UNIT 2 FORMULATION OF RESEARCH PROBLEM

Structure

2.1 INTRODUCTION

If you examine most research papers or project reports/theses/dissertations, there are a number of components in each research article. These include:

a) Introduction wherein the researchers give a statement of the purpose in the form of a problem that needed to be tackled

b) A description of the design of the study

c) Methods used and how the sample was selected and the sample size (the number of subjects studied)

d) How the data was analyzed, and

e) The results, their interpretation and conclusions

This common thread among all studies clearly points out to us that research comprises a definite sequence of procedures, about which we studied in the last unit, which are undertaken one after the other. However, you must bear in mind that research involves a number of interrelated activities and the various operations are inter-dependent. What we mean by this is that what and how well you do in one step influences those, which follow.

The most important and crucial step are (b) and (c) mentioned above followed by (d).

In this unit our aim is to orient you to the major steps or phases in the research process. We will discuss each step or phase separately. It would be worthwhile to remember that each step or phase will consist of a number of operations or activities and decision-making. Even small omissions or errors can affect the quality of a research study in as much as the taste of a curry in which you put less salt or forgot to add salt.

Objectives

After studying this unit, you will be able to:

- discuss the procedures for selection of a suitable problem,
- enumerate the characteristics of a good research problem – important sources of selection, and identification of a problem,
- define/formulate the problem,
- specify the objectives of the research problem/hypothesis, and
- describe the study design including sample size and power.

2.2 SELECTION OF A SUITABLE PROBLEM

The first step in research is for the researcher to decide upon the general area or aspect of a subject area i.e. the general area of interest. For example, one researcher may be interested in working in the area of diabetes, someone else in HIV/AIDS. This by itself is not sufficient because at this stage our decision gives a broad area and indicates the road we choose from a number. But the specific details are lacking hence it does not tell us how to plan and organize the study, what procedures to use for data collection or analysis. Why is this so? Because we do not know what specific questions need to be answered in this general area?

The general area of interest defines only the range of subject matter within which you have to identify a specific problem. This implies that the general area tells us where to look for the problem. Therefore we need to formulate a specific problem. You will find that choosing a general area of interest is fairly easy but formulating a specific problem is more difficult.

Generally it is recommended that one should read in the area of interest to decide the specific problem one would like to work in. In some cases, experience, observations, practical concerns, curiosity or one's own previous research work helps us to decide. In case of many students, the counselor/guide suggests the topic.

A researcher may be interested to work on some topic in an area, about which not much is known. Or there may be phenomena that have been studied to some extent but there is still a lot that can be done in the area. If there is a highly developed theoretical system, the researcher may want to test specific predictions based on the theory. If the theory is tentative, one may want to verify the theory under the same or differing sets of conditions.

Practical issues and concerns also help to identify topics for research e.g. we may want to evaluate a programme e.g. the midday meal programme or find a feasible solution to problem(s) faced by a community. Personal values/interests also can be the basis for choosing a topic. In case of senior scientists, policies or availability of research funds may also determine one's choices. In a few cases the prestige and status associated with an area may determine which area one chooses to work in.

Just by selecting a topic one cannot plunge into research work. We need to have a good vision about the critical components i.e. what data is to be collected, by which methods and how to organize and analyze the data.

Formulating the problem helps us to define the goal in clear terms. Ultimately no researcher would like his/her work to be a meaningless exercise. But there is a risk if the goal is not known or ill defined. Three questions need to be addressed when formulating a problem:

i) What do I want to know? What are the questions/answers?

ii) Why do I want these particular questions to be answered?

iii) What are the possible solutions or answers to the questions/problems in terms that satisfy the rationale

The questions may differ from in their degree of specificity. Questions may need to be broken down into several specifying questions. For example, if the question is "Why is the mortality rate of infants and children high?" This now needs to be broken down into several questions related to particular aspects so that they are simple, pointed, limited and empirically verifiable questions.

So we may ask:

1) Do feeding practices influence mortality?
2) Do hygiene practices influence mortality?
3) Does birth weight influence mortality?
4) Does gestational age influence mortality?

You may realize that the four specified questions make it clearer as to what you need to study.

At this juncture, it is necessary to highlight that study of relevant literature in the field is a basic requirement. Also, before finalizing the topic one may consider discussing it with people who are experienced and have expertise in the field.

The discussion above focused on how to select a suitable problem of study. Once this basic task is done, we next move on to specify the objectives of the research problem. Let us get to know how to go about formulating the objectives.

2.3 SPECIFYING THE OBJECTIVES OF THE RESEARCH PROBLEM

Almost all studies can be regarded as exercises in measurement. Once we constructed the hypothesis we then need to formulate the objectives. Let us understand this with the help of an example.

