4th Edition

IAPSM's Textbook of COMMUNITY MEDICINE

Dr. Wise
Innovative & Genius
AI Chatbot
access code inside

As per the Revised Competency-based Medical Education Curriculum (NMC)

HIGHLIGHTS

- **Current Public Health Priorities:** Leprosy updates, Polio Endgame Strategy (2022–26), Measles–Rubella Elimination (2025–26), Ayushman Aarogya Mandir, and One Health
- **Latest National Health Data:** Updated epidemiological indicators from NFHS-5, SRS (2024–25), and recent national surveys
- **Revised National Programs and Policies:** Key updates in NTEP, NP-NCD, mental health, geriatric care, and Biomedical Waste Management Rules
- **Exam-Focused Content:** Streamlined, exam-oriented structure emphasizing concepts and clinical application

Editor-in-Chief
AM Kadri

Editors
Rashmi Kundapur
Amir Maroof Khan
Rakesh Kakkar
Ankit Sheth
Nidhi Mangrola

Foreword
Ashok Kumar Bhardwaj

JAYPEE

Contents

CHAPTER

11 Research Methodology and Biostatistics

Paragkumar Chavda, Mihir Rupani, Kalpita Shringarpure, Kedar Mehta

CM6.1	Formulate a research question for a study.
CM6.2	Describe and discuss the principles and demonstrate the methods of collection, classification, analysis, interpretation and presentation of statistical data.
CM6.3	Describe, discuss and demonstrate the application of elementary statistical methods including test of significance in various study designs.
CM6.4	Enumerate, discuss and demonstrate common sampling techniques, simple statistical methods, frequency distribution, measures of central tendency and dispersion.
CM7.3	Enumerate, describe and discuss the sources of epidemiological data.
CM7.9	Describe and demonstrate the application of computers in epidemiology.

INTRODUCTION TO HEALTH RESEARCH

While reading the history of medical science in this book, you would have come across the fact that in ancient times physicians tested a variety of treatments by trial-and-error methods. Some of them worked, while others did not. This was known as empirical practice. As medical science evolved, physicians started following experience-based practice; practices that worked well for patients became established. Lately, medical science has moved to a more rationalist approach, the evidence-based medical practice. This evidence comes from research studies which are scientifically robust. So, the journey of medical practice has passed through three "Es"—*empirical, experience, and evidence-based practice.*

Doctors, while treating patients, come across newer methods/modalities of treatment for diseases. A surgeon discovers a new technique of performing a surgery and feels that it is a better method compared to the standard way of performing that surgery. How does this surgeon convert his experience into evidence? It can be done only if the new method is tested and proven to be better using a scientifically planned and methodologically detailed research study.

In simple terms, research in medical science is nothing but a *scientific enquiry*. *Enquiry* means an investigation to find out the facts. The word "*scientific*" means that our investigation is carried using some scientific method.

In this chapter, we will go through the steps of this scientific process. But before that, let us understand broad domains of medical or health research. The scope of research is increasing in multiple fields related to medical science. **Table 11.1** outlines a simple grid to understand classification of health research.

Depending on the object of analysis, the research might be focused on the existing health problems or healthcare responses to such health problems. Depending on the level of analysis the research might be focused on analyzing information from individuals or from populations. Thus, the research that focuses on the description of health problem or disease in individuals falls under the domain of biomedical research.

Table 11.1: Domains of health research.

Object of analysis		*Health problems*	*Healthcare responses*
Level of analysis	Individual or subindividual	Biomedical research: • Biological processes • Body structure and function • Pathological mechanisms	Clinical research: • Natural history of diseases • Efficacy of preventive, diagnostic or therapeutic interventions
	Population—community health	Epidemiological research: Frequency, distribution, and causes of diseases	Health systems research: • Policy research • Operational research

When research focuses on the description of diseases at population level, it falls under the domain of epidemiological research. Research trying to find out the prevalence of goiter in a district would fall under this category. While research testing the efficacy of new drug or therapy on individuals falls in the domain of clinical research. When the research focuses on studying the organized response to health problems at population level it falls under domain of health systems research. An example of research in this domain would be trying to find out whether active case finding through home visits is more economical than passive case finding through clinics for detection of tuberculosis cases in a district, under National Tuberculosis Control Program. You will learn more about the health systems research in a separate chapter in this book.

Whatever be the domain of research, it has a potential to change *policy* or *practice*. For example, new research providing evidence that a new medication helps to reduce blood pressure

among hypertensive patients may influence the practice of the doctors as they start prescribing this new medication. Similarly, under National Tuberculosis Control Programme, research evidence showed that two sputum smear examinations were nearly as good compared to three sputum smears. This evidence changed the national policy from conducting three sputum smears to two sputum smears for diagnosis of tuberculosis.

Where will I Use it in My Professional Life?

Medicine is an ever-evolving science, and each interaction with the patient adds to a doctor's knowledge. When structured through research, these learnings become meaningful evidence that benefits the wider medical community. This is why every medical practitioner is also a potential researcher.

Beyond contributing to research, doctors also rely on it daily. Clinical decisions—from choosing diagnostic tests to selecting treatment options—are guided by research evidence. In this role as consumers of research, it is vital that doctors critically evaluate published studies rather than accepting them at face value. Understanding how to assess the quality, validity, and applicability of research is essential for safe and effective practice. Thus, research is not just an academic requirement but a lifelong component of medical professionalism, supporting both personal growth and improved patient care.

Thus, going through this chapter, you will be better equipped to deal with your role as *contributor* as well as *consumer* of research.

QUANTITATIVE AND QUALITATIVE APPROACHES TO RESEARCH

Dr Curious Goes on a Research Expedition

Dr Curious, a resident in the pulmonary medicine department, noticed an increasing number of drug-resistant tuberculosis (TB) cases in his outpatient department. Concerned, he discussed the issue with his professor, who pointed out that discontinuation of TB treatment—referred to as "loss to follow-up"—could be a key factor.

Determined to explore the issue, Dr Curious collected data from primary health centers in his district. He reviewed records of patients initiated on TB treatment in the past 6 months and found that out of 200 patients, 40 had discontinued treatment—indicating a 20% loss to follow-up, far exceeding the expected rate of less than 5%.

To understand the reasons behind this high default rate, he visited the homes of the patients who had dropped out of treatment. He conducted in-depth interviews with 18 such individuals, continuing until no new reasons emerged. The interviews were audio recorded, transcribed, and analyzed for themes.

Dr Curious found several key reasons for treatment discontinuation: patients stopped medication once symptoms subsided, experienced intolerable side effects, faced difficulties traveling to health centers, had conflicting work and clinic hours, or could not afford to miss work.

Through this research journey, Dr Curious gained valuable insights into both the magnitude and causes of loss to follow-up in TB treatment—highlighting the need for patient-centered interventions to improve adherence and outcomes.

As you read the story of Dr Curious, you must have observed that in the first scenario he was interested in answering the question "*What* is the proportion of patients lost to follow-up on treatment?" In the next scenario he was trying to find out the answers to the question, "*Why* do patients discontinue the treatment?" In the first instance Dr Curious has used a quantitative approach to research, while in the second scenario, he has used a qualitative approach to research. Let us briefly go through the salient features of both the approaches **(Table 11.2)**.

Table 11.2: Differences between qualitative and quantitative research.

	Quantitative research	*Qualitative research*
The purpose	Here the purpose is often to "estimate" a number (e.g., prevalence of patients defaulting on treatment)	Here the purpose is often to "explore" the different perspectives of the problem (e.g., to explore the reasons for default)
Basic process	We use numbers to quantify the problem. We have collection of numbers and data, e.g., we quantify problems such as treatment default	We use words to understand the problem in detail. We have narrative descriptions which are collection of words, e.g., we have the patient's story of why he or she defaulted from treatment
Primary information unit	The information collected is in form of numbers	The collected information is in form of words
Viewpoint	While deciding in whose perspective data is to be collected, the researcher's viewpoint (etic perspective) is given priority	While deciding in whose perspective data is to be collected, the respondent's viewpoint (emic perspective) is given priority
Design of research	The research design is very structured. The epidemiological study designs that you learned in previous chapter are used here	The research design is not fixed but it is flexible. The design evolves as the researcher collects data
Sample	This approach often involves collecting small amount of data from large number of people	This approach involves collecting large amount of data but from a small sample
Methods of data collection	The methods of data collection include surveys, interviews, and experiments	The methods of data collection include observation and interviews. Experiments are usually not used here
Analysis	Since we are dealing with large amount of data, we often make use of statistical software to analyze the data	We deal with narratives and hence software has limited role in analysis of such qualitative data. Much of the processing of such data happens within researcher's mind
Chronology of data management	Data analysis is possible only after the data is collected from all study participants	Data collection and analysis go hand in hand, e.g., researcher starts analyzing the reasons for default in her mind right from the time she takes interview of the first patient
Training required	The researcher is required to be trained in intricacies of research design and statistical analysis	The researcher is required to be trained in the art of qualitative data collection and analysis
To whom can the findings be applied	The findings of research are often generalizable to other populations too	The findings are specific to the individuals studied and often not generalizable to other populations

Each approach has its own strength and hence no approach is superior to the other. Depending on the type of research question you want to uncover you would differ in your chosen approach.

Researchers over the time have realized that both the approaches are complementary to each other. Hence, recently there is a new approach which makes use of both these approaches. Such approach is known as mixed methods approach. The research by Dr Curious described above is an example of mix-method approach.

INTRODUCTION TO QUANTITATIVE RESEARCH

Quantitative research is a systematic process of collecting, analyzing, and interpreting data to answer health-related questions or solve problems. "Systematic" implies that research follows an organized method, not a random sequence of actions. To maintain this structure, researchers apply epidemiological concepts and study designs, broadly classified into descriptive, analytical, and experimental types.

Steps in Conducting Research

We will now take you through the steps followed while conducting any quantitative research study **(Fig. 11.1)**. We will discuss each of these steps in detail in the subsequent sections of this chapter.

Selecting a Topic for Research

It is important to understand at this stage how our general curiosity about a topic is different from scientific curiosity. As part of our general curiosity, we may be interested in many aspects of a particular topic, but to convert it to scientific curiosity, we need to narrow our focus on a single aspect. Quantitative research, being a systematic detailed process, does not allow us to study multiple aspects of a topic simultaneously. If we attempt to do that, we will end up studying multiple problems superficially and may not be confident about the findings of our research. This is the reason why quantitative research expects the researchers to study only one aspect of a health problem with scientific rigor so that the findings and evidence generated is usable.

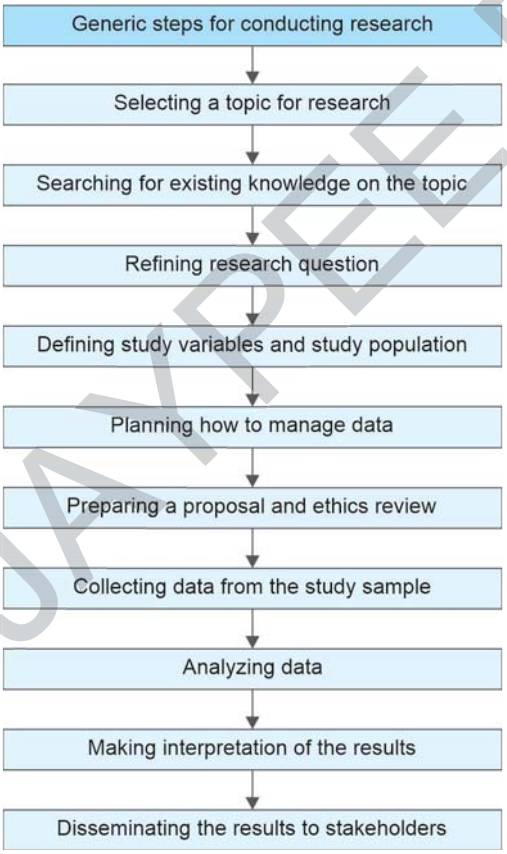


Fig. 11.1: Steps in the research process.

Searching for Existing Knowledge on the Topic

When we narrow down to a specific issue for research, we need to search if other people have worked in this area. It is possible that the topic that was interesting for us might have interested others also who might have already done research on this topic. So, the first step is to search for existing literature to know the current status of evidence in a particular issue of our interest (termed as review of literature). Earlier, researchers spent hours in libraries sifting through the library catalogues to find out related research. Fortunately, there are databases available in electronic format nowadays that make this Herculean task of finding the "relevant" literature an easy one. Google Scholar and PubMed are two of the commonly used tools for literature search.

Refining Research Question

Once we are done with reviewing the existing literature on the topic of our interest, we would be able to find out the gaps in the existing body of knowledge. It is prudent to refine our topic of inquiry towards the direction of this gap. While refining the research question, we will clarify what exactly we want to study and on whom.

Defining Study Variables and Study Population

Having a systematic approach expects us to be accurate in our measurements. So, defining what we want to measure in our study (output variables) is important. It is also important to select the group of people from whom we will gather our data.

Planning for Data Management

Many novice researchers skip this step unknowingly. For being systematic in our approach, we not only have to be specific about what information is to be collected, but also how we are going to collect such information. We have to plan in advance on what would be the sources of data, as well as methods and tools that will be used for data collection.

This step also involves preparing a plan for analysis. This is the advantage of quantitative research.

Preparing a Proposal and Ethics Review

Having completed all the earlier steps, we are ready with most of the ingredients for writing a proposal. A *study protocol* is a formal document containing the "plan for research". A "protocol" to a researcher is similar to that of a "blue print" to an architect. All scientific research involving human participants needs to process the study protocol through an ethics review committee before starting data collection.

Contents of a Study Protocol

- Project title
- Project summary
- Project description:
  - Background explaining need for study
  - Study objectives
  - Study methodology
  - Data management and analysis plan
- Ethical issues
- References

Collecting Data from the Study Sample

Having completed all the above steps, the researcher is now set to collect the data. If the data collection is done on paper, the collected data will have to be entered in a database later on.

Analyzing Data

In quantitative research, we take help of the methods of statistics to analyze the data. Fortunately, statistical software are now available and are commonly used by researchers for statistical analyzes.

Interpreting the Results

Making interpretation of the findings is one of the most crucial steps in the research process. The preceding step of statistical analysis takes care of chance errors in interpretation of results. Taking care of the systematic errors (bias) is also equally important. It is here that the researcher applies wisdom to examine the results in its entirety and not in isolation. The thought process which goes in this step often finds a place in the discussion section of the research report or published article.

Disseminating the Results to Stakeholders

The research process does not end at writing a research report. The findings of the study are commonly shared by researchers in form of a peer reviewed publication or presentation at a scientific conference. If the project involves receiving funding from an agency, a research report also needs to be shared with such an agency. When the research involves immediate policy implications, it should be shared with the policy makers. The research findings should also be shared with the study participants and public at large, since they are the beneficiaries of research.

With this background information on the steps to conduct research we are now ready to learn how to frame a research question.

FIRST STEP: FRAMING A GOOD RESEARCH QUESTION

The impact of good research goes to improving clinical practice and health of the community. Hence, it is vital that the first step of research is done correctly.

Converting a Research Topic into Answerable Research Question

A research question is the starting point of the entire research project. It provides a roadmap for implementing the study.

Fig. 11.2: Converting a research topic into answerable research question.

Generally, the *study protocol* is written only after the investigator finalizes the research question.

Any research usually starts with an idea or a topic of interest of the investigator. These ideas originate from clinical practice, teaching experience, meetings, journal articles, newspapers, talks with colleagues, and new technologies. The initial idea or topic is generally broad to start with. For example, the investigator is interested in addressing the issue of rising trend of heart disease in India. Heart disease is a very broad topic for research and the investigator, based on her interest, she needs to choose a domain—screening, treatment or prognosis of heart diseases—in order to make the topic more specific **(Fig. 11.2)**.

Next step would be to search the existing literature for what is already known on the topic. Research usually being a team work; at the third step, the inputs from a guide or a colleague are valuable in making our scientific curiosity focused and relevant.

Elements of a Research Question

There are two primary elements in a research question: the variable of interest and the population under study. In analytical research designs, there are two variables of interest: a predictor variable and an outcome variable.

Predictor variable is the risk factor (in observational studies) or the intervention (in experimental studies), which has its effect on the outcome variable. In a clinical trial, it is the drug that is administered or the procedure that is performed.

The outcome variable is the effect (disease) caused by the predictor (risk factor) or the consequence of the intervention. In analytical studies, it is the disease caused by the risk factor, while in clinical trials, it is the effect of the drug administered or the result of the performed procedure.

The **PICO** framework is a tool used in research to structure a clear and searchable question. It breaks down a research problem into four key components:

1. **P-Population:** Who are you studying? (e.g., adults with asthma)
2. **I-Intervention:** What is the treatment or exposure? (e.g., a new inhaler)
3. **C-Comparison:** What are you comparing it to? (e.g., a standard inhaler)
4. **O-Outcome:** What is the result you're measuring? (e.g., improved lung function)

The study population comprises of the research participants. It is the population to which the investigator wants to generalize the study findings to. Depending on the research topic, this can be children of a certain age, adults of a specific gender or a geriatric population with specific disease.

