**KENYA MEDICAL TRAINING COLLEGE**

**HEALTH STATISTICS**

**Prepared by: Lect. Ephantus Maina Mwangi**

**BTh, HND Epidemiology, Dip. Health Records & Information Technology**

# Table of Contents

[Table of Contents i](#_Toc466978913)

[HEALTH STATISTICS 1](#_Toc466978914)

[SEMESTER I 2](#_Toc466978915)

[UNIT I (CHS 1.1) Introduction to Health Statistics and Data Collection 2](#_Toc466978916)

[Objectives 2](#_Toc466978917)

[Definition of Statistics 3](#_Toc466978918)

[Characteristics of Statistics 4](#_Toc466978919)

[Main Divisions of Statistics 5](#_Toc466978920)

[Functions of Statistics 6](#_Toc466978921)

[Steps in statistical experiment 6](#_Toc466978922)

[Value and Importance of Statistics 7](#_Toc466978923)

[Uses or importance of statistics 7](#_Toc466978924)

[Limitations 8](#_Toc466978925)

[Distrust of Statistics 8](#_Toc466978926)

[Types of Statistics 9](#_Toc466978927)

[Identification of statistical/research problem 9](#_Toc466978928)

[Objectives 9](#_Toc466978929)

[What is research? 9](#_Toc466978930)

[Definition of a Research Problem 10](#_Toc466978931)

[Why Research? 11](#_Toc466978932)

[Qualities of a Good Research Problem 11](#_Toc466978933)

[Source of a Research Problem 11](#_Toc466978934)

[Aspects of a Research Problem 13](#_Toc466978935)

[Considerations in Selecting a Research Problem 13](#_Toc466978936)

[Research Topic 14](#_Toc466978937)

[Characteristics of a Good Topic 14](#_Toc466978938)

[Problems of Topic Selection 14](#_Toc466978939)

[Problem Statement 15](#_Toc466978940)

[FORMULATION OF RESEARCH OBJECTIVES/HYPOTHESIS 16](#_Toc466978941)

[RESEARCH OBJECTIVES 17](#_Toc466978942)

[Why should research objectives be developed? 18](#_Toc466978943)

[How should you state your objectives? 18](#_Toc466978944)

[HYPOTHESIS 19](#_Toc466978945)

[Characteristics of a good hypothesis 19](#_Toc466978946)

[Definition of Literature Review 20](#_Toc466978947)

[Purposes of Literature Review 21](#_Toc466978948)

[Planning a Literature Search 21](#_Toc466978949)

[Some tips about conducting the literature review. 22](#_Toc466978950)

[Content of the Literature Review 23](#_Toc466978951)

[SEESS: (Relevance of Carrying out Literature Review) 24](#_Toc466978952)

[What is critical literature review? 24](#_Toc466978953)

[Research methods 25](#_Toc466978954)

[Terms related to research question 26](#_Toc466978955)

[What is a sample? 26](#_Toc466978956)

[Sampling 27](#_Toc466978957)

[What is a sampling frame? 27](#_Toc466978958)

[Sample Size 27](#_Toc466978959)

[Sampling unit or element 27](#_Toc466978960)

[Sampling interval 28](#_Toc466978961)

[Sampling Fraction 28](#_Toc466978962)

[What is a representative sample? 28](#_Toc466978963)

[What is the difference between a statistic and a parameter? 28](#_Toc466978964)

[Sampling Error 28](#_Toc466978965)

[Selection bias 29](#_Toc466978966)

[Statistical Inference: 29](#_Toc466978967)

[Strata: 29](#_Toc466978968)

[Structured Question 29](#_Toc466978969)

[SAMPLING TECHNIQUE 29](#_Toc466978970)

[Why sample? 30](#_Toc466978971)

[Sampling methods 30](#_Toc466978972)

[*(a) Probability sampling methods* 30](#_Toc466978973)

[1. Simple random sampling 31](#_Toc466978974)

[2. Stratified sampling 32](#_Toc466978975)

[3. Systematic sampling 33](#_Toc466978976)

[4. Cluster sampling 35](#_Toc466978977)

[Steps of carrying out cluster sampling 35](#_Toc466978978)

[Determining how many clusters to choose in a sample 36](#_Toc466978979)

[5. Multistage sampling 37](#_Toc466978980)

[*(b) Non-probability sampling* 38](#_Toc466978981)

[1. Quota sampling 38](#_Toc466978982)

[2. Convenience sampling 39](#_Toc466978983)

[3. Purposive Sampling or judgment samples 39](#_Toc466978984)

[4. Snowball sampling 40](#_Toc466978985)

[Advantages of probability sampling 40](#_Toc466978986)

[Disadvantages of probability sampling 40](#_Toc466978987)

[Advantages of non- probability sampling 41](#_Toc466978988)

[Disadvantages of non- probability sampling 41](#_Toc466978989)

[Conclusion 41](#_Toc466978990)

[Determining Sample Size 41](#_Toc466978991)

[ERRORS Vs MISTAKES 43](#_Toc466978992)

[**Sources of error** 43](#_Toc466978993)

[ Error of origin 43](#_Toc466978994)

[ Error of manipulation 43](#_Toc466978995)

[ Error of inadequacy 43](#_Toc466978996)

[**Kinds of error** 43](#_Toc466978997)

[ Sampling errors 43](#_Toc466978998)

[ Non sampling errors 43](#_Toc466978999)

[ Biased error 43](#_Toc466979000)

[ Unbiased error 43](#_Toc466979001)

[**Sampling Errors And Biases** 44](#_Toc466979002)

[**Non-sampling error** 44](#_Toc466979003)

[**Bias in Information Collection** 44](#_Toc466979004)

[**Possible sources of bias during data collection:** 45](#_Toc466979005)

[DATA COLLECTION 48](#_Toc466979006)

[Sources of health data 48](#_Toc466979007)

[Methods of Data Collection 49](#_Toc466979008)

[Questionnaires (self-administered /postal, researcher administered/telephone, web based) 50](#_Toc466979009)

[Design of postal questionnaires 51](#_Toc466979010)

[Interviews 59](#_Toc466979011)

[Types of interview 60](#_Toc466979012)

[Observation 69](#_Toc466979013)

[Focus Group Discussion 72](#_Toc466979014)

[**What are FGDs?** 72](#_Toc466979015)

[ The data which results from a focus group requires skill and experience to analyze 73](#_Toc466979016)

[**How to plan and prepare for focus groups** 73](#_Toc466979017)

[**Running focus groups** 73](#_Toc466979018)

[**Managing risks** 74](#_Toc466979019)

[Useful tips to encourage discussion 74](#_Toc466979020)

[Functions of the facilitator 75](#_Toc466979021)

[Functions of the recorder 77](#_Toc466979022)

[Number and duration of sessions 78](#_Toc466979023)

[Processing and analysis of results 78](#_Toc466979024)

[Report writing 79](#_Toc466979025)

[*Pilot Survey/study/Pilot – testing/Feasibility study* 79](#_Toc466979026)

[*Steps of conducting pilot survey* 81](#_Toc466979027)

[*Field work* 82](#_Toc466979028)

[*Editing and cleaning of data* 83](#_Toc466979029)

[*Coding* 83](#_Toc466979030)

[*Rules of coding* 84](#_Toc466979031)

[*Classification* 84](#_Toc466979032)

[Tabulation 84](#_Toc466979033)

[Common Stages of survey or Data Collection 84](#_Toc466979034)

[UNIT II (CHS 1.2) Data Presentation 87](#_Toc466979035)

[Definition of Tables 88](#_Toc466979036)

[Parts of a Table 88](#_Toc466979037)

[Types of Tables 89](#_Toc466979038)

[Simple and Complex tables 89](#_Toc466979039)

[General and Special Purpose Tables 90](#_Toc466979040)

[Principles of Table Construction 91](#_Toc466979041)

[Importance of Tabulation 91](#_Toc466979042)

[SEMESTER III 92](#_Toc466979043)

[UNIT III (CHS 1.3) Measures of Central Tendency 92](#_Toc466979044)

[Data 92](#_Toc466979045)

[Types of data 92](#_Toc466979046)

[Categorical Data 92](#_Toc466979047)

[Frequency Distribution 96](#_Toc466979048)

[MEASURES OF CENTRAL TENDENCY 97](#_Toc466979049)

[Formula for the arithmetic mean: 98](#_Toc466979050)

[**Finding the median where there are an even number of values** 100](#_Toc466979051)

[**USAGES** 111](#_Toc466979052)

[How to Group Data 111](#_Toc466979053)

[The Two Firm Rules in Grouping Data: 112](#_Toc466979054)

[The Means, Medians and Modes of Grouped Data 115](#_Toc466979055)

[Arithmetic Mean of Grouped Data 116](#_Toc466979056)

[The Median of Grouped Data 118](#_Toc466979057)

[**Median for grouped data (another symbol used)** 119](#_Toc466979058)

[UNIT IV (CHS 1.4) Measures of Dispersion 130](#_Toc466979059)

[Measures of Dispersion 130](#_Toc466979060)

[**Normal or Gaussian distribution** 147](#_Toc466979061)

[**Coefficient of variation** 148](#_Toc466979062)

[**Various formulas** 153](#_Toc466979063)

[UNIT V (CHS 1.5) Standard Hospital Statistics and Formulae 158](#_Toc466979064)

[HEALTH ADMINISTRATIVE STATISTICS (IN-PATIENT STATISTICS) 158](#_Toc466979065)

[The Daily Bed Return 159](#_Toc466979066)

[SEMESTER IV 170](#_Toc466979067)

[UNIT VI (CHS 1.6) Vital and Health Statistics 170](#_Toc466979068)

[References i](#_Toc466979069)

# HEALTH STATISTICS

At the end of the module the learner will be able to:

1. Describe the importance and value of health statistics
2. Identify health problems requiring investigation
3. Describe research problem clearly
4. Demonstrate ability to carry out research using appropriate research methods
5. Demonstrate ability to prepare schedules and questionnaires required in obtaining information
6. Describe different sampling methods
7. Demonstrate ability to use various sampling methods
8. Define pilot survey
9. Describe methods of conducting pilot survey
10. Calculation measures of central tendency and dispersion
11. Demonstrate ability to classify and tabulate data
12. Present information in a narrative form
13. Present information in graphs and charts
14. Demonstrate ability to interpret analyzed data
15. Report research findings succinctly
16. Demonstrate ability edit information
17. Show concern for confidentiality of health information
18. Define standard formulae used in calculating hospital administrative statistics
19. Utilize standard formulae to compute the various statistics
20. Describe various vital and health statistics rates
21. Demonstrate ability to calculate vital and health statistics
22. Demonstrate ability to interpret vital and health statistics
23. Display accuracy, and thoroughness in calculating, analyzing and interpreting medical statistics
24. Be innovative and creative in reporting vital and health statistics
25. Appreciate the importance of health statistics

Describe ability to store and retrieve health information

# SEMESTER I

# UNIT I (CHS 1.1) Introduction to Health Statistics and Data Collection

## Objectives

1. Definition of statistics,
2. Outline characteristics of statistics
3. Main divisions of statistics
4. Functions of statistics
5. Steps in statistical experience
6. Identify value and importance
7. Uses or importance of statistics
8. Limitations of statistics
9. Distrusts of statistics
10. Types of statistics
11. Identification of a statistical/research problem
12. Definition of a problem
13. Review of literature
14. Sources of information
15. Critical review of literature
16. Need for the study
17. Formulation of objective/hypothesis of the study
18. Identification of the population from which the problems pertains
19. Review relevant medical terms related to research problems
20. Definition and description of the population
21. Research methods
22. Methods of data collection
23. Direct observation
24. Interviewing
25. Obstruiction from published statistics
26. Preparation of schedules and questionnaires
27. Sampling methods
28. Definition of a sample
29. Random samples
30. Systematic sampling methods
31. Stratified sampling methods
32. Multi-stage sampling
33. Cluster sampling
34. Quota sampling
35. Pilot survey
36. Definition of a pilot survey
37. Purpose of a pilot survey
38. Procedure of conducting the sampling surveys
39. Editing information
40. Definition – purpose, importance, procedures
41. Different methods of editing information
42. Classification of information
43. Definition and description of different methods of data classification
44. Purpose, importance and procedures

## Definition of Statistics

W.I. King “the science of statistics is the method of judging collection, of natural or social phenomena from the results obtained from the analysis or enumeration or collection of estimates”. This definition is close to the modern statistics. But it does not cover the entire scope of modern statistics. It is a branch of mathematics dealing with gathering, analyzing, and making inferences from data.

Although there are many definitions the agreed definition is that statistics is a scientific approach of information presetting itself in numerical form which enables us to maximize our understanding of such information. It is delivered from numerical figures which result from statistical analysis, also referred to as statistics.

Statistics is therefore the branch of scientific method comprising of collection, presentation, analysis and interpretation of data which are obtained by measuring some characteristics. It entails analyzing data which has been collected for decision making and taking appropriate action. Thus we can say that it is concerned with collection, organization, analyses, interpretation and generalization of data. It is important to note that statistical data is numerical figures which are effect of causes only. For instance a single train accident is not a statistical data, but the total number of train accidents during a year constitutes the statistical data.[[1]](#footnote-1)

Statistics is also the mathematics of the collection, organization, and interpretation of numerical data. It is also a branch of applied science concerned with scientific methods for collecting, organizing, summarizing, and analyzing data. Statistics therefore involves both numbers and the techniques and procedures followed in collecting, organizing, analyzing, interpreting and presenting information in numerical form. Having looked at the definition of statistics we need to understand what data is.[[2]](#footnote-2)

## Characteristics of Statistics

1. Statistics are aggregates of facts meaning a single figure pertaining to age, birth, death, production, employment does not constitute statistics because they are not comparable or related but a series of ages, births, deaths, production and employment are called statistics because they are related and comparable. It is also possible to study them in relation to time, place, and frequency of occurrence. For example the number of deaths in the Narok accident is 47 in 2013 though a numerical statement of facts is not statistics. But this number is higher compared to the 2013 Bungoma accident where 20 students died. This is called statistics because is related and comparable.
2. Statistics are affected to a marked extent by multiplicity of causes meaning statistical facts and figures are not generally traceable to a single cause. They are affected to a considerable extent by a number of factors working together. For example the cause of accidents in Kenya cannot be pegged on complacency of police but rather a number of factors working together such as poor roads, reckless driving and many other factors.
3. Statistics are numerically expressed in the sense that they are expressed in numbers. All statistics are numerical statements of facts but numerical statements of facts are not statistics.
4. Statistics are enumerated or estimated according to reasonable standards of accuracy meaning statistical figures and facts can be estimated by expert enumerators, hence estimated figures cannot be absolutely precise or accurate. The degree of accuracy largely depends on the purpose for statistics collected and nature of study.
5. Statistics are collected in a systematic way so that they may confirm to some reasonable standards of accuracy.
6. Statistics are collected for a predetermined purpose in the sense that the purpose should be decided in advance. The purpose should be well defined.
7. Statistics are placed in relation to each other/are comparable/homogenous. For instance the age of husbands should be compared to the corresponding ages of their wives and not trees.[[3]](#footnote-3)

## Main Divisions of Statistics

1. Descriptive Statistics – involve the use of statistical methods. Statistical methods are concerned with the formulation of general rules and principles which are commonly used in collection, processing, presentation, analysis and interpretation of data under inquiry. Helps in summarizing and describing the main features of the data.
2. Applied Statistics – deal with the application of these rules and principles to concrete problems such as population, unemployment, poverty, national income, price rise, trade etc[[4]](#footnote-4).

## Functions of Statistics

1. Statistics simplifies the complexity of mass of data in that human mind is unable to absorb a mass of complicated facts and figures in a stretch.
2. Statistics enlarge individual human experience. This is achieved by exercising one’s best ability and power of judgment to view the quantitative significance of a phenomenon such a desire to understand the cause of rise in cost of living in Kenya.
3. Statistics facilities the comparison of data by comparing the simplified data and measure their relationships.
4. Statistics helps in formulating and testing hypothesis. Hypotheses are tentative statements showing facts and figures being tested.
5. Statistics tests the laws of physical and social science.
6. Statistics helps in the formulation of policies since no decision making process in an organization or country can start without scientific analysis of statistical data forms.
7. Statistics helps in forecasting future events by making use of past and present data to predict of future trends of a phenomenon.
8. Statistics reveals economic and business trends and tendencies
9. Statistics helps in realization of magnitudes of situations.
10. Statistics presents facts in a definite form[[5]](#footnote-5).

## Steps in statistical experiment

1. Data collection
2. Tabulation – comprises of organization and presentation
3. Organization begins with editing data so that omissions, inconsistencies, irrelevant answers and wrong computations in the returns from the survey may be corrected or adjusted.
4. Presentation – in order to facilitate statistical analysis data needs to be presented in an orderly manner in form of graphs and diagrams.
5. Statistical inference which means applying statistical methods on the tabulated data to draw conclusions. Entails analysis and interpretation of data.
6. Analysis means digging out information useful for decision making.
7. Interpretation means drawing conclusions from data collected and analysed.[[6]](#footnote-6)

## Value and Importance of Statistics

Mastery of statistics will help the learner to:

1. Read and understand scientific and professional literature
2. Understand advance causes
3. Acquire professional skills
4. Undertake research[[7]](#footnote-7)

## Uses or importance of statistics

1. Routine administration is guided by statistical data compiled by the department of statistics such as Central Bureau of Statistics.
2. Planning of annual activities and longterm activities in an institution (development plans are based on statistical analyses of compiled data.
3. In politics aspirants justify their political programs and endevour to mobilize popular support at election and allocation of resources by means of reference to selective statistical data.
4. Business and industry rely on market research which are based on surveys that provide extra “market intelligence”. Data obtained in surveys proves to be more accurate than figures based on experience and opinions.
5. In keeping stocks or inventory of items for sale or use inventory control is needful.
6. Institutions operate on budgets which are a product of compiled data in the institution from past experience.
7. Quality control requires data so as to set standards upon which mass production is to derive quality of the desired output.
8. Statistics is used in science and research for improvement of health service delivery in the country.[[8]](#footnote-8)

## Limitations

1. Does not deal with individual measurements since statistics deals with aggregate of facts, thus the study of individual measurements lies outside the scope of statistics.
2. Deals only with quantitative characteristics since statistics are numerical statements of facts. This means characteristics that cannot be expressed in numbers are incapable of statistical analysis, qualitative characteristics like honesty, efficiency, intelligence, blindness
3. Statistical results are true only on an average conclusions obtained statistically; are not universally true. They are true only under certain conditions.
4. Statistics is only one of the methods of studying a problem. Statistical tools do not provide the best solution to a problem. Statistical tools do not provide the best solution under all circumstances
5. Statistics can be misused.

## Distrust of Statistics

Statistics are like a miniskirt; they cover up the essentials but give you the ideas. The main reasons of mistrust are:

1. Figures are convincing and therefore, people are easily led to believe them
2. They can be manipulated in such a manner as to establish foregone conclusions
3. Even if correct figures are used they may be presented in such a manner the reader is misled.

## Types of Statistics

1. Descriptive which is used to describe a whole group or population generally by averaging or using central points such as mean, median, mode
2. Inferential statistics which is used to induce certain characteristics from a small group to generalize the whole group.
3. Correlation statistics is used to show that situations or relationships or associations exist such as improvement of sanitation can lead to less diseases.
4. Predictive statistics is used to predict the kind of associations or relationships that exist[[9]](#footnote-9)

## Identification of statistical/research problem

## Objectives

1. Define a research problem
2. Describe sources and aspects of a research problem
3. Describe considerations of a research problem
4. Define review of literature
5. Describe sources of information
6. Describe critical review of literature
7. Describe Need for the study
8. Describe formulation of objective/hypothesis of the study

## What is research?

* "Systematic inquiry or a quest for knowledge characterized by disciplined inquiry”.
* Exhaustive study, investigation or experimentation following some logical sequence.
* Continued search for new knowledge and understanding of the world around us
* A process of arriving at effective solution to problems through systematic collection, analysis and interpretation of data

## Definition of a Research Problem

Boldly speaking any question that you want to answer and any assumption or assertion that you want to challenge or investigate can become a research problem or a research topic for your study. However, it is important to note that not all questions can be transformed into research problems and some may prove to be extremely difficult to study. Formulation of a research problem requires considerable knowledge of both the subject area and research methodology.[[10]](#footnote-10)

Research can also be defined as a systematic and organized effort to investigate a specific problem encountered in a community or an organization that needs a solution. It is organized, systematic, data based, objective, rigorous analysis, scientific enquiry into a specific problem carried out with a purpose of finding a solution or answer to a problem. Research can also be defined as systematic, controlled, empirical and critical investigation of hypothesized preposition about the presumed relationship among natural phenomenon or different variables. For that reason, there are basically two types of research based on purpose namely:

1. **Applied Research (Action Oriented Research)** – deals with concrete or specific problems currently being experienced in a society or organization e.g. severe outbreak of disease in a community or poor performance of an organization. Action oriented solves immediate and practical problems using scientific methods to give immediate solutions. It comprises of 5 phases:
2. Problem identification
3. Consideration of alternatives
4. Action taking
5. Evaluation of the consequences of the action
6. Lessons learnt
7. **Basic Research also called Pure Research or Fundamental Research** – carried out to enhance the understanding of a certain problem, or to enhance understanding of existing theory. An example of this kind of research is research done in chemistry, agriculture etc. whose findings contribute to the building of existing pool of knowledge. It does not require only solving problems but focuses on increase of knowledge, addressing theoretical dilemma, academic knowledge and experiments.