For examples

A nutritionist is interested to undertake research in the broad area of *OBESITY.*

She now narrows down (after reading literature) to *'Obesity and Lipid Profiles of Children'*. This is her *Specific Topic*

Now she formulates the *Hypotheses*. Hypothesis, we learnt, is a tentative proposition suggested as a solution to a problem or as an explanation of some phenomenon. In this example, the hypothesis formulated is:

Hypothesis: Lipid Profiles (i.e. Serum Cholesterol, HDL, LDL and Triglycerides) are not influence by obesity.

This hypothesis guides the nutritionist but does not specifically identify what she will study.

So she needs to formulate *specific objectives* of the study. Specific objectives of the study should involve obtaining estimates of the measure or effect.

Sometimes we may express the objectives *qualitatively.*

For example: To determine whether obesity in childhood is associated with dyslipidemia

Or

To learn whether moderate daily alcohol consumption during pregnancy causes low birth weight

Or

To examine whether the severity of dyslipidemia is associated with grade of obesity in children

Alternatively we may have a *quantitative objective.*

Example: To determine the proportion of obese children who have elevated serum LDL, total cholesterol

If you compare the objectives presented above, you will find that the latter objective differs from the first three. In the first three objectives (which are qualitative) the answer basically "yes" or "no" but in the fourth objective we are trying to measure the proportion – so there is a quantitative dimension.

Next, it is important to know how to frame the objectives. Let us find out next.

How should we state objectives?

The study objectives should be stated in such a manner that the specific parameter to be measured should be made clear – explicitly or implicitly – so that the parameter to be measured in certain. The object of measurement (i.e. variable) will differ from one study to another. For example:

- One may study the difference or rates of incidence between those exposed to or unexposed to the study factor
- If the study is complicated i.e. there are multiple factors (multifactorial) one needs to explicitly mention these factors. Example, if we are looking at smoking and alcohol intake on cancer we should not only take into account both factors but also that there may be interaction between both.

Remember that if you start a study with poorly conceived objectives, you are more than likely to have a poor study. Also it is important to refine the objectives after writing them.

In the discussion so far we have learnt that by formulating the objectives, we identify what we want to study. But we also need to formulate our hypothesis which would guide us in meeting the objectives. The next section focuses on hypothesis.

2.4 FORMULATING HYPOTHESIS

Once the scientist makes the problem specific and manageable, he/she now proceeds to the next several steps, which are interrelated. This includes formulation of hypothesis (although this may not apply to all studies), operational definitions of concepts that have been incorporated into the hypothesis and the methods that need to be used for the study.

So what do we mean by formulating a hypothesis? In fact what is a hypothesis. Let us find out.

2.4.1 What is Hypothesis?

As already mentioned earlier the researcher formulates tentative answers or solutions to the research questions/problems. These proposed solutions or explanations constitute the hypothesis.

Hypotheses are suppositions that are tested by collecting facts that lead to their acceptance or rejection. They are not assumptions to be taken for granted neither are they beliefs that the investigator sets out to prove. They are "refutable predictions".

Many persons undertaking research, especially first time researchers do not formulate hypotheses. You may ask – Is this step necessary? Yes it is. Even if it is done implicitly it needs to be done. Why? The hypothesis guides the researcher to identify and select on those variables that need to be studied. Hypothesis gives the direction; it helps to determine what is the kind of data we need to collect and the way to organize it.

Hypothesis can be called provisional formulations or tentative solutions/answers to the problem we have chosen. Thus, we anticipate certain logical consequences after we collect the data and analyze it. When we get the results of our data we examine them against the hypothesis we had formulated. If they match i.e. our expectations materialize we accept the hypothesis. We say the hypothesis is proved. If not we reject the hypothesis and accept the alternative hypothesis. Sometimes we may need to formulate alternative hypothesis and test them.

A good hypothesis has several basic characteristics. We discuss some of them as follows:

i) *Providing direction*: Hypotheses provide direction to research and prevent review of irrelevant literature and collection of useless or excessive data. They enable you to classify the information from the stand point of both 'relevance' and 'organization'. This is necessary because, a given fact may be relevant with respect to one hypothesis and irrelevant with respect to another, or it may belong to one classification with regard to first hypothesis or to an entirely different classification with regard to the second. Thus, hypotheses ensure the collection of relevant data necessary to answer questions arising from the statement of the problem. For example, in a research problem, 'Study habits and achievement of Distance Education Learners', the researcher may frame the hypothesis – learners putting in more study hours achieve more in the examination. The researcher will collect data about the number of hours being put in by learners for study and their achievement in the examination.

ii) *Hypothesis should be testable*: Hypotheses should be stated in such a way as to indicate an expected difference or an expected relationship between the measures used in the research. The researcher should not state any hypothesis that she/he does not have reason to believe that it can be tested or evaluated by some objective means. Hypotheses are the propositions about the relationships between variables. These can be tested empirically. There is no relationship between attendance to personal contact programmes in a distance education course and achievement in examination. Such propositions can be tested by means of empirical data.

iii) *Hypothesis should be brief and clear*: Hypothesis should be stated clearly and briefly. It makes problems easier for the reader to understand and also for the researcher to test. The statement should be a concise statement of the relationship expected.