For analytical studies, putting together these three elements into a question form, a research question can be framed as: "Does the *predictor* cause the *outcome*, among the *study population*?" A research question should be phrased as a question; should be brief, clear, and focused. A few examples of analytical research questions are as follows:

- Does *brisk walking for at least 1 hour daily* reduce *fasting blood sugar level* among *adult type 2 diabetes mellitus patients*?
- Does *labetalol* decrease *blood pressure* as *compared with methyldopa* among *pregnant women with hypertension*? (*Note*: In clinical trials, it is prudent to add a comparator)

For descriptive studies measuring the prevalence of predictor variables (risk factors) or prevalence of outcome variables (disease), the research question would be slightly different in order to include two of the three elements described above. A few examples of descriptive research questions are as follows:

- What is the prevalence of high blood pressure [defined as systolic blood pressure (SBP) >120 or diastolic blood pressure >90 mm Hg] among adults above the age of 30 years residing in an urban slum?
- What is the prevalence of anemia [defined as hemoglobin (Hb) <12 g/dL] among adolescents between 10 and 19 years of age residing in a village?

Criteria for a Good Research Question

For a research question to be useful, there are a few criteria defined in the form of an acronym: **FINER**. The FINER criteria stand for feasible, interesting, novel, ethical, and relevant.

Feasible: For all researchers, the research question selected should be manageable within their scope, time, expertise, and funds available.

Interesting: The research question should be interesting to colleagues and experts working in the research area as well as to the funding agency.

Novel: A key characteristic of any research is that it brings out findings that were otherwise not known.

Ethical: A research study is ethical when the risk of research is acceptable in relation to the likely benefits.

Relevant: It is important to do research on a topic that is relevant to the current times. This ensures that the findings will be useful.

Framing Objectives of the Research

The objective/s of any research is/are translated from the research question itself. The difference is that the objectives are framed in scientific/epidemiologic terms making use of no more than one verb for each objective. Usually, the objectives are divided into primary (main) objective and secondary objectives (when applicable). The objectives differ according to the type of research questions: descriptive or analytical, and accordingly, the verb used will differ. In descriptive research questions, the prevalence of a risk factor or a disease is estimated. Therefore, the correct verb to be used is "estimate". In analytical studies, the association between a predictor and an outcome is to be determined. Therefore, the correct verb to be used is "determine". The objectives of the research should be SMART (specific, measurable, achievable, relevant, and time-bound).

For some of the research questions described above, the primary objectives can be framed as:

- To *determine* the effect of brisk walking for at least 1 hour daily on fasting blood sugar level of patients with adult type 2 diabetes mellitus.
- To *determine* the effect of labetalol on blood pressure of pregnant women with hypertension compared with methyldopa.
- To *determine* the effect of a high-fat diet on the development of dementia among adults older than 65 years.
- To *estimate* the prevalence of high blood pressure (defined as SBP >120 or diastolic blood pressure >90 mm Hg) among adults above the age of 30 years residing in an urban slum.

DEFINING STUDY VARIABLES

Variable of interest is one of the two primary elements in a research question.

Concept	*Example*
Variable	Systolic blood pressure
Observational unit	The person whose blood pressure is measured
Observation	The value of measurement (e.g., 122 mm Hg)

Variable is that attribute which varies. If we take systolic blood pressure (SBP) as example, its value varies from person-to-person. The person on whom blood pressure is measured is an *observational unit*. The value of SBP measurement on a person is known as *observation*. Thus, we might have several observational units in our study each one of whom would give us an observation on our variable of interest. SBP is a variable whose value varies not only from person-to-person, but also from time-to-time for the same person. When we measure blood pressure of a person twice a day we have two observations coming from the same observational unit. Such types of observations are called "paired observations". Later in this chapter we will explore the concept of paired observations in detail.

If there are 50 patients who participate in a research, each patient's weight would be different. Hence, patient's weight is a variable. Thus, we have information on weight of 50 patients participating in a research study. A collection of such values in kilogram is known as "data".

The data of a variable of interest is commonly collected in one of the two forms; quantitative and qualitative **(Fig. 11.3)**. (This is different from the quantitative and qualitative approaches for research that we discussed earlier in this chapter.)

Qualitative (Categorical) Data

Gender of the patient (data collected from those participating in research) is qualitative data. Unlike quantitative data, gender cannot be measured on a numbered range. But it is expressed as different categories such as males, females, transgender, etc. Thus, rather than numbers, qualitative data segregates people into different categories. The categorical data can be measured using two scales of measurement; nominal and ordinal. Data is said to be measured on an ordinal scale when the categories can be arranged in a meaningful order. For example, a patient's weight may be classified as normal, overweight or obese. Here order matters and hence it is said to be measured on an ordinal scale. Let us assume that we collect the data of patient's area of residence in categories such as urban, rural, and tribal. Here when the order is not important it is said to be measured on nominal scale. Thus, variables measured on nominal scale have

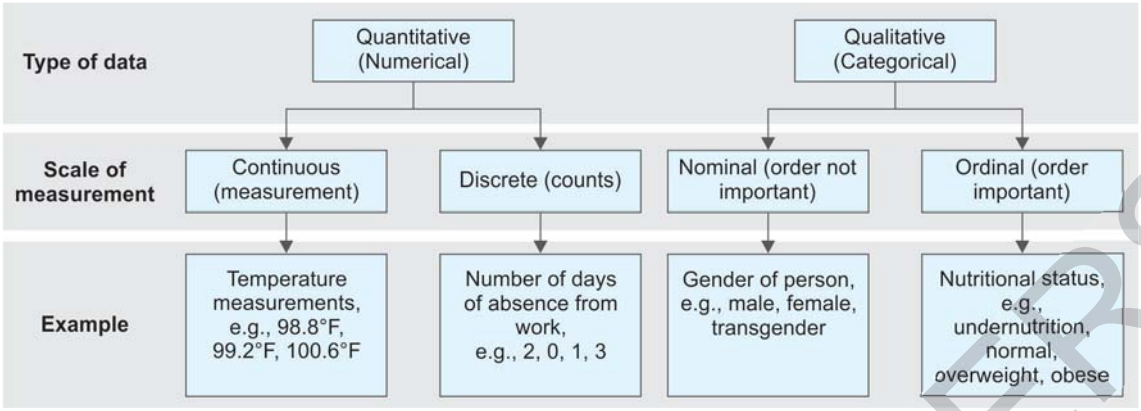


Fig. 11.3: Types of data and measurement scales.

no hierarchy. The choice of a statistical test to be applied will depend on whether the data is measured on nominal or ordinal scale. Qualitative data is summarized using proportions of people belonging to different categories.

Quantitative (Numerical) Data

Systolic blood pressure of patients measured in mm Hg is an example of quantitative data. Here, blood pressure can take any value across a range of numbers such as 80–200 mm Hg. Depending on whether the quantitative data is measurements or counts; they are put on a continuous or discrete scale. Later, we will see that quantitative data is summarized using mean or median.

In certain situations, we may want to convert the quantitative data into qualitative data for ease of understanding. For example, data on SBP which is otherwise measured on a continuous scale (numerical data) may be converted to categories such as normal, high normal, and hypertension which is ordinal scale (categorical data). Such conversion makes it easy for us to understand the data.

CHOOSING STUDY POPULATION

In real-life situations, the researcher does not have access to the whole population of interest. Thus, the researcher ends up conducting study with only a small group of people. In this regard, let us understand three concepts: (1) reference population, (2) accessible population, and (3) study sample **(Fig. 11.4)**. Let us take the following research question: Does brisk walking for at least 1 hour daily reduce fasting blood sugar level among patients with adult type 2 diabetes mellitus? Here type 2 diabetes mellitus patients of the whole world could be our reference population. This means that if this research proves that walking exercise reduces blood sugar, this evidence would be possibly useful for all the type 2 diabetics of the world. However, in real life we do not have access to all type 2 diabetics of the world. The researcher might have access to only those diabetic patients coming to her clinic. Such patients form what is known as a set of accessible population. Further, not the entire accessible population would participate in research. For example, some of them might not be willing to start a walking exercise. Thus, only those patients actually enrolled in the study would form the study sample.

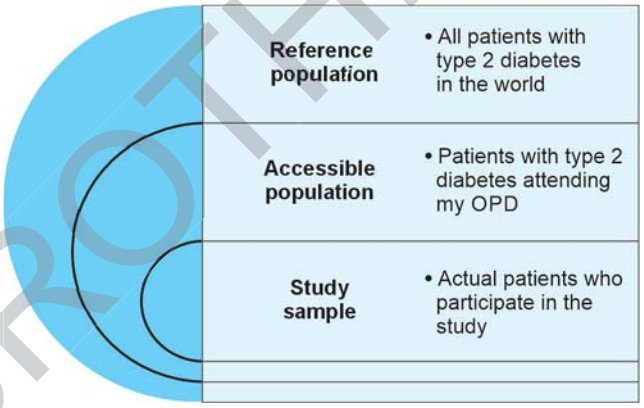


Fig. 11.4: Choosing study population.

For each of following research questions, identify which is the reference population.

- What is the prevalence of anemia among adolescents of a district X?
- What percentage of secondary schools are routinely procuring iron and folic acid supplements for their students in district X?
- What proportions of households in city Z are consuming green leafy vegetables at least thrice a week?

As you have observed in the first research question, all adolescents of the district are the reference population. In the second question, the secondary schools are reference population, while in the third question the households form a reference population. The population can consist of not only human beings but also hospitals, schools, or household, and many more such entities depending on the research question.

Sampling

To gain an idea of the next concept that we are going to cover, let us visit a grocery shop. From your past experience of visiting a grocery shop, you would realize that the shopkeeper keeps a sample of different qualities of grains for our inspection. We make the buying decision based on our assessment of quality of the sample of grains. Imagine if the shopkeeper cheats and mixes half quantity of poor-quality grains to the lot you purchase from this shop. This means that the sample you were shown was not representative of the lot you received. For

quantitative research, there are many reasons that makes the study sample not representative of the reference population. For the research question, estimating the prevalence of anemia among adolescents residing in a village, what if adolescents coming to our OPD are enrolled for the sake of convenience? In such scenario, our estimate of anemia prevalence might be falsely high since adolescents coming to OPD are more likely to be anemic. Thus, we will falsely overestimate the anemia prevalence.

When the sample is representative of the reference population, the researchers would be confident that the findings also apply to whole reference population. Fortunately, there are methods available to ensure that the sample we choose is representative of the reference population. Such methods are known as probability sampling methods.

Probability Sampling Methods

In probability sampling methods, every unit of the accessible population has a known probability of being selected. We briefly describe different probability sampling methods below **(Fig. 11.5)**.

Simple Random Sampling

Theoretically, this is the simplest method. The logic behind all type of random sampling methods is to eliminate the human choice from selecting study participants. If human choice is allowed, there are chances that some bias would make the study sample nonrepresentative of the population. In simple random sampling, we eliminate this human selection by making use of a computer program or random number tables to select which numbers from the list get selected in sample. A prerequisite for using this method is that a complete enumeration list of all members of the set of accessible population should be available. This is sometimes difficult. Imagine conducting the study of effect of walking on fasting blood sugar among diabetics. If the accessible population consists of patients coming to our OPD, we may not be sure who will come for OPD visit on a particular day. Thus, it becomes difficult to implement this method when we do not have a list of entire accessible population.

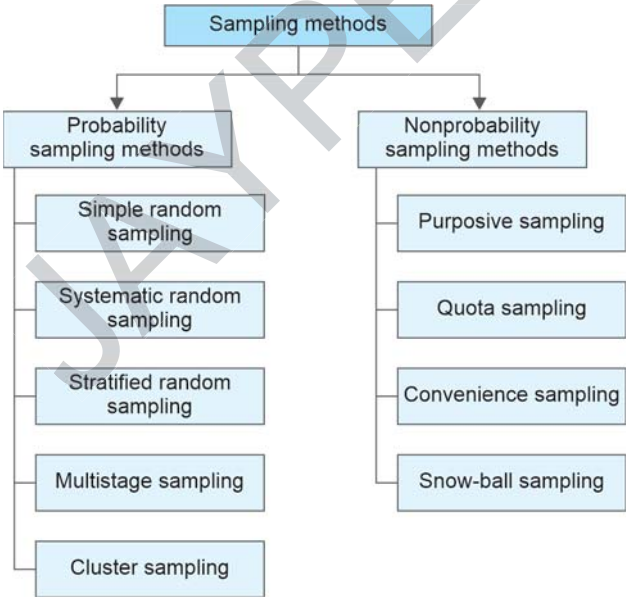


Fig. 11.5: Sampling methods.

Systematic Random Sampling

In this method, the name-wise enumeration of all the people of accessible population is not needed; a count of the accessible population is sufficient. Suppose we denote the required number of people in our study sample as "n". If we divide the number of people in accessible population by n, what we get is a sampling interval commonly denoted as "k". Suppose for our study on prevalence of hypertension among adults above 30 years of age in an urban slum, we estimate that there are 100 households in the slum. Let us assume that the number of households required to be enrolled as study sample is 20. The calculation would be as follows: size of accessible population/size of study sample = k. Substituting the numbers, we get: 100/20 = 5. We take every kth (every 5th in this example) household in our study sample. We select the first unit in our study sample by generating a random number between 1 and k. In this case, if that random number comes out to be 4, the first unit in our sample will be house no. 4 and then every fifth house will be taken in study sample till we reach the last (20th) unit of study sample which will be 99th house in this slum. Advantage of this method is that it uniformly covers all parts of accessible population.

Stratified Random Sampling

At times, we may want to separately estimate the variable of interest in some subgroups of the reference population. If we are conducting research on estimating the prevalence of obesity among patients having acute myocardial infarction, prevalence of obesity might be different in male and female patients. Thus, a researcher may choose to stratify the patients into the two gender groups and then select the study sample from each of these two groups separately. This will enable us to estimate the obesity prevalence separately among the patients of both the genders.

Multistage Sampling

This method is useful in situations where a large population is to be covered. Let us take an example of research involving a statewide survey of patient satisfaction with the medical care received from government primary health centers. For the first stage, we may choose some districts from the list of districts in the state using simple random sampling. In the second stage, we choose primary health centers from the list of all primary health centers (for each of the selected districts) using simple random sampling. This method adds convenience, since field visits for taking patient interviews need to be conducted in selected districts only.

Cluster Sampling

This is another popular method used in field research, when reference population is aggregated in naturally occurring clusters such as villages. In simple cluster sampling, 30 clusters are selected using simple random sampling from the list of all clusters. For example, if a district X has 300 villages 30 villages

will be selected using simple random sampling. In the next stage, the study sample will be selected from each of these clusters. The advantage of cluster sampling is that the list of accessible population is not needed, but the list of only the clusters is enough. Less travel is required since fixed (e.g., 30) number of clusters are to be visited. This method is commonly used to study the immunization coverage of particular area.

How is Cluster Sampling Different from Multistage Sampling?

Cluster sampling selects entire groups (clusters) from a population, while multistage sampling is an extension where you first randomly select clusters and then randomly sample within those selected clusters, adding more stages of selection for efficiency and precision, often in large populations. The key difference is that cluster sampling includes all members of chosen clusters, while multistage sampling only samples some members from selected clusters, making it more complex but flexible.

Nonprobability Sampling Methods

There are other methods which are known as nonprobability samples. Here, the probability of being selected is not known for the members of accessible population. These methods are used more frequently in qualitative research as we will discuss in later part of this chapter.

Adequacy of the Study Sample: Sample Size

The researcher is restricted by the resources and time she has for conducting research. Hence, she cannot include the entire reference population in the study. But, if the number of people studied (sample size) is very small it would be difficult for the researcher to be confident that the findings are true. So how much should be the size of sample for a study? There are different methods for calculating the sample size depending on the type of research design. For descriptive research design, the sample size can be calculated using simple formula with a paper and pen. For analytical research designs the formula for calculating the sample size are a little complex and hence we use computer programs to calculate the sample size for such analytical studies. We discuss here both the methods one by one.

Sample Size for Descriptive Studies

The first task in calculating sample size for descriptive studies is to determine if the variable of interest is measured on quantitative or qualitative scale.

Calculating sample size when variable is qualitative:

$$n = \frac{Z^2pq}{L^2}$$

Here,

- n = sample size
- z = confidence level
- p = proportion of variable of interest (%)
- q = compliment of p, [1-p] (%)
- L = allowable error on either side of the estimated 'pi'

Confidence level	Z score
80%	1.28
90%	1.64
95%	1.96
98%	2.32
99%	2.57

Let us take an example. In earlier research, the researcher wanted to estimate prevalence of anemia among the adolescent boys residing in a district. Based on literature review, suppose the prevalence of anemia is reported to be around 30% among adolescent boys (p = 30). We decide that we want to detect anemia prevalence in this district with precision of 5% on either side of the expected prevalence (L = 5). As researchers we want to be 95% confident about our results so we keep the confidence level at 95%.