## Why Research?

1. Short term reasons:
   1. Add new knowledge to existing pool of knowledge
   2. Identify and solve problems
   3. Identify factors which influence a certain situation
   4. To take calculative risks in decision making
   5. To influence policy
   6. To present a personal interest
   7. To develop new research methodology
   8. To create new methods of doing things, practices, ideas and products
2. Long term reasons:
   1. Confirm existing theories
   2. Testing hypothesis and theories
   3. Generating and advancing objective and subjective knowledge
   4. Solving long term problems

## Qualities of a Good Research Problem

1. Should be concise and clearly stated
2. Should impact on the topic and objectives being investigated
3. Should be researchable in terms of data collection and empirical analysis
4. Should be supported by theoretical or observed recent evidence
5. Should be simple and objective language devoid of comical, poetic, emotional or intimidating language

## Source of a Research Problem

Most research in humanities revolve around four Ps

1. People
2. Problems
3. Programmes
4. Phenomena

## Aspects of a Research Problem

|  |  |  |  |
| --- | --- | --- | --- |
| **Aspects of Study** | **About** | **Study of** |  |
| Study Population | People | Individuals, organizations, groups, communities | They provide you with the relevant information or you collect information from or about them |
| Subject Area | Problem  Programme  Phenomenon | Issues, situations, associations, needs, population composition, profiles etc.  Contents, structure, outcome, attributes, satisfaction, consumers, providers etc.  Cause and effect, relationships, the study of the phenomenon itself, etc. | Information that you need to collect to find answers to your service research questions |

## Considerations in Selecting a Research Problem

When selecting a research problem/topic there are a number of considerations to keep in mind which will help to ensure that your study will be manageable and that you remain motivated. They are:

1. Interest because research endeavor is usually time consuming, involves hard work, and possibly unforeseen problems, hence to sustain the required motivation and put enough time and energy to complete it must be something of interest to you.
2. Magnitude meaning you should have sufficient knowledge about the research process to be able to visualize the work involved in the completing the proposed study. Narrow the topic to something manageable, specific and clear. Always select a topic that you can manage within the time and with the resources at your disposal.
3. Measurement of concepts meaning you should be clear about indicators and measurements involved. For example if you plan to measure the effectiveness of a health promotion programme, you must be clear as to what determines effectiveness and how it will be measured.[[11]](#footnote-11)

## Research Topic

Is the focus of the study and its process which entails:

1. Identify area of interest in a study
2. Identify key words in the topic that would guide the study itself
3. Define the study by analyzing the key words. This assists in identifying the materials and variables in the study area

## Characteristics of a Good Topic

1. Should be researchable i.e. easy to formulate instruments, identify population and measurable objectives (variables)
2. Brief, straight forward
3. Should stimulate the interest of the researcher
4. Should contribute to the existing pool of knowledge
5. Should be clear and focused
6. Should be subject to various views and interpretations i.e. should provoke the researcher and the audience

## Problems of Topic Selection

1. Choosing a topic which is too wide and complex
2. Choosing a topic with limited access to research materials
3. Choosing a topic which you might have a failure to identify respondents
4. Choosing a topic with limited access to research materials
5. Choosing a topic with limited time to collect necessary data
6. Sailing in unfamiliar waters
7. Writing a wordy and bombastic topic

## Problem Statement

The next step after topic selection is the statement of the problem. This provides the context for your research. The problem statement is a specific statement that clearly conveys the scope, magnitude and purpose of the research study*.* It defines clearly the problem you propose to examine. It explains why it is important to carry out the research.

A brief review of literature will be required in delineating and defining the research problem. A well-defined research problem statement leads naturally to the statement of

* 1. Research objectives,
  2. Hypotheses;
  3. Definition of key variables, and
  4. Selection of a methodology for measuring the variables

A poorly defined research problem leads to confusion.

## FORMULATION OF RESEARCH OBJECTIVES/HYPOTHESIS

**NB:** Development of a research process is a cyclical process. The double-headed arrows indicate that the process is never linear.

****

## RESEARCH OBJECTIVES

The **objectives** of a research project summarize what is to be achieved by the study.

Objectives should be closely related to the statement of the problem. For example, if the problem identified is low utilization of child welfare clinics, the general objective of the study could be to identify the reasons for this low utilization, in order to find solutions.

The **general objective** of a study states what researchers expect to achieve by the study in general terms. It is possible (and advisable) to break down a general objective into smaller, logically connected parts. These are normally referred to as **specific objectives.**

Specific objectives should systematically address the various aspects of the problem as defined under ‘Statement of the Problem’ and the key factors that are assumed to influence or cause the problem. They should specify **what** you will do in your study, **where** and **for what purpose**.

For a study into the cost and quality of home-based care for HIV/AIDS patients and their communities in Msambweni, may have the below statement as its general objective:

To explore to what extent community home-based care (CHBC) projects in Msambweni provide adequate, affordable and sustainable care of good quality to people with HIV/AIDS, and to identify ways in which these services can be improved.

It may be split up in the following specific objectives:

* 1. To identify the full range of economic, psychosocial, health/nursing care and other needs of patients and their families affected by AIDS.
  2. To determine the extent to which formal and informal support systems address these needs from the viewpoint of service providers as well as patients.
  3. To determine the economic costs of CHBC to the patient and family as well as to the formal CHBC programmes themselves.
  4. To relate the calculated costs to the quality of care provided to the patient by the family and to the family/patient by the CHBC programme.

## Why should research objectives be developed?

The formulation of objectives will help you to:

* **Focus** the study (narrowing it down to essentials);
* **Avoid** the collection of data which are not strictly necessary for understanding and solving the problem you have identified; and
* **Organize** the study in clearly defined parts or phases. Properly formulated, specific objectives will facilitate the development of your research methodology and will help to orient the collection, analysis, interpretation and utilization of data.

## How should you state your objectives?

Take care that the objectives of your study:

1. Cover the different aspects of the problem and its contributing factors in a **coherent** way and in a **logical sequence**;
2. Are **clearly phrased** in **operational terms,** specifying exactly what you are going to do, where, and for what purpose;
3. Are **realistic** considering local conditions; and
4. Use **action verbs** that are specific enough to be evaluated.

**Examples** of action verbs are: to determine, to compare, to verify, to calculate, to describe, and to establish. Avoid the use of vague non-action verbs such as: to appreciate, to understand, or to study. Keep in mind that when the project is evaluated, the results will be compared to the objectives. If the objectives have not been spelled out clearly, the project cannot be evaluated.

Using the previous example on cost and quality of CHBC, we may develop more specific **research questions** for the different objectives, such as:

* 1. Do rural and urban CHBC projects differ with respect to the adequacy, quality, affordability and sustainability of HBC provided?
  2. How satisfied are AIDS patients, relatives and service providers with the care provided? Are there differences in perceptions between those groups?
  3. Is the stigma attached to being HIV+ the same strong for women as for men? Or are there gender differences in stigma?
  4. What impact does the care provided to AIDS patients have on the economy of the homestead? Is there competition with other basic needs (e.g. schooling of children, purchases of food)?

## HYPOTHESIS

While a research problem consists of an unanswered question, a research hypothesis is a proposition, which is assumed to offer a possible and reasonable solution to the problem (a tentative answer). Based on your experience with the study problem, it might be possible to develop **explanations** for the problem, which can then be **tested**. If so, you can formulate hypotheses in addition to the study objectives.

A **hypothesis** is a prediction of a relationship between one or more factors and the problem under study that can be tested. In our example concerning the cost and quality of HBC in Msambweni it would have been possible to formulate and test the following hypotheses:

* 1. The role of first-line relatives in the provision of care to AIDS patients is more substantial in rural than in urban areas.
  2. The silence and stigma surrounding AIDS makes the formation of self-help groups of AIDS patients and their relatives next to impossible, which in turn maintains the high level of stigma on HIV/AIDS.

**N.B** Not all research problems may have a formally stated hypothesis. In studies, which are essentially exploratory or in an area of knowledge that has very little previous research, it might not be possible to formulate any reasonable hypothesis. In such situations, the researcher may omit the hypothesis and the study is guided by the stated objectives (Mugenda and Mugenda, 1999; Leedy, 1980).

## Characteristics of a good hypothesis

1. It should be capable of being expressed as a question e.g. Does improved access to potable water contribute to reduction of water borne diseases
2. It may stand in a negative way e.g. slum clearance does not lead to increased urban crime
3. Should be capable of being answered a ‘yes’ and ‘no’ i.e. capable of being accepted, rejected or not rejected.

## Definition of Literature Review

Literature review is a systematic examination of knowledge available on a topic. It is an examination of scholarly information and research-based information on a specific topic. In other words, it’s a review of what’s known, not suspected or assumed about a specific subject. Its goal is to create a complete, accurate representation of knowledge and research-based theory available on a topic. Literature review is an integral part of research process, in that it provides a platform to gain an understanding of what is and isn’t known about the reality of a situation, event, or circumstance[[12]](#footnote-12).

As an academic task literature review is where you show that you are both aware of and can interpret what is already known and where eventually you will be able to point out contradictions and gaps in existing knowledge. This means you will have to explain why your review is important, why is it different and what it adds to knowledge. In research we seek to be original and to make an original contribution to knowledge. Hence it is creating a new dimension or fresh perspective that makes a distinct contribution.

A literature review is ***an account*** of what has been published on a topic by ***accredited scholars and researchers.***

A **literature review** is a body of text that aims to review the critical points of current knowledge including substantive findings as well as theoretical and methodological contributions to a particular topic. There are many reasons why literature review is rendered as a significant part of any **research** or dissertation paper.

## Purposes of Literature Review

Literature review serves the following purposes.

1. Distinguishing what has been done from what needs to be done
2. Discovering relevant variables relevant to the topic
3. Synthesizing and gaining a new perspective
4. Identifying relationships between ideas and practice
5. Establishing the context of the topic or problem
6. Rationalizing the significance of the problem
7. Enhancing and acquiring the subject vocabulary
8. Understanding the structure of the subject
9. Relating ideas and theories to applications
10. Identifying the main methodologies and research techniques that have been used
11. Placing the research in a historical context to show familiarity with state-of-the-art developments

## Planning a Literature Search

1. Define the topic – this should begin with general reading to familiarize yourself with the topic. Take note of the concepts, terms and authors used and begin thinking of the way to shape your topic in the future.
2. Think about the scope of the topic. Ask questions about which language or languages it might be necessary to search, what time frame (how far back you might need to search, and what subject areas might be relevant). Make a list of terms and phrases you need to search normally known as search vocabulary.
3. Think about outcomes – your proposal for your search will have stated an aim pertinent to the search and review of literature. Think therefore of what is it that you want to get out of the search and why you are undertaking the search in the first place.
4. Plan about the housekeeping – design a means by which you will record what you find and how you will cross-reference materials. It is important to keep consistent records of both what you have searched and how you have searched it. This is because you may need to go back to undertake a further search of the same sources using different terms like your search requiring to be written as part of the methods by which you did your research.
5. Plan the sources to be searched – prepare a list of likely relevant resources of information such as encyclopedias, OPAC and indexes. An interview with the subject librarian can be very useful in guiding you to the relevant materials.
6. Search the sources listed – work through the list of sources you have made. Start with general sources as you move to abstract and indexes. Be systematic and thorough, making consistent references as you go along. Make notes on possible further leads and ideas to be followed up.

## Some tips about conducting the literature review.

First, **concentrate your efforts on the *scientific* literature**. Try to determine what the most credible research journals are in your topical area and start with those. Put the greatest emphasis on research journals.

1. You might be able to find a study that is quite similar to the one you are thinking of doing. Since all credible research studies have to review the literature themselves, you can check their literature review to get a quick-start on your own.
2. Check whether someone else have carried out a similar investigation and where such has been done, the findings about the problem
3. Prior research will help assure that you include all of the major relevant constructs in your study. You may find that other similar studies routinely look at an outcome that you might not have included. If you did your study without that construct, it would not be judged credible if it ignored a major construct.
4. Literature review will help you to find and select appropriate measurement instruments. You will readily see what measurement instruments researchers use themselves in contexts similar to yours.
5. Literature review will help you to anticipate common problems in your research context. You can use the prior experiences of other to avoid common traps and pitfalls.
6. It gives readers easy access to research on a particular topic by selecting high quality articles or studies that are relevant, meaningful, important and valid and summarizing them into one complete report.
7. It provides an excellent starting point for researchers beginning to do research in a new area by forcing them to summarize, evaluate, and compare original research in that specific area.
8. It ensures that researchers do not duplicate work that has already been done.
9. It can provide clues as to where future research is heading or recommend areas on which to focus.
10. It highlights key findings.
11. It identifies inconsistencies, gaps and contradictions in the literature.
12. It provides a constructive analysis of the methodologies and approaches of other researchers.
13. Whether the findings existing can be used to solve the current problem
14. Have more knowledge about the problem

**Sources** of literature review includes

1. Scientific journals
2. Research papers
3. Books
4. Records
5. Internet
6. Abstracts
7. Thesis/dissertations
8. Conference papers
9. Periodicals
10. Archives
11. Reports
12. Interviews
13. Observations
14. MOH. NGOs etc

A well-structured literature review is characterized by a logical flow of ideas; current and relevant references with consistency, appropriate referencing style; proper use of terminology; and an unbiased and comprehensive view of the previous research on the topic.

## Content of the Literature Review

**Introduction -** The introduction explains the focus and establishes the importance of the subject. It discusses what kind of work has been done on the topic and identifies any controversies within the field or any recent research which has raised questions about earlier assumptions.

**Body -** Often divided by headings/subheadings, the body summarizes and evaluates the current state of knowledge in the field. It notes major themes or topics, the most important trends, and any findings about which researchers agree or disagree.

**Conclusion** - The conclusion summarizes all the evidence presented and shows its significance. If the review is an introduction to your own research, it highlights gaps and indicates how previous research leads to your own research project and chosen methodology.

## SEESS: (Relevance of Carrying out Literature Review)

1. Summarize relevant research,
2. Evaluate research,
3. Establish relationships between different work/research,
4. Show connections between the research and your work
5. Select parts of research – methodology & findings (compare and contrast with your own work, make connections to your own work).

If the finding of the literature review does not suggest there is solution to the current problem that one has identified, then investigation is required. Having determined that an investigation has to be done and needs money, one makes the ***next step of proposal writing*** and submitting it to a donor/partner for support.

The following should be included in the **proposal**:-

1. Statement of the problem (clear, who is the target)
2. Objectives of the study
3. Justification of the study (why the study is necessary/important)
4. Details of critical literature review
5. Study design/methodology
6. How data analysis and interpretation will be done
7. Budget
8. Bibliography/references

## What is critical literature review?

Research is made in order to **inform** people with **new** knowledge or discovery. However, it is not to be expected that everybody would willingly believe what you are tackling in your whole research. Thus, what you can do to make your research more credible will be to support them with other works which have spoken about the same topic that you have for your research. This is where literature review comes in.

One of the most important early steps in a research project is the conducting of the literature review. This is also one of the most humbling experiences you're likely to have. Why? Because you're likely to find out that just about any worthwhile idea you will have has been thought of before, at least to some degree. A literature review is designed to identify related research, to set the current research project within a conceptual and theoretical context.

1. Identification of the population from which the problems pertains
2. Review relevant medical terms related to research problems
3. Definition and description of the population
4. Research methods
5. Methods of data collection

## Research methods

Research methods are conventionally divided into quantitative, qualitative and participatory each with differing underlying approaches, tools and techniques.

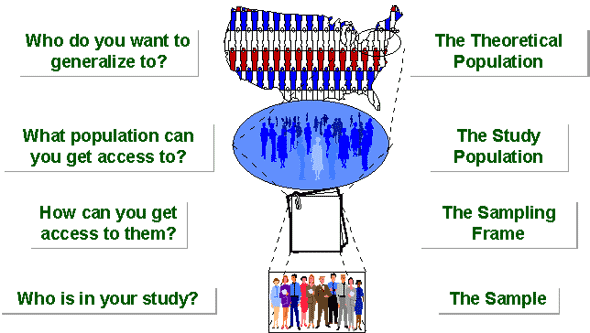
In development research quantitative methods has typically been the main focus, with qualitative and participatory methods often relegated to desirable 'frills'. This is partly because of the overwhelming emphasis in many development agencies concern with quantification. Pressures for quantification have further intensified to demonstrate progress on Millennium Development Goals, 'scaling up' of impacts and macro-level change.

## Terms related to research question

**Population** is a group of individual persons, objects, or items from which samples are taken for measurement for example a population of presidents or professors, books or students.

**Population** is the entire finite or infinite aggregate of individuals or items from which samples are drawn. A collection of items of interest in research. The population represents a group that you wish to generalize your research to.

Probably you should make a distinction between the population you would like to generalize to, and the population that will be accessible to you. We'll call the former the **theoretical population** and the latter the **accessible population**.



## What is a sample?

A group that is selected to study as a representative of the true population for a given experiment. A sample is a finite part of a statistical population whose properties are studied to gain information about the whole (Webster, 1985). When dealing with people, it can be defined as a set of respondents (people) selected from a larger population for the purpose of a survey.

## Sampling

Sampling is the act, process, or technique of selecting a suitable sample, or a representative part of a population for the purpose of determining parameters or characteristics of the whole population. Sampling is the process of selecting participants for a research project. A process using a segment (sample) of a population to represent the entire population’s activities, attitudes, opinions, and interests and the results from the sample study can be inferred upon the population.

## What is a sampling frame?

Sampling frame (synonyms: **"survey frame"**) is the actual set of units from which a sample has been drawn. In the ideal case, the sampling frame should coincide with the population of interest.

It’s the list of all the elements or all the people in a population.  Examples of sampling frames are phone books, college student directories, directories of members of an association, a list of all the teachers in your county, etc.

The description of the characteristics that define a particular population from which a sample is taken.

Even though the set of maps of enumeration areas is not a list of individuals in the population, it is still considered a sample frame. In that case, however, it is a sample frame of individuals that reside in housing units, not of the total population. Any individual, who does not live in a housing unit, for example, a homeless person, is not covered by the sample frame.

## Sample Size

Number of sample units to be included in the sample.

## Sampling unit or element

Those units that qualify to be sampled and are available during the sampling process to be selected.

This refers to a single thing selected for inclusion in a research project. For example if you sample students from a college, one student would be your sample unit or element.

If all the workers in a factory make a population, a single worker is a unit of the population. If all the factories in a country are being studied for some purpose, a single factory is a unit of the population of factories. The sampling frame contains all the units of the population. It is to be defined clearly as to which units are to be included in the frame.

**Starting Point** - The initial number chosen when utilizing an Nth sampling interval.

## Sampling interval

Sampling interval is the process of taking a list, and selecting each nth unit as a participant in a study. The interval creates random selection throughout the population and is decided upon by dividing the total population by the number of sample units desired. Also referred to as interval or Nth selection.

## Sampling Fraction

The ratio comparing sample size to population size.

## What is a representative sample?

A representative sample is a sample that resembles the total population. It is important to use a sampling method that produces representative samples when your goal is to understand the characteristics of a population based on study of a sample (i.e., when you want to directly generalize from your sample to your population).

## What is the difference between a statistic and a parameter?

A calculated numerical quantity derived from the number of observations and occurrences in a sample. A parameter is a numerical characteristic of a population.

## Sampling Error

An assumed inaccuracy associated with using the sample results as an indication of the behavior of the population.

**Secondary data** - Data that was collected previously and not for the particular study at hand.

## Selection bias

A bias that is present when choosing between the test group and the control group due to logical differences in the units.

## Statistical Inference:

Making conclusions about a population from the results of a sample.

## Strata:

A population segment based on stated characteristics.