Now that we know what is a hypothesis and its characteristics, let us move on to review the different forms of hypothesis.

2.4.2 Forms of Hypothesis

We can state hypotheses in three ways. Let us understand this with the help of an example.

Example: If we are concerned as to whether infant mortality rates are different in the various geographic regions. The hypothesis could be:

1) Positive declaration (research hypothesis) The infant mortality rate is higher in one region than another

2) Negative declaration (null hypothesis) There is no difference between the infant mortality ratio of two regions

3) Implicit question: To study the association between infant mortality and geographic region

To arrive at some conclusions pertaining to a particular research problem, you would realize, a hypothesis is generally stated in testable form for its proper testing. It may

be stated either in *declarative* form, the *null* form or the *question* form. What do these three forms mean? Let us find out.

Declarative hypothesis

When a researcher makes a positive statement about the outcome of the study, we get a declarative hypothesis. For example, the hypothesis *'The performance of the non-anaemic healthy children on problem solving tasks is significantly higher than the anaemic children'* is stated in the declarative form. Here, the researcher makes an attempt to predict the future outcome. This prediction is based on the theoretical formulation of what should happen in a particular situation if the explanations of the behaviour (performance on problem solving tasks) which the researcher has given in his/her theory are correct.

Null hypothesis

A null hypothesis is a non-directional hypothesis that proposes no difference or no relationship. The usual form of such hypothesis is: *"There is no significant difference between the performance of two groups of students, one following the conventional system of education and the second following distance-mode of education.* Since a null hypothesis can be statistically tested, it is also known as "statistical hypothesis" or "testing hypothesis". The notation used for this is HO? This is the hypothesis under scrutiny and sought to be refuted by conducting a study. The proponents of null hypothesis emphasize that the researcher must remain unbiased throughout his/her research efforts. This view is defended on the basis of the fact that in this case the researcher neither predicts a result nor indicates a preconceived attitude that may influence his/her behaviour during the conduct of the study. On the other hand, those who criticize the use of null hypothesis argue that the researcher should indicate the direction of the outcomes of the study, wherever possible. It is further argued that predicting the results of a study is less awkward in phrasing a relationship, than in using the 'no different' phrase that is usual in the null form.

A null hypothesis challenges the assertion of a declarative hypothesis and also denies it altogether. It says even where it seems to hold good, it is so due to mere coincidence. It is for the researcher to reject the null hypothesis by showing that the outcome mentioned in the declarative hypothesis does occur and the quantum there of is so significant that it cannot easily be said to have occurred by chance. The reasons for rejecting the null hypothesis may differ. Sometimes the null hypothesis is rejected only when the probability of its having occurred by a mere chance is 1 out of 100 or .01 out of 1. In such instances, we consider the probability of its having occurred by chance to be too little to be considered, and we reject the chance component of the null hypothesis and take the occurrence to be due to a genuine tendency.

If the null hypothesis is found false, what alternate would be true? The *alternate hypothesis*, denoted by H1 is the opposite of HO that must be true when HO is false.

Hypothesis in question form

In the question-form-hypothesis, instead of stating what outcome is expected, a question is asked as to what the outcome will be e.g. if you are interested to find out whether instructions through video programmes have any positive effect on the learning of the students of the Master's Programme in Dietetics and Food Service Management. The declarative form of the hypothesis will be: *'Will Instruction through video programmes affect the learning of student of Distance Education?'* This statement shows that instruction through video programmes is not related to learning.

It is easier to state a hypothesis in question form because it appears to be quite useful to write down all the questions that one wants to answer in a particular research study. On the other hand, a researcher faces difficulties in predicting the outcome of the study and stating the hypothesis in declarative form. But it is worth noting that the question form is less powerful than the declarative or null form as a tool for

obtaining valid information, and it is generally advisable to state a hypothesis in directional i.e., declarative form to arrive at valid conclusions and generalizations. However, this last statement should not be taken as if it were a law in the practice and theory of research.

At this point it may be worthwhile to remember that there are many kinds of hypotheses. These are highlighted next.

Types of Hypotheses

There are many kinds of hypotheses that scientists work with:

a) We may consider the concept of '*association*' between exposure to a specific variable and a disease.

   Association refers to the statistical dependence between the two variables i.e. the degree to which the rate of disease in persons with a specific exposure is either higher or lower than the rate of disease among those without exposure. However it must be emphasized that an association does not mean that there is a cause and effect relationship. For example, we may find that those who eat vegetables and fruits have lower incidence of cancer. But that does not mean that eating vegetables and fruits can totally prevent or cure cancer.

b) Often want to see if there is a *causal association* i.e. whether a change in the frequency or quality of the variable we are exposing our subjects to (independent variable) there is a corresponding change in the frequency of the disease or the outcome we are interested in.

c) May also want to determine whether there is a *cause and effect* relationship.