The corresponding Z value is 1.96 (Z = 1.96). We will learn more about Z later in this chapter. Now replacing the numbers in formula, we get the following result.

$$n = \frac{Z^2pq}{L^2} = \frac{(1.96)^2 \times (30) \times (70)}{5^2} = \frac{3.84 \times 2{,}100}{25} = 322.5 = 323$$

Thus, we need to enroll 323 adolescent boys from the district to be able to estimate the anemia prevalence with 95% confidence level and 5% precision on either side of the estimated prevalence. Here, we assume that the sample will be chosen using simple random sampling.

What does allowable error at 5% on either side of p actually mean? It means that if the prevalence of anemia among adolescent boys in this study comes at 32% then the true prevalence of anemia among adolescent boys in this district is likely to be in the range of 32 + 5, i.e., anywhere between 27% and 37%. 95% confidence level means that if this study were to be repeated 100 times, for 95 of such studies the sample mean would be in the range of 27–37%. Now, let us try calculating sample size while changing the allowable error.

$$n = \frac{Z^2pq}{L^2} = \frac{(1.96)^2 \times (30) \times (70)}{3^2} = \frac{3.84 \times 2{,}100}{9} = 896$$

$$n = \frac{Z^2pq}{L^2} = \frac{(1.96)^2 \times (30) \times (70)}{7^2} = \frac{3.84 \times 2{,}100}{49} = 164.6 \approx 165$$

$$n = \frac{Z^2pq}{L^2} = \frac{(1.96)^2 \times (30) \times (70)}{20^2} = \frac{3.84 \times 2{,}100}{400} = 20.1 \approx 20$$

You will observe that as we try to become more precise in our estimate (keeping L small) the required sample size increases. Inversely, if we relax our allowable error, the required sample size would decrease. We saw in this last example that by increasing allowable error we can complete the research by studying a sample as small as 20 participants. But the drawback of such a choice would be that the estimate of true prevalence will be very wide, i.e., if prevalence of anemia in the study sample is 30%, the true prevalence in the district could be anywhere from 10% to 50%. Such a wide range of the estimated prevalence of anemia for the district is practically unusable for any policymaker. The allowable error chosen for a study has to be small enough to give a reasonably precise estimate and large enough that it is feasible for the researcher to collect data within the limited resources.

Sample size when variable is quantitative

$$n = \frac{Z^2\sigma^2}{L^2}$$

Here,

- n = sample size
- z = confidence level
- σ = standard deviation (SD)
- L = allowable error on either side of the estimated mean

Let us take a hypothetical study which aims to estimate the SBP of adults in an urban slum. At 95% confidence level, Z = 1.96. We can find out the estimated mean and SD of SBP among adults from previous similar published studies or through a small pilot study. Suppose, a previous study from another city reports the mean SBP among adults at 120 mm Hg with a SD of 15 mm Hg. Let us keep the allowable error at 2 points on either side of the estimated mean. What do these 2 points on either side of the mean signify? It means that, if at the end of our study the mean SBP in our study sample came out to be 124 mm Hg, the true mean SBP of the adults of urban slums of this city would be anywhere from 122 to 126 mm Hg.

We have Z = 1.96, σ = 15, and L = 2.

$$n = \frac{Z^2\sigma^2}{L^2} = \frac{(1.96)^2 \times (15)^2}{2^2} = \frac{3.84 \times 225}{4} = 216$$

With a SD of 15 mm Hg and keeping an allowable error of 2 points on either side of mean, the calculated sample size for estimating the mean SBP of adults of urban slums of this city comes at 216 at 95% level of confidence. As we tried altering the allowable error for qualitative variable, here also we can alter the allowable error.

Choosing Sample Size for Analytical Studies

Since the formulae for calculation of sample size for analytical studies are complex ones, we make use of computer programs. OpenEpi is one such free to use online tool that helps in calculating the sample size. Demonstration of sample size calculation with use of software is beyond the scope of this book. However, a list of broad parameters to keep handy to feed in the sample size calculating software is provided in the box here.

DATA COLLECTION: SOURCES, METHODS, AND TOOLS

Once the study variables and study population are defined, the researcher has to plan how the collected data will be managed. This has to be decided before the actual data collection starts. The following are the steps in management of collected data:

- Defining the source of data
- Defining data collection method
- Developing data collection tool
- Developing data documentation sheet
- Developing dummy tables for analysis
- Data collection process
- Data entry
- Data analysis

Sources of Data

The source of data for a research study can be primary or secondary. ***Primary data*** is when the researcher collects the data afresh for the specific purpose of the research study. ***Secondary data*** is one which is already collected for some other purpose which the researcher uses for her study. Commonly, researchers can use the data available from the census, national sample surveys such as National Family Health Survey (NFHS), data from the national health program records, hospital records, etc. The advantage with primary data is that it is more reliable since the researcher clearly defines the variables under study and data collection method. In our country, there are ample number of clinical records generated by the hospitals. There lies a great opportunity in making use of this data to generate meaningful evidence. As we have already started moving towards digitization (electronic health records) in India, it will also be a very good source of data for research purpose in the days to come.

Data Collection Methods

Interviews

Interviews are nothing but a one-to-one verbal dialog with the respondent for collecting data required for the study. Interviews can be face-to-face or telephonic. An interview guide is a document containing instructions for the interviewer and a list of questions with space for recording answers. The questions and answers are made available in the language that the respondent understands.

Questionnaires

A questionnaire is a specially designed set of questions which the respondent herself/himself is expected to answer. Traditionally, questionnaires were made available to the respondent on paper. Nowadays electronic questionnaires are becoming popular. Researchers can use the generic survey tools such as Google forms (www.forms.google.com) and Survey Monkey (www.surveymonkey.com) or especially designed tools such as EpiCollect (www.five.epicollect.net). Questionnaire has a limitation that it can be used only when the respondents are able to read/write.

Observation/Examination/Investigations

Health research often involves collecting data on biological variables through patient examination by doctors or healthcare professionals. Sometimes, the variable under study may require laboratory investigations or radiological investigations. Whenever such measurements are used, it is important to pay attention to the accuracy of measurement.

Data Collection Tool

Whatever be the method of data collection, a researcher will always need a data collection tool. This is nothing but a set of variables on which the data will need to be collected. A questionnaire is a commonly used data collection tool.

Anatomy of a Questionnaire

A questionnaire is the heart of the research study. It commonly contains the following elements.

- ***Title*** of the study and name and contact details of the investigator.

- A paragraph on ***background information:*** This will include the purpose of the study, an indication of what kind of information is being sought, and approximate duration it will take to complete the questionnaire.
- ***Directions for answering:*** Directions can be of two types. General directions for answering can be put at the start of the questionnaire. Directions for specific questions can be put at the start of that particular set of questions.
- ***Set of questions:*** This is the list of questions with space for recording answers organized in a logical order. This set of questions would be primarily covering information related to predictor, outcome, and confounding variables as well as some background information of the study participants. Two types of questions commonly used in questionnaires are:
  - **Open-ended questions:** These are questions where the scope of answering is kept open for the respondent to decide. Here the respondent is given freedom of length to provide an answer. Take example of this question:
    - "What causes your psoriasis skin rashes to flare up?"
  - **Close-ended questions:** Here options available for the respondent when answering this question are limited. The example we used for open-ended question can be converted to a close-ended question like this:
    - "From the list of following options please choose what causes your psoriasis skin rashes to flare up. [Tick all that apply]

      i. Winter ☐
      ii. Stress ☐
      iii. Alcohol ☐
      iv. Smoking ☐
      v. Drugs ☐
      vi. Pregnancy ☐
      vii. Others ☐

  In quantitative research questionnaires, most questions are close ended questions.
- ***Ending note:*** Do remember to thank the respondent for responding to your questions.

Data Documentation Sheet

The purpose of the study tool is to collect the required data on the variables under the study. To make our questionnaire focused and analysis-ready we need to prepare a plan for data management. This plan is referred to as data documentation sheet. We present here a sample data documentation sheet for a tobacco usage survey among adult OPD attendees of a primary health center **(Table 11.3)**.

We observe in this sample sheet that for each of the variable, it is important to define the measurement scale. The sheet also helps in deciding in advance the *possible answers* and the proposed methods for summarizing or analysis. While using paper-based forms with large samples, it is advisable to use codes for each possible answers (*codes* are the numbers you see against the option categories in **Table 11.3**). This makes the job of data entry easier and more efficient.

Once the data documentation sheet is ready it would be possible to prepare *dummy tables* for analysis. Dummy tables are nothing but a framework of how the analyzed data will be summarized, what statistical tests will be applied, and how they will be presented. Preparing dummy tables is a good practice in quantitative research. Once this is done, the researcher is ready to *collect data* for the study.

Data Entry Tools

When data is collected on paper-based questionnaire it will have to be entered in a database file on the computer. When data is collected through the electronic means the data entry step is skipped since the data is directly available in a database. The data capture features of tools such as EpiInfoTM, Microsoft

Table 11.3: Sample data documentation sheet for a tobacco usage survey among adults.

Sl. No.	*Question*	*Measurement scale Nominal Ordinal Numerical*	*Possible answers/answer range (including assigned codes if any)*	*Comments (skip/ jumps)*	*Plan for summarization Mean + SD Median + IQR Percent*
1.	Age	Numerical	18–99		Mean + SD
2.	Gender	Nominal	Male Female		Proportion
3.	Education of respondent	Ordinal	Primary schooling Secondary/higher secondary schooling Graduation/Postgraduation completed Beyond postgraduation		Proportion
4.	Did you smoke tobacco in any form in last 30 days?	Nominal	No Yes No answer	If Yes go to Q. No. 5; otherwise go to Q. No. 7	Proportion
5.	Beedi/cigarette smoked per day	Numerical	1–50		Mean + SD/median + IQR
6.	Since how many years smoking	Numerical	0–99		Mean + SD/median + IQR
7.	Would you like to use services of a tobacco cessation clinic if we start one?	Nominal	No Yes Cannot say		Proportion

(IQR: interquartile range; SD: standard deviation)

Access or EpiData Manager provide a range of options to add checks on what can be entered in a field for a particular question. This helps in minimizing the data entry errors.

Data Analysis and Interpretation

Summarizing and analyzing the data using statistical tools is the step of data analysis. These will be discussed in detail in subsequent sections of this chapter. However, the process of data analysis is not limited to statistical processing through software. When we summarize the data and make use of relevant statistical tests, we convert the data into usable information. But this information is not sufficient. When we use the technical knowledge of the relevant field of research topic and apply clinical wisdom to make meaningful inference out of this information, it is known as intelligence. In the chapter on epidemiologic methods, you learned about the difference between association and causation; that all associations that we observe when we plot data in a 2 × 2 table may not necessarily be causation. Multiple errors can possibly influence our data. You also learned about these errors in the section that dealt with bias (systematic error). When we apply our wisdom to identify such possible errors and interpret the results in this context, we convert the information into intelligence or evidence.

PRESENTATION OF DATA

Importance of Data Presentation

It is important to summarize the data to give clear message from the results of the research study. So, data should be presented in a simple form to make it easy for the readers to understand. It should be concise without losing the important information and need few words to explain. Overcomplicated presentation is often ignored by the readers.

For example, Hb measurements (g/dL) of 20 patients of a research study are as below:

9.0, 11.3, 10.6, 11.8, 13.2, 9.2, 11.6, 10.8, 12.4, 14.7, 12.2, 11.5, 12.1, 12.4, 10.7, 13.1, 15.0, 12.4, 10.5, 9.6.

It is very difficult to understand such raw data. So, if the same values are presented in a simple tabular form as given below it becomes easy to understand.

Hemoglobin (g/dL)	*No. of patients*
9–9.9	3
10–10.9	4
11–11.9	4
12–12.9	5
13–13.9	2
14–14.9	1
15–15.9	1
Total	20

Methods of Data Presentation

Data can be presented in three different forms as described in **Flowchart 11.1**.

Text

It is not always necessary to have a table or diagram to represent the data. Sometimes, even a simple sentence in the text is sufficient to present essence of the data. For example, the temperature in ice-lined refrigerator for five primary health centers is as given in table below:

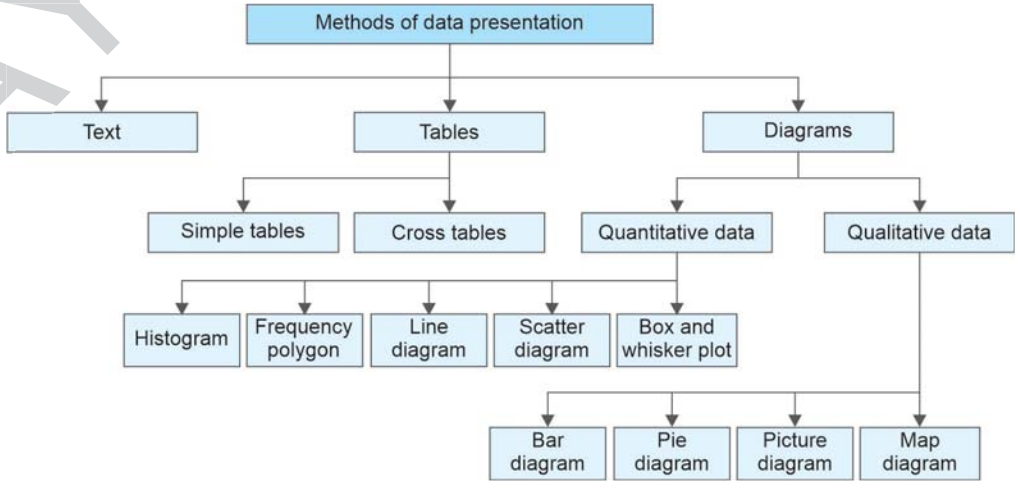
Primary health center (PHC)	*Temperature in ice-lined refrigerator (°C)*
PHC 1	4
PHC 2	5
PHC 3	3
PHC 4	5
PHC 5	4

This data can be presented in simple text, "the temperature in the ice-lined refrigerator of all five primary health centers was in the range from 3 to 5°C". It is not necessary to have a separate table for such study findings.

Table

It is a basic form of data presentation. A table should have following components to be able to give a complete picture: table title, column titles, table body (data), and footnotes. For instance, hypothetical example of a table on clinical profile of dengue patients is given below:

Flowchart 11.1: Methods of data presentation.



Clinical profile of dengue patients admitted in a hospital, 2017 (N = 50) → Table title

Clinical variables	Frequency	Percentage
Fever	45	90
Rash	20	40
Vomiting	10	20
Headache	8	16
Bleeding tendency*	4	8

→ Column title

Table body

*Bleeding from nose or ear or hematemesis or melena → Footnote

Table title should be clear and stand-alone. It should mention time, place, and person details of the data it contains. Every column of the table should have a title. Table body includes data or text. Short forms should be avoided in a table. Conventionally, data is usually presented to the precision of only one digit after decimal, unless more precise information is needed in a specific table. As shown in the example, when arranging the categories in order, try to follow descending or ascending order of frequency of the data to highlight the common findings of the study. Footnotes are used to give that additional information related to data which is difficult to incorporate in the table body.

Depending on the complexity involved, the tables can be simple table or cross table.

Simple table: This is also known as frequency table. A simple table can be used to present qualitative as well as quantitative data.

Simple table presenting qualitative data: The above-mentioned example of clinical profile of dengue patients is an example of such table. The clinical profile is presented in the form of categories of symptoms. In such tables usually the first column lists the categories. The second column gives frequency against each category.

Simple table presenting quantitative data: The example at the start of this section providing the Hb values of 20 patients in a tabular form is an example of such a table. Here the range of possible Hb values is divided into different categories in the first column. The second column gives frequency against each category.

While preparing such a table, it is advised to keep the number of categories in the range of 4–8. More than eight categories make it complex for the reader to understand data.

We observe here that each category covers a range of 1 g%. This is known as class interval. To decide the class interval, we find out the maximum and minimum values in our data set. In this given data set maximum Hb is 15 g% while minimum is 9 g%. We calculate the range using formula: maximum value – minimum value. Thus, here 15 – 9 = 6. The range divided by desired number of classes gives the class interval. In this case, 6/6 = 1. This is how we decided that each class will cover 1 g%. The first class chosen is one which covers the minimum value.

If we take example of the first class the value 9 is known as lower limit of the class and 9.9 as upper limit of the class.

Cross table: A cross table is used when we want to compare two or more different groups or want to see association between two variables. The simplest form of a cross table is a 2 × 2 table. The following 2 × 2 table contains first variable (obesity) split into two categories occupying two rows and second variable (depression) split into two categories occupying two columns **(Table 11.4)**.

Table 11.4: Relationship between obesity and depression among adolescent boys of XYZ school.

	Depression	No depression	Total
Obese	5	25	30
Nonobese	10	60	70
Total	15	85	100

Diagram/Graph

It is a visual form of data presentation. The advantage is that it gives a quick picture of the data. Similar to table construction, diagram should also have components such as title, X-axis, label to X-axis, Y-axis, label to Y-axis, key to symbols, and footnotes.

For example, the immunization coverage of different states in India, 2017 (as per NFHS-4 survey) is given below **(Fig. 11.6)**.

Diagrams can be prepared easily in a computer using Microsoft Excel program. Choice of type of diagram depends on the type of data as shown below.