## Structured Question

A questionnaire that already includes fixed answers. The interviewer reads the question and answers and records which answer the respondent selected.

## SAMPLING TECHNIQUE

Collecting data is time consuming and expensive, even for relatively small amounts of data. Hence, it is highly unlikely that a complete population will be investigated. Because of the time and cost elements the amount of data you collect will be limited and the number of people or organizations you contact will be small in number. You will, therefore, have to take a sample.

**Sampling** is that part of statistical practice concerned with the selection of an unbiased or random subset of individual observations within a population of individuals intended to yield some knowledge about the population of concern, especially for the purposes of making predictions based on statistical inference. Sampling is an important aspect of data collection.

Sampling theory says a correctly taken sample of an appropriate size will yield results that can be applied to the population as a whole. There is a lot in this statement but the two fundamental questions to ensure generalization are:

1. How is a sample taken correctly?
2. How big should the sample be?

The answer to the second question is ‘as large as possible given the circumstances’. It is like answering the question ‘How long is a piece of string’? It all depends on the circumstances. The theory of sampling is based on random samples – where all items in the population have the same chance of being selected as sample units.

## Why sample?

1. Sampling reduces the time of a study
2. Sampling cuts research cost
3. Sampling allows for better supervision, recordkeeping and training of researchers
4. Sampling produces more accurate results

## Sampling methods

## *(a) Probability sampling methods*

Probability sampling is a sampling technique wherein the samples are gathered in a process that gives all the individuals in the population equal chances of being selected (Joan Joseph Castillo, 2009). Probability sampling methods are based on ***probability theory***, a mathematical concept based on accepted statistical principles that refers to the ability to predict the statistical likelihood that a random event will occur. Probability sampling methods require that: Every program element has a chance of being chosen (note: elements do not have to have an equal chance, only a chance), and it is possible to calculate the probability of each element being selected, and a random chance determines which elements are chosen. Examples of probability sampling are:

1. Simple random sampling
2. Stratified random sampling
3. Systematic random sampling
4. Cluster random sampling
5. Mixed/multi-stage random sampling

The distinguishing characteristic of probability samples is that the researchers can specify for each unit in the population the likelihood that will be included in the sample. Probability sampling therefore uses randomization process of element selection to reduce or eliminate sampling bias.

## Simple random sampling

Is the basic probability sampling design. SRS is one in which every member in the population has an equal and independent chance of being selected.

In a simple random sample ('SRS') of a given size, all such subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection: the frame is not subdivided or partitioned. Furthermore, any given *pair* of elements has the same chance of selection as any other such pair (and similarly for triples, and so on). This minimizes bias and simplifies analysis of **results**. In particular, the variance between individual results within the sample is a good indicator of variance in the overall population, which makes it relatively easy to estimate the accuracy of results.

However, SRS can be vulnerable to sampling error because the randomness of the selection may result in a sample that doesn't reflect the makeup of the population. For instance, a simple random sample of ten people from a given county will *on average* produce five men and five women, but any given trial is likely to over represent one sex and under represent the other. Systematic and stratified techniques, discussed below, attempt to overcome this problem by using information about the population to choose a more representative sample.

SRS may also be cumbersome and tedious when sampling from an unusually large target population. In some cases, investigators are interested in research questions specific to subgroups of the population. For example, researchers might be interested in examining whether cognitive ability as a predictor of job performance is equally applicable across racial groups. SRS cannot accommodate the needs of researchers in this situation because it does not provide sub-samples of the population. Stratified sampling, which is discussed below, addresses this weakness of SRS.

Simple random number sampling is used as the basis for many other sampling methods, but has three disadvantages;

* A sampling frame is required. This may not be available, exist or be incomplete.
* The procedure is unbiased but the sample may be biased. For instance, if the 90 people are a mixture of men and women and all men were selected this would be a biased sample.
* If population is large randomization process might be quite time consuming

To overcome this problem a stratified sample can be taken.

## Stratified sampling

A stratified sample is obtained by independently selecting a separate simple random sample from each population stratum/strata/homogeneous subgroups. A population can be divided into different groups may be based on some characteristic or variable. For example you may categorize ten years of education into groups, group A, between 10 and 20 group B and between 20 and 30 group C. These groups are referred to as strata. You can then randomly select from each stratum a given number of units which may be based on proportion like if group A has 100 persons while group B has 50, and C has 30 you may decide you will take 10% of each. So you end up with 10 from group A, 5 from group B and 3 from group C.

Where the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected.

Stratified sampling involves the following steps

1. Identify the population
2. Define criterion for stratification
3. List the population according to the defined strata or subgroups
4. Determine the required sample size and appropriate representation in each stratum. This can be proportionate or equal
5. Select using SRM or systematic the appropriate subject for each stratum

A stratified sampling approach is most effective when three conditions are met

1. Variability within strata are minimized
2. Variability between strata are maximized
3. The variables upon which the population is stratified are strongly correlated with the desired dependent variable.

**Advantages** over other sampling methods

1. It ensures inclusion in the sample subgroup, which otherwise would be omitted entirely by other sampling methods because of their small numbers in the population
2. Dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample
3. Allows use of different sampling techniques for different subpopulations, since each stratum is treated as an independent population.
4. Utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples).
5. Permits greater balancing of statistical power of tests of differences between strata by sampling equal numbers from strata varying widely in size.

**Disadvantages**

1. Requires selection of relevant stratification variables which can be difficult.
2. Is not useful when there are no homogeneous subgroups.
3. Can be expensive and complex to implement (obtaining satisfactory frame).

## Systematic sampling

Systematic sampling relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every *K*th element from then onwards. In this case, K= (population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the *K*th element in the list. A simple example would be to select every 10th name from the telephone directory (an 'every 10th sample, also referred to as 'sampling with a skip of 10). In simple words you begin with a listing of all the elements in the designated population then determine the desired sample size and divide it into the population size to give you incremental value. First element is selected by random process in order to avoid bias. As long as the starting point is randomized, systematic sampling is a type of probability sampling.

**Steps in systematic sampling**

1. List all the units in the population in random order e.g. all first year students (0001 -5,000)
2. Determine the sample size e.g. 500
3. Determine the sampling interval (i.e. population/sample size) = 10
4. Randomly select any number between 1 and 10. Suppose the number you have picked is 8, that will be your starting number. So student number 8 has been selected.
5. From there you will select every 10th number until you reach the last one. You will end up with 500 selected students.

**Example1:** If there are 100 students in your class. You want a sample of 20 from these 100 and you have their names listed on a piece of paper may be in an alphabetical order. If you choose to use systematic random sampling, divide 100 by 20, you will get 5. Randomly select any number between 1 and five. Suppose the number you have picked is 4, that will be your starting number. So student number 4 has been selected. From there you will select every 5th name until you reach the last one, number one hundred. You will end up with 20 selected students.

**Example2:**Suppose we wish to sample people from a long street that starts in a poor district (house #1) and ends in an expensive district (house #1000). A simple random selection of addresses from this street could easily end up with too many from the high end and too few from the low end (or vice versa), leading to an unrepresentative sample. Selecting (e.g.) every 10th street number along the street ensures that the sample is spread evenly along the length of the street, representing all of these districts. (Note that if we always start at house #1 and end at #991, the sample is slightly biased towards the low end; by randomly selecting the start between #1 and #10, this bias is eliminated.)

However, systematic sampling is especially vulnerable to periodicities in the list. If periodicity is present and the period is a multiple or factor of the interval used, the sample is especially likely to be *un*representative of the overall population, making the scheme less accurate than simple random sampling.

***Example:***Consider a street where the odd-numbered houses are all on the north (expensive) side of the road and the even-numbered houses are all on the south (cheap) side. Under the sampling scheme given above, it is impossible' to get a representative sample; either the houses sampled will all be from the odd-numbered, expensive side, or they will all be from the even-numbered, cheap side*.*

As described above, in systematic all elements have the same probability of selection (in the example given, one in ten). It is *not* 'simple random sampling' because different subsets of the same size have different selection probabilities.

## Cluster sampling

Cluster sampling is the most widely used type of probability sampling. A cluster sample is obtained by selecting clusters from the population on the basis of simple random sampling. Cluster sampling is an example of 'two-stage sampling' or 'multistage sampling': in the first stage a sample of areas is chosen; in the second stage a sample of respondents *within* those areas is selected. In such case the method may be referred as **multi-stage cluster sampling**.

The sample comprises a census of each random cluster selected. For example, a cluster may be something like a village or a school, a state. So you decide all the elementary schools in Kenya are clusters. You want 20 schools selected. You can use simple or systematic random sampling to select the schools, and then every school selected becomes a cluster. If your interest is to interview teachers on their opinion of some new program which has been introduced, then all the teachers in a cluster must be interviewed. Though very economical cluster sampling is very susceptible to sampling bias. Like for the above case, you are likely to get similar responses from teachers in one school due to the fact that they interact with one another.

This can reduce travel and other administrative costs. It’s less costly as that one does not need a sampling frame listing all elements in the target population. Instead, clusters can be chosen from a cluster-level frame, with an element-level frame created only for the selected clusters. Cluster sampling generally increases the variability of sample estimates above that of simple random sampling, depending on how the clusters differ between themselves, as compared with the within-cluster variation.

## Steps of carrying out cluster sampling

**Step 1:** Define the clusters to be used

A ***cluster***is a clearly defined group of sampling elements from which you can select a smaller sub-sample. Some examples of clusters are: geographic areas with fixed boundaries, schools, health facilities, and youth clubs.

**Step 2:** Develop the sampling frame.

In cluster sampling, a ***sampling frame***is a list of clusters. In some cases, a list of the clusters you have defined may already be available, although you should make sure that the list is complete and up-to-date. Or, you may have to create a sampling frame.

**Step 3:** Select a sample of clusters

Once you have developed your sampling frame, the next step is to select a sample of clusters by using either simple random or systematic sampling.

**Step 4:** Select a sample of elements

Finally, you will need to choose a sample of elements from the clusters you have selected. This is often done using either simple random or systematic sampling, but you could use other sampling schemes, depending on the context.

**Note:** Deciding whether to sample clusters or sample elements depends on two things:

1. **How many clusters you have:** For small-scale programs covering only a handful of program sites (for example, communities, schools or facilities), you may be able to include all sites in the evaluation
2. **How many elements exist within each cluster:** If the number of elements at each site is not too high, it may be simpler logistically to include all sample cluster elements in the sample.

## Determining how many clusters to choose in a sample

When choosing clusters for a sample, select as many clusters (e.g., communities, schools or facilities) as your resources will permit.

* Generally, you should choose at least 30 clusters, especially for large programs.
* A sample of more clusters of smaller size is always preferable to one with fewer clusters of larger size.
* Choosing a smaller number of clusters will reduce the precision or reliability of your data, thus making it more difficult to detect real changes in indicators.
* For example, it is better to choose 30 schools with a population of 200 students each, than to choose 20 schools with a population of 300 students each.
* If you choose fewer than 20 clusters, there is little advantage in randomly selecting clusters, because the desirable statistical properties of probability sampling do not apply to samples of fewer than 20.

Nevertheless, some of the disadvantages of cluster sampling are the reliance of sample estimate precision on the actual clusters chosen. If clusters chosen are biased in a certain way, inferences drawn about population parameters from these sample estimates will be far off from being accurate. The more clusters are used and more randomly they are chosen, the more representative the results are likely to be.

**When might a researcher want to use cluster sampling?**

When the population is widely dispersed and you must visit all the people selected to be in your sample (e.g., for in-person interviews).

## Multistage sampling

Multistage sampling is a complex form of cluster sampling in which two or more levels of units are embedded one in the other. This is like cluster sampling, but with several stages of sampling and sub-sampling. This method is usually used in large-scale population surveys. This technique, thus, is essentially the process of taking random samples of preceding random samples. It is not as effective as true random sampling, but it probably solves more of the problems inherent to random sampling. Moreover, it is an effective strategy because it banks on multiple randomizations. As such, it is extremely useful.

Multistage sampling is used frequently when a complete list of all members of the population does not exist and is inappropriate. Moreover, by avoiding the use of all sample units in all selected clusters, multistage sampling avoids the large, and perhaps unnecessary, costs associated traditional cluster

## *(b) Non-probability sampling*

Non-probability sampling is a sampling technique where the samples are gathered in a process that does not give all the individuals in the population equal chances of being selected. This entails that the sample may or may not represent the entire population accurately. Therefore, the results of the research cannot be used in generalizations pertaining to the entire population. Most qualitative studies use non-probability samples because the focus is on the in-depth information and not making inference or generalizations. Examples of non-probability sampling are:

* Quota sampling
* Purposive sampling
* Snow ball sampling
* Convenience sampling.

It is not quite possible to introduce randomization into this type of sampling. Non probability sampling cuts on cost and time requirements.

## Quota sampling

In **quota sampling**, the population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion

In this sampling method the interviewer are requested to interview **all** the people they **meet** up to a given number which is called their quota. The method is not a random one because there isn’t an actual predesigned sampling frame. Interviewer decides to interview anybody not predetermined but who conforms with a given characters.

Despite being cheaper (quicker and less expensive) and the interviewer **need not** to s to call back again as it happens with other methods, this method has the following disadvantages:-

* Some people may be asked to be interviewed and refused to be interviewed
* Interviewer may be late to go to their quotas thus not meeting their expected characters hence some quotas not being representative

## Convenience sampling

Convenience sampling (sometimes known as **grab** or **opportunity or accidental sampling**) is a type of non-probability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, a sample population selected because it is readily available/accessible and convenient. It may be through meeting the person or including a person in the sample when one meets them or chosen by finding them through technological means such as the internet or through phone. This saves time, money and effort. What is lost in accuracy is gained in efficiency. The researcher using such a sample **cannot scientifically make generalizations** about the total population from this sample because it would not be representative enough. For example, if the interviewer was to conduct such a survey at a shopping center early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey was to be conducted at different times of day and several times per week. This type of sampling is most useful for **pilot testing**. Several important considerations for researchers using convenience samples include:

1. Are there controls within the research design or experiment which can serve to lessen the impact of a non-random convenience sample, thereby ensuring the results will be more representative of the population?
2. Is there good reason to believe that a particular convenience sample would or should respond or behave differently than a random sample from the same population?
3. Is the question being asked by the research one that can adequately be answered using a convenience sample?

## Purposive Sampling or judgment samples

Sometimes a researcher selects a subgroup which can judged to be representative of the population e.g. picking a typical village to represent a national rural population is an example of a purposive sample. This sampling allows researcher to use cases that have required information. Cases of subjects are therefore hand picked because they are informative or they posses the required characteristics. A Purposive sampling selects information rich cases for in-depth study. Size and specific cases depend on the study purpose. Some researcher use purposive sampling as part of multistage sampling procedure. In such cases; purposive sampling is applied to get the location or district in which the units of observation have the required characteristics. Within the selected location or district random sampling may then be applied to obtain the actual sample of cases.

## Snowball sampling

Also known as **chain referral sampling**. Participants or informants with whom connect has already been made use their social network to refer the researcher to other people who could be potentially until the researcher gets the number of cases he or she requires. For example if a researcher wants to study some aspects of Kenya’s nationalist, he or she may identify a few people who participated in the Mau Mau uprising. The few thus identified may assist in identifying others that they know of. The same may apply to commercial sex workers identification.

## Advantages of probability sampling

* Less prone to bias
* Allows estimation of magnitude of sampling error
* The sample size is based on statistical measure based on the population and can be measured and estimated in terms of precision required
* It provides a more accurate method of drawing conclusion about the characteristics of the population expressed as parameters

## Disadvantages of probability sampling

* Requires that you have a list of all sample elements.
* Time consuming
* Costly
* Requires a level of skills which the program may not have.
* No advantage when small number of elements are selected

## Advantages of non- probability sampling

* Flexible
* Less costly to execute
* Requires lesser level of skills and experience
* Less time consuming
* Simplicity, no need for sample frames.
* Speedy no need for call backs
* Judgmentally representative samples may be preferred when small numbers of elements are to be chosen

## Disadvantages of non- probability sampling

* Greater risk of bias
* May not be possible to generalize to program target population
* Subjectivity can make it difficult to measure changes in indicators over time
* No way to assess precision

## Conclusion

There no method which is better than the other. It depends on design of the study, resource available and skill of the evaluator

## Determining Sample Size

**Sample size**refers to the number of sample elements from which you will need to collect data in order for your evaluation findings to be statistically significant. Where time and resources allow, a researcher should take as big a sample as possible. With a large sample size, the researcher is confident that if another sample of the same size were to be selected, findings from the two samples would be similar to a high degree. The danger of small samples is that they do not reproduce the salient characteristics of the accessible population to an acceptable degree. The discrepancy between the sample characteristics and the population characteristics is referred to as **sampling error**. The small the sample the bigger the sampling error.

The size of your sample depends on a number of factors, including:

* the indicators chosen,
* the baseline value of the indicators in the study population, and
* the amount of change you want to be able to measure accurately

In social science research, the following formula (**Fisher et al**) can be used to determine the sample size.

**n = Z2pq/d2**

where:

**n** = the desired sample size (if target population is greater than 10,000)

**Z** = the standard normal deviate at the required confidence level (1.96)

**p** = the proportion in the target population estimated to have characteristics being measured

**q** = 1-p

**d** = the level of statistical significance set (0.05)

**Note:** If there is no estimate of the proportion in the target population assumed to have the characteristics of interest, 50% should be used as recommended by Fisher et al.

**Example**

If the proportion of a target population with a certain characteristics is 0.50, the Z-statistics is 1.96 and we desire accuracy at the .05 level, then the sample size is:

n = (1.96)2(0.50) (0.50)/(0.05)2  = 384

If the target population is less than 10,000, the required sample size will be smaller. In such cases, calculate a final sample estimate (nf) using the following formula:

**nf= n/{1 + (n/N)}**

Where: **nf** = the desired sample size (when population is less than 10,000)

**n** = the desired sample size (if target population is greater than 10,000)

**N** = the estimate of the target population size

# ERRORS Vs MISTAKES

A mistake is incorrect presentation or calculation due to human factor while an error is difference between the actual figure and estimate figure. The deviation is just by chance and not due to carelessness of human being.

**Sources of error**

* Error of origin
* Error of manipulation
* Error of inadequacy

**Kinds of error**

* Sampling errors
* Non sampling errors
* Biased error
* Unbiased error

**Sampling Errors And Biases**

Sampling errors and biases are induced by the sample design. They include:

1. **Selection bias**: When the true selection probabilities differ from those assumed in calculating the results.
2. **Random sampling error**: Random variation in the results due to the elements in the sample being selected at random.

**Non-sampling error**

Non-sampling errors are caused by other problems in data collection and processing. They include:

* **Over coverage**: Inclusion of data from outside of the population.
* **Under coverage**: Sampling frame does not include elements in the population.
* **Measurement error**: E.g. when respondents misunderstand a question, or find it difficult to answer.
* **Processing error**: Mistakes in data coding.
* **Non-response**: Failure to obtain complete data from all selected individuals.

**Bias in Information Collection**

Bias in information collection is a distortion in the collected data so that it does not represent reality.

**Methods of selection that give rise to bias**

1. Deliberate selection of a representative sample- items picked at will
2. Selection using improper random methods- proper random process is not strictly adhered to.
3. Substitution of items in a sample.
4. Failure to cover all the item in a sample

**Possible sources of bias during data collection:**

**1. Defective instruments, such as:**

* Questionnaires with:
  + fixed or closed questions on topics about which little is known (often asking the ‘wrong things’);
  + open-ended questions without guidelines on how to ask (or to answer) them;
  + vaguely phrased questions;
  + leading questions’ that cause the respondent to believe one answer would be preferred over another; or
  + questions placed in an illogical order.
  + weighing scales or other measuring equipment that are not standardized

These sources of bias can be prevented by carefully planning the data collection process and by pre-testing the data collection tools.

**2. Observer bias:**

Observer bias can easily occur when conducting observations or utilizing loosely structured group- or individual interviews. There is a risk that the data collector will only see or hear things in which (s) he is interested or will miss information that is critical to the research.

Observation protocols and guidelines for conducting loosely structured interviews should be prepared, and training and practice should be provided to data collectors in using both these tools. Moreover it is highly recommended that data collectors **work in pairs** when using flexible research techniques and discuss and interpret the data immediately after collecting it. Another possibility - commonly used by anthropologists - is using a tape recorder and transcribing the tape word by word.

