**6. Selecting Samples (Sampling)**

**A sample is a subgroup of the target population that the researcher plans to study for generalising about the target population.**

The reasons for using a sample rather than collecting data from the entire population:

1. It would be practically impossible to collect data several hundreds or thousands of people

 in a population. Even if it were possible, it would be prohibitive in terms of cost, time and

 other human resources.

2. Studying a sample rather than the entire population is more likely to lead to more reliable

 results, mostly because fatigue is reduced, resulting in fewer errors in collecting data.

Representativeness of samples

Probabilistic and Nonprobabilistice Sampling

In Quaantitative research, a researcher can employ either probability or nonprobability sampling approaches/strategies.

 Quantitative Sampling Strategies

 Probability Sampling Nonprobability Sampling

 Simple Stratefied Multistage Convenience Snowball

 random sampling cluster sampling sampling

 sampling sampling

In probability sampling, the researcher selects individuals from the pppulation who are representative of that population. This is the most rigorous form of sampling in quantitative rsearch because the investigator can clain that the sample is representative of the population and, as such, can make generalizations to the population.

**Sampling Designs (Techniques)**

There two types of sampling designs:

1. Probability sampling - the persons in the population have the known chance or probability

 of being selected. This design is used when the representativeness of the sample is of

 importance in the interests of widdr generalizability.

2. Nonprobability sampling - not every person in the population has the equal chance of

 being selected. This design is not for generalizability but critical for other factors.

Each of these two designs has different sampling strategies.

***Probability Sampling***

Types:

1. Simple random
2. Systematic
3. Stratefied random
4. Cluster

1. Simple random - Every person in the population has a ‘known-and-equal’ chance of being selected to make up a sample. E.g. there are 1000 people in the population and we need a sample of 100 persons. We can pieces of paper, each bearing the name of one of the persons into a hat and we can draw 100 of them from the hat. Each of those names has a 100/1000 chance of being drawn. The probability of any one of them being chosen as a subject is 1 and that everyone has the same or equal probability of being chosen. The sampling design is known as ***simple random sampling***. It has the least bias and offers the most generalizability. However, it is cumbersome and expensive and updating of the listing of the population is not possible. As a consequence complex probability sampling designs are used such as:

1. Systematic sampling
2. Stratified random samplinng
3. cluster sampling
4. Area sampling
5. Double sampling

2. Systematic sampling - This involves drawing every nth element in the population starting

 with a randomly chosen element between 1 and n.

 E.g. There are 260 houses and we want to have 35 houses for a sample. We can draw

 every 7th house starting from a random number from 1 to 7. Let us say that the random

 number is 7, then houses numbered 7, 14, 21, 28 and so on would be sampled until the 35

 houses are selected.

 However this systematic sampling design may suffer from systematic biasness. E.g. every 7th house is a corner house.

 For market surveys, consumer attitude surveys and the like, the systematicsampling design

 is often used.

3. Stratified random sampling - there are subgroups of elements existing within the

 population. E.g. the HR manager wants to train everyone employee in the organisation. Within the population of the organisation there different levels of people such as top management managers, middle-level managers and lower-level managers, first live supervisors, computer analysts, clerical workers and so on. The training requirements for different levels will be different. Data will need to be collected from each subgroup level in the population to assess the needs at each subgroup. In such a situation, stratified random sampling is used.

 This stratified random sampling is used to study buying habits of customers on the basis of life stages, income levels or geographical areas. Stratification ensures homogeneity within each subgroup or stratum.

4. Cluster sampling - intact grops, not individuals are randomly selected. All the members of selected groups hhave similar characteristics. E.g. instead of randomly selecting all fifthgraders in a large school district, you can randomly select fifthgrade classrooms and use all the students in each classroom.

 Cluster sampling is more convenient when the population is very large or spread out over a wide geographical area.

 It may be the only possible method of selecting a sample when the researcher is unable to obtain a list of all members of the population.