What ever be the hypothesis, remember there are some important aspects to be looked into to judge the worth of a hypothesis in research. A good hypothesis must be:

i) consistent with known facts and theories, and might be even expected to predict or anticipate previously unkonwn data,

ii) able to explain the data in simpler terms,

iii) stated in the simplest possible terms, depending upon the complexity of the concepts involved in the research problem, and

iv) stated in a way that it can be tested for its being probably true or probably false, in order to arrive at conclusions in the form of empirical or operational statements.

So once the hypothesis has been formulated, we move our attention to the research design. We will get to know more about this aspect in the next section But first let us answer the questions given in check your progress exercise 1.

Check Your Progress Exercise 1

1) List the different components of a research article/dissertation.

   .....

   .....

   .....

2) What is a quantitative objectives? Explain, giving examples.

   .....

   .....

   .....

3) Define hypothesis. What are the different ways in which the hypothesis can be stated?

.....

.....

.....

.....

Now let us move on to the study of research designs.

2.5 THE DESIGN OF RESEARCH

Every activity runs smoothly when planned. In research like in architecture the design is to be made before you undertake the work. Why is it necessary? If we have the design with us, we can either correct any mistakes we make or improve upon the plan. This helps us to minimize waste of time, as well as, money and energy. If we look at the explanation of the term design – it explains almost everything – *"designing is the process of making decisions before the situation arises in which the decision has to be carried out. It is a process of deliberate anticipation directed toward bringing an expected situation under control".*

Some persons may say that although it is not put on paper, the design has been planned mentally. This is not enough because in academic settings and for projects, which require funding, no decisions can be made without submitting the plan.

When developing the research design, we need to take into account the following:

a) What the study is about and what are the types of data needed?

b) Why the study is being made?

c) Where the data needed can be found?

d) Where or in what area the study will be carried out?

e) What periods of time the study will include?

f) How much material or how many cases (sample) will be needed?

g) What bases of selection (sample) will be used?

h) What techniques of gathering data will be adopted?

i) How will the data be analyzed?

j) How best can these questions be decided upon so that the research purpose is achieved with minimum expenditure of money, time and energy?

Thus at this stage the decisions we make about what data is to be collected, and how the sample is to be selected, and how the data is to be organized. Thus *a research design is the arrangement of conditions for collections and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure.*

Formulating the research design has several advantages. These include:

1) To determine how much inaccuracy can be tolerated

2) The time required so that a time plan can be made

3) Costing of the project

4) Sample size and selection – how and where to obtain samples/subjects and how many?

5) What is the data and type of data to be collected – will the data you plan to collect help in testing your hypothesis and achieving your objectives.

A research design must comprise of the following:

a) The *sampling design*: this deals with the methods of selecting subjects for the study

b) The *observational design*: this relates to the conditions under which the observations are to be made

c) The *statistical design*: this deals with how many subjects are to be observed and how the observations (results) are to be analyzed

d) The *operational design*: this deals with the specific techniques by which the procedures specified in the sampling, statistical and observational designs can be carried out

It must be remembered that none of the above are independent. A decision about any one phase may influence a decision in any other phase. Also please remember that these phases overlap.

Last but not the least; a practical research design represents a compromise dictated by a number of practical considerations that are related to the actual conduct of social research.

Research designs differ depending on the research purpose. The *research purposes* may be grouped broadly under four general categories. These include:

a) To *gain familiarity with a phenomenon* or to *achieve new insights* into it, often in order to formulate a more precise research problem or to develop hypotheses. Studies having this purpose are known generally as *Exploratory* or *Formulative Studies.*

b) To *portray accurately the characteristics of a particular situation or group or individual* (with or without specific initial hypotheses about the nature of these characteristics). Studies having this purpose are known as *Descriptive Studies.*

c) To *determine the frequency with which something occurs* or with which it is associated with something else (usually but not necessarily, with a specific initial hypotheses). Studies having this purpose are known as *Diagnostic Studies.*

d) To *test a hypothesis of a causal relationship* between variables. Studies characterized by this purpose are called *Experimental Studies.*

In the *formulative or exploratory studies* the premium is on discovery of ideas and insights. Therefore, the research design corresponding to such studies must be flexible enough to permit consideration of different aspects of a phenomenon.

In the *descriptive and diagnostic studies*, the major consideration is accuracy. Therefore, the research design for such studies must be such that bias will be minimized and the reliability of the evidence collected maximized (we shall have occasion to consider 'bias' and 'reliability' in a later unit). These two studies, namely, descriptive and diagnostic, though somewhat different in purpose, present similar requirements for research design.

Studies aiming at testing causal hypotheses, i.e. *the experimental studies*, require procedures that will not only reduce bias and increase reliability but will also permit inferences about causality. Experiments are particularly suited to this end. In practice, however, these different types of study are not always sharply separable. Any given research may have in it elements of two or more of the functions we have described. In a single study, however, the primary emphasis is usually on only one of these

functions, and the study can be thought of as falling into the category corresponding to the major function emphasized by it.

We hope the discussion above must have oriented you to the different types of research studies and the purpose they are used for. The next important decision we take in research is linked with the sample size i.e. what should be the sample size that we need to study to be able to get accurate results. The next section focuses on this aspect.