**3. Effect of the interview on the informant:**

This is a possible factor in all interview situations. The informant may mistrust the intention of the interview and dodge certain questions or give misleading answers**. For example:** in a survey on alcoholism you ask school children: ‘Does your father sometimes get drunk?’ Many will probably deny that he does, even if it is true. Such bias can be reduced by adequately introducing the purpose of the study to informants, by phrasing questions on sensitive issues in a positive way, by taking sufficient time for the interview, and by assuring informants that the data collected will be

It is also important to be careful in the selection of interviewers. In a study soliciting the reasons for the low utilization of local health services, for example, one should not ask health workers from the health centres concerned to interview the population. Their use as interviewers would certainly influence the results of the study.

**4. Information bias:**

Sometimes the information itself has weaknesses. Medical records may have many blanks or be unreadable. This tells something about the quality of the data and has to be recorded. For example, in a TB defaulter study the percentage of defaulters with an incomplete or missing address should be calculated. Another common information bias is due to gaps in people’s memory; this is called ***memory* or *recall bias****.* A mother may not remember all details of her child’s last diarrhoea episode and of the treatment she gave two or three months afterwards. For such common diseases it is advisable to limit the period of recall, asking, for example, ‘Has your child had diarrhoea over the past two weeks?’

**Note:**

All these potential biases will threaten the validity and reliability of your study. By being aware of them it is possible, to a certain extent, to prevent them. If the researcher does not fully succeed, it is important to report honestly in what ways the data may be biased.

**Ethical considerations**

As we develop our data collection techniques, we need to consider whether our research procedures are likely to cause any physical or emotional harm. Harm may be caused, for example, by:

* violating informants’ right to privacy by posing sensitive questions or by gaining access to records which may contain personal data;
* observing the behaviour of informants without their being aware (concealed observation should therefore always be crosschecked or discussed with other researchers with respect to ethical admissibility);
* allowing personal information to be made public which informants would want to be kept private, and
* Failing to observe/respect certain cultural values, traditions or taboos valued by your informants.

Several methods for dealing with these issues may be recommended:

* Voluntary and informed consent before the study or the interview begins;
* not exploring sensitive issues before a good relationship has been established with the informant;
* Ensuring the confidentiality and privacy of the data obtained; and
* Learning enough about the culture of informants to ensure it is respected during the data collection process.
* Anonymity – to the identity of individuals being protected either by using numbers or third parties or pseudo names.

If sensitive questions are asked, for example, about family planning or sexual practices, or about opinions of patients on the health services provided, it may be advisable to omit names and addresses from the questionnaires.

# DATA COLLECTION

# Sources of health data

**Primary source - (**Entire population/sample**)**

Primary data are original data/raw facts. This category of data is collected from the field through survey and/or experimentation. The researcher designs the methods and instruments of primary data collection which suit the nature of the data to be collected. Data obtained from this source will be complete and accurate than any other source.

**Advantages**

1. Gives very reliable data
2. Give data that includes those who attend and those who do not attend health facility

**Disadvantages**

1. Very expensive to conduct
2. Difficult to follow-up individual who may not be available during the study period
3. Impossible to show trends
4. Possibility of sampling errors
5. Has a short term value- value end immediately after the completion of the investigation

**Secondary sources**- consists of existing information collected by others and available from secondary sources e.g. (existing medical and civil registration records and other published/unpublished documents).

**Advantages**

1. There is accumulation of data over time hence you may show trends
2. Can be accurate when it is compulsory to report
3. Less expensive
4. Easy to follow-up
5. Has long term value

**Disadvantages**

1. Population to which we relate data is not known and therefore impossible to compute further statistics
2. Excludes cases who do not attend health facility or registered by the source

# Methods of Data Collection

* Direct Observation
* Interviews – a disadvantage of interviewing is that inaccurate or false data may be given to the interviewer due to
  + Forgetfulness
  + Misunderstanding the question
  + Deliberate intent to mislead
  + Failure of interviewers recording the answers in the same way as the investigator himself would do.
* Abstraction from published statistics – secondary statistics
* Questionnaires
* Focus group discussions

In primary data collection, you collect the data yourself using methods such as interviews and questionnaires. The key point here is that the data you collect is unique to you and your research and, until you publish, no one else has access to it.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | http://brent.tvu.ac.uk/dissguide/images/spacer.gif | http://brent.tvu.ac.uk/dissguide/images/spacer.gif | http://brent.tvu.ac.uk/dissguide/images/spacer.gif | http://brent.tvu.ac.uk/dissguide/images/spacer.gif |
| [Interviews](http://brent.tvu.ac.uk/dissguide/hm1u3/hmu3int/LRNViewer.htm) | [Questionnaires](http://brent.tvu.ac.uk/dissguide/hm1u3/hmu3qnr/LRNViewer.htm) | [Observations](http://brent.tvu.ac.uk/dissguide/hm1u3/hmu3obs/LRNViewer.htm) | [Historical Methods](http://brent.tvu.ac.uk/dissguide/hm1u3/hmu3hist/LRNViewer.htm) | http://brent.tvu.ac.uk/dissguide/images/spacer.gif |

The primary data, which is generated by the above methods, may be qualitative in nature (usually in the form of words) or quantitative (usually in the form of numbers or where you can make counts of words used).

# Questionnaires (self-administered /postal, researcher administered/telephone, web based)

Questionnaires are a popular means of collecting data, but are difficult to design and often require many rewrites before an acceptable questionnaire is produced.

**Advantages:**

* Can be used as a method in its own right or as a basis for interviewing or a telephone survey.
* Can be posted, e-mailed or faxed.
* Can cover a large number of people or organizations.
* Wide geographic coverage.
* Relatively cheap.
* Avoids embarrassment on the part of the respondent.
* Possible anonymity of respondent.
* Free from interviewer bias, answers are in the respondent’s words
* Respondents have enough time to give their well though answers

**Disadvantages:**

* Design problems.
* Questions have to be relatively simple.
* Historically low response rate (although inducements may help).
* Time delay whilst waiting for responses to be returned.
* Questionnaire might be lost when sent
* Require a return deadline.
* Several reminders may be required.
* Assumes no literacy problems, can only used when the respondents are educated and co-operating
* No control over who completes it.
* Not possible to give assistance if required.
* Problems with incomplete questionnaires.
* Replies not spontaneous and independent of each other.
* Respondent can read all questions beforehand and then decide whether to complete or not. For example, perhaps because it is too long, too complex, uninteresting, or too personal.

# Design of postal questionnaires

**Theme and covering letter**

The general theme of the questionnaire should be made explicit in a covering letter. You should state who you are; why the data is required; give, if necessary, an assurance of confidentiality and/or anonymity; and contact number and address or telephone number. This ensures that the respondents know what they are committing themselves to, and also that they understand the context of their replies. If possible, you should offer an estimate of the completion time. Instructions for return should be included with the return date made obvious. For example: ‘It would be appreciated if you could return the completed questionnaire by... if at all possible’.

**Instructions for completion**

You need to provide clear and unambiguous instructions for completion. Within most questionnaires these are general instructions and specific instructions for particular question structures. It is usually best to separate these, supplying the general instructions as a preamble to the questionnaire, but leaving the specific instructions until the questions to which they apply. The response method should be indicated (circle, tick, cross, etc.). Wherever possible, and certainly if a slightly unfamiliar response system is employed, you should give an example.

**Appearance**

Appearance is usually the first feature of the questionnaire to which the recipient reacts. A neat and professional look will encourage further consideration of your request, increasing your response rate. In addition, careful thought to layout should help your analysis. There are a number of simple rules to help improve questionnaire appearance:

* Liberal spacing makes the reading easier.
* Font-reduction can produce more space without reducing content.
* Consistent positioning of response boxes, usually to the right, speeds up completion and also avoids inadvertent omission of responses.
* Choose the font style to maximize legibility.
* Differentiate between instructions and questions. Either lower case and capitals can be used, or responses can be boxed.

**Length**

There may be a strong temptation to include any vaguely interesting questions, but you should resist this at all costs. Excessive size can only reduce response rates. If a long questionnaire is necessary, then you must give even more thought to appearance. It is best to leave pages unnumbered; for respondents to flick to the end and see ‘page 27’ can be very disconcerting!

**Order**

Probably the most crucial stage in questionnaire response is the beginning. Once the respondents have started to complete the questions they will normally finish the task, unless it is very long or difficult. Consequently, you need to select the opening questions with care. Usually the best approach is to ask for biographical details first, as the respondents should know all the answers without much thought. Another benefit is that an easy start provides practice in answering questions.

Once the introduction has been achieved the subsequent order will depend on many considerations. You should be aware of the varying importance of different questions. Essential information should appear early, just in case the questionnaire is not completed. For the same reasons, relatively unimportant questions can be placed towards the end. If questions are likely to provoke the respondent and remain unanswered, these too are best left until the end, in the hope of obtaining answers to everything else.

**Coding**

If analysis of the results is to be carried out using a statistical package or spreadsheet it is advisable to code non-numerical responses when designing the questionnaire, rather than trying to code the responses when they are returned. An example of coding is:

|  |  |
| --- | --- |
| Male [   ] | Female [   ] |
| 1 | 2 |

The coded responses (1 or 2) are then used for the analysis.

**Thank you**

Respondents to questionnaires rarely benefit personally from their efforts and the least the researcher can do is to thank them. Even though the covering letter will express appreciation for the help given, it is also a nice gesture to finish the questionnaire with a further thank you.

**Questions**

* Keep the questions short, simple and to the point; avoid all unnecessary words.
* Use words and phrases that are unambiguous and familiar to the respondent. For example, ‘dinner’ has a number of different interpretations; use an alternative expression such as ‘evening meal’.
* Only ask questions that the respondent can answer. Hypothetical questions should be avoided. Avoid calculations and questions that require a lot of memory work, for example, ‘How many people stayed in your hotel last year?’
* Avoid loaded or leading questions that imply a certain answer. For example, by mentioning one particular item in the question, ‘Do you agree that Colgate toothpaste is the best toothpaste?’
* Vacuous words or phrases should be avoided. ‘Generally’, ‘usually’, or ‘normally’ are imprecise terms with various meanings. They should be replaced with quantitative statements, for example, ‘at least once a week’.
* Questions should only address a single issue. For example, questions like: ‘Do you take annual holidays to Spain?’ should be broken down into two discreet stages, firstly find out if the respondent takes an annual holiday, and then secondly find out if they go to Spain.
* Do not ask two questions in one by using ‘and’. For example, ‘Did you watch television last night and read a newspaper?’
* Avoid double negatives. For example, ‘Is it not true that you did not read a newspaper yesterday?’ Respondents may tackle a double negative by switching both negatives and then assuming that the same answer applies. This is not necessarily valid.
* State units required but do not aim for too high a degree of accuracy. For instance, use an interval rather than an exact figure:

‘How much did you earn last year?’

Less than £10,000 [   ]

£10,000 but less than £20,000 [   ]

Avoid emotive or embarrassing words – usually connected with race, religion, politics, sex, money.

**Types of questions**

**Closed /structured questions**

A question is asked and then a number of possible answers are provided for the respondent. The respondent selects the answer which is appropriate. Closed questions are particularly useful in obtaining factual information:

Sex:    Male [   ] Female [   ]

Did you watch television last night?     Yes [   ] No [   ]

Some ‘Yes/No’ questions have a third category ‘Do not know’. Experience shows that as long as this alternative is not mentioned people will make a choice. Also the phrase ‘Do not know’ is ambiguous:

Do you agree with the introduction of the KMTC?

Yes [   ] No [   ] Do not know [   ]

What was your main way of traveling to the hotel? Tick one box only.

|  |  |
| --- | --- |
| Car | [   ] |
| Coach | [   ] |
| Motor bike | [   ] |
| Train | [   ] |
| Other means, please specify | Top of Form    Bottom of Form |

With such lists you should always include an ‘other’ category, because not all possible responses might have been included in the list of answers.

Sometimes the respondent can select more than one from the list. However, this makes analysis difficult:

Why have you visited the historic house? Tick the relevant answer(s). You may tick as many as you like.

|  |  |
| --- | --- |
| I enjoy visiting historic houses | [   ] |
| The weather was bad and I could not enjoy outdoor activities | [   ] |
| I have visited the house before and wished to return | [   ] |
| Other reason, please specify | Top of Form  Bottom of Form |

**Advantages of closed - ended questions**

1. Easier to analyze since they are in immediate usable form
2. Easier to administer
3. Economical to use in terms of time and money

**Disadvantages of closed - ended questions**

1. More difficult to construct because categories must be well thought out
2. Responses are limited and compelled to answer questions according to researcher’s choices

**Open ended questions/ unstructured questions**

An open question such as ‘What are the essential skills a manager should possess?’ should be used as an adjunct to the main theme of the questionnaire and could allow the respondent to elaborate upon an earlier more specific question. Open questions inserted at the end of major sections, or at the end of the questionnaire, can act as safety valves, and possibly offer additional information. However, they **should no**t be used to introduce a section since there is a high risk of influencing later responses. The main problem of open questions is that many different answers have to be summarized and possibly coded.

**Advantages of open - ended questions**

1. They permit a greater depth of response
2. Are simply to formulate, researcher does not need to labor to come up with appropriate choices
3. Respondent’s response may give insight into his feelings, background, hidden motivation, interests and decisions
4. Can stimulate a person to think about his feelings or motives and express what he considers to be most important

**Disadvantages of open - ended questions**

1. There is tendency to provide information which does not answer the stipulated question
2. Response given may difficult to categorize hence difficult to analyze
3. Time consuming and may put off some respondents

**Contingency questions**

Subsequent questions that are asked after initial questions are called contingency questions. The purpose of these kinds of questions is to probe for more questions. There are various formats to formulate contingency questions e.g.

**Format one**

Have you ever voted in a political election?

Yes [   ] No [   ]

If yes, how many times

Once [   ] 2-4 times [   ] 5-10 times [   ] over 10 times [   ]

**Format two**

Have you ever used birth control measures?

[   ] Yes (please answer question 3-5)

[   ] No (please skip question 3-5 and go to question 6)

**Matrix questions /Attitude questions**

Frequently questions are asked to find out the respondents’ opinions or attitudes to a given situation. A **Likert type scale** provides a battery of attitude statements. The respondent then says how much they agree or disagree with each one:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Strongly agree | Agree | Disagree | Strongly disagree |
| My visit has been good value for money |  |  |  |  |

Read the following statements and then indicate by a tick whether you strongly agree, agree, disagree or strongly disagree with the statement.

There are many variations on this type of question. One variation is to have a ‘middle statement’, for example, ‘Neither agree nor disagree’. However, many respondents take this as the easy option. Only having four statements, as above, forces the respondent into making a positive or negative choice. Another variation is to rank the various attitude statements; however, this can cause analysis problems:

Which of these characteristics do you like about your job? Indicate the best three in order, with the best being number 1.

|  |  |
| --- | --- |
| Varied work | [   ] |
| Good salary | [   ] |
| Opportunities for promotion | [   ] |
| Good working conditions | [   ] |
| High amount of responsibility | [   ] |
| Friendly colleagues | [   ] |

Other types of questions to determine peoples’ opinions or attitudes are:

Which one/two words best describes...?

Which of the following statements best describes...?

How much do you agree with the following statement...?

**Distribution and return**

The questionnaire should be checked for completeness to ensure that all pages are present and that none is blank or illegible. It is usual to supply a prepaid addressed envelope for the return of the questionnaire. You need to explain this in the covering letter and reinforce it at the end of the questionnaire, after the ‘Thank you’.

Finally, many organizations are approached continually for information. Many, as a matter of course, will not respond in a positive way.

**Response rate**

This refers to the percentage of subjects who respond to questionnaires. If the response rate is low, the researcher must question the representativeness of the sample.

**What is an acceptable response rate?**

Over 70% is recommended but the researcher should aim very high to increase the response rate in order to have representative sample for meaningful generalizations. Non- respondents are subjects who do not respond to questionnaire

# Interviews

Interviewing is a technique that is primarily used to gain an understanding of the underlying reasons and motivations for people’s attitudes, preferences or behaviour. Interviews can be undertaken on a personal one-to-one basis or in a group. They can be conducted at work, at home, in the street or in a shopping centre, or some other agreed location

**Face -to -face interviews** have a distinct advantage of enabling the researcher to establish rapport with potential participants and therefore gain their cooperation. These interviews yield highest response rates in survey research.

**Telephone interviews** are less time consuming and less expensive and the researcher has ready access to anyone on the planet who has a telephone. Disadvantages are that the response rate is not as high as the face-to- face interview as but considerably higher than the mailed questionnaire. The sample may be biased to the extent that people without phones are part of the population about whom the researcher wants to draw inferences.

**Computer Assisted Personal Interviewing (CAPI):** is a form of personal interviewing, but instead of completing a questionnaire, the interviewer brings along a laptop or hand-held computer to enter the information directly into the database. This method saves time involved in processing the data, as well as saving the interviewer from carrying around hundreds of questionnaires. However, this type of data collection method can be expensive to set up and requires that interviewers have computer and typing skills.

# Types of interview

**Structured:**

In a structured interview, the researcher asks a standard set of questions and nothing more (Leedy and Ormrod, 2001).

* Based on a carefully worded interview schedule.
* Frequently require short answers with the answers being ticked off.
* Useful when there are a lot of questions which are not particularly contentious or thought provoking.
* Respondent may become irritated by having to give over-simplified answers.

**Semi-structured**

The interview is focused by asking certain questions but with scope for the respondent to express him or herself at length.

**Unstructured**

This also called an **in-depth interview**. The interviewer begins by asking a general question. The interviewer then encourages the respondent to talk freely. The interviewer uses an unstructured format, the subsequent direction of the interview being determined by the respondent’s initial reply. The interviewer then probes for elaboration – ‘Why do you say that?’ or, ‘That’s interesting, tell me more’ or, ‘would you like to add anything else?’ being typical probes.

**Advantages of interviews**

1. Provide in-depth data that is not possible to get using a questionnaire
2. Interview can clarify hence helping the respondent giving the relevant answers
3. More flexible than questionnaire as interview can adapt the situation and get much information as possible
4. Respondents can give honest results after interviewer clarifying the purpose of the study
5. Interviewer can get more information by using probing questions
6. It yields a high response rate as it is difficult for e respondent to refuse completely to answer questions or ignore the interviewer

**Disadvantages of interviews**

1. More expensive
2. Can be misused to get factual responses which could be obtained more accurately through other methods
3. Requires high level of skills – communication and interpersonal skills
4. Interviews need to be trained to avoid bias
5. Often they introduce bias and subjectivity to the study- respondent may want to please the interviewer or the interviewer may ask questions which tend to support his perceived notion. In this case, there exist a difference between the respondent’s response and the true answer. This is called ”**the response effect**.”
6. More applicable for small samples than bigger samples
7. Responses may be influenced by the respondent’s reaction to the interviewer e.g. way of dressing

**Personal interview**

**Advantages:**

* Serious approach by respondent resulting in accurate information.
* Good response rate.
* Completed and immediate.
* Possible in-depth questions.
* Interviewer in control and can give help if there is a problem.
* Can investigate motives and feelings.
* Can use recording equipment.
* Characteristics of respondent assessed – tone of voice, facial expression, hesitation, etc.
* If one interviewer used, uniformity of approach.
* Used to pilot other methods.

**Disadvantages:**

* Need to set up interviews.
* Time consuming.
* Geographic limitations.
* Can be expensive.
* Normally need a set of questions.
* Respondent bias – tendency to please or impress, create false personal image, or end interview quickly.
* Embarrassment possible if personal questions are asked.
* Transcription and analysis can present problems – subjectivity.
* Expensive as many interviewers, training required.

**Planning an interview:**

* List the areas in which you require information.
* Decide on type of interview.
* Transform areas into actual questions.
* Try them out on a friend or relative.
* Make an appointment with respondent(s) – discussing details of why and how long.
* Try and fix a venue and time when you will not be disturbed.

**Conducting personal interview:**

|  |  |
| --- | --- |
| Personally | Arrive on time be smart smile employ good manners find a balance between friendliness and objectivity. |
| At the start | Introduce yourself re-confirm the purpose assure confidentiality – if relevant specify what will happen to the data. |
| The questions | Speak slowly in a soft, yet audible tone of voice control your body language know the questions and topic ask all the questions. |
| Responses | Recorded as you go on questionnaire written verbatim, but slow and time-consuming summarized by you taped – agree beforehand – have alternative method if not acceptable consider effect on respondent’s answers proper equipment in good working order sufficient tapes and batteries minimum of background noise. |
| At the end | Ask if the respondent would like to give further details about anything or any questions about the research thank them. |

**Telephone interview**

This is an alternative form of interview to the personal, face-to-face interview.