 Any location within which we find an intact group of similar characteristics is a cluster. Examples of clusters are classrooms, schools, city blocks, hospitals and department stores.

Benefits of cluster sampling:

 1. Involves less time and expense

 2. Generally it is more convenient than other techniques.

 3. It is easier to get permission to work with all members (students) in several classrooms

 than to work with a few students in many classrooms or to survey all the people in a

 limited number of city blocks than a few people in many city blocks.

Drawbacks of Cluster Sampling

1. The chances are greater of selecting a sample that is not representative of the population.
2. The smaller the sample size, the more likely that the sample selected may not represent the population.
3. Cluster samples offer more heterogeneity within groups and more homogeneity among groups.
4. Cluster sampling lends itself to greater biases and is the least generalizable of all the probability sampling designs (techniques) because most naturally occurring clusters in the organisational context do not contain heterogeneous elements. In other words, the conditions of intracluster heterogeneity and inter-homogeneity are often not met. For this reason cluster sampling is not well accepted in organisational research.

Single Stage and MultiStage Cluster Sampling

Cluster sampling can also be done in several stages and is then called ‘multistage cluster sampling. E.g. A national survey is to be done to determine the monthly bank deposits. Cluster sampling would first be used to select the urban, semiurban and rural geographical locations for study.

At the next stage particular areas in each of these locations would be chosen.

At the third stage banks in each area would be chosen

***Nonprobability Sampling***

In nonprobability sampling designs,the members in the population do not have the known and equal chance of being selected as sample subjects. This implies that the findings from the study of the sample cannot be confidently generalized to the population. This is the reason for using the nonprobability sampling designs. However, some nonprobability sampling designs are more dependable than others and could offer some important leads to potentially useful information with regard to the population.

Types:

1. Convenience sampling
2. Snowball sampling
3. Quota sampling
4. Purposive sampling
5. Self-selection sampling

1. Convenience sampling

Convenience or haphazard sampling involves selecting haphazardly those cases that are easiest to obtain for your sample, such as the person interviewed at random in a shopping centre for a television programme. The sample selection process is continued until your required sample size has been reached. This technique may be widely used but it is prone to bias and influneces that are beyond your control, as the individuals only appear in the sample because of the ease of obtaining them. Often the sample is intended to represent the total population, for example managers taking an MBA course as a surrogate for all managers! In such instances the choice of sample is likely to have biased the sample, meaniing that subsequent generalisations are likely to be at best flawed. These problems are less important where there is little variation in the population, and such samples often serve as pilots to studies using more structured samples.

2. Snowball sampling

This is an alternative to convenience sampling. In snowball sampleing, the researcher asks participants to identify others to become members ofr the sample. This is normally used when it is difficult to identify members of the desired populationl e.g. people who are working while claiming unemployment benefit. You need to:

1. Make contact with one or two persons in the population.
2. Ask these persons to identify further people.
3. Ask these new persons to identify further new people and so on.
4. Stop wheneither no new ones are givenor the sampe is as large as is manageable.

The main problem in the sampling method is making initial contact. It is used when there are difficulty to identify the people for the sample.

3. Quota Sampling

It is a form of proportionate stratified sampling in which a predetermined proportion of people are sampled from differentgroups, but on a convenience basis.

E.g. It is assumed that the work attitudes of blue-colar workers are quite different from those of white-collar workers. If there are 60% blue-colar workers and 40% white-collar workers in the organisation and that a total of 30 people are to be interviewed to find the answer to the research question. Then a quota of 18 blue-collar workers and 12 white-collar workers will form the sample because these numbers represent 60 and 40 percent of the sample size.

It is obvious that the sample will not be totally representative of the population and therefore the generalizability of the findings will be restricted.