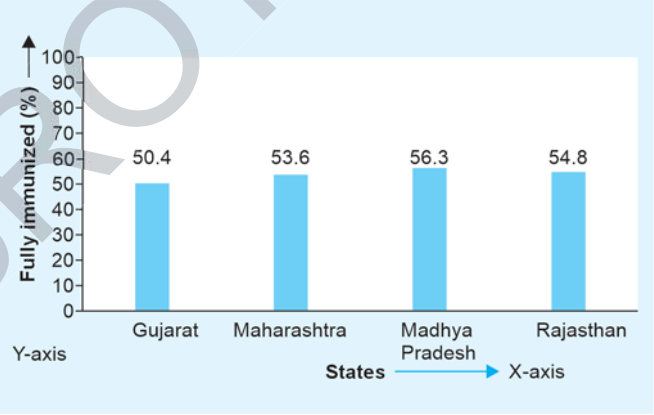


Fig. 11.6: Immunization coverage* of different states in Western India, 2017.
**Source:* National Family Health Survey-4 Data.

Quantitative

Histogram: Frequency distribution of continuous variables like age, height, weight, and Hb can be presented using a histogram **(Fig. 11.7)**. Variable is represented on X-axis while frequency is plotted on Y-axis. The frequency of each group/range will construct a column graph without the spaces between columns. It is also known as an area diagram since the area of the column varies with the frequency.

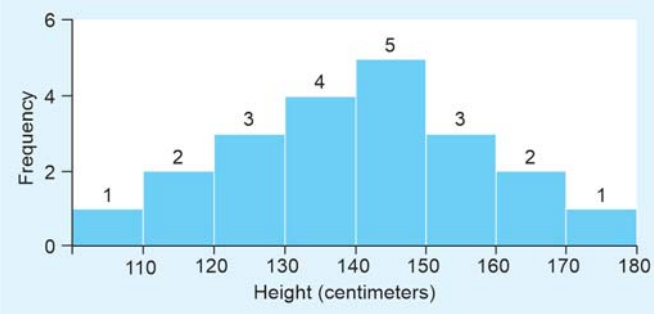


Fig. 11.7: Frequency distribution of height using a histogram.

Frequency polygon: Frequency polygons are analogous to line graphs, and just as line graphs, they also make continuous data visually easy to interpret. They can be used to graph large data sets with data points that repeat **(Fig. 11.8)**. The literal meaning of polygon is figure with many (poly) angles (gon). It is developed over histogram. When the midpoints of the class interval of the variables are joined together at the height of their frequencies by straight lines, a frequency polygon is developed.

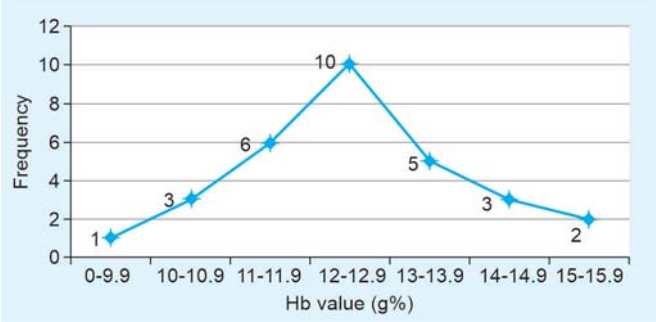


Fig. 11.8: Frequency distribution of hemoglobin using a frequency polygram.

Line diagram: This form of diagram is widely used to depict the trend of an event over a period of time. Time is shown on X-axis while frequency of variable is shown on Y-axis. It is not necessary to start with zero on the Y-axis.

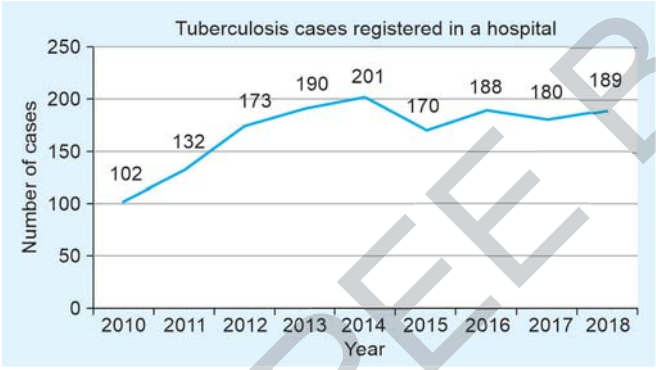


Fig. 11.9: Line diagram showing the trend of tuberculosis patient registered in a hospital over years.

Scatter/dot diagram: This type of figure is used when we want to show the relationship between two quantitative variables. It is used to show the correlation between two variables like age and weight, so it is also known as correlation diagram. Perpendicular lines are drawn for relationship between two variables and the point at which those lines would meet is represented by a dot. The different frequencies of the variables would give many such dots, which is a scatter. Finally, a line is drawn to show the type of correlation at one glance.

Box and whisker diagram: When numerical data is to be compared between two groups, a box and whisker chart is preferred over a histogram. The middle line inside the box is median, the ends of box are 1st and 3rd quartiles, and ends of whiskers are minimum and maximum values in data range. Sometimes, data not included between the whiskers is plotted as an outlier with dots.

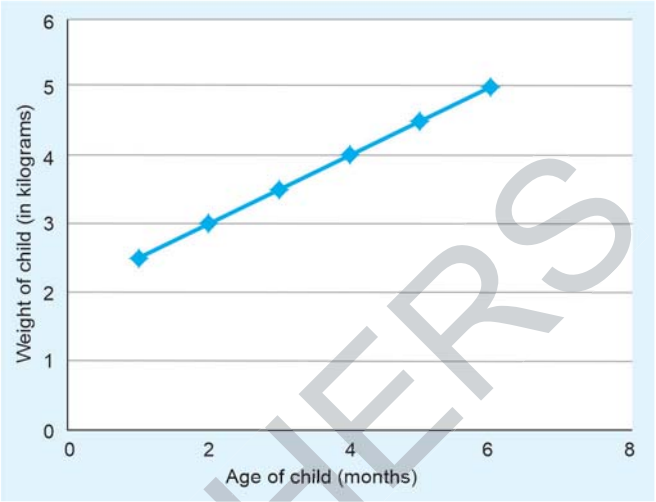


Fig. 11.10: Correlation diagram between weight and age of child.

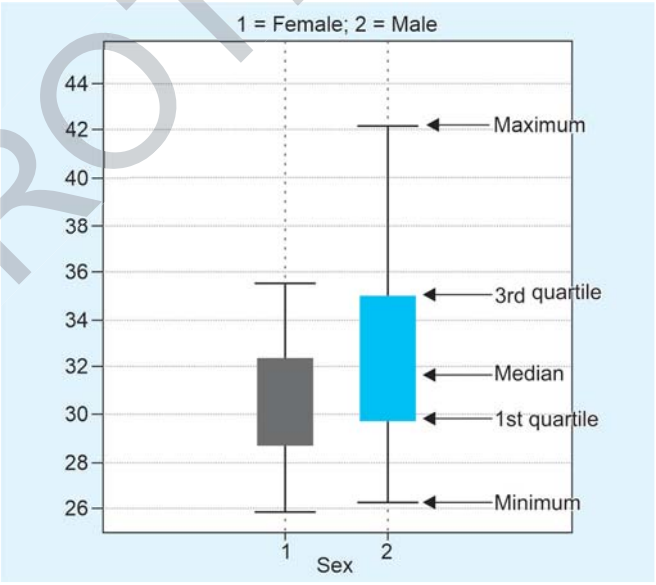


Fig. 11.11: Box and whisker diagram showing the distribution of BMI among male and female patients.

For example, the box and whisker diagram below shows the distribution of BMI among male and female patients **(Fig. 11.11)**.

Qualitative

Bar diagram: Bars can be drawn horizontal or vertical. Length of bar is proportional to the frequency of the variable. Width of each bar is same. Distance between two bars is at least half the width of the bar. There are three types of bar diagrams: simple, multiple, and component bar diagram.

1. *Simple bar diagram:* As shown above, coverage of children fully immunized in different states of India is compared according to NFHS-4 data **(Fig. 11.12)**.
2. *Multiple bar diagram:* Multiple bars can be drawn within the same category. For example, comparison of coverage of fully immunized children as per NFHS-3 and NFHS-4 for selected states of India is shown in the multiple bar diagram **(Fig. 11.13)**.

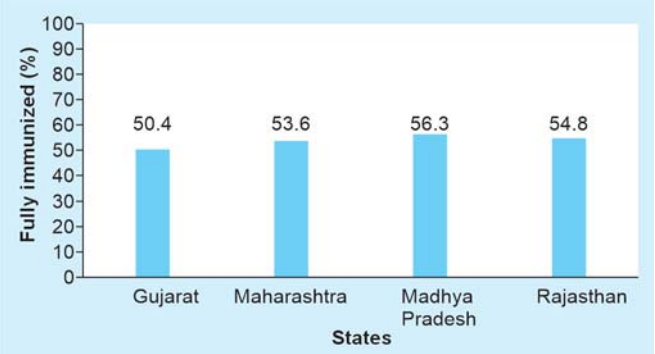


Fig. 11.12: Bar diagram showing coverage of children fully immunized in different states of India.

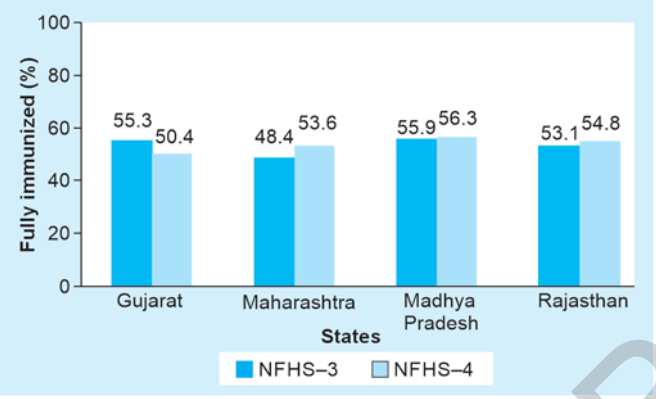


Fig. 11.13: Multiple bars showing comparison of coverage of fully immunized children as per NFHS-3 and NFHS-4 for selected states of India.

3. *Component bar diagram:* It is also known as proportional bar diagram. Total height of bar indicates the frequency of the category. Subcategories within these categories can be shown by splitting the height of the bar into different colored components. This bar diagram is used when we want to focus on the relative proportion of subcategories out of the total frequency.

For example, average number of patients per day in various OPDs of the hospital is shown in component bar diagram above **(Fig. 11.14)**. First bar indicates on an average 100 patients per day attends medicine OPD, out of which the proportion of males and females is 60 and 40, respectively.

Pie diagram: Qualitative data can also be presented by pie chart. It is commonly used to present causes of morbidity or mortality in a population. The frequency of each group is shown in a circle, depicted by the degrees of the angle. The group with more frequency will have more degrees of angle. Each angle size is calculated by dividing the class frequency with total observations and then multiplying with 360° **(Fig. 11.15)**. It can be easily prepared in the Microsoft Excel worksheet without manual calculations of such degrees for angle size.

Map diagram: Such diagram is prepared to show geographical distribution of frequencies of some characteristic traits. Some symbols or colors in the map denote the frequency of the characteristic. For example, state-wise estimation of new HIV

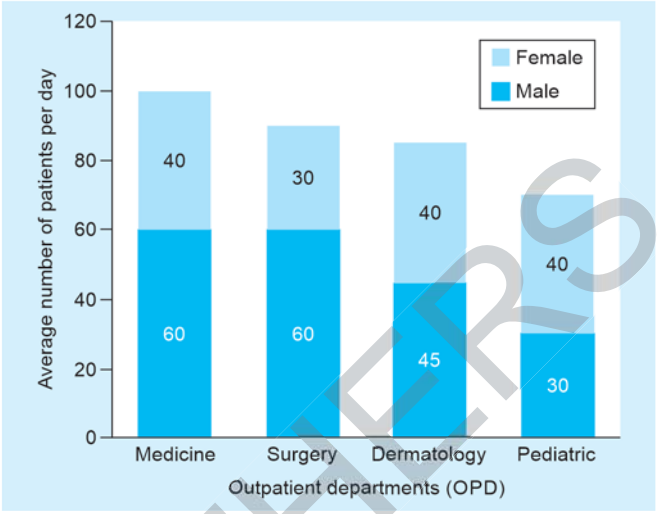


Fig. 11.14: Component bar diagram showing average number of patients per day in various OPDs of the hospital.

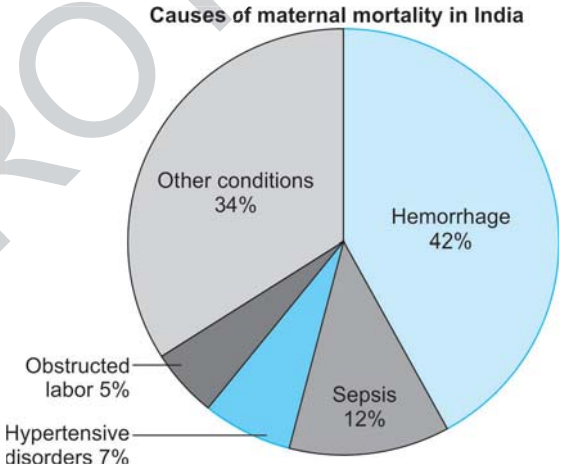


Fig. 11.15: Pie diagram showing different causes of maternal mortality in India.

infections in India during 2015 is shown in the map below using different colors **(Fig. 11.16)**.

BIOSTATISTICS: THE CONCEPT

Biostatistics is the science of dealing with numbers related to biomedical phenomena. These numbers, when put into a set of columns and rows, form a "dataset" or simply referred to as "data". Data is useful only when it is inferred in the form of information by counting, dividing, and comparing (CDC) the numbers, which helps in decision making.

Usage of statistical methods and their interpretation have guided important health policies worldwide. A health professional also deals with numbers in daily practice. A large part of this chapter deals with the usage of biostatistics in health research. Apart from health research, biostatistics is also useful on day-to-day basis for management of medical practice as well as community health programs. Both the medical practice and community health programs generate a lot of data on routine basis. Health professionals are expected to analyze and interpret

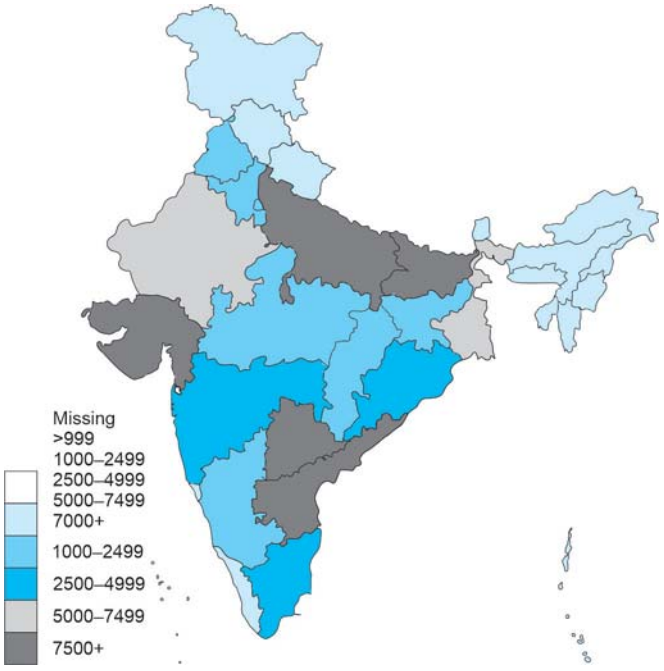


Fig. 11.16: Map diagram showing state-wise estimation of new HIV infections in India during 2015.

these data to make meaning and data-informed decisions periodically. For example, a health professional in charge of a large hospital would be interested in knowing the average bed-occupancy rate of different departments or average length of stay for different diseases in his hospital. A doctor running a clinic would be interested in monitoring the stock of medicines in pharmacy or finding out the average waiting period for patients in OPD. In community health programs, one would be interested to monitor the monthly blood examination rate for malaria program or monitor the quarterly treatment outcome for tuberculosis. Thus, every doctor/healthcare professional need to familiarize themselves with the art and science of dealing with data.

Application of Biostatistics in Health Research

The science of biostatistics has manifold applications in health research, as below:

- To *describe* the biological characteristics/health outcomes (e.g., proportion of patients of a hospital suffering from hypertension)
- To find the *significance of differences* in values of biological characteristics/health outcomes between two groups (e.g., to compare the mean number of hours of exercise in two groups of patients with normal cholesterol and elevated cholesterol)
- To know the *association/relationship* between two biological characteristics/health outcomes (e.g., to know the association between family history of allergy and development of asthma among children)
- To *predict* the biological characteristics/health outcomes (e.g., to predict the value of blood pressure for a given value of BMI and salt consumption in a day).

Broadly, the use of statistical methods in health research can be divided in two parts—descriptive statistics and inferential/analytical statistics. It also depends on the type of study design that is covered under the epidemiological study designs. For most of descriptive studies only descriptive statistics is used, wherein, the variable of interest is summarized as measured among the study participants. In case of analytical studies, descriptive statistics is used to describe the background of the study participants, whereas inferential statistics is used to find out if the difference in variable of interest among groups is statistically significant. In these inferential statistics, we make use of various statistical tests.