**Advantages:**

* Relatively cheap.
* Quick.
* Can cover reasonably large numbers of people or organizations.
* Wide geographic coverage.
* High response rate – keep going till the required number.
* No waiting.
* Spontaneous response.
* Help can be given to the respondent.
* Can tape answers.

**Disadvantages:**

* Often connected with selling.
* Questionnaire required.
* Not everyone has a telephone.
* Repeat calls are inevitable – average 2.5 calls to get someone.
* Time is wasted.
* Straightforward questions are required.
* Respondent has little time to think.
* Cannot use visual aids.
* Can cause irritation.
* Good telephone manner is required.
* Question of authority.

**Getting started**

* Locate the respondent:
  + Repeat calls may be necessary especially if you are trying to contact people in organizations where you may have to go through secretaries.
  + You may not know an individual’s name or title – so there is the possibility of interviewing the wrong person.
  + You can send an advance letter informing the respondent that you will be telephoning. This can explain the purpose of the research.
* **Getting them to agree to take part:**
  + You need to state concisely the purpose of the call – scripted and similar to the introductory letter of a postal questionnaire.
  + Respondents will normally listen to this introduction before they decide to co-operate or refuse.
  + When contact is made respondents may have questions or raise objections about why they could not participate. You should be prepared for these.

**Ensuring quality**

* **Quality of questionnaire** – follows the principles of questionnaire design. However, it must be easy to move through as you cannot have long silences on the telephone.
* **Ability of interviewer** – follows the principles of face-to-face interviewing.

**Interview schedule** – each interview schedule should have a cover page with number, name and address. The cover sheet should make provision to record which call it is, the date and time, the interviewer, the outcome of the call and space to note down specific times at which a call-back has been arranged. Space should be provided to record the final outcome of the call – was an interview refused, contact never made, number disconnected, etc.

* **Procedure for call-backs** – a system for call-backs needs to be implemented. Interview schedules should be sorted according to their status: weekday call-back, evening call-back, weekend call-back, specific time call-back.

**Comparison of postal, telephone and personal interview surveys**

The table below compares the three common methods of postal, telephone and interview surveys – it might help you to decide which one to use.

|  | **Postal survey** | **Telephone survey** | **Personal interview** |
| --- | --- | --- | --- |
| Cost (assuming a good response rate) | Often lowest | Usually in-between | Usually highest |
| Ability to probe | No personal contact or observation | Some chance for gathering additional data through elaboration on questions, but no personal observation | Greatest opportunity for observation, building rapport, and additional probing |
| Respondent ability to complete at own convenience | Yes | Perhaps, but usually no | Perhaps, if interview time is prearranged with respondent |
| Interview bias | No chance | Some, perhaps due to voice inflection | Greatest chance |
| Ability to decide who actually responds to the questions | Least | Some | Greatest |
| Impersonality | Greatest | Some due to lack of face-to-face contact | Least |
| Complex questions | Least suitable | Somewhat suitable | More suitable |
| Visual aids | Little opportunity | No opportunity | Greatest opportunity |
| Potential negative respondent reaction | ‘Junk mail’ | ‘Junk calls’ | Invasion of privacy |
| Interviewer control over interview environment | Least | Some in selection of time to call | Greatest |
| Time lag between soliciting and receiving response | Greatest | Least | May be considerable if a large area is involved |
| Suitable types of questions | Simple, mostly dichotomous (yes/no) and multiple choice | Some opportunity for open-ended questions especially if interview is recorded | Greatest opportunity for open-ended questions |
| Requirement for technical skills in conducting interview | Least | Medium | Greatest |
| Response rate | Low | Usually high | High |

**Note taking during interviews**

It refers to the method of recording in which the interviewer records the respondent’s response during the interview. The interviewer should record the respondent’s answer exactly as expressed (no summarizing or paraphrasing). It is also important to have some comments on the margin regarding observed gestures.

**Advantages of note taking**

1. It facilitates data analysis since the information is readily accessible and already classified by the interviewer
2. No information will be left out owing to forgetfulness or any other kind of omission

**Disadvantages of note taking**

1. Interfere with communication between respondent and interviewer. This is worse if the interview guide has many unstructured questions
2. It may be difficult to establish rapport due to recording interruptions
3. It may upset the respondent if answers are personal and sensitive
4. It makes the interview lengthy
5. If note taking is delayed, important details may be forgotten

**Tape recording**

Interviewer questions and the respondent’s answer are recorded either using tape recorder or video tape. Tape recorder is frequently used.

**Advantages of tape recording**

1. Reduces the tendency for the interviewer to make unconscious selection of data in the course of recording
2. Can be played back and studied more thoroughly
3. Other person other than the interviewer can evaluate and categorize the responses
4. Communication is not interrupted
5. It speeds the interview as there is no recording involved

**Disadvantages of tape recording**

1. May make respondent to be nervous. Some researcher however, argue that soon after the interview starts the respondents forget about the presence of the tape
2. Transcribing the tape and then analyzing the information is time consuming and tedious

**Some rules of communication during interviews**

1. The language used must ensure effective communication. Respondent must fully understand the language
2. Technical terms must be avoided when talking to lay people as they will not understand the question
3. The purpose of the study should be explained. Honesty and communication is enhanced when respondent appreciate the purpose of the study
4. Establish rapport to put respondent at ease and also inspire confidence
5. When communicating avoid harsh and discriminatory remarks
6. Interviewer should monitor the situation (pick non-verbal communication and sensitive question can be withdrawn or revisited differently)
7. If respondent is carried away, interviewer should be tactfully the respondent back to the point without embarrassing them

**Difference between questionnaire and schedules**

1. Questionnaire are generally sent through mail to be answered without further assistance while schedule are filled out by the research worker/interviewer
2. Questionnaire is relatively cheap and economical and schedule is relatively more expensive
3. Non-response is usually high in questionnaire and very low in schedules
4. It is not clear as who replies in questionnaire, but in schedule the identity of respondent is known but remains the danger of interviewer bias and cheating
5. Questionnaire is likely to be slow but schedule data collected will be timely
6. Questionnaire can only used when respondents are literate but with schedule information can be gathered to illiterate respondents
7. Wider and more representative sample is possible with questionnaire, but in respect to schedules it remains difficulty
8. Incomplete and wrong information with questionnaire whole schedule will provide complete and accurate information
9. The success of questionnaire lies more on the quality of the questionnaire but for the schedules much depends on the honesty and competence of enumerators
10. Along with schedules observation method can be used but impossible while collecting data through questionnaire
11. To attract attention to the respondent the questionnaire must be attractive while for the schedule the enumerator must have good interpersonal communication

# Observation

Observation involves recording the behavioral patterns of people, objects and events in a systematic manner. Observational methods may be:

* structured or unstructured
* disguised or undisguised
* natural or contrived
* personal
* mechanical
* non-participant
* participant, with the participant taking a number of different roles.

**Structured or unstructured**

In **structured** observation, the researcher specifies in detail what is to be observed and how the measurements are to be recorded. It is appropriate when the problem is clearly defined and the information needed is specified.

In **unstructured** observation, the researcher monitors all aspects of the phenomenon that seem relevant. It is appropriate when the problem has yet to be formulated precisely and flexibility is needed in observation to identify key components of the problem and to develop hypotheses. The potential for bias is high. Observation findings should be treated as hypotheses to be tested rather than as conclusive findings.

**Disguised or undisguised**

In **disguised** observation, respondents are unaware they are being observed and thus behave naturally. Disguise is achieved, for example, by hiding, or using hidden equipment or people disguised as shoppers.

In **undisguised** observation, respondents are aware they are being observed. There is a danger of the Hawthorne effect – people behave differently when being observed.

**Natural or contrived**

**Natural** observation involves observing behavior as it takes place in the environment, for example, eating hamburgers in a fast food outlet.

In **contrived** observation, the respondents’ behavior is observed in an artificial environment, for example, a food tasting session.

**Personal**

In personal observation, a researcher observes actual behavior as it occurs. The observer may or may not normally attempt to control or manipulate the phenomenon being observed. The observer merely records what takes place.

**Mechanical**

Mechanical devices (video, closed circuit television) record what is being observed. These devices may or may not require the respondent’s direct participation. They are used for continuously recording on-going behaviour.

**Non-participant**

The observer does not normally question or communicate with the people being observed. He or she does not participate.

**Participant**

In participant observation, the researcher becomes, or is, part of the group that is being investigated. Participant observation has its roots in ethnographic studies (study of man and races) where researchers would live in tribal villages, attempting to understand the customs and practices of that culture. It has a very extensive literature, particularly in sociology (development, nature and laws of human society) and anthropology (physiological and psychological study of man). Organizations can be viewed as ‘tribes’ with their own customs and practices.

**Points to bear in mind when using observation forms/checklist**

* Observation form must be tested in situations similar to those which are expected during the data collection – helps to correct any mistake that may be discovered
* The behavior to be observed should be defined in sufficient details to make it easier to the observer to determine whether a particular behavior as occurred or not
* The number of subjects must be controlled. Not too many behaviors should be observed. At most ten
* Where more than one observer is used in a study, training is necessary in order to master the observation form/checklist and standardize observation procedures

# Focus Group Discussion

**What are FGDs?**

A focus group involves encouraging an invited group of participants to share their thoughts, feelings, attitudes and ideas on certain subject. Organizing focus groups within an organization can also be very useful in getting buy-in to a project from within that company.

(FGD) is a group discussion of approximately 6 - 12 persons guided by a facilitator, during which group members talk freely and spontaneously about a certain topic.

A FGD is a qualitative method. Its purpose is to obtain in-depth information on concepts, perceptions and ideas of a group. A FGD aims to be more than a question-answer interaction. The idea is that group members discuss the topic among themselves, with guidance from the facilitator.

**Advantages of focus groups include:**

* Quick, cheap and relatively easy to assemble
* Good for getting rich data in participants' own words and developing deeper insights
* People are able to build on one another's responses and come up with ideas they might not have thought of in a 1-on-1 interview
* Good for obtaining data from children and/or people with low levels of literacy
* Provides an opportunity to involve people in data analysis (e.g. "Out of the issues we have talked about, which ones are most important to you?")
* Participants can act as checks and balances on one another - identifying factual errors or extreme views

**Limitations of focus groups include:**

* The responses of each participant are not independent
* A few dominant focus group members can skew the session
* Focus groups require a skilled and experienced moderator
* The data which results from a focus group requires skill and experience to analyze

**How to plan and prepare for focus groups**

Invite around 6 to 8 people to participate for a session to last for about an hour. Then, prepare an agenda including a list of the top-level issues to be tackled (if appropriate).

Prepare an introduction script explaining the purpose of the day and how the day will be run. This can include issues of consent and fire regulations (if relevant). Be sure to always use a quiet room with few distractions and **arrange people in a circle** (possibly around a table).

**Introduction**

* Welcome participants and introduce yourself.
* Explain the general purpose of the discussion and why the participants were chosen.
* Discuss the purpose and process of focus groups
* Explain the presence and purpose of recording equipment and introduce observers.
* Outline general ground rules and discussion guidelines such as the importance of everyone speaking up, talking one at a time, and being prepared for the moderator to interrupt to assure that all the topics can be covered.
* Review breaks schedule and where the restrooms are.
* Address the issue of confidentiality.
* Inform the group that information discussed is going to be analyzed as a whole and those participants' names will not be used in any analysis of the discussion.
* Read a protocol summary to the participants.

**Running focus groups**

If appropriate, ask the participants to introduce themselves and/or wear name tags. Most importantly, **all questions you ask should be open and neutral**. It's also important for the moderator to be aware of participants' energy and concentration levels and provide short breaks if necessary. The moderator should encourage free-flowing discussion around the relevant issue(s).

**Other tips** for running focus groups include:

* Start on an issue people have strong feelings about and are familiar with
* Phrase issues in terms people will be familiar with
* Let participants know their contributions are valuable (both through what you say and also your body language)

It's also important that the moderator realizes that:

* It may be necessary for them to step in and keep the session on-track
* Disagreements and debates are useful when they lead to new and interesting ideas, but have to be managed carefully
* Issues of power and privacy need to be managed sensitively

Focus groups should end with the moderator winding-up the session by stressing all that has achieved and casting it in a positive light.

**Managing risks**

A number of potential problems could arise during focus groups, which will all need addressing:

* If one participant tries to dominate the session, the moderator should invite each person to speak in turn
* Avoid interviewing friends in the same group as they can form cliques - if cliques do form, suggest taking a break and changing seating positions upon returning from the break
* Avoid personal confrontation - allow the group to police itself (e.g. "do others in the group agree?")
* Respect someone's right to be quiet, but do give them a chance to share their ideas 1-to-1 (e.g. during a break)
* Use differences of opinion as a topic of discussion - the moderator should avoid taking sides

## Useful tips to encourage discussion

To facilitate useful, free-flowing discussion during the focus group, follow some of these tips:

* Ask participants to think about an issue for a few minutes and write down their responses
* Ask each participant to read, and elaborate on, one of their responses
* Note the responses on a flipchart/whiteboard
* Once everyone has given a response, participants will be asked for a second or third response, until all of their answers have been noted
* These responses can then be discussed

## Functions of the facilitator

The facilitator should NOT act as an expert on the topic. His or her role is to stimulate and support discussion.

* **Introduce the session**

Introduce yourself as facilitator and introduce the recorder. Let participants introduce themselves with whatever names they wish to use. Put the participants at ease and explain the purpose of the FGD, the kind of information needed, and how the information will be used (for the planning of a health programme, an education programme, etc). Ask permission to use a tape-recorder, let people hear their own voices before the session starts. You might offer drinks and allow some informal discussion before the actual session starts.

* **Encourage discussion**

Be enthusiastic, lively, and humorous and show your interest in the groups’ ideas. Formulate questions and encourage as many participants as possible to express their views. Remember there are **no** ‘right’ or ‘wrong’ answers. **React neutrally** to both verbal and non-verbal responses.

* **Encourage involvement**

Avoid a question-and-answer session. Some useful techniques include:

—Asking for clarification: ‘Can you tell me more about. . . ?’

— Reorienting the discussion when it goes ‘off the track’: Saying: ‘Wait, how does this relate to. . . ?’   
Saying: ‘Interesting point, but how about. . . ?’   
Using one participant’s remark to direct a question to another, for example, ‘Mrs. X said . . . , but how about you, Mrs. Y?’

— When dealing with a dominant participant, avoiding eye contact or turning slightly away to discourage the person from speaking, or thanking the person and changing the subject.

— When dealing with a reluctant participant, using the person’s name, requesting his/her opinion, making more frequent eye contact to encourage his/her participation.

* Deal correctly with **sensitive issues**. If you notice that the discussion stops when dealing with a sensitive topic, you could ask participants (if literate) to anonymously write down their responses or opinions on the topic. Alternatively, you could summarize for the group some of the opinions from previous focus group discussions, focusing on one or two major contrasting opinions. Still another strategy is to form *sub-groups*, and to get a member of the sub-group to summarize and present the opinions of their sub-group members after which the whole group can still discuss these opinions.
* **Build rapport, empathize**

Observe non-verbal communication. Ask yourself, ‘What are they saying? What does it mean to them?’ Be aware of your own tone of voice, facial expressions, body language, and those of the participants.

* **Avoid being placed in the role of expert**

When asked for **your** ideas or views by a respondent, remember that you are not there to educate or inform. Direct the questions back to the group by saying: ‘What do you think’, ‘What would you do?’ Set aside time, if necessary, after the session to give participants the information they have asked for.

Do not try to comment on everything that is being said. Don’t feel you have to say something during every pause in the discussion. Wait a little and see what happens.

* **Control the rhythm of the meeting, but in an unobtrusive way**

Listen carefully, and move the discussion from topic to topic. Subtly control the time allocated to various topics so as to maintain interest. If participants spontaneously jump from one topic to another, let the discussion continue for a while since useful additional information may surface; then summarize the points brought up and reorient the discussion.

* Take time at the end of the meeting to summarize, check for agreement and thank the participants

Summarize the main issues brought up, check whether all agree and ask for additional comments. Thank the participants and let them know that their ideas have been a valuable contribution and will be used for planning the proposed research, intervention, or health education materials.

* Listen for **additional comments** and spontaneous discussions which occur after the meeting has been closed.

## Functions of the recorder

The recorder should keep a record of the content of the discussion as well as emotional reactions and important aspects of group interaction. Assessment of the emotional tone of the meeting and the group process will enable you to judge the validity of the information collected during the FGD.

Items to be recorded include:

* Date, time, place
* Names and characteristics of participants
* General description of the group dynamics (level of participation, presence of a dominant participant, level of interest)
* Opinions of participants, recorded as much as possible in their own words, especially for key statements
* Emotional aspects (e.g., reluctance, strong feelings attached to certain opinions)
* Vocabulary used - particularly in FGDs that are intended to assist in developing questionnaires or health education materials
* Spontaneous relevant discussions during breaks or after the meeting has been closed

It is highly recommended that a tape-recorder be used to assist in capturing information. Even if a tape-recorder is used, notes should be taken as well, in case the machine malfunctions and so that information will be available immediately after the session for discussion.

If there is no reliable tape-recorder available, it is advisable to have **two recorders**.

A **supplementary role** for the recorder could be to assist the facilitator (if necessary) by drawing his or her attention to:

* missed comments from participants
* missed topics (the recorder should have a copy of the discussion guide during the FGD)

If necessary, the recorder could also help resolve conflict situations within the group that the facilitator finds difficult to handle on her own.

## Number and duration of sessions

* **Number of sessions**

The number of focus group sessions to be conducted depends upon project needs, resources, and whether new information is still coming from the sessions, (that is, whether contrasting views within and between various groups in the community are still emerging). If not, you may stop.

One should plan to conduct at least two FGDs for each sub-group (for example, two for males and two for females). Otherwise you have no way of assessing whether the information you get from the first FGD is representative for that group.

* **Duration**

A focus group session typically lasts up to an hour and a half. Generally the first session with a particular type of group is longer than the following ones because all of the information is new. Thereafter, if it becomes clear that all the groups have a similar opinion on particular topics, the facilitator may be able to move the discussion along more quickly to other topics which still elicit new points of view.

## Processing and analysis of results

* After each focus group session the facilitator and recorder should meet to review and **complete the notes** taken during the meeting. This is the right moment to **evaluate** how the focus group went and what changes might be made in the topics when facilitating the next focus group.

Immediately afterwards a full report of the discussion should be prepared which reflects the discussion as completely as possible, using the participants’ own words. List the key statements, ideas, and attitudes expressed for each topic of discussion.

* After the transcript of the discussion is prepared, **code**, following your topics, the participants’ statements right away, using the left margin. Make finer sub-codes. **Write comments** (your first interpretation of the data) in the right margin. Formulate additional questions if certain issues are still unclear or controversial and include them in the next FGD. Further categories the statements for each topic, if required.
* When you have all the data, **summarize** it in a **compilation sheet** organizing the findings per topic for each. Number the FGD interviews and use key words to summarize group statements in the compilation sheet so that you can always go back to the full statement. If you have different categories of informants, e.g., male and female, you can summarize the information from the male and female groups on two **separate** compilation sheets. You should then to do a **systematic comparison** between groups on all topics. Use your **objectives** and problem analysis diagram as a framework for analysis and comparison.
* The next step could be to put the major findings for different study populations on one sheet. You may want to use some of these sheets in your research report.
* Sometimes you may also wish to use diagrams when summarizing the causes or components of the problem understudy.
* Only now can you report the major findings of the FGDs in a narrative.

## Report writing

Start with a description of the purpose of the FGDs, the selection and composition of the groups of FGD participants and a commentary on the group process, so the reader can assess the validity of the reported findings.

Present your findings, following your list of topics and guided by the objective(s) of your FGD.

Include quotations whenever possible as illustrations, particularly for key statements.

## *Pilot Survey/study/Pilot – testing/Feasibility study*

Is a small experiment designed to test logistics and gather information prior to a larger study to improve the latter’s quality and efficiency. A preliminary piece of research conducted before a complete **survey** to test the effectiveness of the research methodology. It is a preliminary survey carried out to see if a full survey would be worth while. This should be completed before the final survey commences. The intention is to alert the surveyor to any difficulties that were not anticipated at the survey proposal stage.

The action plan created during the planning phase will help the organization conduct the pilot. Other consideration in pilot include:-

* [Set Timeline/Schedule](http://progressoutofpoverty.org/en/conducting-pilot#timeline)
* [Train Field Staff](http://progressoutofpoverty.org/en/conducting-pilot#train)
* [Determine Intake Methodology](http://progressoutofpoverty.org/en/conducting-pilot#determine)
* [Ensure Quality Control](http://progressoutofpoverty.org/en/conducting-pilot#ensure) - supervision

Pilot testing involves conducting a preliminary test of data collection tools and procedures to identify and eliminate problems, allowing programs to make corrective changes or adjustments before actually collecting data from the target population.