However it is a convenience that the type of sampling technique can provide in terms of effort, costs and time. This technique becomes necessary when a subset of the population is underrepresented in the organisation e.g. minority groups, foremen and so on. Its good point is that it ensures that all the subgroups in the population are adequately represented in the sample. It also indicates that *quota samples are basically stratified samples from which subjects are selected nonrandomly.*

As the workplace or the society becomes more heterogeneous because of the changing demographics, quota sampling can be expected to be used more frequently in the future. E.g. quota sampling can be used to have some idea of the buying behaviour of various ethnic groups, for getting an understanding as to how employees from different nationalities perceive the organisational culture and so on.

Although quota sampling is not as generalizable as stratified random sampling, it does offer some information based on which further investigation can be done. This means that at the first stage of research the nonprobability design of quota sampling can be used and once some information has been attained, a probability design will follow. The opposite approach is also possible, by starting the research with probability sampling design and from the information received a new area for researh is indicated and a nonprobability sampling design might be used to explore further findings.

4. Purposive sampling

This approach is to obtain from specific target groups. Here, the sampling is confined to specific types of people who can provide the desired iinformation, either because they are the only ones who possess it or conform to some criteria set by the researcher. The two major types of purposive sampling are judgement sampling and quota sampling.

In judgement sampling, the choice of subjects are those who are in the best position to provide the information required. E.g. if a research wants to find out what it takes for women mangers to make it to the top, the only people who can provide firsthand information are the women who have risen to the positions of presidents, vice presidents and important top-level executives in work organisations. Therefore this sampling design is used when a limited number or category of people have the information that is sought. Obviously judgement sampling may curtail the generalizability of the findings. It is, however, the only viable sampling method for obtaining the type of information that is required from very specific pockets of people who alone possess the needed facts and can give the information sought. It requires the researcher to locate them.

5. Self-selection sampling

This occurs when you allow an individual to identify his/her desire to take part in the research. You therefore:

1. Publicise your need for cases, either by advertising through appropriate media or by asking them to take part.
2. Collect data from those who respond.

Cases that self-select often do so because of their feelings or opinionsabout the research question(s) or stated objectives.

E.g. Sila’s research was concerned with teleworking. She had decided to administer hor questionnaire using the Internet. She published her research on a range of bulletin boards and through the teleworkers’ association asking for volunteers to fill in a questionnaire. Those who responded were sent a short questionnaire by email.

Sample Size

* When selecting participants for a study, it is important to determine the size of the sample.
* A general rule of thumb is to select as large a sample as possible from the population. the larger the sample, the less the potential error that the sample will be different from the population. this difference between the sample estimate and the true population score is called ***sampling error***.
* If you were to select one sample after another, the average score of each sample would likely differ from the true average score for the entire population. For example, if we could obtain scores from sixth graders across the country about the importance of student-parent relationships, the average might be a 30 on a 50-point scale. If we obtain a sample from one school district and get an average score of 35 on the scale. The next time we might obtain a score of 33, and the next time a 36, because our sample will change from one schol district to another, this means that our average score is five points, three points and six points respectively, away from trhe “true” population average (i.e.30). This difference between the sample estimate and the true population socre is sampling error.
* Since it is usually not possible to know the true population score, it is important to select as large a sample as possible from the population to minimize sample error.
* However there are circumstances and situations, by which the number of participants may be limited, such as access, findings, the overall size of the population and the number of variables.
* One way to determine the sample size is to select a sufficient number of participants for the stattistical procedures you plan to use. This means that you have identified the statistic to use in analysis. As a rough estimate, a researcher may need:
* Approximately 15 participants in each group in an experiment
* Approximately 30 participants for a correlational study that relates variables.
* approximately 350 individuals for a *survey study* but this size will vary depending on several factors.

**The Importance of High Response Rate**

A sample that is selected must be a perfect representative sample of the population. This means that every member in the sampe is able to make the same response. If a sample is made up of 100 members in the small service business, it means the 100 members must be really involved in the small service business and they could make the necessary responses. This implies a high response rate because the sample provides a true representative of the population. It will then reduce the biasness because of the absence of non-responses and the cost of the survey.