DESCRIPTIVE STATISTICS: MEASURES OF CENTRAL TENDENCY AND VARIATION

Concept of Central Tendency and Dispersion

When we collect a large amount of data, it is always a good practice to concise it and express it in a manner which will be easily understood by the statistician, clinician as well as a lay person. A good method to express data is through tables and figures.

However, if we are dealing with a very large amount of quantitative data, viz. Hb levels of all the adolescent girls in a district, the average Hb level of adolescent girls in a district would be more useful information rather than presenting all the individual values. This summarized single information is called the central tendency. It is also useful to know how much the individual data varies from the single summary measure. This measure is called the measure of dispersion.

Measures of Central Tendency

Measures of central tendency give an idea about the value around which the observations are concentrated. Mean, median, and mode are measures of central tendency.

Mean

This is a measure of central tendency used for normally distributed data.

It is the sum of all observations divided by the number of observations. Mean for a sample is denoted as x, while that for a population is denoted by μ.

Mean $(\bar{x})$ = Sum of all observations ÷ number of observations

$$= \frac{\Sigma x}{n}$$

Though mean is the simplest measure of central tendency, it is affected by extremes of observations; so sometimes it may not give the central value correctly.

Example: Calculation of mean

Hemoglobin levels (g%) of 10 boys was found to be 10.8, 12, 11.8, 13, 13.2, 14, 13.4, 13, 12.6, and 14.2. Calculate the mean Hb.

The above data is quantitative. Mean is sum of observations divided by the number of observations viz. 10

x = (10.8 + 12 + 11.8 + 13 + 13.2 + 14 + 13.4 + 13 + 12.6 + 14.2)/10
= 128/10 = 12.8 g%

Example: Calculation of mean from grouped data. In order to calculate the mean from grouped data:

- Find the midpoint of the class interval (denoted by x)—add the upper and the lower limit of the class interval and divide by 2 for the midpoint.

- Multiply the class frequency with the midpoint.
- Find the sum of all these multiplied values.
- Divide it by the number of observations.

The serum cholesterol levels (mg/dL) of 10 patients were found to be as follows: 192, 242, 203, 212, 175, 284, 256, 218, 182, and 228.

This can be converted into grouped data by preparing a frequency table with class intervals as shown below:

Serum cholesterol level (mg/dL)	*Midpoint (x)*	*Frequency (f)*	*x*f = fx*
175–199	(175 + 199)/2 = 187	3	561
200–224	212	3	636
225–249	237	2	474
250–274	262	1	262
275–299	287	1	287
Total		10 = ∑*f*	2,220 = ∑*fx*

Mean is calculated as = 2,220/10 = 222 mg/dL

Median

Median is that value which divides the complete data set into two equal parts; when the data is arranged in ascending or descending order. It is the middle most value of the data when arranged in ascending or descending order. When total observations are an odd number, there is single middle value. When total observations are an even number, there are two middle values. The median is then calculated by taking mean of these two middle observations.

Median = (n + 1)/2 when the number of observations (n) is odd
= mean of n/2th and [(n/2) + 1]th observation when the number of observations (n) is even

Continuing with example of serum cholesterol levels, the ascending order of these observations is 175, 182, 192, 203, 212, 218, 228, 242, 256, and 284.

The median = mean of the 5th and 6th value = (212 + 218)/2 = 215

Advantages of median are as follows:

- Median is not affected by extreme high-and low-values.
- Median is often used for nonnormal distributions wherein; it helps to convey the middlemost value.

Mode

The most commonly occurring value or the most often repeated value is mode. In case two values repeat themselves same number of times, the distribution may be bimodal or multimodal. Mode is a less commonly used measure in health research.

Relation between Mean, Median, and Mode

- For a symmetric curve (normal curve): Mean = Median = Mode
- For positively skewed curve: Mean >Median >Mode
- For negatively skewed curve: Mean <Median <Mode
- For skewed data, median and mode are better indicators of central value as compared to the mean.

Measures of Dispersion/Variability

Variability is an inherent biological phenomenon. As we saw in previous examples if we measure Hb of 10 boys, each of the boy's Hb value would be different. We can calculate the mean Hb level but each boy's Hb would be a little different than this central value. Thus, when we have a data of observations on a biological variable (e.g., Hb) not all observations fall on a certain point, but they are spread across a range. This spread is called variability or dispersion of data. As there are measures to calculate the central tendency in single value, there are measures to calculate this variation in data in a single measure. Following is a list of such possible measures of variability within the data in a sample and variability across samples.

Measures of variability of individual observations within a sample	*Measures of variability of samples (will be discussed in the next section)*
1. Range 2. Interquartile range 3. Mean deviation (MD) 4. Standard deviation 5. Coefficient of variation	1. Standard error of mean 2. Standard error of difference between means 3. Standard error of proportion 4. Standard error of difference between proportions

Measures of Variability of Individual Observations within a Sample

Range: Range is a simplest measure of variation. Range can be calculated as difference between the maximum and the minimum value of observations in a sample (range = maximum value–minimum value). In the previous example of serum cholesterol levels of 10 patients, the range = 284–175 = 109 mg/dL. Thus, it uses only extreme values.

Interquartile range: Centiles are levels which divide the entire dataset into equal parts. Percentiles divide the data set into 100 equal parts. Similarly, quartiles divide the data set into four equal parts. There are three quartiles, viz. Q1, Q2, and Q3. Q1 divides the data set into 25:75 (25% observations are below and 75% observations are above it) observations, while Q3 divides it into 75:25 observations when observations are arranged in ascending order. The interquartile range (IQR) gives observations which are between Q1 and Q3. Thus, it gives the range of middle 50% observations. It is estimated by Q3–Q1. Interquartile range is often used along with median in order to express the data which is nonnormally distributed. Thus, they are not affected by the extremes of values.

Mean deviation: For calculation of MD, for each observation, we measure how far away it is from mean. Thus, we calculate deviation of each observation from mean. So, we have a parallel dataset of deviations of observations. If we take a mean of these deviations, it is known as MD.

$$\text{MD} = \frac{\Sigma[X - \bar{X}]}{n}$$

Mean deviation is not much used in statistical analysis, but its improved version SD is frequently used.

Standard deviation: While calculating MD, we took a modulus to ignore the negative signs for the observations lying on the left side of mean in a distribution. Another mathematical process for overcoming this negative sign is to take a square of the deviations. But when we take square of deviation, the unit of deviation also gets squared (e.g., if the observations are height of individuals in cm, then such deviation would be in cm^2). So, to bring back the deviation in original units (e.g., cm) we take

a square root of the entire calculation. In the denominator we place (n – 1) instead of n in formula for SD.

Thus, formula for SD is:

$$SD = \sqrt{Var} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n}} = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1}}$$

Steps for calculation of SD are as follows:

- Find mean
- Find difference of each observation from the mean (deviation)
- Square this difference (deviation^2)
- Add up all these squared values (sum of squares of deviation)
- Divide this by the number of observations minus one (gives the variance)
- Find out the square root of this variance (gives SD).

Hence, SD is root mean squared variance. Small SD means that the observations are spread closely around the mean, while a wider SD means they are spread farther from the mean.

Uses of SD are as follows:

- It gives the dispersion of observations around the mean in a single unit.
- It helps to decide whether this dispersion from mean is real or by chance (using statistical tests).
- It helps to find out the standard error (SE); whether two samples are different from each other.
- It helps to calculate the sample size for a study when variable is measured on a quantitative scale.

Coefficient of variation: Coefficient of variation (CV) is used when variations in two different datasets have to be compared. For example, to know whether variation is more in the height or weight of individuals. CV converts the variation in a single value, which can then be compared. CV is in fact SD expressed as percentage of the mean.

$$CV = \frac{SD}{Mean} \times 100$$

Example: Mean height of adults is 160 cm and the SD is 10 cm. Mean height of 3 months old children is 60 cm and SD is 5 cm. Which group shows greater variation?

Here, CV of adult is (10/160) × 100 = 6.25%

Coefficient of variation of 3 months old children is (5/60) × 100 = 8.33%

Coefficient of variation of children is more than that of adults. Thus, the height of children shows greater variability than that of adults.

Concept of Standard Error and Confidence Intervals

Imagine that a specific population comprises of 10,000 people. If you can measure the height of all 10,000 and then calculate the mean of all 10,000 you get the population mean. Now, measuring height of 10,000 people is a huge task. So, in real life, it is acceptable to take only a small sample of people and measure their height. We understand that when we take only a small sample, the mean height as derived from this sample will not be exactly same as population mean but it might be a little different. We would be interested in knowing how different it would be from population mean. The concept of SE helps us determine how different this value of sample mean would be from population mean.

Suppose you are going to take sample of 100 people from a population of 10,000. You can take many such samples of 100 people with replacement from this population of 10,000. Each of these sample would have its own mean and SD.

Imagine you have taken many samples and so, now you have many sample means. You might be surprised to note that if the values of these sample means are plotted to form a frequency distribution curve, the curve would be similar to normal distribution curve. It is expected that the mean of all the sample means (if you have considered all possible samples) will be equal to the population mean. Of course, now you would like to know the dispersion of sample means from the population mean and this measure is called standard error of mean (SEM) (denoted by "s") and is calculated by following formula. (Here, n is sample size)

$$SEM = \frac{\text{Standard deviation}}{\sqrt{n}}$$

Concept of Confidence Interval

As discussed earlier, it is not always possible to study the entire population to understand its characteristics or estimate its central tendency or variability. Hence, based on sampling techniques, the investigator studies a sample and tries to estimate the population mean (μ) or population proportion (P) based on the sample mean (x) or sample proportion (p).

When we try to estimate population parameter based on the sample statistic, we may not get the exact value, some error is bound to occur. Hence, we try to estimate a range within which the population parameter is expected to lie. This interval range is known as confidence limits or CI.

Standard deviations describe the spread of individual observations around sample mean. Similarly, SE describes the spread of sample means around population mean. We have seen that, when individual observations in a sample follow normal distribution mean ±1SD covers around 68% observations and mean ±2SD covers around 95% observations. Similarly, we discussed that the distribution of sample means also follow normal distribution. Hence, sample mean +1SE covers 68% of such sample means and sample mean ±2SE covers around 95% of such sample means.

Imagine we take 100 samples each having 100 individuals from the original population of 10,000. All samples combined together, the total number of people studied is possibly 10,000. Thus, with 100 such samples, we are close to studying the entire population. If we had the resources, we could actually do such a study taking 100 samples. The average (mean) of these sample means would be the actual population mean. In real life, we usually do not have resources to take 100 such samples. However, the statistical measure of SE can help us know the range within which 95% of such sample means would possibly lie. In real life, most scientists are satisfied with 95% cutoff level. Thus, to know the population mean in real life, we

do not have to study the entire population. We only study a sample and using the SD within sample and the sample size (n) we can estimate the range (2SE on either side of sample mean) within which 95% of such sample means would lie. We can say that we are 95% certain that the population mean would lie in this range.

This range calculated using 2SE is known as 95% confidence limit or 95% confidence interval (CI). As SE is dependent on sample size (recall: SE = $SD/\sqrt{n}$) we understand that larger the sample size, smaller will be SE and hence more precise will be the confidence limit.

Let us take an example to understand how 95% CI is used in real life in descriptive research studies.

A researcher wanted to estimate the average salt consumption per person per day among adults in a tribal village. She found that the mean salt consumption was 14 g with SD of 3 g from her study among 30 adults. How do we interpret findings?

Here, we want to calculate the 95% CI.

$$SE = SD/\sqrt{n} = 3/\sqrt{30} = 0.54 \text{ hence, } 2SE = 1.08$$

95% CI for population mean = sample mean + 2SE = (14–1.08) to (14 + 1.08) = 12.92–15.08 g.

We are 95% certain that the mean salt consumption among adults per day in this village lies within range of 12.92–15.08 g.

We understood how to calculate the SE of mean. What do we do when the variable is measured on a qualitative scale? Suppose we conduct a survey in a tribal district to know the prevalence of sickle cell trait among children. We carry out a survey among 400 children and find out that 80 of them have sickle cell trait. This converts to a prevalence of 20%. What would be the population prevalence of sickle cell trait?

As we took help of SE of mean in case of quantitative data, we can take help of SE of proportion when we have measured our variable on qualitative scale.

The formula for SE of proportion (SEP) is as follows:

$$SEP = \sqrt{pq/ns}$$

Replacing the values of our survey we get SEP =

$$SEP = \sqrt{20 \times 80 / 400} = \sqrt{4}s = 2\%. \text{ So, } 2 \text{ SEP} = 4\%.$$

95% CI for population proportion = sample proportion + 2 SEP = (20–4) to (20 + 4) = 16–24.

We are 95% certain that the proportion of sickle cell trait among children population in this district would be within the range of 16–24%.

NORMAL DISTRIBUTION

We learnt about the use of frequency polygon to present numerical data collected on a variable of interest. When large number of observations of a quantitative variable are divided into small class intervals and presented as a frequency polygon; it is seen that:

- The highest frequency is around the mean, lowest is at the extremities, and frequency is decreasing on either side of the mean, and

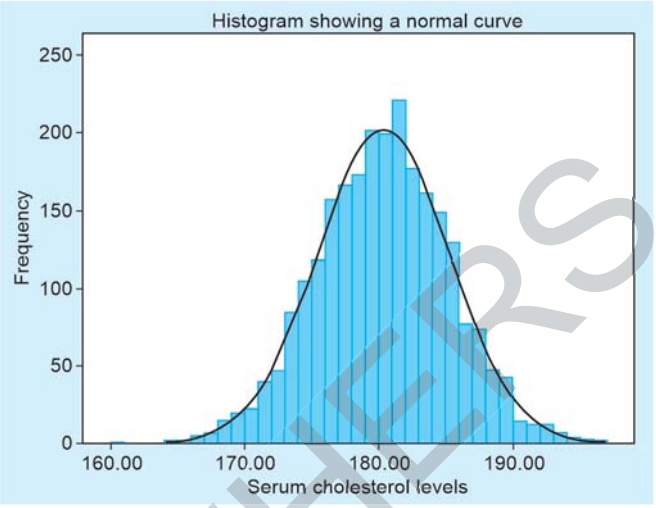


Fig. 11.17: A normal curve drawn over a histogram.

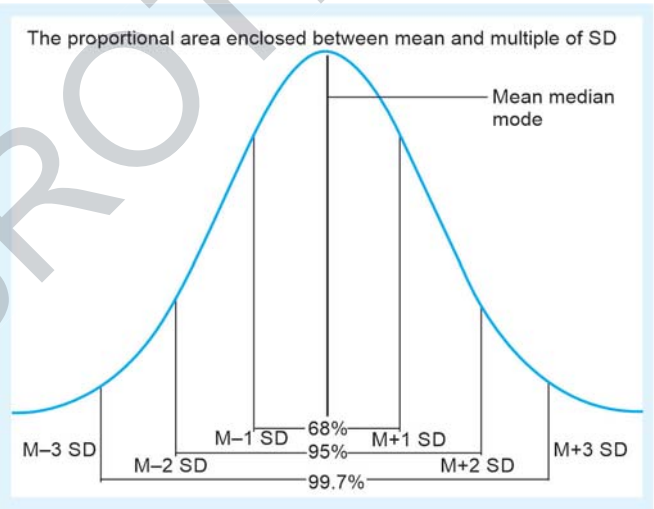


Fig. 11.18: A normal curve showing proportional area enclosed between mean and multiples of standard deviation.

- Half of the observations lie on the right of mean and a half on left, i.e., observations are symmetrically distributed on either side of mean. This curve is called a normal curve and the distribution is called a normal distribution. It is also known as "Gaussian distribution" in honor of the scientist Carl Friedrich Gauss who first described it. **Figure 11.17** shows the normal distribution as a frequency polygon drawn on a histogram (as a curve) for a simulated data of serum cholesterol levels of 2,500 people with a mean of 180 and SD of 5.

Figure 11.18 shows the normal curve without the histogram.

The ***characteristics of a normal curve*** are as follows:

- It is bell-shaped
- It is symmetric around the mean: Two halves of the curve are of the same size (mirror images)
- Mean, median, and mode are at the same value
- Since histogram represents all the values in the dataset, the area under curve (AUC) is 100%. The proportional area

enclosed between mean and multiples of SD is constant and is as described below and shown in **Figure 11.18**:

- Mean ± 1SD = 68.26% of the total area
- Mean ± 2SD = 95.44% of the total area
- Mean ± 3SD = 99.74% of the total area.

It is interesting to note that most of the biological variables that we deal with in medical science follow a distribution called as the normal distribution. In earlier section, we observed that dataset of sample means also follow normal distribution.