A pilot test usually involves simulating the actual data collection process on a small scale to get feedback on whether or not the instruments are likely to work as expected in a “real world” situation. A typical pilot test involves administering instruments to a small group of individuals that has similar characteristics to the target population, and in a manner that simulates how data will be collected when the instruments are administered to the target population.

Pilot testing gives programs an opportunity to make revisions to instruments and data collection procedures to ensure that appropriate questions are being asked, the right data will be collected, and the data collection methods will work. Programs that neglect pilot testing run the risk of collecting useless data. Pilot testing provides an opportunity to detect and remedy a wide range of potential problems with an instrument. These problems may include:

* Questions that respondents don’t understand
* Ambiguous questions
* Questions that combine two or more issues in a single question (double-barreled questions)
* Questions that make respondents uncomfortable

Pilot testing can also help programs identify ways to improve how an instrument is administered. For example, if respondents show fatigue while completing the instrument, then the program should look for ways to shorten the instrument. If respondents are confused about how to return the completed instrument, then the program needs to clarify instructions and simplify this process.

## *Steps of conducting pilot survey*

The following guidelines can be used to conduct a simple pilot test of an instrument.

* 1. Find at least 4 or 5 people from the same group of people whom you will actually measure (the target population).
  2. Arrange for these people to complete the instrument under conditions that match as closely as possible the actual conditions under which the instruments will be administered when you collect performance measurement data for your program. Consider the time of day, the location, and the method. If it is a phone interview, then conduct the pilot test over the phone. If it is a mail survey, then make sure the pilot test is completed via the mail. Whenever possible, record the time it takes for respondents to complete the instrument so that you can inform the data collectors of the approximate time needed for respondents to complete the instrument.
  3. After each respondent completes the instrument, take some time with the respondent to discuss his or her experience. The following are some questions you might want to ask.
     + How long did it take you to complete the instrument?
     + What do you think this instrument is about?
     + For what purposes do you think this information will be used?
     + What problems, if any, did you have completing the instrument?
     + Are the directions clear?
     + Are the instructions clear on what to do with the instrument after completing it?
     + Is there any words/language in the instrument that people might not understand?
     + Did you find any of the questions to be unnecessary or too sensitive?
     + Were any questions difficult to answer?
     + [For specific questions:] What do you think this question is asking?

How would you phrase this question in your own words?

* + - Did the answer choices allow you to answer as you intended?
    - Is there anything you would change about the instrument?
  1. Collect the completed instruments. Read through the responses. Did respondents interpret the questions the way you intended?
  2. Analyze the data and present the results of the pilot test as you would when you actually administer the instrument. Will the results give you the information you need?
  3. Share the results of your pilot test with other stakeholders who will be using the data. Does this instrument provide the data they need to answer their questions?
  4. Modify your instrument based on the information you have gathered.

When reporting performance measurement results, be sure to describe any pilot testing you have done. This gives readers greater confidence in the results you report.

It may be tempting to skip the pilot testing step, but remember that you run the risk of collecting useless data. A pilot test almost always reveals ways to improve an instrument. Once data are collected, it will be too late in the data collection cycle to fix problems that were missed because an instrument was not pilot tested. It may be necessary to carry out a second pilot study to assess the revised main study or in some cases the main study may be abandoned.

**Briefing conference**

Is an informatory meeting attended by all people to be involved in the study. In the meeting:-

* Problem / question for the study is discussed
* Make them understand their responsibility and be accountable
* Discuss the results of pilot survey at length
* Circumstances which may not be overcome
* Issue of tools of study etc

## *Field work*

Here people go the study sites and do the actual study/survey and submit the results

A field checklist is required which shows all the item to be carried along (e.g. camera, pencils questionnaires..) and the procedures/protocol which reached to the field e.g. courtesy call, start with introduction..)

## *Editing and cleaning of data*

Editing and cleaning the data are important steps in data processing that should always precede analysis of collected information. Data editing occur during and after the coding phase. Coders perform some editing by checking for errors and omissions and by making sure that all interview schedules have been completed as required. A supervisor may do this by reviewing each completed question to evaluate reliability of interview and check for consistency of responses e.g checking sure that all filter questions have been correctly marked so that the data will fit into correct skip (or “go to”) pattern.

Editing is basically checking of obvious mistakes, omissions and any other logical entries on the schedules/questionnaire once they begin flowing in the investigation centre.

**Data cleaning** is the proof reading of the data to catch and correct errors and inconsistent codes. Another function of data cleaning is checking for wild codes e.g. the question; “Do you believe there is life after death?” may have legitimate codes of 1 for “yes”. 2 for “no”, 8 for “undecided” and 9 for “no answer”. Any code other than these four would be considered illegitimate.

The interviewer may take back (call back) some of the interviewer to the respondents if any mistake is noted

## *Coding*

Is the arrangement of information to a pre-designed format ready to be processed. Coding is the process by which responses are classified into meaningful categories; the initial rule of coding is that the numbers assigned must make intuitive sense. For example, higher scores on a variable should be assigned to high codes than lower scores. This is most demonstrated in interval-level variables, e.g a person who is older than another should receive a higher code on age. The processing may be mechanical/manual or automated using various coding devices e.g. transfer sheet, edge coding, and optical scanning

## *Rules of coding*

1. Code number should make intuitive sense for variables that can be rank order i.e. highest should be assigned higher codes
2. The coding categories must be mutually exclusive – each unit of analysis should fit into one and only one category
3. The coding scheme must be exhaustive – every response must fit into a category with few responses being classified as “other”
4. Categories must be specific enough to capture differences using the smallest possible number of categories- the criterion of detail

## *Classification*

Is the process of keeping data with common characteristics together e.g. you may need to put data concerning males, females, children separate from each other. This enables to analyze data concerning specific character or items more easily.

## Tabulation

Is the process in which one extracts information either from questionnaire or schedules or the data that has been classified and is entered into separate summary sheet.

The main purpose of tabulation is:-

* Simplification of details from different classification
* To allow the silent/prominent features to be seen clearly
* Facilitate interpretation of the assembled data

## Common Stages of survey or Data Collection

Survey is a process of obtaining data on a problem so that attempts may be made to solve it. It is also referred to as data collection process/investigation. When the data is about the well being of people (health) it is referred to as health survey. There are many data collection process that seek to collect information on specific problems so that they may be solved. The survey would include:

1. Demographic survey
2. Economic survey
3. Social survey
4. Morbidity survey

Stages refer to flow or order under which data is collected (data collection process) follows. There are commonly agreed stages in data collection process (survey/investigation). The depth of each stage depends on data collector, the purpose (use) of which data is made and the knowledge of the data collector.

A good survey is usually done with an aim of collecting compete accurate and timely data on the subject problem/problem under investigation. Most surveys are aimed at accessing individual views on social matter, economic issues, political issues etc.

The commonly agreed stages include:

1. Problem analysis – identification, statement of the problem, literature review, objectives, hypothesis& proposal writing. This is a major subject and stage of the survey and has to be clearly known and understood by the data collector. This is accomplished through literature review which is wide reading on subject (problem) and determining the available data on it (information in data) and the various possible sources of assistance towards survey and solving of the problem at hand.
2. Identifications of data collection sources (entire population/ primary sources or existing records/secondary sources (identify the actual source of data). One should be able to know exactly where and which people to contact to obtain the data i.e. the entire population or existing records.
3. Determination of data collection method, thus choose which method you are going to use. One is required to select or identify a method to be used in data collection. The method used should be compatible with the problem of study, data source and sample selected.
4. Selection of sample(s) where required thus using any of the various sampling methods. This is done where the data to be collected (required) is obtained from part of the population. It is required that any of the sampling method that is suitable for the situation so as to identify the sample.
5. Decide and designing data collection tools/instruments (questionnaire/interview schedule). Involves determining the document (tool) which will assist in tapping or carrying data from the determined source. The tool selected should be compatible with the selected source (sample) and method of data collection. In designing the tool one should observe all the principles required as discussed previously.
6. Pilot survey. This is a small scale survey, designed to test the tool for data collection, method, source and population chosen. It is usually the actual survey but on the small-scale, thus the expected data collection process being tested to be sure that when it is done it will give accurate and reliable results. It is usually recommended that this be done by people who have wider knowledge and skill in data collection process (surveys) as well as experience. Certain surveys demand that expert services be contracted (obtained) for the stage. Usually the No. of items included is not the major aim but the technical taste of the process. It is required that all the detected shortfalls/comings should be noted and discussed and where correction is due it is done during this stage. This stage is also said to be giving clearance for the data collection commencement.
7. Briefing conference is a stage in which the process of the survey is discussed, major requirements and where assistants are to be used in the survey they are trained/informed. It is necessary that all the people being involved in the survey understand what they are expected to do where seminars or short trainings may be applied. All documents to be used in the survey are clarified, payments to be made are discussed and agreed upon, contracts with research assistants are signed, and relevant stationery and materials are issued. It is required that difficulties faced in the pilot survey are discussed (included in the discussion).
8. Field work is the actual data collection process where people would be required to go to their interview points, sending out of questionnaires or looking for individual quotas and doing the actual interview.
9. Editing is the correction of the simple obvious mistakes that may have occurred on the questionnaire or interview schedule i.e. where the name may refer to a female respondent/interviewee and the sex ticked is male. It is recommended that it is done immediately so that where a call back to the interviewee is necessary is done promptly.
10. Classification, coding and processing of information that has be received. This stage seeks to put the collected data into desired class or groups where specific marks or codes are required on the collected data they are inserted or made in this stage. The desired classes may be structured on geographical regions, sex, age, occupation, social classes or citizenship etc.
11. Calculations, presentation and storage of data/information. This stage allows for desired calculations to be made and ends up with specific portions of information such as mean, median, mode, percentages, rate, ratio etc. and where dramatic presentations are required they are made at this stage.
12. Presentation of the results and evaluation whether the aim has been achieved or solved. Also known as final report writing, recommendation and evaluation. This is a comprehensive and a detailed account on what was done on survey (data collection process). In this stage we would explain what was done at every stage and the final findings of the survey. Recommendations are made as a result of findings of the survey and a conclusion made in relation to the problem indicating if there is any solution obtained out of data collection. It is advisable booklets are produced and copies made available to all the people that were considered as major participants in data collection process.

# UNIT II (CHS 1.2) Data Presentation

1. Definition and description of different methods of data tabulation
2. Principles of table construction
3. Simple tables
4. Multiple column tables
5. Contingency tables
6. Array presentation
7. Frequency distribution tables
8. Cumulative frequency tables
9. Definition, purpose, importance of data
10. Description of narrative presentation
11. Different ways of presentation of information through charts and graphs
12. Simple bar charts
13. Component bar charts
14. Percentage component bar charts
15. Multiple bar charts
16. Pie charts
17. Principles of graph construction
18. Histograms
19. Frequency polygon curve
20. Ogive
21. Break-even charts
22. The Z charts
23. Lorenz charts

# Definition of Tables

A table is a group of figures systematically arranged in the form of rows and columns. Rows are horizontal arrangements whereas columns are vertical ones. The purpose of the table is to simplify the presentation and to facilitate comparisons.[[13]](#footnote-13) The form of arrangement the table takes depends on:

* The number of figures
* The object of the table

## Parts of a Table

Number of parts of table varies from case to case depending on data. However the main parts include:

1. Table Number – each table should be numbered
2. Title of the table – each table should be given suitable title which should be a description if the contents of the table. The title should answer the questions what, where, and when in that sequence
   1. What precisely are the data in the table
   2. Where the data occurred
   3. When the data occurred

The title should be clear, brief and self-explanatory

1. Caption – refers to the column headings. It explains what the column represents. It may consist of one or more column headings. Under a column heading there may be sub-heads. The caption should be clearly defined and placed at the middle of the column.
2. Stub – are designations of the rows or row headings. They are at the extreme left and perform the same function for horizontal rows of numbers in the table as the column headings do to the vertical columns of numbers. Stubs are usually wide than column headings but should kept as narrow as possible.
3. Body – it contains the numerical information. Data presented in the body arranged according to description are classifications of the captions and stubs.
4. Headnote – is a brief explanatory statement applying to all or major parts of the material in the table, and is placed below the point centered and enclosed in brackets. It is used to explain certain points relating to the whole table that have not been included in the title nor in the captions and stubs. For instance the unit of measurement is frequently written as a headnote, such as “in thousands” or “in million tonnes” or “in crores”, etc.
5. Footnotes – anything in a table which the reader may find difficult to understand from the title, captions and stubs should be explained in footnotes. They are used:
   1. To point out any exceptions as to the basis of arriving at the data
   2. Any special circumstances affecting data
   3. To clarify anything in the table
   4. To give the source in case of secondary data

## Types of Tables

* Simple and complex tables
* General purpose and special purpose (or summary) tables.

## Simple and Complex tables

Simple table only one characteristic is shown thus the table is known as one-way table. Complex table two or more characteristics are shown. Such tables are more popular in practice because they enable full information to be incorporated and facilitate a proper consideration of all related facts. When two characteristics is shown such a table is known as two-way table (double tabulation). When three characteristics are shown such a table is referred to as treble tabulation. When four or more characteristics are simultaneously shown it is a case of manifold tabulation.

1. **Simple table (one-way)**

STUDENT PERFOMANCE BY SUBJECT

|  |  |  |
| --- | --- | --- |
| S/No | Subject | Mean Score |
| 1 | Health Records Science | 74 |
| 2 | Medical Data Classification | 49 |
| 3 | Health Statistics | 54 |
| 4 | Anatomy, Physiology and Pathology | 55 |
| 5 | Health Information Systems | 94 |
| 9 | Personnel Management | 56 |
| 10 | Community Health | 46 |
|  | **Class Mean Score** | **61** |

1. **Two-way table**

A two-way table shows two characteristics and is formed when either the stub or the caption is divided into two coordinate parts.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Employees | | Total |
| Age (in Years) | Males | Females |  |
| Below 25 | ---- | ----- | ---- |
| 25 – 35 | ---- | ----- | ---- |
| 35 – 45 | ---- | ----- | ---- |
| 45 – 55 | ---- | ----- | ---- |
| Above 55 | ---- | ----- | ---- |
| Total | ---- | ----- | ---- |

1. **Higher order Table**

When three or more characteristics are represented in the same table, such a table is called higher order table.

## General and Special Purpose Tables

General tables are also known as reference tables or repository tables provide information for general use or reference. They serve as repository of information and are arranged for easy reference. Special purpose table also known as summary tables, provide information for particular discussion. When attached to a report they are found in the body of the text.

## Principles of Table Construction

The basic principle for table construction is

Construct the table so that it achieves its object in the best manner possible

1. The table should be simple
2. The table must have a comprehensive, explanatory title
3. The source must be stated
4. Units must be clearly stated
5. The headings to columns and rows should be unambiguous
6. Double counting should be avoided
7. Totals should be shown where appropriate
8. Distinctive rulings should be used appropriately e.g. double lines, heavy single lines
9. Footnotes should be used to qualify or clarify the table.[[14]](#footnote-14)

## Importance of Tabulation

Tables make it possible for the analyst to present a huge mass of data in a detailed orderly manner within minimum space. Thus tabulation:

1. Simplifies complex data
2. Facilitates comparison
3. It gives identity to data
4. Reveals patterns within figures which cannot be seen in the narrative form

# SEMESTER III

# UNIT III (CHS 1.3) Measures of Central Tendency

1. Calculation of different measures of central tendency
2. Simple arithmetic mean
3. Weighed arithmetic mean
4. Arithmetic mean of a frequency distribution
5. Arithmetic mean of a grouped frequency distribution
6. Mean deviations from the mean
7. Finding mode using histogram
8. Obtaining mode of ungrouped/grouped data
9. The geometric mean

## Data

Data is defined as information organized for analysis or used as the basis for a decision; numerical information. They are those facts any particular situation affords or gives to an observer. According to some sources data are raw facts and figures that are meaningless in and of themselves, and refer to information as meaningful data meaning knowledge resulting from processing data.[[15]](#footnote-15)

## Types of data

1. Categorical data (Qualitative)
2. Measurement/Numerical data (Quantitative)

## Categorical Data

* The objects being studied are grouped into categories based on some qualitative trait.
* The resulting data are merely labels or categories.

Examples: Categorical Data

* Hair color - Blonde, brown, red, black, etc.
* Marital status - Married , divorced and separated
* Smoking status - smoker, non-smoker



**Nominal Data-** A type of categorical data in which objects fall into *unordered* categories.

Examples: Nominal Data

* Pregnancy status - Pregnant, not pregnant. .
* Blood type - A, B, AB, O.
* Smoking status - smoker, non-smoker

**Ordinal Data**- A type of categorical data in which *order* is important.

Examples: Ordinal Data

* Lecturer - assistant lecturer, senior lecturer, professor
* Degree of illness - mild, moderate, severe
* Cancer stage - stage I, stage II, stage III

**Binary Data -** A type of categorical data in which there are *only two categories*.

* Binary data can either be nominal or ordinal.

Examples: Binary Data

* Smoking status - smoker, non-smoker
* Attendance - present, absent
* Class - lower classman, upper classman

**Measurement/Numerical Data -** The objects being studied are “measured” based on some quantitative traits. The resulting data are set of numbers.

Examples: Measurement Data

* Cholesterol level
* Height
* Age
* SAT score
* Number of students late for class
* Time to complete a homework assignment



**Discrete Measurement Data**- Only certain values are possible (there are gaps between the possible values).

**Continuous Measurement Data**-Theoretically, any value within an interval is possible with a fine enough measuring device.

Examples of Discrete Measurement Data

* SAT scores
* Number of students late for class
* Number of crimes reported to SC police
* Number of times the word number is used

***Note****: Generally, discrete data are counts.*

Examples of Continuous Measurement Data

* Cholesterol level
* Height
* Age
* Time to complete a homework assignment

***Note****: Generally, continuous data come from measurements.*

Note- The type(s) of data collected in a study determine the type of statistical analysis used and graphical display.

* Categorical data are commonly summarized using “percentages” (or “proportions”).
  + 11% of students have a tattoo
  + 2%, 33%, 39%, and 26% of the students in class are, respectively, freshmen, sophomores, juniors, and seniors
* Measurement data are typically summarized using “averages” (or “means”).
  + Average number of siblings students have is 1.9.
  + Average weight of male students is 173 kg.
  + Average weight of female students is 138 kg.
* **A variable** is the general characteristic being measured on a set of people, objects, or events, the members of which may be on different values. It denotes variation, as opposed to a constant. e.g. height, gender, etc
* **Discrete Vs continuous variables** - variables may be discrete or continuous. A discrete variable is one that can assume only a countable number of values between any two points eg 0, 1, 2, 3, 4 . There is no 1.5 goats etc; goals scored by a team etc.
* **A continuous variable** is one that theoretically can assume an infinite number of values between any two points on the measurement scale e.g. weight can be in any number of decimal places; time can even be in microseconds; etc, making space between two full numbers have an infinite number of sub-numbers, all of which are valid. However, even these variables such as weight we don’t go for the infinite numbers, but define the number of decimal points such that the approach tends to be discrete.
* In statistics, a **variable** has two defining characteristics:
  + A variable is an attribute that describes a person, place, thing, or idea. The value of the variable can "vary" from one entity to another.
  + For example, a person's *hair color* is a potential variable, which could have the value of "blond" for one person and "brunette" for another.
* **A constant** is a quantity that **does not** change its value within a given context, e.g. pie, (22/7); ratio of converting cm to inches being 2.54; etc.

## Frequency Distribution

A ***Frequency Distribution*** organizes data into classes, or categories, with a count of the number of observations that fall into each class

* **Class Limits; -** smallest and largest observed values that can belong to a class
* Boundaries; - actual values that separate successive classes
* Intervals; - the distance spanned by the boundaries of a class
* Class Midpoint - the arithmetic mean of **its class boundaries**

**Steps for Constructing a Frequency Distribution**

1. Array the data values in order by size from lowest to highest (or vice versa);
2. Compute the range;
3. Divide the range into a convenient number of class intervals of equal size;
4. Count the number of observations in each class to determine the total frequency; and
5. Display the class intervals with their frequencies.

There are two main types of measures of frequency distributions: measures of central tendency and measures of variability/dispersion[[16]](#footnote-16). Measures of central tendency measure location, or focus on the typical value of a data set. On the other hand variation emphasizes difference, meaning measures of variability measure distance and/or dispersion around the typical value of data set.