Non-response can be due to 4 reasons:

1. Refused to respond.
2. Ineligibility to respond (not meeting the requirements).
3. Inability to locate respondent.
4. Respondent located but unable to make contact.

Calculation of response rate:

 Total number of responses

 Total response rate =

 Total number in sample - ineligible

Calculation of active response rate

 Total number of responses

 Active response rate =

 Total number in sample - (ineligible + unreachable)

Calculate Actual Sample Size from Estimated Response Rate

The sample size must be big enough to provide the necessay confidence in you data i.e. as a representative of the population.

First estimate the response rate from the sample and then calculate to increase the size of the sample using the formula:

 n x 100

 n2 =

 re%

 n2 = actual sample size

 n = minimum sample size

 re% = estimated response rate in percentage

Example:

John wanted to make a survey of customers and calculated that an adjusted minimum sample size of 439 was required. He estimated the response rate would be 30%. What is the actual sample size John should have?

Applying the formula:

 n x 100

 n2 =

 re%

 439 x 100

 =

 30

 = 1463

John’s actual sample size should be 1463.

It is important to know the response rate when selecting the sampe size. E.g. using post survey, it is found that only 30% response rate is achievable and for interview only 50%.

Questions

1. Identify a suitable sampling frame for each of the following research questions:

1. How do company directors of manufacturing firms of over 500 employees think a specified piece of legislation will affect their companies?
2. Which factors are important in accountants’ decisions regarding working in mainland China?
3. How do employees at MAY Bank in Seremban think the proposed introduction of compulsory Saturday working will affect their working lives?

2. Identify the relevant population for the following ressearch foci, and suggest the

 appropriate sampling design to investigate the issues, explaining why they are appropriate.

 Wherever necessary, identify the population frame as well.

1. A gun manufacturing firm would like to know the types of guns possessed by various age groups in Kuala Lumpur.
2. A hospital administrator wants to find out if the single parents wortking in the hospital have a higher rate of absenteeism than parents who are not single.
3. A researcher would like to assess the extent of pilferage in the materials storage warehouses of manufacturing firms in Selangor.
4. The director of human resources wants to investigate the relationship between drug abuse and dysfunctional behaviour of blue-collar workers in a particular plant.

3. a. Explain why cluster sampling is a probability sampling design.

 b. What are the advantages and disadvantages of cluster sampling?

 c. Describe a situation where you would consider the use of cluster sampling.

4. Use of a sample of 5000 is not necessarily better than using one of 500. How would you

 react to this statement?

5. Nonprobability sampling designs ought to be preferred to probability sampling designs in

 some cases. Explain with examples.

**ADDITIONAL NOTES**

**Level of Precision**

It is about how close is the estimate to the true population. It is dependent on the sample size.

Variability in Data

If a number of samples are taken from a population, their means will be different but they will show a normal distribution and there is a dispersion among them. The smaller the dispersion or variability the closer the sample is to the population.

It is possible to avoid having to take many samples from a population. This is done by taking a sample of at least 30 subjects from the population and this is good enough to estimate the variability of the sampling distribution of the sample mean. This variability is known as the standard error (SEₓ).

S

SEₓ = , where S is the standard of deviation of the sample

 √n and n the sample size.

The standard error (SEₓ) can be reduced, if the sample size is increased at a given standard of deviation of the sample.

**Level of Confidence/Risk Level**

According to the Central Limit Theorem, if a population is repeatedly sampled, the average value of the attribute by those samples will be equal to the true population value. The values obtained by these samples will be normally distributed around the true value. In a normal distribution about 95% of the sample values will be within 2 standard deviations (2α) of the true population.