Standard Normal Distribution

We use multiple units of measurement for variables measured on quantitative scale. For example, height can be measured in centimeters or inches. To avoid discrepancy due to choice of units, we can take help of characteristics of normal distribution to express data in a unit free form (data can be "normalized"). This normalized value is called z-score or standard normal deviate.

$$\text{Standard normal deviate (z)} = \frac{\text{X-mean}}{\text{SD}}$$

z-score indicates how many SDs away from the mean an observation lies. Thus, instead of describing value of an observation in the original units, we can describe it in terms of z-score. z-score is used at many places in real life. The variable "weight of child" follows normal distribution. World Health Organization (WHO) has come up with growth charts with values of weight at 2 and 3SD around mean, for different age groups. Thus, in place of describing the weight of a male child at age of 1 year as 7 kg we can also mention that it is at 2SD point on left side of mean on the curve. Pediatricians define a child to be undernourished when weight is less than 2SD. Thus, when we convert the weight of child from original units (kg) to z-score, it helps in classifying the child as normal weight or underweight.

INFERENTIAL STATISTICS: STATISTICAL TESTS USED IN HEALTH RESEARCH

We differentiated between descriptive and analytical research studies earlier in this chapter. We used means and proportions and CIs for descriptive studies. The analytical studies are different in that they try to find out association between two variables or compare the variable of interest between two groups. In any case, there is a comparison involved in analytical studies. In this situation, a hypothesis comes into picture. It is a statement describing the intent of the researcher to prove a specific finding from the study. A research hypothesis is stated when a research question has terms like: *compared with, greater than, less than, associated with, correlated with, leads to, causes,* etc. The research hypothesis should be focused on the primary objective of the study. A few examples of testable research hypothesis from analytical studies are described below:

- Fasting blood sugar level of type 2 diabetes patients who do brisk walking for at least 1 hour daily is *lower as compared with* those who do not.
- Reduction in blood pressure of pregnant women by labetalol is *greater than* that by methyldopa.
- Prenatal depression is *associated with* infant mortality among pregnant women belonging to the low-income population.
- Mobile phone usage of 6 hours daily for over 30 years causes cancer of the central nervous system.

We take help of statistical significance tests to check whether the researcher's hypothesis is true or not. A number of statistical tests are available for use while analyzing the health research data **(Table 11.5)**. The choice of a statistical test depends on the measurement scale for the variable of interest, the sample size, number of groups in the study, and whether the observations are paired or coming from independent samples.

Null and Alternate Hypothesis in Statistical Significance Testing

In medical research, there are two types of hypotheses. Suppose a researcher is comparing the height of boys and girls. A null hypothesis in this example would be that the height of boys and girls are similar, any difference observed between the heights is just by chance. An alternative hypothesis (which usually the researcher is interested in) would be the height of boys is higher than the height of girls, so the observed difference between heights is real. Thus, the null hypothesis (denoted by H_0) is the hypothesis of no difference and the alternate hypothesis (denoted by H_1) is the hypothesis of difference.

Table 11.5: A simplified guide to choose a statistical test.

Goal	*Quantitative data (numerical)*	*Qualitative data (categorical)*
Compare two different groups	Unpaired t-test of difference of two means (if n <30 in any group); Z-test of difference of two means (if n >30 in both groups)	Chi-square test of proportion (if n <30 in any group); Z-test of difference of two proportions (if n >30 in both groups)
Compare two paired measurements (before and after any intervention)	Paired t-test for difference of means of same sample	McNemar chi-square test *(out of the scope of this book)*
Determine association between two variables	Correlation coefficient	Chi-square test of association
Compare two or more different groups	One-way ANOVA (analysis of variance)	Chi-square test of proportion/ association
Predict value of outcome variable (disease) for a given value of exposure variable (risk factor)	Simple linear regression	Simple logistic regression

Note: The calculations for one-way ANOVA, McNemar chi-square, correlation coefficient, and regression coefficients are out of the scope of this book, hence, only the concepts are explained.

Errors in Medical Research

In any medical research, there are possibilities of committing two types of statistical errors while conducting the study. For explaining these errors, let us take an example of raising an alarm in case of a fire:

Status of fire	You raised an alarm	You did not raise an alarm
Yes, there was a fire	Correct decision	Wrong decision
No, there was no fire	Wrong decision	Correct decision

As given in the table above, there are two correct decisions and two wrong decisions.

In terms of null and alternate hypothesis, the above example can be explained as: null hypothesis (H_0)—there was no fire and alternate hypothesis (H_1)—there was a fire.

How does this relate to analytical studies? The fire in the above example is analogous to a real-life difference in values between two groups (e.g., suppose the height of boys were more than that of girls in reality). The alarm is analogous to researcher finding a difference in study sample (e.g., researcher finds that mean height is more among boys than girls in study sample).

In reality	In study sample you find	
	Height of boys is more than girls (reject H_0)	Height of boys and girls is similar (do not reject H_0)
Height of boys is more than girls (H_0 not true)	Correct decision Power (1-β)	Wrong decision (type 2 error-β)
Height of boys and girls are same (H_0 true)	Wrong decision (type 1 error-α)	Correct decision Level of confidence (1-α)

Suppose in real life, there is a difference in height between boys and girls then there are two possibilities; the probability of researcher finding a difference in his study sample is known as power of study (1-β) and probability of researcher not finding a difference is known as type 2 error (β).

Thus, ideally if there is a real difference the researcher would wish to detect this difference in his sample and so would like the power to be 1 and β to be 0, but practically most scientists are satisfied if we keep the power at 0.8 and β at 0.2.

Suppose in real life there is no difference in height between boys and girls then there are two possibilities; the probability of researcher finding a difference in his study sample is known as type 1 error (α) and probability of researcher not finding a difference is known as level of confidence (1-α).

Thus, ideally if there is no difference in real life then the researcher would wish his study sample does not show a difference and so would like the level of confidence to be 1 and α to be 0, but practically most scientists are satisfied if we keep the level of confidence at 0.95 and α at 0.05.

Thus, the two correct and two wrong decisions can be written as:

- ***Level of confidence (1-α):*** Accepting null hypothesis when it is true—you find no difference, when actually there is no difference.
- ***Type I error (α):*** Rejecting null hypothesis when it is true—you find a difference, when actually there is no difference.
- ***Power (1-β):*** Rejecting null hypothesis when it is false—you find a difference, when actually there is a difference.
- ***Type II (β):*** Accepting null hypothesis when it is false—you find no difference, when actually there is a difference.

Probability (p-value)

Much of the statistical significance testing revolves around this type 1 error. Type 1 error occurs when the study finds a difference when in reality there is no difference. Why would a study find a difference? We studied the concept of variability earlier. When variability is at play, certain amount of variation in data is bound to occur by chance alone. So, the difference in our study sample might be observed just by chance. How much of a difference should be attributed to chance? We have to make use of statistical tests to find out the limit of this chance. Further, a statistical test cannot tell with certainty that the difference is not by chance. It will tell us what the probability of this difference occurring by chance is. This probability of difference being by chance is known as *p*-value. Most scientists are satisfied if this probability is kept less than 5% (in fraction: <0.05), the difference becomes statistically significant. This cutoff mark of 0.05 is also called as level of significance. Although traditionally, it is kept at 0.05 we can also keep it at 0.01 if we want to be more stringent.

If α is predecided at 5% and the *p*-value in a research study is obtained as 0.20, i it implies that the probability of this difference being by chance is 20% which is more than the cutoff level of 5%. Thus, there is high probability that this difference is observed just by chance and hence we call it statistically insignificant.

If α is predecided at 5% and the *p*-value in a research study is obtained as 0.03, it implies that the probability of this difference being by chance is 3% (and the probability of the difference being real is 97%). The probability of difference being by chance is less than the predecided cutoff level of 5%. Thus, the probability of this difference being just by chance is very small and hence this difference is likely to be real. Hence, we call it statistically significant.

p <0.05: Statistically significant
p >0.05: Statistically insignificant

Elements of Statistical Inference Procedure (Hypothesis Testing)

For the purpose of statistical analysis, a logical flow of steps has to be followed. For the purpose of explanation, a research study comparing the mean serum triglyceride levels among cardiovascular disease (CVD) patients and normal individuals is taken. The steps of hypothesis testing and their explanation for this example are as follows:

Step 1: Stating the null and alternate hypothesis

For our example, the null hypothesis is "there is no difference in mean serum triglyceride levels among CVD patients and normal individuals. Whatever difference is found in study is by chance". The alternate hypothesis is "the difference between mean serum triglyceride levels among CVD patients and normal individuals is real".

Step 2: Calculate the summary measures of the data

Next, depending on the measurement scale of variable of interest, we calculate the summary measures such as mean and SD for quantitative variable and proportions for qualitative variable. In our example, we will calculate the mean and SD for serum triglyceride levels in both groups.

Step 3: Choose a statistical test

Depending on the number of groups in the study, the sample size in each group and measurement scale for variable of interest,

appropriate statistical test is chosen by referring to **Table 11.5**. In this example, variable (serum triglyceride levels) is measured on quantitative scale and compared between two different groups. Let us assume that one of the two groups sample size is less than 30 and since the difference of means is to be compared, the unpaired t-test will be applied.

Step 4: Set cutoff value for type 1 error (α)
Alpha is commonly set at 0.05. (i.e., 95% confidence level).

Step 5: Calculation of test statistic (as applicable along with formula)
This is detailed in the subsequent section.

Step 6: Comparison of test statistic with table value
At a particular significance level, table values for test statistics are the maximum points till which the difference is considered by chance. If test statistic is more than this table value, it is less likely to be by chance, more likely to be true.

Thus, if we are checking the table values at 0.05 level of significance and calculated value of test statistic is more than allowable table value of test statistics, the obtained *p*-value is less than 0.05 meaning thereby the probability of difference being by chance is less than 5% and hence difference is likely to be real (accept alternate hypothesis).

Alternately, while checking the table values at 0.05 level of significance, if the calculated value of test statistic comes less than the allowable table value of test statistic, the obtained *p*-value is more than 0.05 meaning thereby the probability of difference being by chance is more than 5% and hence the difference is likely to be by chance and not real one (accept null hypothesis).

Step 7: State the statistical inference
If *p*-value is less than 0.05, we can conclude that at 95% confidence level the difference in mean serum triglyceride levels among CVD patients and normal individuals is found to be real and not by chance.

If *p*-value is more than 0.05, we can conclude that at 95% confidence level the difference found in mean serum triglyceride levels among CVD patients and normal individuals is likely to be by chance and so it is not a real difference.

Step 8: State the conclusion
If *p*-value is more than 0.05, conclusion would be, "mean serum triglyceride levels are *significantly higher* (p <0.05) among CVD patients compared to normal individuals."

If *p*-value is more than 0.05, conclusion would be, "the study failed to detect any significant difference in mean serum triglyceride levels among the CVD patients and normal individuals".

Applied Aspect

When we obtain a *p*-value less than 0.05, we conclude that the difference observed in study is statistically significant. This statistical significance means that researchers could reasonably eliminate the possibility of chance error in their study findings. However, researchers have to check if other types of errors (bias and confounding) are there in study findings. Further, it is important to understand that a difference which is found statistically significant may not always be clinically or epidemiologically relevant. Researchers have to apply their mind to decide the utility of study findings.

Tests of Significance: Significance of Difference in Means and Proportions

We will now describe how different statistical tests are applied in some sample study situations. The steps of statistical inference procedure as described earlier are to be followed in these calculations also.

Concept of Degrees of Freedom

Degrees of freedom are the "space" available for observations to vary. For example, there are three numbers x, y, and z which add up to a total of 100. The first two numbers x and y can be freely chosen, but the third number z will be restricted to z = 100–x – y. Thus, three observations have only two free "spaces" to vary. Similarly, n observations have n – 1 degrees of freedom.

As described previously in the steps of statistical inference, the calculated test statistic is to be compared with the table value. For deciding the table value, the t-distribution and chi-square distribution tables have to be looked at. The tables have degrees of freedom in the first column and table values in subsequent columns according to the level of significance. For instance, in t-distribution table, for degrees of freedom of 5, the corresponding table value for t is 2.571. Similarly, in chi-square distribution table, for degrees of freedom of 1, the corresponding table value for chi-square is 3.84.

For a one-sample t-test of mean and for a paired t-test, the formula for degrees of freedom is n–1. For unpaired t-test, the formula for degrees of freedom is $n_1 + n_2 – 2$. For chi-square test, the formula for degrees of freedom is the number of rows–1 multiplied by the number of columns–1 = (r–1 × c–1).

Concept of One-tailed and Two-tailed Hypothesis

The first two rows in t-distribution table are that for a one-tailed and two-tailed test and the values of t-distribution differ accordingly. Deciding on which row to look at depends on the type of hypothesis in research study.

When the investigator does not know whether mean of any one group will be higher or lower than the other, then it is said to be a two-tailed hypothesis. Many times, based on clinical experience or biological plausibility the investigator would know if mean of one group will be higher than the other. In this situation, it will be a one-tailed hypothesis. Thus, the direction of hypothesis, whether it is two-tailed or one-tailed, is to be decided by the researcher before the study starts.

For z-tests, the table value for two-tailed tests is 1.96 while the table value for one-tailed tests is 1.64. The degrees of freedom concept does not apply to z-tests.

Significance of Difference of Means of Two Different Groups when n <30 (Unpaired t-test)

The unpaired t-test is applied to compare means of two different groups. Before applying the unpaired t-test, the data needs to meet a few requirements and assumptions as follows:

- Variable is being compared between two different groups in the hypothesis
- Variable is measured on quantitative scale
- Variable follows normal distribution
- *Equality of variance*: Data in both groups have the same *variance*
- Sample size (n) in either of the two groups is less than 30.

Please note that while applying t-test, the assumption of normal distribution stands true even when it is an approximately normal data because it is quite robust to deviations from normality.

The formula for applying unpaired t-test is as follows: $t = \frac{\bar{x}_1 - \bar{x}_2}{SE}$

Where, $\bar{x}$ is mean of the first group, $\bar{x}_2$ is mean of the second group, and SE is standard error of difference between two means.

Standard error (SE) = $Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$, where $Sp = \sqrt{\frac{SD_1^2(n_1-1)+SD_2^2(n_2-1)}{(n_1+n_2-2)}}$

The unpaired t-test is explained with the following example:

"A group of 69 CVDs patients had a mean serum triglyceride level of 230 mg/dL with an SD of 4.9 and another group of 29 normal individuals had a mean serum triglyceride level of 180 mg/dL with an SD of 8.4. Apply appropriate statistical test to test the hypothesis that the mean serum triglyceride levels among CVD patients are different than normal individuals. (*Note:* Two-tailed table value of t at degrees of freedom of 96 is 1.985.)"

On doing the calculations, standard error (SE) = 1.36.

Therefore, $t = \frac{230-180}{1.36} = 36.85$

Degrees of freedom for unpaired t-test = $n_1 + n_2 - 2 = 69 + 29 - 2 = 96$; table t-value = 1.985.

At degree of freedom of 96, since the calculated t-value (36.85) is more than the table value of 1.985, null hypothesis is rejected and alternate hypothesis is accepted. Hence, the difference in mean serum triglyceride levels among CVD patients and normal individuals is statistically significant ($p < 0.05$). The study found that mean serum triglyceride levels is significantly higher among CVD patients compared to normal individuals ($p < 0.05$).

Significance of Difference of Means of the Same Group (Paired Data)—Paired t-test

The paired t-test is applied to compare means of the same sample before and after any intervention/procedure. Before applying the paired t-test, the data needs to meet a few requirements and assumptions as follows:

- Data comes from paired observations from only one group (before-after)
- Data to be compared should be quantitative
- Variable follows normal distribution.

The sample size for calculation of paired t-test can be any number as far as the above assumptions stand true.

The formula for applying paired t-test is,

$$t = \frac{\text{Mean of difference}}{\frac{\text{SD of differences}}{\sqrt{n}}}$$

The paired t-test is explained with the following example:

"A new drug for weight loss was being tested and the reduction in weight among 10 patients is given below:

Weight in kilograms before taking the drug	*Weight in kilograms after taking the drug*	*Difference*
52	51	1
109	105	4
112	113	−1
98	95	3
87	80	7
128	120	8
115	114	1
69	65	4
80	82	−2
85	80	5

Apply appropriate statistical test to find out whether the drug is significantly reducing the weight among the 10 individuals. (*Note*: Table value of t at degrees of freedom of 9 is 2.26)."

On doing the calculations, $\frac{\text{Mean of differenes}}{\frac{\text{SD of differences}}{\sqrt{n}}} = \frac{3}{\frac{3.26}{\sqrt{10}}} = 2.9$

Degrees of freedom = n−1 = 9. At degrees of freedom of 9, since the calculated t-value (2.9) is more than the table value of 2.26, the null hypothesis is rejected and alternate hypothesis is accepted. Hence, the difference in mean weight before and after taking the drug is statistically significant. The new drug significantly reduces weight among the study participants ($p < 0.05$).

Analysis of Variance

In previous example, we compared means between two groups. ANOVA test is applied to compare means of more than two different groups. If the ANOVA test is statistically significant ($p < 0.05$), it implies that the mean is significantly different in at least one of then different groups being tested. In order to find out exactly which group is significantly different, a posthoc test is applied. The calculations of ANOVA test and posthoc tests are beyond the scope of this book.