## MEASURES OF CENTRAL TENDENCY

Measures of central tendency summarize the typical value of a variable. They yield information about the center, or middle part, of a group of numbers. Measures of central tendency can yield information on average score, the middle score, and the most frequently occurring score. Worth noting is that MCT do not focus on the span of the data set or how far values are from the middle numbers. When we think of measures of central tendency, we usually think of averages; however, averages or arithmetic mean may not always be the most appropriate way of summarizing the most typical value of a data set. There are three major measures of central tendency, whereby the selection of the relevant measure of central tendency is related to the scale of measurement used in data collection[[17]](#footnote-17). The central tendency of a dataset, i.e. the center of a frequency distribution, is most commonly measured by **the 3 Ms:** i.e. **mean, median and mode**.

**Central Location / Position / Tendency** - a single value that represents (is a good summary of) an entire distribution of data

1. Mode (*M*o) – defined as the most frequently occurring observation in a frequency distribution; it is the only measure of central tendency that is appropriate for nominal data. Easy to determine, but subject to variation & of limited value. In the world of business mode is often used to determine sizes. Where two modes are listed in a data set the data is said to be bimodal, while where data set has more than two the data is said to be multimodal.
2. Median(*M*D) – is the midpoint of a frequency distribution; it is appropriate for ordinal and metric data. It is the middle value in an ordered array of numbers. For an array with an odd number of items, the median is the middle number, while in an array with even numbers median will be average of two middle numbers. In dealing with large amount of data it is advisable to use the formula (n+1)/2. The median is unaffected by the magnitude of extreme values. Half the variables have values greater than the median and the other half values which are less. The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions, e.g. family income.
3. Mean ( ) (pronounced “x-bar”) – is arithmetic average; it is appropriate for metric data. It is the average of a group of numbers and is computed by summing all numbers and dividing by number of numbers. The arithmetic mean is also referred to as the mean. The population mean is represented by the Greek letter mu (*µ*) and the sample mean is represented by ( ). The formulas for calculating the population mean and the sample mean are:

*X)*

## Formula for the arithmetic mean:

*Population Mean µ* = ∑ (xi) = x1 + x2 + x3 + x4 ……… + xN

*N* *N*

*Sample Mean =* ∑ (xi) = x1 + x2 + x3 + x4 ……… + xn

*X)*

*n* *n*

img2 Formula for ***Frequency Distribution***:

Mean is therefore, the average value of a dataset, i.e. the sum of all the data divided by the number of variables. **Average:** An average is a single value which is a representative of a given set of values. The **arithmetic mean** is commonly called the **"average"**. When the word "mean" is used without a modifier, it usually refers to the arithmetic mean. The mean is a good measure of central tendency for **symmetrical (e.g. normal) distributions** but can be misleading in skewed distributions since it is influenced by outliers. In general, the mean is larger than the median in positively skewed distributions and less than the median in negatively skewed distributions.

****

Consider a group of people with ages ranging from the teens to the 80s. This histogram displays the frequency distribution by 10-year age group. Well, if you had to present a single number that best characterized this age distribution, what would it be? That is the function of a measure of central location. Later we will also talk about measures that reflect the spread of the distribution.

**Example # 1**

**The Mean**

To calculate the mean, we need to add all the values up and divide by the number of values.

|  |  |  |  |
| --- | --- | --- | --- |
| 5 + 9 + 12 + 4 + 5 + 14 + 19 + 16 + 3 + 5 + 7 | = | 99 | = 9 |
| p11 | p11 |
| 11 | 11 |

In this case the *mean* is 9 which is one of the values in the list. Sometimes the mean will not appear in the original list. It might even be a decimal value.

**The Median**To calculate the median, we need to put the numbers in order and find the middle value.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3 | 4 | 5 | 5 | 5 | |  | | --- | | **7** | | 9 | 12 | 14 | 16 | 19 |

Here the *median* is 7 because this is the middle value. Half of the other values in the list are below 7 and half are above 7.

**The Mode**   
To calculate the mode, we need to look at which value appears the most often. It can help if the numbers are in order.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3 | 4 | |  |  |  | | --- | --- | --- | | **5** | **5** | **5** | | 7 | 9 | 12 | 14 | 16 | 19 |

In this list the *mode* is 5, because it appears most often. Sometimes there will be more than one mode, because two or more values appear the same number of times.

For the list: **"5, 9, 12, 4, 5, 14, 19, 16, 3, 5, 7"**, the mean is 9, the median is 7 and the mode is 5.

**Finding the median where there are an even number of values**

When there is an even number of values, there is no clear middle value.

For example, what is the median of: 3, 6, 7, 8, 11, 15?

In this case, there are two middle values.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 3 | 6 | |  |  | | --- | --- | | **7** | **8** | | 11 | 15 |

The median is the *mean* of these two middle numbers.

|  |  |
| --- | --- |
| 7 + 8 | = 7.5 |
| p11 |
| 2 |

So the median for this set of values is **7.5**.

|  |  |
| --- | --- |
| **Mean:** | **Average.** The sum of a set of data divided by the number of data. (Do not round your answer unless directed to do so.) |
| **Median:** | The **middle** value, or the mean of the middle two values, when the data is arranged in numerical order.   Think of a "median" being in the middle of a highway. |
| **Mode:** | The value (number) that appears the **most**. It is possible to have more than one mode, and it is possible to have no mode.  If there is no mode-write "no mode", do not write zero (0). |

Consider this set of test score values:

|  |  |
| --- | --- |
| meanpic1 Normal listing of scores. | meanpic2 Scores with the lowest score replaced with outlier. |

The two sets of scores above are identical except for the first score.  The set on the left shows the actual scores.  The set on the right shows what would happen if one of the scores was WAY out of range in regard to the other scores.  Such a term is called an **outlier.** *With the outlier, the mean changed. With the outlier, the median did NOT change.*  **How do I know which measure of central tendency to use?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | **MEAN** Use the mean to describe the middle of a set of data that *does not* have an outlier. **Advantages:    • Most** popular measure in fields such as business, engineering and computer science.    • It is unique - there is only one answer.    • Useful when comparing sets of data.  **Disadvantages:    •** Affected by extreme values (outliers) | |  | | |  | | --- | | **MEDIAN** Use the median to describe the middle of a set of data that *does* have an outlier. **Advantages:    • Extreme** values (outliers) do not affect the median as strongly as they do the mean.    • Useful when comparing sets of data.    • It is unique - there is only one answer. **Disadvantages:    • Not** as popular as mean. | |

|  |
| --- |
| **MODE** Use the mode when the data is non-numeric or when asked to choose the most popular item.  **Advantages:    • Extreme** values (outliers) do not affect the mode.  **Disadvantages:**Not as popular as mean and median.   Not necessarily unique - may be more than one answer   When no values repeat in the data set, the mode is every value and is useless.   When there is more than one mode, it is difficult to interpret and/or compare. |

**Advantages of Arithmetic Mean:**•Easy to understand   
•Simple to compute

•Based on all the observation

•Uniquely defined

**Disadvantages of Arithmetic Mean:**

•Affected by extreme value

•Unable to compute mean for open-ended classes

•Tedious to compute

|  |  |
| --- | --- |
| thinkingguy | **What will happen to the measures of central tendency if we add the same amount to all data values, or multiply each data value by the same amount?** |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Data | Mean | Mode | Median | | Original Data Set: | 6, 7, 8, 10, 12, 14, 14, 15, 16, 20 | 12.2 | 14 | 13 | | Add 3 to each data value | 9, 10, 11, 13, 15, 17, 17, 18, 19, 23 | 15.2 | 17 | 16 | | Multiply 2 times each data value | 12, 14, 16, 20, 24, 28, 28, 30, 32, 40 | 24.4 | 28 | 26 |   **When added:**  Since all values are **shifted** the same amount, the measures of central tendency all shifted by the same amount.  If you add 3 to each data value, you will add 3 to the mean, mode and median.  **When multiplied:** Since all values are affected by the same multiplicative values, the measures of central tendency will feel the same affect.  If you multiply each data value by 2, you will multiply the mean, mode and median by 2. | |

**Example #2**   
Find the **mean, median and mode** for the following data:  5, 15, 10, 15, 5, 10, 10, 20, 25, 15.   
                             (You will need to organize the data.)  
                                      **5, 5, 10, 10, 10, 15, 15, 15, 20, 25**

|  |
| --- |
| **Mean:**      measur1  **Median:**     5, 5, 10, 10, 10, 15, 15, 15, 20, 25            Listing the data in order is the easiest way to find the median.           The numbers 10 and 15 both fall in the middle.   Average these two numbers to get the median.      10 + 15 = **12.5**                                                                               2  **Mode:** Two numbers appear **most often:**  **10 and 15.**                   There are three 10's and three 15's.                   In this example there are two answers for the mode. |

**Example #**   3     
For what value of  ***x*** will  8 and *x* have the same mean (average) as 27 and 5?

|  |  |
| --- | --- |
| First, find the mean of 27 and 5:  27 + 5 = 16      2 | Now, find the *x* value, knowing that the average of *x* and 8 must be 16:  *x* + 8 = 16                   2  32 = *x* + 8     cross multiply -8         - 8 24 = *x*             and solve |

**Example #4 :**                                                         
On his first 5 biology tests, Bob received the following scores:  **72, 86, 92, 63,** and **77**.  What test score must Bob earn on his sixth test so that his average (mean score) for all six tests will be 80?  Show how you arrived at your answer.   
**Possible solution:** Set up an equation to represent the situation.  Remember to use all 6 test scores:    
                                             72 + 86 + 92 + 63 + 77 + ***x***   =  80  
                                                                6  
 cross multiply and solve:     (80)(6) = 390 + *x*  
                                                          480 = 390 + *x*  
                                                        - 390   -390  
                                                            90 =          *x  
                                         Bob must get a 90 on the sixth test.*

**Example #5**   
The mean (average) weight of three dogs is 38 pounds.  One of the dogs, Sparky, weighs 46 pounds.  The other two dogs, Eddie and Sandy, have the same weight.  Find Eddie's weight.  
  
Let *x* = Eddie's weight        (they weigh the same, so they are both represented by "*x*".)  
Let *x* = Sandy's weight                            
  
Average:   sum of the data divided by the number of data.   
                  *x* + *x* + 46 = 38                 cross multiply and solve  
                      3(dogs)                                    
                 (38)(3) = 2*x* + 46   
                      114 = 2*x* + 46   
                       -46          -46    
                        68 = 2*x*    
                         2      2   
                        34 = *x*     Eddie weighs 34 pounds.

|  |
| --- |
| **Weighted Arithmetic Mean**  In calculation of arithmetic mean, the importance of all the items was considered to be equal. However, there may be situations in which all the items under considerations are not equal importance. For example, we want to find average number of marks per subject who appeared in different subjects like Mathematics, Statistics, Physics and Biology. These subjects do not have equal importance. If we find arithmetic mean by giving Mean. |

Thus, ***arithmetic mean computed by considering relative importance of each items is called weighted arithmetic mean.*** To give due importance to each item under consideration, we assign number called weight to each item in proportion to its relative importance. Weighted Arithmetic Mean is computed by using following formula:   
            **clip_image002**

The **weighted mean** is similar to an [arithmetic mean](http://www.answers.com/topic/arithmetic-mean) (the most common type of [average](http://www.answers.com/topic/average)), where instead of each of the data points contributing equally to the final average, some data points contribute more than others. The notion of weighted mean plays a role in [descriptive statistics](http://www.answers.com/topic/descriptive-statistics) and also occurs in a more general form in several other areas of mathematics.

If all the weights are equal, then the weighted mean is the same as the [arithmetic mean](http://www.answers.com/topic/arithmetic-mean). The term **weighted average** usually refers to a weighted arithmetic mean.

**Example # 1**

A student obtained 40, 50, 60, 80, and 45 marks in the subjects of Math, Statistics, Physics, Chemistry and Biology respectively. Assuming weights 5, 2, 4, 3, and 1 respectively for the above mentioned subjects.

Find Weighted Arithmetic Mean per subject.

**Solution:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Subjects** | **Marks Obtained** (**x)** | **Weight** (**w)** | **Wx** |
| **Math** | 40 | 5 | 200 |
| **Statistics** | 50 | 2 | 100 |
| **Physics** | 60 | 4 | 240 |
| **Chemistry** | 80 | 3 | 240 |
| **Biology** | 45 | 1 | 45 |
| **Total** |  | clip_image042 | clip_image044 |

Now we will find weighted arithmetic mean as:   
            clip_image046 marks/subject.

If w1, w2, w3, …, wn are the weights assigned to the values x1, x2, x3, …, xn respectively, then the weighted average is defined as:

 Weighted Arithmetic Mean = w1x1+w2x2+...+wnxn / w1+w2+...+wn.

***Example # 2:*** A group of 10 items has mean 6. If the mean of 4 of these items is 7.5, then the mean of remaining items is

***Solution:***         Sum of all the 10 items = 10 x 6 = 60

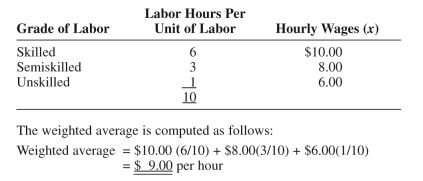
                         Sum of four of these items 4 x 7.5 = 30

                         ∴ sum of the remaining six items = 60 –30 = 30

                         hence, the mean of remaining six items = 30/6 = 5

***Example # 3:***

The formula for a weighted average is Weighted Average = S*wx,* where *x* = the data values and *w* = relative weight assigned to each observation, expressed as a percentage or relative frequency. For example, assume the XYZ Company uses three grades of labor to produce a finished product as follows:

****

The weighted average is computed as follows:

Weighted average = $10.00 (6/10) + $8.00(3/10) + $6.00(1/10) = $ 9.00 per hour

***Example # 4:***

Given two school classes, one with 20 students, and one with 30 students, the grades in each class on a test were:

Morning class = 62, 67, 71, 74, 76, 77, 78, 79, 79, 80, 80, 81, 81, 82, 83, 84, 86, 89, 93, 98

Afternoon class = 81, 82, 83, 84, 85, 86, 87, 87, 88, 88, 89, 89, 89, 90, 90, 90, 90, 91, 91, 91, 92, 92, 93, 93, 94, 95, 96, 97, 98, 99

The straight average for the morning class is 80 and the straight average of the afternoon class is 90. The straight average of 80 and 90 is 85, the mean of the two class means. However, this does not account for the difference in number of students in each class, and the value of 85 does not reflect the average student grade (independent of class). The average student grade can be obtained by averaging all the grades, without regard to classes:

\bar{x} = \frac{4300}{50} = 86.

Or, this can be accomplished by weighting the class means by the number of students in each class (using a weighted mean of the class means):

\bar{x} = \frac{(20)80 + (30)90}{20 + 30} = 86.

Thus, the weighted mean makes it possible to find the average student grade in the case where only the class means and the number of students in each class are available.

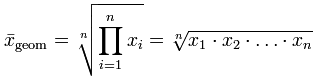
**Geometric Mean**

Means are mathematical formulations used to characterize the central tendency of a set of numbers. Most people are familiar with the "arithmetic mean", which is also commonly called an average. Geometric mean has the specific definitions below, and has utility in science, finance, and statistics.   
  
**Mathematical definition:** The nth root of the product of n numbers.   
**Practical definition:** The average of the logarithmic values of a data set, converted back to a base 10 number.

Geometric mean is a kind of average of a set of numbers that is different from the arithmetic average. The geometric mean is well defined only for sets of positive real numbers. This is calculated by multiplying all the numbers (call the number of numbers “n”), and taking the nth root of the total. A common example of where the geometric mean is the correct choice is when averaging growth rates.

The **geometric mean**, in [mathematics](http://en.wikipedia.org/wiki/Mathematics), is a type of [mean](http://en.wikipedia.org/wiki/Mean) or [average](http://en.wikipedia.org/wiki/Average), which indicates the central tendency or typical value of a set of numbers. It is similar to the [arithmetic mean](http://en.wikipedia.org/wiki/Arithmetic_mean), which is what most people think of with the word "average", except that the numbers are multiplied and then the [*n*th root](http://en.wikipedia.org/wiki/Nth_root) (where n is the count of numbers in the set) of the resulting [product](http://en.wikipedia.org/wiki/Product_%28mathematics%29) is taken.

For instance, the geometric mean of two numbers, say *2* and *8*, is just the [square root](http://en.wikipedia.org/wiki/Square_root) of their product; that is *2√2 × 8 = 4*. As another example, the geometric mean of three numbers *1*, *½*, *¼* is the [cube root](http://en.wikipedia.org/wiki/Cube_root) of their product (1/8), which is *1/2*; that is *3√1 × ½ × ¼ = ½* .

**Formula**:  
**Geometric Mean**   
  
**Arithmetic Mean**   
Formula arithmetic mean - sengpielaudio

**Geometric Mean :**   
Geometric Mean = ((X1)(X2)(X3)........(XN))1/N   
where   
              X = Individual score   
              N = Sample size (Number of scores)   
The geometric mean of a data set \{a_1,a_2 , \ldots,a_n\}is given by:

\bigg(\prod_{i=1}^n a_i \bigg)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}.

The **geometric mean** is the nth root of the product of the scores, e.g: the geometric mean of the scores: 1, 2, 3, and 4 is the 4th root of 1 \* 2 \* 3 \* 4 which is the 4th root of 24 = 2.21. The geometric mean is less affected by extreme values than the arithmetic mean and is useful for some positively skewed distributions.

**USAGES**

In social sciences, we frequently encounter this in a number of ways. For example, the human population growth is expressed as a percentage, and thus when population growth needs to be averaged, it is the geometric mean that is most relevant.

In surveys and studies too, the geometric mean becomes relevant. For example, if a survey found that over the years, the economic status of a poor neighborhood is getting better, they need to quote the geometric mean of the development, averaged over the years in which the survey was conducted. The arithmetic mean will not make sense in this case either.

In economics, we see the percentage growth in interest accumulation. Thus if you are starting out with a sum of money that is compounded for interest, then the mean that you should look for is the geometric mean. Many such financial instruments like bonds yield a fixed percentage return, and while quoting their “average” return, it is the geometric mean that should be quoted.

## How to Group Data

**How to Select a Class Interval? Some Rules of Thumb!**

On your exam, you may have to construct a frequency distribution. Constructing a frequency distribution is the same thing as grouping data. The first step in grouping data is deciding how large of a class interval to use.

**(Class interval = Class size)**

There are 2 formulas for determining the appropriate class interval. **You must be able to choose which one would be appropriate for any given problem**

1. **Class interval = Highest value – Lowest value**

**Number of classes**

***(Use when the problem states the number of classes to be used)***

1. **Class interval = Highest value – Lowest value**

**1+ 3.322 Log (N)**

***(Use when the problem does not state the number of classes to be used)***

Therefore from formula 2 above we can say **desired classes = 1+ 3.322 Log (N)**

**Example**

A distribution with 169 items and a range of 72

Desired classes = 1+ 3.322 Log (N) =1+ 3.322 Log (169)

= 1+ 7.401039 = 8 or 9 classes

**Note**

* Select a class interval that allows from 6 to 15 classes. Too many classes can destroy the summary effect of the grouping; too few classes can produce oversimplification of the data and result in inaccuracies from subsequent calculations
* The number of classes, ***k***, should be the smallest integer such that 2*k* > n, where n is the number of observations.

## The Two Firm Rules in Grouping Data:

* The All-Inclusive Rule: classes must be All-Inclusive. All-inclusive classes are classes that together contain all the data.
* The Mutually-Exclusive Rule: classes must be mutually exclusive. Classes must be arranged such that every piece of data can be placed in only one class

There are two types of frequency distribution namely:

* Discrete series – various units are capable of exact measurement and each data unit is separate and complete
* Continuous series – statistical units arranged in groups or classes because they are not exactly measurable and are only approximations
  + Inclusive form – both limits are included while taking the items in a group e.g. 10 – 20 both values (10 and 20) are included
  + Exclusive form – the lower limit is included but upper limit is excluded while taking the items in a group e.g. 10 – 20, 20 – 30 in the first group 10 or more than 10 but below 20 will be included, and in the second group 20 or more than 20 but below 30 will be included.

**Class Midpoint**

Each class has a lower limit and an upper limit. Class midpoint, Mi, is the arithmetic mean of the two limits.

Mi = (lower limit + upper limit) / 2

**Sample Mean**



Where, fi is the frequency of the ith class, and Mi is the midpoint of the ith class.