 68.26%

 95.66%

99.73%

 **- 3α -2α -1α  *x***   **1α 2α 3α**

**Sample Data, Precision and Level of Confidence in Estimation**

For example, we want to estimate the daily mean (average) assembled PCBoard for the entire population of 200 workers. We take a sample of 40 workers. Let the sample mean *x* = 50 and the sample standard deviation S = 8. *x* the sample mean is the point estimate of µ, the population mean.

Next we want to construct the confidence interval around *x* to estimate the range within which µ would fall.

The standard error SE*x* and the level of confidence determine the width of the interval K is calculated as follows using the formula: µ = *x* ± KSE*x ,* where K = the t statistics for the level of confidence desired.

For example, SE*x*  = S/√n = 8/√40 = 8/6.324 = 1.265

From the table of critical values for t as given in the T-distribution table, we know that :

 For 90% level of confidence, the K value is 1.645

 For 95% level of confidence, the K value is 1.96

 For 99% level of confidence, the K value is 2.576

Therefore at 90% level of confidence, µ = 50 ± 1.645 (1.265) = 47.919 and 52.081. This indicates that a sample size of 40, we have 90% confidence that the true population mean (average number of assembled PCBoards) for all workers would fall between 47.919 and 52.081.

Similarly, if we want to be relatively more confidence at 99%, without increasing the sample size, we need to sacrifice the precision. In this case µ = 50 ± 2.576 (1.265) i.e. µ = 50 ± 3.259. The true value of µ would lie between 46.471 and 53.259. The width of the estimation has increased and therefore we are relatively less precise in estimating the population parameter. Therefore there is a trade-off between precision and level of confidence.

**Sample Size**

Suppose the manager of a bank wants to be 95% confidence that the expected monthly withdrawals in the bank will be within the confidence interval of ± RM400.

A study of a sample of customers indicate that the average withdrawal made by them has a standard deviation of RM2800. What sample size is needed under the above criteria and information given?

From the formula, population mean can be estimated: µ = *x* ± KSE*x*.

Since the confidence level needed is 95%, the corresponding K value is 1.96 (t table). The interval estimate of RM400 will have to encompass a dispersion of (1.96 x Standard error).

That is, 400 = 1.96 x SE*x*

 Therefore SE*x* = 400/1.96 = 204.08

 But SE*x*  = S/√n

 Therefore, 204.08 = 2800/√n

 √n = 2800/204.08

 = 13.72

 n = (13.72)2 = 188.23 ≈ 188

 Therefore the sample size needed is 188.

Suppose the bank has only 185 customers. In such a case, we need to apply the correction formula in order to estimate the sample size under the same conditions of precision (±RM400) and confidence level (95%).

The correction formula is given as SE*x* = S/√n x √N-n/N-1 , where N = size of population

and n = sample size.

Applying the correction formula, we get:

 204.8 = 2800/√n x √185 - n/184

 √n/√185-n = 2800/204.8 x √184 = 1.01146

 n/185-n = (1.01146)2 = 1.023051

 n = 1.023051 (185-n)

 = 189.2644 - 1.023051n

 n + 1.02305n = 189.2644

 2.02305n = 189.2644

 n = 189.2644/2.02305

 = 93.55 ≈ 94

Therefore, we need approximately 94 customers out of total 185 customers.

**Factors affecting the Decision on Sample Size**

Sample size does not just based on formula but it is also an art. **Rosco** (1975) proposes the following rules of thumb for determining the sample size:

1. Sample size larger than 30 but less than 500 are appropriate for most research.
2. Where samples are to be broken into subsamples such as male/female. Malay/Chines/Indian, etc., a minimum sample size of 30 for each category is necessary.
3. In multivariate data analysis (including multiple regression analysis) the sample size should be several times (preferably 10 times or more) as large as the number ov variables in the study.
4. For simple experimental research with tight experimental controls (matched pairs, etc.) successful research is possible with samples as small as 10 or 20 in size.

(Source: Mukesh Kumar, Salim Abdul Talib & T. Ramayah (2013). Business Research Methods, C.9 pp. 121-133.)