2.6 SAMPLE SIZE CONSIDERATIONS

The overall goal of any study is accuracy in measurement i.e. the value of the phenomenon we measure should be with little error. However it is wise to remember that there are always errors in measurement. These sources of error are classified as either *random* or *systematic*. In any study our effort should be to reduce both types of error.

Precision corresponds to the reduction of random error. Precision can be improved by in two ways: the primary means is by increasing sample size but it can also be improved by modifying the design of the study. By doing this (modifying the design) we are able to increase the efficiency with which we obtain data from a given number of subjects.

What is systematic error?

Systematic error is validity consisting of two components:

a) the validity of the inferences drawn from the observations of the individuals we have studied (internal validity), and

b) validity of the inferences as they pertain to people outside the study population (external validity or generalizability).

What is random error?

It is believed that chance plays an important role in all biological phenomenons. Random error is that part of our experience that we cannot predict. Random error has many components but the most important one is the process of selecting the subjects for the study (sample). The process is called *sampling* and the error therein is *sampling error*. We will learn more about this later in Unit 6.

The primary way to reduce random error is to increase precision by having larger sample size. How do we know what should be the size of the sample? There are formulae that help us. These formulae relate the size of the sample to the following:

1) *Level of statistical significance (alpha – error).* The values observed in the sample are used either to reject or not reject a HO. But these values are subject to sampling fluctuation and may or may not lead to a correct conclusion. Two types of error can occur. Wrongly rejecting a true null hupothesis (HO) is called Type-I error as can be seen in Table 2.1. The probability of this error is popularly known as P-Value.The maximum P-value allowed in a study/problem is called the level of significance or sometimes α level.

Table 2.1: Error in statistical decision

Statistical Decision	Null Hypothesis	
	True	False
Rejected	Type-I error	✓
Not Rejected	✓	Type-II error

The P-value is kept at a low level, mostly less than 5% or $P<0.05$. We call a result statistically significant when $P<0.05$ and highly significant when $P<0.01$. The

complement of this is 95% used in confidence interval (CI) abut which we will learn later in Unit 13 in this course.

2) *Chance of missing a real effect (beta – error).* In the Table 2.1, you notice the second type of error is failure to reject HO when it is actually false. The probability of this error is denoted by β (beta) and this is also referred to as β-error. The complementary of probability of Type II error is called statistical power denoted by (1–β) about which we will read in a little while from now. Statistical power becomes a specially important consideration when the researcher does not want to miss a specified difference.
3) Magnitude of the effect
4) Disease rate in the absence of exposure or exposure prevalence in the absence of diseases
5) Relative size of the compared groups (i.e. ratio of exposed to unexposed subjects, or of cases to controls)

The formula specifies the size necessary to detect an effect at a given significance level, beta-error, etc. However it is important to remember that just using these formula do not lend rigour to the study because the significance level and beta error are arbitrary. Even estimating the exposure prevalence or diseases rates may be guess mates.

Then should we not look for something better? A more informative method for assessing desirable sample size is to use the power calculations.

What is Power?

As mentioned earlier, statistical power is the complement to the beta error i.e. the probability of Type II error or detecting as "statistically significant" a postulated level of effect. Thus power of a statistical test is the probability of correctly rejecting HO when it is false. This is the probability of getting a statistically significant result. The power of a test is high if it is able to detect small differences and reject HO.

For a non-zero effect the power of a study increases towards 1.0 as the sample size increases. We can plot power curves (as shown in the Figure 2.1) where we can plot power versus size for different levels of a postulated effect. Curves are provided for various combinations of the number of populations (k), the degrees of freedom (n = n–1), and the level of significance (α).

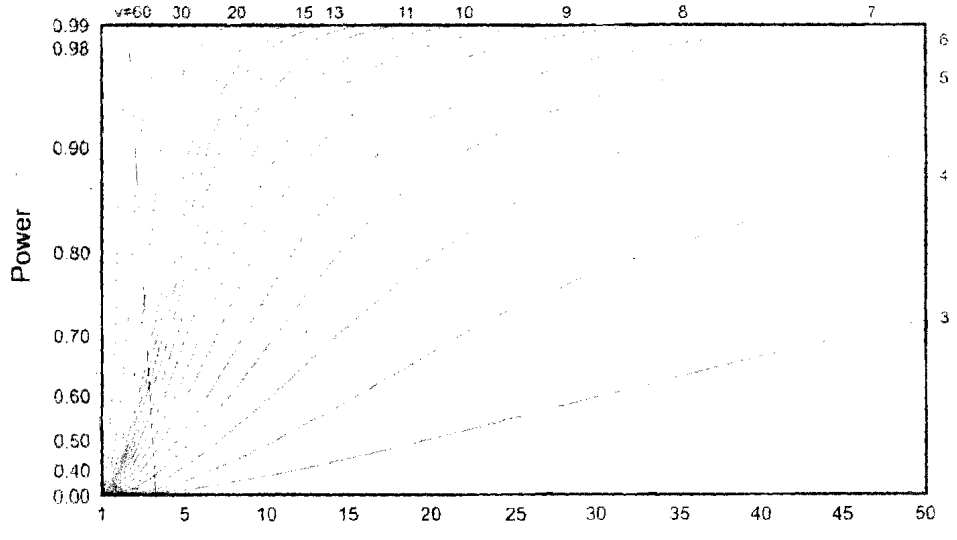

Figure 2.1: Power curve for α = 0.1 and k = 4

Source: Power curves for the analysis of means of variance. UNF Centre for Research and Consulting in Statistics. Wludyaka P, Nelson P, Silva P. 1999.