Chi-square Test

Chi-square test is a nonparametric test mainly used to test the difference between two proportions or for testing associations between two categorical variables. There are three types of chi-square tests:

1. Chi-square test of association
2. Chi-square test of proportion
3. Chi-square test of goodness of fit.

Chi-square Test of Association

Chi-square test of association is applied to test the association between two categorical variables. Before applying the chi-square test, the data needs to meet a few requirements and assumptions as follows:

- Both the variables between which association is to be tested should be categorical

- At least 80% of expected cell values should be more than 5
- None of the expected cell values should be less than 1.

The calculation of expected cell values is explained below. The minimum sample size required for applying chi-square test of association can be any number as far as the above assumptions stand true. Also, as chi-square is a nonparametric test, the assumption of normal distribution does not apply here.

Classically, for a chi-square test, a 2 × 2 contingency table between an exposure and a disease as given below is to be constructed.

Exposure	Disease present	Disease absent	Total
Present	a	b	a + b
Absent	c	d	c + d
Total	a + c	b + d	a + b + c + d

The cells a, b, c, and d are observed values or values from findings of the study. The first step in calculating a chi-square test is to calculate expected values for each of the observed values.

In the above table, expected value for cell "a" =

$$\frac{(\text{Row total}) \times (\text{Column total})}{\text{Table total}} = \frac{(a+b) \times (a+c)}{a+b+c+d}$$

Similarly, the expected value for "b" = $\frac{(b+d) \times (a+b)}{a+b+c+d}$ and so on for "c" and "d".

For a 2 × 2 contingency table, the observed values are denoted by O_1, O_2, O_3, and O_4 and the expected values are denoted by E_1, E_2, E_3, and E_4. While the formula for calculation of chi-square (χ^2) is

$$\chi^2 = \chi_1^2 + \chi_2^2 + \chi_3^2 + \chi_4^2 \text{ where } \chi_1^2 = \frac{(o_1 - E_1)^2}{E_1}, \chi_2^2 = \frac{(o_2 - E_2)^2}{E_2}$$

and so on.

The chi-square test is explained with the following example:

"A study was conducted to find out the association between parity and development of breast cancer. A group of 450 women with less than 3 parity and another group of 400 women with more than or equal to 3 parity were followed up for the development of breast cancer. Out of 450 women with less than 3 parity, 26 developed breast cancer and out of 400 women with more than or equal to 3 parity, 10 developed breast cancer. Apply appropriate statistical test to check whether parity less than 3 is associated with the development of breast cancer. (*Note*: Table value for χ^2 at degrees of freedom of 1 at 0.05 level of significance is 3.84.)"

The 2 × 2 contingency table prepared from the provided data is as below:

Parity groups	Developed breast cancer	Did not develop breast cancer	Total
<3 parity	26	424	450
≥3 parity	10	390	400
Total	36	814	850

On doing the calculations, $\chi^2 = \chi_1^2 + \chi_2^2 + \chi_3^2 + \chi_4^2$ = 2.6 + 0.11 + 2.9 + 0.13 = 5.74

At degree of freedom of 1, the calculated ξ^2 value (5.74) is more than the table value (3.84). Hence, the null hypothesis is rejected and alternate hypothesis is accepted. Thus, at 95% confidence level, parity less than 3 is associated with the development of breast cancer (p <0.05). Women with parity less than 3 are at higher risk of developing breast cancer than women with more than or equal to 3 parity (p <0.05).

Chi-square Test of Proportion

Chi-square test of proportion is applied to compare the difference between two proportions. Before applying the chi-square test of proportion, the data needs to meet a few requirements and assumptions as follows:

- There should be two different groups in the hypothesis
- Data to be compared between the two groups should be qualitative (categorical, i.e., either nominal or ordinal)
- Sample size (n) in either of the two groups is less than 30
- At least 80% of expected cell values should be more than 5
- None of the expected cell values should be less than 1.

As chi-square test is a nonparametric test, the assumption of normal distribution does not apply here. The formula for applying chi-square test of proportion is same as described for chi-square test of association above.

The calculation of chi-square test of proportion is explained with the following example:

"Out of 20 hypertensive adolescent patients in a hospital, 10 had obesity and out of 45 normotensive adolescent patients in the hospital, 10 had obesity. Apply appropriate statistical test to compare whether the proportion of adolescent obesity is different among hypertensive and normotensive patients. (*Note*: Table value for χ^2 at degrees of freedom of 1 at 0.05 level of significance is 3.84.)"

The 2 × 2 contingency table prepared from the provided data is as below:

Obesity	Hypertensive patients	Normotensive patients	Total
Present	10	10	20
Absent	10	35	45
Total	20	45	65

Looking at the table, we can see that among hypertensives 50% are obese while among normotensives 22.2% are obese. Let us apply test to see if this difference is statistically significant.

On doing the calculations, = 2.41 + 1.07 + 1.07 + 0.48 = 5.03 At degree of freedom of 1, the calculated $\chi 2$ value (5.03) is more than the table value (3.84). Hence, the null hypothesis is rejected and alternate hypothesis is accepted. Thus, at 95% confidence level, the difference in proportion of obese among hypertensive and normotensive patients is statistically significant (p <0.05). Obesity is significantly higher among hypertensive patients than normotensive patients (p <0.05).

Chi-square Test of Goodness of Fit (2 × 1 Chi-square)

Chi-square test of goodness of fit is used to compare the difference between two proportions of the same sample. Before applying the chi-square test, the data needs to meet a few requirements and assumptions as follows:

- There should be only one variable, which should be categorical (i.e., nominal or ordinal)
- At least 80% of the expected cell values should be more than 5
- None of the expected cell values should be less than 1.

The minimum sample size required for applying chi-square test of goodness of fit can be any number as far as the above

assumptions stand true. In this test, the observed values are compared with the "normally" expected values for that variable.

The calculation of chi-square test of goodness of fit is explained with the following example:

"In a random sample of 100 people in a city, it was found that 70 persons were male and 30 were female. Apply an appropriate statistical test to find out whether the proportion of males is higher than the proportion of females in this sample. (*Note*: Table value for χ^2 at degrees of freedom of 1 at 0.05 level of significance is 3.84.)"

Number of males (O_1) = 70, number of females (O_2) = 30. It is "normally" expected that in a sample of 100 people, there should be 50 males and 50 females. Therefore, expected values $E_1 = 50$ and $E_2 = 50$.

On doing the calculations, $\chi^2 = \chi_1^2 + \chi_2^2 = 8 + 8 = 16$

At degrees of freedom of 1, the calculated χ^2 value (16) is more than the table value (3.84). Hence, the null hypothesis is rejected and alternate hypothesis is accepted. Thus, at 95% confidence level, the difference in proportion of males versus females in the sample is statistically significant (p <0.05). Proportion of males is significantly higher than that of females in the sample (p <0.05).

CORRELATION AND REGRESSION

Correlation

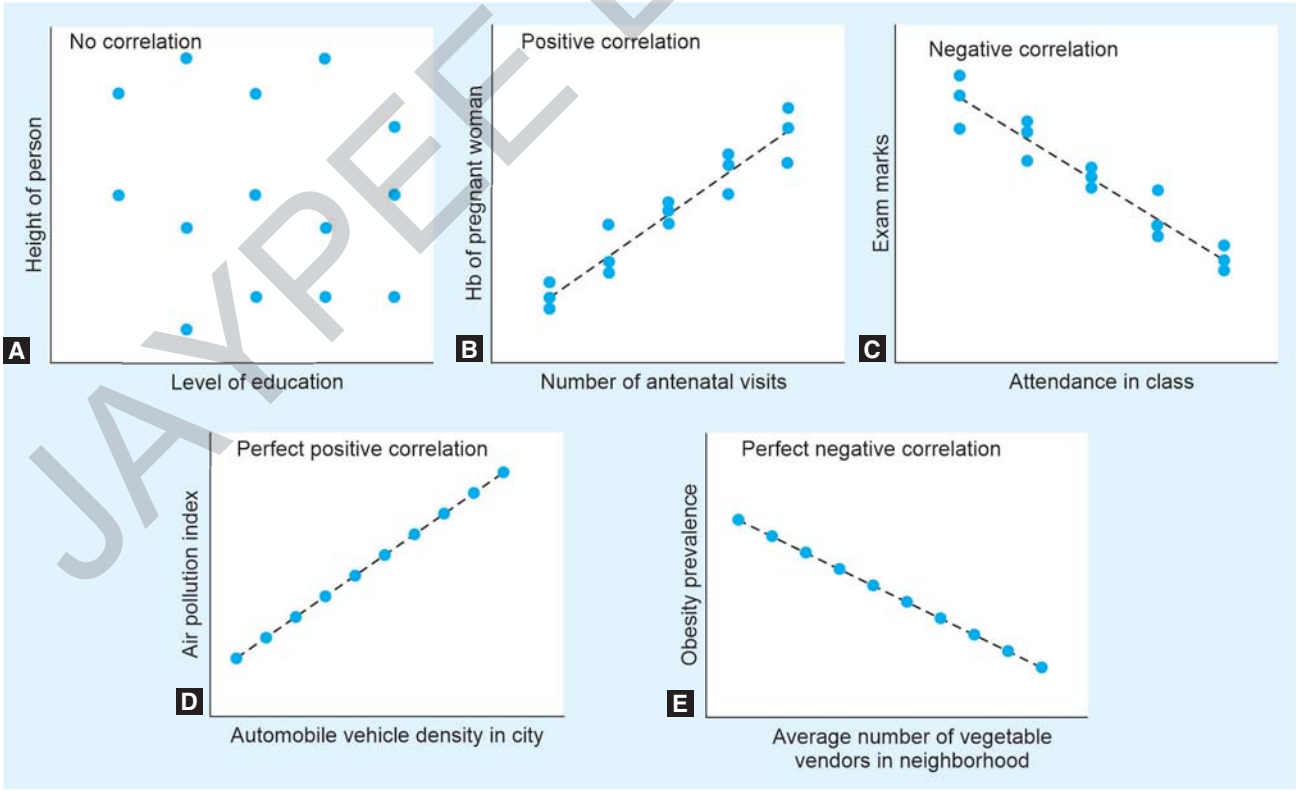
Scatter diagram is a visual method of checking if there is a relationship between two quantitatively measured variables. If a relationship exists between the two variables; with changes in value of one variable the other variable will also change.

Types of Correlation

In the figure, chart A shows *no correlation* between two variables **(Fig. 11.19)**. Chart B shows a *positive correlation*, which means that with increase in value of variable put on X-axis the value of variable on Y-axis also increases. Chart C shows *negative correlation*, where with increase in one variable the value of other variable decreases. Chart D shows a *perfect positive correlation* where there is a linear relationship between the two variables. Chart E shows such *perfect negative correlation*.

Correlation Coefficient (r)

The degree of relationship between the two variables is measured by correlation coefficient (denoted by "r"). The value of "r" ranges from –1 for a perfect negative correlation (chart E) to +1 for a perfect positive correlation (chart D). Value of "r" at 0 indicates that there is no relationship between the two quantitative variables (chart A). There are two methods for calculation of correlation coefficient. When the variables follow normal distribution the formula of Karl Pearson's product-moment correlation coefficient is used and when they do not follow normal distribution formula of Spearman correlation coefficient is used. Irrespective of the formula used for calculation of "r", its interpretation is same which is as follows.



Figs. 11.19A to E: (A to E) Types of correlation.

Value of "r"	Interpretation
0–0.3	Negligible correlation
0.3–0.5	Low positive correlation
0.5–0.7	Moderate positive correlation
0.7–0.9	High positive correlation
0.9–1	Very high positive correlation

Same way interpretation of negative correlation goes from 0 to –1.

Regression

While correlation is used to know the strength of association between two quantitative variables, regression goes a step further. Regression is used to predict the value of one variable (dependent variable) with each unit change in another variable (independent variable) **(Fig. 11.20)**.

Note

Correlation gives degree and direction of relationship between two variables
Regression predicts the values of one variable on the basis of the other variable.

The regression equation is y = a + bx where; y is dependent variable, "a" is intercept, "x" is independent variable, and "b" is the regression coefficient.

If the regression equation for height (cm) as independent variable (x) and weight (kg) as dependent variable (y) is y = –133.1 + 1.16 x; then its interpretation is as follows:

For each 1 cm increase in height there will be 1.16 kg increase in weight. In medical science the intercept 'a' does not provide any useful interpretation.

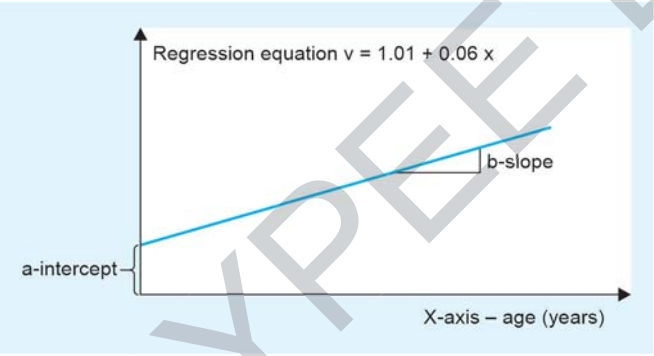


Fig. 11.20: Regression curve.

Multivariate Regression Analysis

Earlier we saw how change in one independent variable can predict the change in a dependent variable. In real life many of the outcome variables are dependent on more than one independent variables. In such situations, if we look at the equation, we will have one y and more than one x.

$$Y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

For example, SBP might be dependent on age (x_1), gender (x_2), waist circumference (x_3), and height (x_4).

Types of Multivariate Regression

The type of multivariate regression depends on the nature of dependent variable.

- Multiple linear regression is used when the dependent variable is measured on quantitative scale (e.g., serum cholesterol level).
- Multiple logistic regression is used when the dependent variable is measured on binary categorical scale (e.g., treatment outcome measured in two categories success or failure).
- Cox regression model is used when dependent variable is survival data.

INTERPRETING THE STUDY FINDINGS

Are all statistically significant results also clinically significant? The following table **(Table 11.6)** shows findings from a study comparing the antihypertensive effect of drugs A and B on SBP. Observe the result of statistical significance testing when sample size is increased from 100 per group to 300 per group.

Table 11.6: Findings from a study comparing the antihypertensive effect of drugs A and B on SBP.

Situation 1: With original sample size			*Situation 2: Using very high sample size*		
Groups	***Mean ± SD of SBP***	Statistical test	***Groups***	***Mean ± SD of SBP***	Statistical test
Group A (n = 100)	140 ± 10 mm Hg	Unpaired t-test $p = 0.15$	Group A (n = 300)	140 ± 10 mm Hg	Unpaired t-test, $p = 0.01$
Group B (n = 100)	138 ± 10 mm Hg		Group B (n = 300)	138 ± 10 mm Hg	

We see that when sample size is increased from 100 per group to 300 per group, the result of statistical significance changes from nonsignificant to significant difference. More important to understand here is that the difference is only of 2 mm Hg which may not be clinically relevant. Thus, the clinical utility of the study results should also be examined.

Let us take another example where antibiotics A and B are being compared **(Table 11.7)**.

Table 11.7: Findings from a study comparing two antibiotic drugs A and B.

With small sample				*With a small increase in sample size*			
Status	*Group A*	*Group B*		*Status*	*Group A*	*Group B*	
Cured	6 (30%)	12 (60%)	$\chi^2 = 3.6$	Cured	9 (30%)	18 (60%)	$\chi^2 = 5.4$ p = 0.019
Not cured	14 (70%)	8 (40%)	$p = 0.056$	Not cured	21 (70%)	12 (40%)	
Total	20	20		Total	30	30	

With a sample size of 20 patients in each group, there is a large difference found in the study where only 30% patients in group A are cured versus 60% in group B. In spite of the large difference the statistical significance testing through chi-square test shows a nonsignificant difference ($p > 0.05$). Looking at the large difference when the researcher increases the sample size to 30 in each group now, the difference is statistically significant ($p < 0.05$).

Thus, going through these two examples we understand that at times a small difference may come as statistically significant and large difference may become statistically nonsignificant

because of the sample size involved, hence checking whether the findings are clinically meaningful or not is also important to check while interpreting study results.

Descriptive Studies

Descriptive studies are done to estimate the average value of variable of interest in given reference population. It is important to make use of CIs rather than only stating point estimates.

Analytical Studies

The analytical studies could be a risk factor study that tries to find association between exposure and outcome variable or an intervention study trying to find out the difference in outcome variable between two groups. Below is an example of findings from an analytical study trying to see if owning a mobile phone among school students is associated with low physical activity. Commonly the results of such study are presented as below **(Table 11.8)**.

Table 11.8: findings from an analytical study trying to see if owning a mobile phone among school students is associated with low physical activity.

	Low physical activity	*Moderate physical activity*	*Total*	*Relative risk (95% CI of RR)*	*p-value*
Owns mobile phone	40 (80%)	10 (20%)	50	2 (1.5–2.6)	$p = 0.000003$ Chi-square test
Does not own mobile phone	40 (40%)	60 (60%)	100		

Ask these questions while interpreting analytical studies
1. What is the value of measure of strength of association?
2. What is the range of confidence interval?
3. What is the likelihood of chance error (*p*-value)?
4. What is the likelihood of systematic error?
5. Utility of results:
   a. Is association causative?
   b. Intervention's safety, acceptability, and cost.

