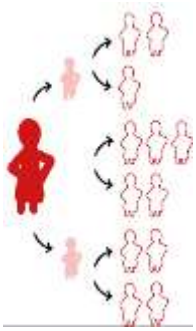

Figure 6.8: Illustration of Snowball Sampling

d. Volunteer sampling

The fourth non-random sampling method is volunteer sampling. This sample is made up of people who self-select into the survey. These people usually have personal reasons (e.g. interested in the topic, concerned about their health, need financial incentive offered) for volunteering to be in the study.

Figure 6.9: Illustration of Volunteer Sampling

6.5 Conclusion

The following is a summary of the steps in sampling. First, you must define your population precisely: person, place and time. You then determine your sampling frame; these are all available units from which you can select your sample. The next step is to determine the sampling method you want to use. For results that will be referred back to the original population, you will choose a random sampling method because each member of the population has a known non-zero probability of being selected into the sample, which ensures representativeness, and you can easily make generalisations.

Suppose the above is not crucial to your study; in that case, you will choose non-random sampling methods where members of the population have different chances of being selected into the sample; usually, this is when you want to focus on a specific group of people, e.g. poultry farmers. Once you have decided on your method, you determine your sample size and execute sampling.

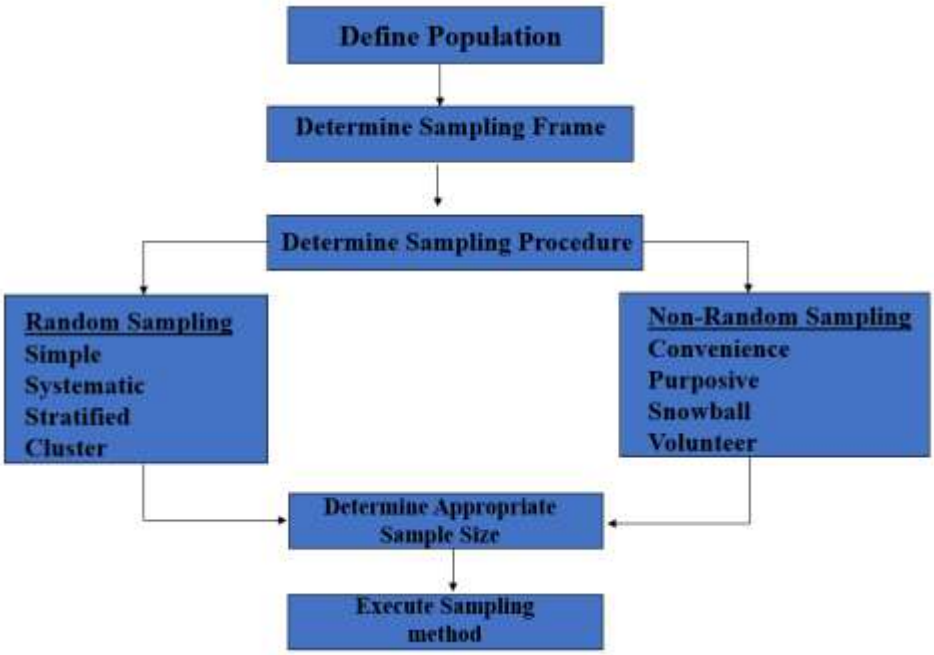

Figure 6.10: Summary of steps involved in sampling

References

Etikan, I. and Bala, K., 2017. Sampling and sampling methods. *Biometrics & Biostatistics International Journal*, *5(6), p.00149.*

Taherdoost, H., 2016. Sampling methods in research methodology; how to choose a sampling technique for research. *How to Choose a Sampling Technique for Research (April 10, 2016)*.

Casella, G. and Berger, R.L., 2021. *Statistical inference*. Cengage Learning.
Kadam, P. and Bhalerao, S., 2010. Sample size calculation. *International Journal of Ayurveda research*, *1*(1), p.55.