Suppose we plan an experiment in which samples of size 7 are to be used to test whether the variances of 4 populations are equal. Let the probability of rejecting the equal variance hypothesis when it is true be 0.1. Using Figure 2.1 ($\alpha = 0.1$, $k = 4$) and the curve for n = 6 one determines that the probability of rejecting the equality of variance hypothesis when the ratio of the largest variance to the smallest variance is 20 is approximately 0.86.

As shown in the figure, the curves indicate the power of a study of given size to detect the effect to which the sample size must be increased to achieve greater power at various levels of relative risk (RR). It is better to examine families of curves than a single number, since it reduces the arbitrariness. But it does not eliminate the arbitrariness completely since we still have an arbitrary aspect i.e. the reliance on a given level of statistical significance.

So what is an optimum sample?

An optimum sample is one, which fulfills the requirements of efficiency, representativeness, reliability and flexibility. Thus, the sample should be small enough to avoid unnecessary expense but large enough to avoid sample error and yield statistically representative and significant results. However we should not go for such a large sample that entails wastage of funds, or retard the project. Thus the size should be such that it yields the desired information with the required level of reliability at the least possible cost, also keeping in mind the limitations of time and personnel. Also it should be possible when and if necessary to increase or decrease the sample size, to meet unforeseen exigencies in the course of the study (without affecting reliability and efficiency).

The researcher must know statistics, e.g., percentages, averages, standard deviation etc. for such an estimation. We will learn about these measures in greater details later in Unit 12. This is important because different kinds of statistics for a particular degree of precision are provided by different sample sizes. Percentages or averages are the more commonly desired statistics.

Since the sample drawn by the researcher is only one of the many possible samples (of the universe) that she might have happened to choose, she needs to know how much reliance she can place on the sample as a representative of the 'universe' about which she wants to know something. She needs to know how large it should be to give her a satisfactory level of precision. This calculation is possible by recourse to mathematics since in random sampling (probability sampling design), the precision of the prediction or estimate is related to the square root of the number of items in the sample.

Before proceeding with the calculation of the required size of the sample for a given study, it is necessary in practice, to secure some preliminary information about the population or universe. If the researcher intends to use the sample to make an estimate of the average measure of a particular characteristic in the universe, he needs to have some preliminary estimate of the standard deviation in the distribution of the values of items in the universe with respect to the given characteristic.

The researcher who happens to know the range of values in respect of a particular characteristic in the universe can get a preliminary estimate of the standard deviation by dividing this range by 6, since the standard deviation of the (finite) universe may for all practical purposes be taken to be around 1/6 of the full range of variation. The preliminary information about the universe may be had by means of a pilot study, results of past surveys, experience of experts in the field, etc. Some other aspects we need to consider while calculating the sample size include:

a) *Margin of Error or Limit of Accuracy*: The basic question here is: 'How much is the percentage or average to be secured from the sample likely to vary from the true mean (for the population) and still be tolerated?' The researcher might tolerate 5% error or he might require accuracy within 2%.

b) *Probability or Confidence Level*: In addition to limit of accuracy, the researcher must also decide with reference to the study, how much confidence he would like to place in the sample estimates being so close to the true estimate as to be within the limits of tolerance or accuracy set for the study. In research, two degrees of probability or confidence are very well known and often used. One of theses is *0.95 level of probability*, i.e. there are 95 chances out of 100 that the sample estimate will not exceed the limits of tolerance or margin of error, and the second level is *the 0.99 level of probability*, i.e., there are 99 chances in 100 of not exceeding the margin of error aimed at. The level of confidence would not deviate from the true value (of universe) beyond the limits of tolerance. For certain purposes, the researcher may aim low and set the probability level *at 0.67* (i.e. 2 out of 3).

Having considered 1) the margin of error and 2) the probability or confidence level, the researcher can proceed with the calculation of a desired sample-size for different estimations as given in Table 2.2 and 2.3. Sample size calculation for different testing of hypothesis situations is given in Table 2.3.