In this table, the relative risk indicates the strength of association. With help of computer applications, we can also easily calculate the 95% CI of relative risk. This CI in this example means that students possessing mobile phones are 1.5–2.6 times more likely to have low physical activity level. In other words, CI indicates how precise our estimate of the relative risk is. Larger the sample size more precise would be this estimate. Results of chi-square test here show that the association is statistically significant. The low or high *p*-value does not indicate anything about the strength of association; it only tells that the probability of chance error in our results. In this case since p is very small, researchers have reasonably eliminated the chance error in study results. Apart from chance errors the systematic errors (selection and measurement bias and confounders) are also important. So, in this example also the researchers should discuss how the findings of this table are free from these systematic errors. Finally, not all associations are causations. So, applying the criteria for determining causation (described in epidemiology) is necessary to check if this apparent association is really causative.

If the research study is an intervention study assessing the efficacy of one intervention with the other on the outcome variable, the steps of interpretation would be largely the same. Here also, for assessing efficacy we would interpret the relative risk reduction, its 95% CI, chance error (*p*-value), and systematic error. In last step, the researchers should also discuss the acceptability (tolerability) of the intervention by the patients, its safety, and costs associated with such investigational new intervention.

CONCEPT OF ETHICS IN HUMAN RESEARCH

What is Medical Ethics?

Ethics refers to "a set of moral principles or values which determines the code of conduct in medical profession so as to serve in the best interest of individual and society".

Why Ethics in Research?

Similar to the expectation of ethical conduct by a health professional in medical practice, it is also expected while performing health research. Ethics is about safeguarding the dignity, rights, safety, and well-being of the human research participants.

History

In past, medical research witnessed few cruel forms of exploitation of human beings for research purpose.

During World War II, medical experiments were performed on camp prisoners without their consent leading to death and permanent disability. The *Nuremberg trial* considered it as violation of human rights. In *Tuskegee Syphilis study* which was planned to understand the natural progression of syphilis disease, around 600 male participants were enrolled but they were never told they had syphilis. Researchers became aware of penicillin for treatment in 1947, but knowingly did not treat patients and prevented them for taking treatment. Thus, many men, their wives, and children suffered from dreaded consequences of syphilis. A series of brutal medical experiments with freezing temperatures, high altitude, head injury, sea water submersion, and mustard gas were performed during World War II.

After these series of unethical conduct in medical research, a series of meetings and guidance reports were generated. The Belmont Report (1979) drafted by the National Commission for the Protection of Human Participants of Biomedical and Behavioral Research in United States is widely used as a guiding document on ethical conduct of health research. It focused on three important principles of ethics.

Note

Three important principles of ethics:
1. Respect for persons
2. Beneficence and nonmaleficence
3. Justice

1. ***Respect for persons*:** This principle tells that each person should be treated as autonomous individual capable of taking decisions for him/herself. Hence each human being

participating in research should be given an opportunity to choose what would or would not happen to them. This opportunity is given in form of informed consent process explained below. This underscores that participation in research study should be voluntary and not forced. This principle also mentions that special protections be provided to those individuals who have reduced capacity to take decisions for themselves such as children or vulnerable people such as prisoners when they participate in research.

2. ***Beneficence and nonmaleficence*:** It means that research should maximize the possible benefits to the subjects and minimize the possible harm.
3. ***Justice*:** This principle means that there should not be exploitation of some vulnerable population group or community for research purpose. The community that runs the risk by participating in research should also get the benefits of the research outcome.

Informed Consent

The word "informed" in this term means informing the research participants of the risks and benefits of participating in research. Consent means human beings are not forced to participate in research, but they participate in it voluntarily after knowing the risks and possible benefits. The recent guidelines suggest audiovisual recording of this entire consent process in case of clinical trials involving human participants. It includes two main components: (1) participant information sheet (PIS) and (2) informed consent form.

1. ***Participant information sheet*:** This is a document containing all relevant information of the study that a possible participant needs to know. This document should include answers to a list of questions provided in the box in simple language which can be easily understood by the study participants. A copy of the PIS is to be provided to all the study participants preferably in the local vernacular language.
2. ***Informed consent form*:** It is the statement of declaration by the study participants and the researcher as shown in the sample form in the box below. This written informed consent form has to be signed or thumb impressed by the study participant for literate or illiterate participants, respectively. It has to be signed by the participant in presence of the third-party witness, who is neither related to the study participant nor to the researcher. It has to be preserved by the researcher.

A participant information sheet answers the following questions:

- What is this study about?
- Why are you being invited to participate?
- What will you be asked to do?
- Will there be any benefits to you?
- Will there be any risks to you?
- Will there be recordings or photos?
- Who will get the results?
- Where can you get the results?
- Are there any costs?
- Any questions?
- Who to contact if there are questions later?

Why there is a Need for an Ethics Committee?

If the opinion on ethical aspects of a proposed research study is left to the individual researcher, there is a risk of running a biased decision. Researchers might feel that their study is ethical and there is no harm to their study participants.

Most of the academic institutions have their local Institutional Ethics Committee (IEC). Researchers are required to submit the detailed study protocol to IEC in the prescribed format. Member secretary of IEC screens the received proposals and depending on the possible risk involved, the type of review will be decided as: (1) *exemption from review* (proposals with less than minimal risk like educational surveys), (2) *expedited review* (proposals with minimal risk like review of secondary records or minor deviations from previously approved proposals) or (3) *full review* (proposals with more than minimal risk like blood collection, investigations or intervention trials). Meetings of IEC are held at regular intervals and full reviews of the proposals are conducted. Based on the discussion during the meetings, the study proposal is either approved in the same state or approved with some major/minor suggestions or disapproved. Approval from IEC is mandatory before conducting any research study.

DISSEMINATION OF RESEARCH FINDINGS: WRITING A RESEARCH REPORT

Importance of Writing a Research Report

Dissemination of the research findings is an important responsibility of the researcher. Without dissemination of the research findings, the research process is incomplete.

Format of Writing a Research Report

The most commonly used format is known by the acronym "IMRAD".

Note

I—Introduction	Why did you study?
M—Methodology	How did you study?
R—Results [A—and]	What did you find?
D—Discussion	What does it mean anyway?

Introduction

In this section, the background of the study is mentioned. We answer three questions in this section. (1) Why this topic or problem is important to study? What is already known about this topic or problem through existing evidence? (2) What are the gaps in existing knowledge or what is still not known about this topic or problem? (3) How this study plans to answer this unanswered question or fill the gap in knowledge? This section usually ends with specific objectives of the study.

Methodology

Also described as "materials and methods", this section explains "how the study was done?" Dictum is that it should be sufficiently detailed to enable other interested researcher to conduct a similar study in her setting by reading this section of the research project.

Components of methodology section:
- Study design
- Study setting
- Study duration
- Study population
- Sample size
- Inclusion criteria
- Exclusion criteria
- Study procedure
- Data collection
- Outcome measures
- Data entry and analysis
- Quality assurance
- Ethical considerations

The following are common headings for describing methodology:

- ***Study design:*** Mention the type of study design used like cross-sectional or case-control or cohort.
- ***Study setting/site:*** Where was the study conducted? Hospital or community-based study?
- ***Study duration:*** When was the study conducted? Mention the period of data collection.
- ***Study population:*** Among whom was the study conducted? Whether among patients or healthy population or pregnant women or children or adolescents? Sampling method is also mentioned here.
- ***Inclusion and exclusion criteria:*** What criteria were used to enroll the study participants? For instance, adult patients above age of 18 years only were enrolled in the study and severely ill patients were excluded.
- ***Sample size:*** How many study participants were enrolled? How was sample size calculated and whether calculated sample size was achieved?
- ***Study procedure:*** We mention the flow of activities for this study here. It is usually starting from participant enrolment till the completion of all possible follow-up with the study participant.
- ***Data collection:*** How was the data collected from the study participants? Whether any validated tool or study questionnaire or any equipment was used? Here, we also mention about various variables in study questionnaire. Who collected the data? When and where was the data collected? Any challenges faced during data collection?
- ***Outcome measures:*** The main outcome variable including its scale of measurement and operational definition is explained.
- ***Data entry and analysis:*** Which software was used for data entry and analysis? How were data entry errors handled? How was data analyzed? Which statistical tests were applied? What was the level of significance accepted?
- ***Quality assurance:*** What were the quality assurance measures taken during this study? For instance, conducting double data entry with validation to minimize data entry errors; use of standard validated tool for data collection; blinding/random allocation of participants in randomized control trial (RCT) are examples of quality assurance measures.
- ***Ethical considerations:*** Whether the study was approved by the IEC? Whether written informed consent was obtained from the study participants? How privacy and confidentiality were maintained?

Results

This section presents the findings of the research study. Findings are presented in text, tables, and figures. The statistical tests with *p*-value will be found in this section. We should try to avoid the repetition of all the result findings mentioned in the tables and figures.

Discussion

Discussion is meant to write the interpretation of the study findings and their implications. Here, we summarize the important findings of the study and explain possible reasons for such study findings. We compare our findings with other research studies. We need to discuss similarities and contradictions with other studies by appropriate logical reasoning. We also state the strengths and limitations of the study. We can also mention the directions for future research. Lastly, it ends with conclusion summarizing the most important study findings. Conclusion should always be aligned with the objectives of the research study. This section gives the researchers a liberty to write in a story telling fashion to maintain the flow of reading.

References

It is the list of all resources, books or reports or previous research article used as reference in the research study. Reference list is placed at the end after the discussion section in appropriate format such as the Vancouver style.

USES OF COMPUTER IN HEALTH RESEARCH

Computer applications are helpful in different stages of health research starting from protocol writing till the dissemination of research findings. Following is a brief on uses of computer at different stages.

During Protocol Writing

Literature search includes finding of reports, health statistics, and existing evidence in the topic of research. Google scholar (http://scholar.google.com) is a generic search engine while PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) is specific for field of medicine. Softwares for sample size calculation, such as Epi Info™ (https://www.cdc.gov/epiinfo/index.html) as well as Statulator (http://statulator.com/SampleSize/ss1P.html) and (http://powerandsamplesize.com) are also available. Epi Info™ also helps to create a structured questionnaire.

During Research Project Implementation

EpiCollect (https://five.epicollect.net/) is a free electronic data collection tool especially designed for health research. Microsoft Excel can be used for generating random numbers in simple random sampling or for randomized clinical trials.

For Data Management

Although Microsoft Excel is used for data entry by many researchers, EpiData (http://www.epidata.dk) is a more

appropriate tool since it provides facility of more stringent quality checks and double data entry and validation. Basic and some descriptive statistics, e.g., mean, SD and proportion, and t-test can be easily calculated in Microsoft Excel. For other statistical tests EpiData, EpiInfo, and OpenEpi (https://www.openepi.com) are the freely available packages while SPSS, Stata are available to only licensed users. For presentation of data, tables and charts can be easily prepared in Microsoft Excel.

For Dissemination of Research Findings

Nowadays, the entire process of processing of a research manuscript for publication in a scientific journal is online through the journal's online manuscript submission system. PowerPoint is commonly used for presenting the research findings in conferences meetings for oral or poster presentation.

SUMMARY

- Purpose of research is to add to the scientific knowledge, change policy or medical practice.
- The quantitative approach of research is done to estimate a problem using numbers, while the qualitative approach is used to explore the different perspectives of the health problem.
- Variable of interest and study population are two important components to define while conducting research. Type of variable and scale of measurement decide how data will be collected and analyzed.
- Probability sampling methods allow the study sample to be representative of reference population.
- Sample size calculation helps us to keep study sample big enough for our study to be scientifically robust while being small enough to be feasible.
- Defining the data collection tool, and analysis plan at the start of the study makes it scientifically robust.
- Methods of data presentation help the researchers communicate research findings in clear and concise manner.
- The tools of biostatistics are useful as descriptive statistics to describe the data in concise form and as inferential statistics to find the chance errors in analytical study findings.
- Interpretation of the research results is as important as applying correct statistical tool. Ensuring ethical conduct in research is important.
- Research report should be sufficiently detailed to allow a scientific critique by another researchers and possible users of the results.

SUGGESTED READING

1. Bewick V, Cheek L, Ball J. Statistics review 8: Qualitative data—tests of association. Crit Care. 2004;8(1):46-53.
2. Brereton RG. The normal distribution. J Chemom. 2014;28(11):789-92.
3. Centers for Disease Control and Prevention (CDC). Epi Info™ [Internet]. Atlanta, GA: CDC; 2025 [cited 2025 Sep 14]. Available from: https://www.cdc.gov/epiinfo/index.html(Note: CDC will phase out Epi Info™ support after Sept 2025; users are advised to migrate to alternative tools.)
4. Centers for Disease Control and Prevention (CDC). Top 10 Great Public Health Achievements in the 20th Century [Internet]. Atlanta, GA: CDC; 2025 [cited 2025 Sep 14]. Available from: http://www.cdc.gov/about/history/tengpha.htm
5. Cummings SR, Browner WS, Hulley SB. Conceiving the research question and developing the study plan. In: Hulley SB, Cummings SR, Browner WS, Grady D, Newman TB, editors. Designing clinical research: an epidemiologic approach. 4th ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2013. pp. 17-19.
6. Daniel WW, Cross CL. Biostatistics: A foundation for analysis in the health sciences. 10 th ed. Hoboken, NJ: John Wiley and Sons; 2013. pp. 614-16.
7. David B. Medical statistics from scratch: An introduction for health professionals. 2nd ed. West Sussex, UK: John Wiley and Sons Ltd; 2008.
8. EpiData. EpiData software [Internet]. 2013 [cited 2025 Sep 14]. Available from: http://www.epidata.dk
9. Fathalla MF, Fathalla MMF. A practical guide for health researchers. Cairo: World Health Organization Regional Office for the Eastern Mediterranean; 2004.
10. Frenk J. The new public health. Annu Rev Public Health. 1993;14:469-90.
11. Gauss CF. Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Crawley, UK: ABC Books; 1809. pp. 1-266.
12. Gupta P, Singh N. How to write the thesis and thesis protocol: A primer for medical, dental, and nursing courses. 1st ed. New Delhi: Jaypee Brothers Medical Publishers (P) Ltd; 2014.
13. Hulley SB, Cummings SR, Browner WS, Grady D, Newman TB. Designing clinical research: an epidemiologic approach. 4th ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2013.
14. Indian Council of Medical Research (ICMR). Short Term Studentship [Internet]. 2018 [cited 2025 Sep 14]. Available from: http://14.139.60.56:84/Homepage.aspx
15. Indian Council of Medical Research. National ethical guidelines for biomedical and health research involving human participants. New Delhi: ICMR; 2017.
16. International Union Against Tuberculosis and Lung Disease, South East Asia Office. Efficient, quality-assured data capture and analysis using EpiData: Course manual. New Delhi: IUATLD; 2013.
17. International Union Against Tuberculosis and Lung Disease. Structured operational research course material. New Delhi: IUATLD South-East Asia Office; no date.
18. National AIDS Control Organization (NACO). National strategic plan for HIV/AIDS and STI 2017–2024. New Delhi: NACO; 2017.
19. OpenEpi. Open source epidemiologic statistics for public health [Internet]. 2013 [cited 2025 Sep 14]. Available from: https://www.openepi.com/Menu/OE_Menu.htm

IAPSM's Textbook of COMMUNITY MEDICINE

Salient Features

- **Competency-Based Orientation:** Fully aligned with the Competency-Based Medical Education (CBME) curriculum prescribed by the National Medical Commission, with emphasis on measurable learning outcomes and applied understanding.
- **Structured Organization:** Divided into five logically sequenced sections, progressing from foundational concepts of health and disease to the organization and management of community health in India, thereby catering to students across different phases of medical training.
- **Updated Epidemiological Data:** Epidemiological indicators have been revised in accordance with the latest available data from NFHS-5, recent SRS reports (2024–25), other national health surveys, ICMR-NIN (2020 onward), WHO global updates, and other credible sources.
- **Comprehensive Coverage of Contemporary Topics:** Includes emerging and re-emerging diseases, noncommunicable diseases, environmental and occupational health, urban, rural and tribal health, geriatric health, travelers' health, health management, health financing and economics, research methodology, and social medicine.
- **Programmatic Integration:** Provides detailed and updated coverage of major national health programs, including RMNCAH+N, NP-NCD, NTEP, NLEP, NACP-V, NPHCE, nutritional programs, initiatives addressing universal health coverage (PM-JAY, Ayushman Bharat), and relevant health legislations.
- **Scholarly Authorship:** Developed by experienced teachers and public health professionals affiliated with IAPSM, ensuring academic depth, contextual relevance, and pedagogical clarity.
- **Applied Learning Emphasis:** Concepts are reinforced through illustrations, flowcharts, case scenarios, and summary boxes to facilitate analytical thinking and real-world application.

Shelving Recommendation
COMMUNITY MEDICINE

ISBN 978-93-7806-251-3
9 789378 062513

JAYPEE BROTHERS