Table 2.2: Sample size calculation for different estimations (valid for large *n* only)

Problem	Formula for Computing *n*	Description of the Notations
a) Estimating a population proportion with specified absolute precision	$\frac{z^2_{\alpha/2}\pi(1-\pi)}{d^2}$	π = anticipated value of the propotion in the population d = absolute precision required on either side of the proportion
b) Estimating a population proportion with specified relative precision	$\frac{z^2_{\alpha/2}\pi(1-\pi)}{(\varepsilon\pi)^2}$	π = anticipated value of the proportion in the population ε = relative precision in term of fraction
c) Estimating a population mean with specified precision	$\frac{z^2_{\alpha/2}\sigma^2}{L^2}$	σ = population SD (can be estimated from a pilot study) L = specified precision of the estimate on either side of mean
d) Estimating a difference between two population proportions with specified absolute precision - equal *n* in the two groups	$\frac{z^2_{\alpha/2}[\pi(1-\pi_1)+\pi_2(1-\pi_2)]}{d^2}$	π_1,π_2 = anticipated value of the proportions in the two populations d = absolute precision required on either side of the difference in proportions
e) Estimating a differences between means of two populations with specified precision - equal *n* in the two groups	$\frac{z^2_{\alpha/2}(\sigma^2{}_1+\sigma^2{}_2)]}{L^2}$	σ_1,σ_2= population, SD of the two populations (can be estimated from a pilot study) L = specified precision of the estimated difference on either side of the mean difference.

Large *n* is needed so that the distribution of *p* can be approximated by Gaussian form. For $\alpha = 0.10$, $z_{\alpha/2} = 1.645$; for $\alpha=0.05$, $z_{\alpha/2} = 1.96$. For other values, consult a Gaussian table. The formulas are based on two sided CI. In case of one-sided confidence bounds, replace $z_{\alpha/2}$ by z_{α}.

Table 2.3: Sample size calculation for different testing of hypothesis situations (valid for large *n* only) two-side H_1

Problem	Formula for Computing *n*	Description of the Notations
a) Hypothesis testing for a population proportion	$\frac{\left[z_{\alpha/2}\sqrt{\pi_0(1-\pi_0)}+z_{\beta}\sqrt{\pi_{\alpha}(1-\pi_{\alpha})}\right]^2}{(\pi_0-\pi_{\alpha})^2}$	π_0 = value of p under H_0 π_a = medically important value of population proportion under H_1 that is proposed to be detected.
b) Hypothesis testing for population mean	$\frac{\alpha^2(z_{\alpha/2}+z_{\alpha})^2}{(\mu_0-\mu_a)^2}$	σ = population SD (can be estimated from a pilot study)

		μ_0 =	value of population mean under H_0
		μ_a =	medically important value of population mean under H_1 that is proposed to be detected.
c)	Hypothesis testing for difference between two population proportions - equal n in the two groups.	$\frac{\left[z_{\alpha/2}\sqrt{2\pi(1-\pi)} + z_\beta\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_1)}\right]^2}{(\pi_1-\pi_2)^2}$	$\pi_1\pi_2$= anticipated proportions in the two populations π = $(\pi_1 + \pi_2)/2$ H_0:$\pi_1 = \pi_2$
d)	Hypothesis testing for difference between two population means eqaul n in the two groups	$\frac{(\sigma^2_1+\sigma^2_2)(z_{\alpha/2}+z_\beta)^2}{d^2}$	$\sigma_1\sigma_2$= population SD of the two populations (can be estimated from a pilot study) d = medically important difference between means under H_1 that is proposed to be detected.

α = is the level of significance and $(1-\beta)$ is the statistical power.
$z_{\alpha/2}$ is such that $\rho(Z > z_\alpha) = \alpha$. For $\alpha = 0.05$, $z_{\alpha/2} = 1.96$ and $z_\alpha = 1.645$.
For $\beta = 0.10$, $z_\beta = 1.28$; for $\beta = 0.05$, $z_\beta = 1.645$. For other values, consult a Gaussian table. If the alternative hypothesis is one-sided, replace $z_{\alpha/2}$ by z_α.

Source: Indrayan A, Satyanarayana L. Essentials of Biostatistics. & Basic Philosophy of Statistical Tests, Confidence Interval and Sample size determination. Indian Pediatr 2000, 37:739-750. Use of some of these formulae are illustrated here with the help of few examples.

Example 1: Suppose you as a researcher wishes to estimate the prevalence of anaemia among under three children. What should be the sample size for this investigation when simple random sampling is used?

For this first we need to consider the following:

a) What is the anticipated prevalence of anaemia among under three children? ----- Let us anticipate it to be around 70%.

b) What is the minimum degree of precision required? ------ The estimate should not be different by more than 5 per cent from the actual prevalence i.e. absolute precision d = 0.05, and

c) What least confidence in the estimate would be tolerable? ---- The chance of (ii) above should be atleast 95 per cent.

Thus we have p = 0.70, and d = 0.05 and α = 0.05 using formula (a) of Table 2.2 gives

$$n = \frac{(1.96)^2\, 0.70(1-0.70)}{(0.05)^2} = \frac{3.8416 \times 0.21}{0.0025}$$

$$n = 327$$

Example 2: Suppose you want to know the sample size required for detecting a mean difference of 0.3, mg/dl in haemoglobin levels in children. This difference is supposed to be considered clinically important. Let the required confidence level be 95%. Suppose the mean (SD) haemoglobin level available from a pilot study/literature is 10.2(1.9).

By substituting the population SD as 1.9 required precision L = 0.30 and z = 1.96 in formulae (c) of Table 2.2. We obtain the required sample size as follows:

$$n = \frac{(1.96)^2(1.9)^2}{(0.3)^2} = \frac{3.8416 \times 3.61}{0.09} = \frac{13.87}{0.09}$$

$$n = 154$$

Example 3: Suppose you want to know the sample size for detecting a mean difference of 0.3 mg/dl in haemoglobin level in children with two different interaction strategies. Let the required confidence level be 0.95%. Suppose the mean (SD)

haemoglobin levels available from a pilot study/literature in the two groups are 10.2(19) and 11.0(1.7) mg/dl. respectively.

Now by substituting the observed standard deviation as estimates of S_1 and S_2, required precision $L = 0.52$ and $z_{\alpha/2} = 1.96$ in the formula (e) of Table 2.2, we obtain the required sample size as follows:

$$n = \frac{(1.96)^2\left((1.9)^2+(1.7)^2\right)}{(0.3)^2} = \frac{3.8416\times(3.61+2.89)}{0.09}$$

$$= \frac{3.8416(6.5)}{0.09} = \frac{10.34}{0.09} = 114.88 = 115$$

The required sampe is 115 in each group. In order to to compensate for 10 per cent loss in followup, the sample size must be inflated by a factor of 100/90 = 1.11. Thus the final sample size needed in each group is 128.

Check Your Progress Exercise 2

1) List the advantages of formulating a research design.

.....

.....

.....

.....

2) How does a random error differ from a systematic error?

.....

.....

.....

.....

3) Explain the following in 2-3 sentences only:

a) Power: .....

.....

b) Level of confidence: .....

.....

c) Limit of accuracy: .....

.....

2.7 LET US SUM UP

In this unit we studied the components which constitute a research article. We learnt that the process of research begins with formulating a research problem. The process starts with the selection of a broad area of study and narrowing down to the specific problem of study. Various sources of selection and identification of a problem are available. Once the research topic has been finalized the objectives of the research problem need to be specified and hypothesis which would guide us in meeting the objectives need to be formulated. The unit further elaborated on the types of hypothesis and how to state these hypotheses.

Studies in various forms are carried out to acquire better and wider knowledge. The unit focused on the various research designs that can be used in research highlighting the advantages of formulating the research design and the various purposes and a brief overview of various types of designs.

The concepts and the constructs linked with sample size determination namely the power of the study, confidence level, level of accuracy etc. were described and the method of determining the sample size was explained.

2.8 GLOSSARY

Bias : in statistics, bias is a prejudice in a general or specific sense, usually in the sense for having a preference to one particular sample, perspective, external influence etc.

Null Hypothesis : the statistical hypothesis that states that there are no differences between observed and expected data.

Reliability : the amount of credence placed in a result. The precision of a measurement, as measured by the variance of repeated measurements of the same object.

2.9 ANSWERS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress Exercise 1

1) Different components in each research article include:
   a) Introduction wherein the researchers give a statement of the purpose in the form of a problem that needed to be tackled
   b) A description of the design of the study
   c) Methods used and how the sample was selected and the sample size (the number of subjects studied)
   d) How the data was analyzed
   e) The results, their interpretation and conclusions
2) When an objective is trying to measure the proportion it is a quantitative objective. Give example based on your own understanding.
3) *Hypotheses* are suppositions that are tested by collecting facts that lead to their acceptance or rejection. Hypothesis can be stated as;
   a) Positive declaration (research hypothesis) The infant mortality rate is higher in one region than another.
   b) Negative declaration (null hypothesis) There is no difference between the infant mortality ratio of two regions.
   c) Implicit question: To study the association between infant mortality and geographic region.

Check Your Progress Exercise 2

1) The various advantages of formulating a research design include:
   - To determine how much inaccuracy can be tolerated
   - The time required so that a time plan can be made
   - Costing of the project
   - Sample size and selection – how and where to obtain samples/subjects and how many?
   - What is the data and type of data to be collected – will the data you plan to collect help in testing your hypothesis and achieving your objectives.

2) Random error is that part of our experience that we cannot predict. It is based on the fact that chance plays an important role. Systematic error is the validity of the inferences drawn from the observations of the individuals we have studied (internal validity), and validity of the inferences as they pertain to people outside the study population (external validity).

3) a) Power is the probability of detecting as "statistically significant" a postulated level of effect.

   b) Confidence level refers to the confidence a researcher would like to place in the sample estimates being so close to the true estimate as to be within the limits of tolerance or accuracy set for the study.

   c) Limit of accuracy is the margin of error a researcher is willing to accept. The researcher might tolerate 5% error or he might require accuracy within 2%.