**Arithmetic mean of a frequency distribution**For a frequency distribution,

 (a)        A.M = Σfx/Σf (Mean of grouped data)

(b)        **Short Cut Method/Assumed mean method**

             A.M. = A + Σd/n, where d = x – A, where A is the assumed mean **or**

A.M. = A + Σfd/Σf, where d = x – A, where A is the assumed mean

(c)        **Step Deviation Method**

A.M. = A + (Σfu/Σf) h, where A is the assumed mean, h is the classs interval and u = (x-A)/h.

Given a list, by **frequency distribution** with a **class interval** of  we mean a table where the numbers are sorted by class intervals of .

* The **frequency**  of a class interval is the number of numbers in the list that belong to the class interval.
* The **Midpoint ** of a group: for example the midpoint of the class interval  is

**Example # 1**



It may sound complicated but it is in fact rather straightforward. Let’s have a look at the frequency distribution of our example, with a class interval of, starting with 12.0 – 12.9:

| Class interval of 1% | Frequency | Midpoint of the group |  |
| --- | --- | --- | --- |
| 12.0 – 12.9 | 1 | 12.45 | 12.45 |
| 13.0 – 13.9 | 0 | 13.45 | 0 |
| 14.0 – 14.9 | 1 | 14.45 | 14.45 |
| 15.0 – 15.9 | 1 | 15.45 | 15.45 |
| 16.0 – 16.9 | 4 | 16.45 | 65.8 |
| 17.0 – 17.9 | 6 | 17.45 | 104.7 |
| 18.0 – 18.9 | 8 | 18.45 | 147.6 |
| 19.0 – 19.9 | 3 | 19.45 | 58.35 |
| 20.0 – 20.9 | 3 | 20.45 | 61.35 |
| 21.0 – 21.9 | 1 | 21.45 | 21.45 |
| 22.0 – 22.9 | 1 | 22.45 | 22.45 |
| 23.0 – 23.9 | 1 | 23.45 | 23.45 |
| **Total** | **30** |  |  |

* ***Arithmetic mean*** from a frequency distribution:



Using the formulae above, the mean in our example is



So that 

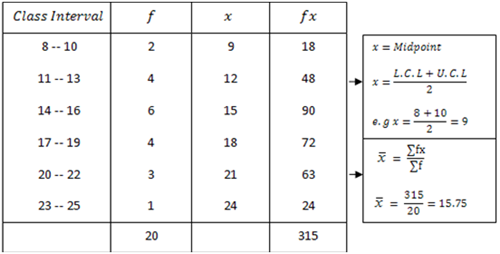
It is essential to know how to interpret your answers.

**Example #2**

**Problem:** Find the arithmetic mean of the data given in the table below.

image

 In frequency distribution actual values of the data are not known therefore an assumption is made that the values of the data are mid-points of the intervals. The calculation of these mid points and arithmetic mean is given in the table below.

[](http://mba-lectures.com/wp-content/uploads/2010/05/image15.png)

 In the above table we have calculated the mid points (variable x) by adding the lower class limit (L.C.L) of each class interval with upper class limit (U.C.L) and divided the sum by 2.  Similarly mean is obtained by multiplying x with corresponding frequencies f and dividing the sum of fx with total frequency.

## The Means, Medians and Modes of Grouped Data

**Introduction**

Let's look briefly at the measures of central tendency associated with grouped data. What does that mean? Well, sometimes we find data that is presented in grouped form, not as individual data points. The following table contains grouped data:

|  |  |
| --- | --- |
| Classes | Frequency |
| 0 ave235 | 6 |
| 5 ave2310 | 12 |
| 10 ave2315 | 19 |
| 15 ave2320 | 3 |
| Total | 40 |

Let's assume that the classes refer to the ages of our sample: 0 ave235 means that we have gathered data here for anyone who is older than 0 but younger than or exactly equal to 5 years old. 5 ave2310 means that we have looked at someone who is in the class or group of people who are older than 5 years but younger than or exactly equal to 10 years … and so on.

In the frequency column, we find that 6 people are between 0 and 5 years of age, 14 people are between 5 and 10 years of age … and altogether, we have data on 40 people spanning the entire age range 0 to 20 years.

## Arithmetic Mean of Grouped Data

|  |  |  |  |
| --- | --- | --- | --- |
| Classes | Frequency | Mid-Point of Class |  |
| Age (years) | F | X | fX |
| 0 ave235 | 6 | 2.5 | 15 |
| 5 ave2310 | 12 | 7.5 | 90 |
| 10 ave2315 | 19 | 12.5 | 237.5 |
| 15 ave2320 | 3 | 17.5 | 52.5 |
| Total | 40 |  | 395 |

To calculate the arithmetic mean of grouped data, we need to extend the above table as follows:

The arithmetic mean is

ave16

**Further examples**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Age** | **F** | **Midpoint(x)** | **fx** | **d=x-A (where A =32)** | **fd** | **u=(x-A)/h** | **fµ** |
| 15-19 | 17 | 17 | 289 | -15 | -255 | -3 | -51 |
| 20-24 | 55 | 22 | 1210 | -10 | -550 | -2 | -110 |
| 25-29 | 45 | 27 | 1215 | -5 | -225 | -1 | -45 |
| 30-34 | 31 | 32 | 992 | 0 | 0 | 0 | 0 |
| 35-39 | 18 | 37 | 666 | 5 | 90 | 1 | 18 |
| 40-44 | 11 | 42 | 462 | 10 | 110 | 2 | 22 |
| 45-59 | 9 | 47 | 423 | 15 | 135 | 3 | 27 |
| 50-54 | 15 | 52 | 780 | 20 | 300 | 4 | 60 |
| **Total** | **201** |  | **6037** |  | **-395** |  | **-79** |

**Using different methods**

(a)        A.M = Σfx/Σf (Mean of grouped data)

= 6037/201 = 30.035

(b)        **Short Cut Method/assumed mean method**

A.M. = A + Σfd/Σf, where d = x – A, where A is the assumed mean

= 32+ (-395/201) = 32 - 1.965 =30.035

(c)        **Step Deviation Method**

A.M. = A + (Σfu/Σf) h, where A is the assumed mean, h is the class interval and u = (x-A)/h.**=** 32+ (-79/201)5 = 32 - 1.965 =30.035

## The Median of Grouped Data

The median is the middle value of a data set and for grouped data, we can find the class that the median resides in relatively easily. In the case of the example we used for the arithmetic mean of grouped data, we can see that the median value is the average of the 20th and 21st values … there are 40 data points, an even number of data points. The median class is highlighted in the following table:

|  |  |
| --- | --- |
| Classes | Frequency |
| Age (years) | F |
| 0 ave235 | 6 |
| 5 ave2310 | 12 |
| 10 ave2315 | 19 |
| 15 ave2320 | 3 |
| Total | 40 |

The calculation of the median of grouped data is based on the following formula

**ave17**

Where

L = the lower limit of the class containing the median   
n = the total number of frequencies   
f = the frequency of the median class   
CF = the cumulative number of frequencies in the classes preceding the class containing the median   
i = the width of the class containing the median

Putting the numbers from the example into the formula now, we see that the median value is 10.53:

ave18

**The Median of Grouped Data**

The median of a sample of data organized in a frequency distribution is computed by the following formula:

**Median = L + [(n/2 - CF)/f] (i)**

**Where**

L is the lower limit of the median class,

CF is the cumulative frequency (class above) the median class,

f is the frequency of the median class, and

i is the median class interval.

|  |  |  |
| --- | --- | --- |
| Determine which class will contain this value. For example, if n=50, 50/2 = 25, then determine which class will contain the 25th value - the median class. Movies showing | Frequency | Cumulative Frequency |
| 1-2 | 1 | 1 |
| 3-4 | 2 | 3 |
| 5-6 | 3 | 6 |
| 7-8 | 1 | 7 |
| 9-10 | 3 | 10 |

To determine the median class for grouped data:

1. Construct a cumulative frequency distribution.
2. Divide the total number of data values by 2.

**Median for grouped data (another symbol used)**

The median for grouped data may not necessarily be computed precisely because the actual values of the measurements may not be known. In that case, you can find the particular interval that contains the median and then approximate the median.

|  |  |  |
| --- | --- | --- |
| |  | | --- | | 25708 | |  | |

Where

|  |  |  |
| --- | --- | --- |
|  | ***L*** | lower class limit of the interval that contains the median |
|  | ***n*** | total number of measurements |
|  | ***w*** | class width |
|  | ***f med*** | frequency of the class containing the median |
|  | ***Σ f b*** | sum of the frequencies for all classes before the median class |

Consider the information in Table

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | |  | **Distribution of Prices of Items Sold at a Garage Sale** |  | ***Class Boundaries*** | ***Frequency (f)*** | | --- | --- | | $.995-$5.995 | 8 | | $5.995-$10.995 | 6 | | $10.995-$15.995 | 4 | | $15.995-$20.995 | 2 | | $20.995-$25.995 | 4 | | $25.995-$30.995 | 6 | | $30.995-$35.995 | 2 | |  | *n* = 32 | |

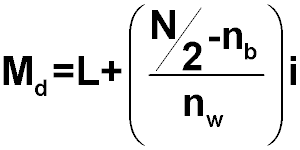
As we already know, the median is located in class interval $11-$15. So *L* = 11, *n* = 32, *w* = 4.99, *f**med*= 4, Σ *f**b*= 14.

|  |
| --- |
| 25709 |

Substituting into the formula:

|  |
| --- |
|  |

**Median - 3rd symbol used**



**Where:**

|  |  |
| --- | --- |
| **L** | = **L**ower exact limit of the interval containing Md. |
| **nb** | = **n**umber of scores **b**elow L. |
| **nw** | = **n**umber of scores **w**ithin the interval containing Md. |
| **I** | = the width of the **i**nterval (for ungrouped data i=1). |
| **N** | = the **N**umber of scores. |

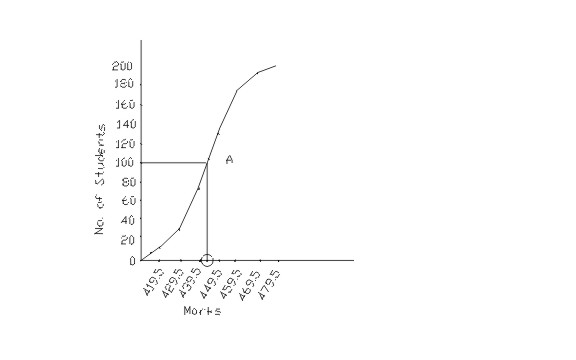
**The diagrammatic method- The cumulative frequency curve (O’give)**

The cumulative frequency curve also known as the o’give is used to obtain the median graphically.

**Steps**

1. Plot the cumulative frequency against the upper class boundary on the x-axis.
2. Join this points using free hand curve
3. Identify the median item/position using n/2
4. Determine the median position on the cumulative frequency curve and plot it. Usually the y-axis is used in determining the position
5. Locate the median value by drawing a perpendicular line from the plotted point on the x-axis. The resultant value on the x-axis is the median of the distribution

**Example**

****

**The Mode of Grouped Data**

The mode is, very simply, the mid-point for the class containing the largest number of class frequencies.

•Midpoint of the modal class

•Modal class has the greatest frequency

**To find the mode, we use the formula**

**Formulae 1**

**X = L + [d1/ (d1+d2)] i**

Where:

**x**- mode of the data

**L**- lower limit of the modal class

**d1** - difference between the frequency of the modal class and below the modal class

**d2** - difference between the frequency of the modal class and above the modal class

**i**- the size of the class interval

**Formulae 2**

**X = L + [f1- f0/ (2f1- f0-f2)] i**

Where:

**x**- mode of the data

**L**- lower limit of the modal class

**f0** - frequency of the class below the modal class

**f1** - the frequency of the modal class

**f2** - the frequency of the class above the modal class

**i** - the size of the class interval

Examples for calculating median and mode

|  |  |  |
| --- | --- | --- |
| **Age** | **F** | **Cf** |
| 15-19 | 17 | 17 |
| 20-24 | 55 | 72 |
| 25-29 | 45 | 117 |
| 30-34 | 31 | 148 |
| 35-39 | 18 | 166 |
| 40-44 | 11 | 177 |
| 45-59 | 9 | 186 |
| 50-54 | 15 | 201 |
| **Total** | **201** |  |

**ave17**

Median =25 + {(100.5-72)/45}5

= 25 + 3.1666 = 28.167

a) **Mode = L + [d1/ (d1+d2)] i**

= 20 +{(55-17)/[(55-17)+(55-45)]}5

= 20 + (38/48)5 = 23.958

**Alternatively**

b) **Mode =** **L + [f1- f0/ (2f1- f0-f2)] i**

**=** 20 **+** {(55-17)/(2\*55-17-45)}5

= 20 + (38/48)5 = 23.958

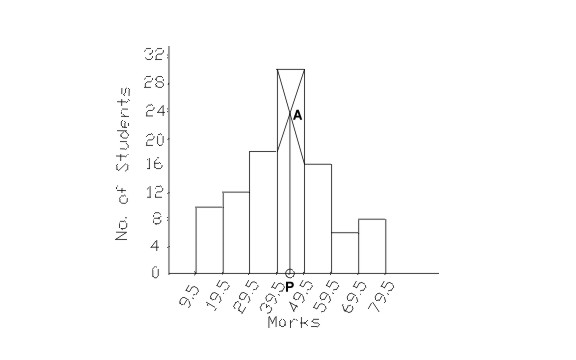
**The mode graphically- Histogram**

The mode can be determined graphically using a histogram.

**Steps**

1. Plot the frequencies on the y-axis against the class on the x-axis
2. Identify the highest bar and draw a diagonal from the top right hand corner to the point where the top of the adjacent bar to the left touches x. Draw a corresponding diagonal on the opposite from top left to the bar on the right
3. Where the two diagonal intercept draw a perpendicular line to the x-axis, the value is the mode

**Example**

****

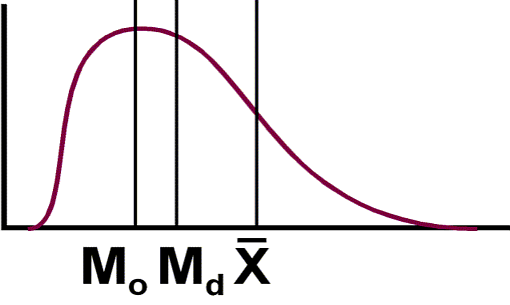
**Measures of Central Location – Summary**

* Measure of Central Location – single measure that represents an entire distribution
* Mode – most common value
* Median – central value
* Arithmetic mean – average value
* Mean uses all data, so sensitive to outliers
* Mean has best statistical properties
* Mean preferred for normally distributed data
* Median preferred for skewed data

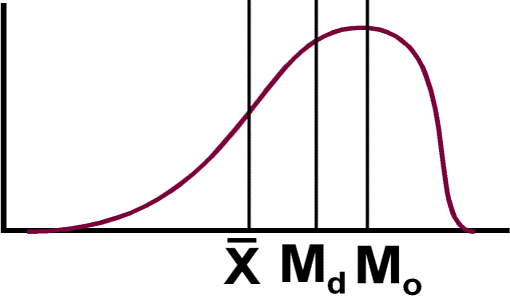
**MEASURES OF SHAPE**

**Note that the presence and direction of skew in the distribution can be determined from the mean and median.** The key to understanding this is to be aware that the mean is sensitive to all scores, while the median is not. **There are three rules:**

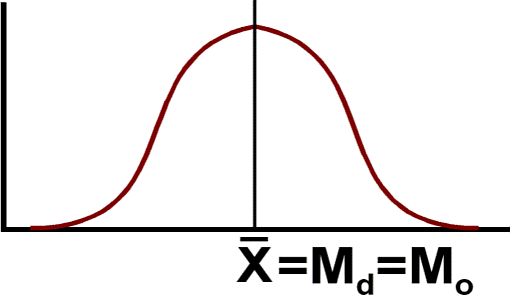
1. **If Eq-14- Md > 0 then +skew**



1. **If Eq-14- Md < 0 then -skew**



1. **If Eq-14- Md = 0 then the distribution is normal  
   and all three measures of central tendency coincide.**



**Home work assignments**



Assemble the data given below in the form of a grouped frequency distribution

40 20 50 61 30 58 51 75

58 70 93 59 49 55 63 38

87 57 23 41 60 57 52 77

37 62 53 83 48 73 28 31

21 76 32 57 53 25 42 63

95 54 64 39 82 54 33 45

48 20 53 65 26 65 87 43

51 66 34 78 55 44 27 74

89 46 67 45 30 57 97 81

43 28 99 47 79 56 86 35

Using the above data calculate the following and give your answers to **1 decimal point**:

1. Mean **(4 mks)**
2. Mode **(4 mks)**
3. Median **(4mks**
4. Standard deviation **(8 mks)**
5. 4th Decile **(5 mks)**

Below are anatomy marks (%) subject for first semester for HRIT students in Msambweni MTC:-

Marks (%): 0-10 10-20 20-30 30-40 40-50

No. of Stds: 2 7 11 6 4

Using the data above calculate the following and give your answer to **nearest whole number**:

1. Arithmetic mean **(5 mks)**
2. Standard deviation **(8 mks)**
3. Coefficient of variation **(2 mks**
4. The range of marks the students scored at 95% confidence limit assuming the distribution was normal/symmetrical **(5 mks)**

# UNIT IV (CHS 1.4) Measures of Dispersion

**Objectives**

1. Definition of the term measures of dispersion
2. Calculation of different measures of dispersion e.g. standard deviation, quartiles, deciles, and range

# Measures of Dispersion

**Spread / Dispersion / Variability** - how much the distribution is spread or dispersed from its central location

Measures of dispersion (spread), refers to the extent to which the data values scatter about their central location value. Dispersion is a measure of data variability, which influences the confidence that an analyst can have in the representativeness and reliability (stability) of central location measures. Widely dispersed data values indicate low reliability and less confidence in the central location as a representative measure for a sample of data. Conversely, high concentration of data values about central location indicates high reliability and greater confidence in the representativeness of the central location value[[18]](#footnote-18).

Whereas measures of central tendency describe the typical average, or central score, measures of dispersion describe the variety, diversity or heterogeneity a set of scores. *Variability* refers to the extent to which the scores in a distribution differ from each other. An equivalent definition (that is easier to work with mathematically) says that variability refers to the extent to which the scores in a distribution differ from their mean. If a distribution is lacking in variability, we may say that it is *homogenous* (note the opposite would be *heterogenous*). Measures of dispersion are descriptive statistics that describe how similar a set of scores are to each other

* + The more similar the scores are to each other, the lower the measure of dispersion will be
  + The less similar the scores are to each other, the higher the measure of dispersion will be[[19]](#footnote-19)
  + In general, the more spread out a distribution is, the larger the measure of dispersion will be

**Which of the distributions of scores has the larger dispersion?**



**Answer**-The upper distribution has more dispersion because the scores are more spread out. That is, they are less similar to each other

Measure of central tendency give us good information about the scores in our distribution. However, we can have very different shapes to our distribution, yet have the same central tendency. Measures of dispersion or variability will give us information about the spread of the scores in our distribution. Are the scores clustered close together over a small portion of the scale, or are the scores spread out over a large segment of the scale?

**Significance of Measures of Dispersion**

* Test the reliability of an average. Measures of variation are used to test to what extent an average represents the characteristic of a data set. If the variation is small, that is, extent of dispersion or scatter is less on each side of an average, than it indicates high uniformity of values in the distribution and the average represents an individual value in the data set. On the hand, if the variation is large, then it indicates a lower degree of uniformity in values in the data set, and the average may be unreliable.
* Control the variability. Measuring variation helps to identify the nature and causes of variation. In social science, the measurement of ‘inequality’ of distribution of income and wealth requires the measurement of variability
* Compare two or more sets of data with respect to their variability[[20]](#footnote-20)

**Types of Dispersion Measures**

The measures that are commonly used to describe data dispersion are:

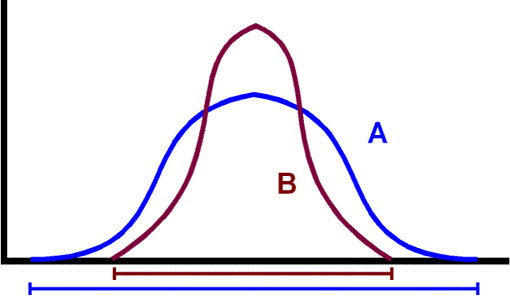
1. Range
2. Interquartile Range
3. Quartile Deviation
4. Variance
5. Standard Deviation[[21]](#footnote-21)

**Range**

Range is the difference between the highest and the lowest data values in a data set. It identifies the interval of values between the highest and the lowest data values. The greater the interval, the wider the range; and the smaller the interval, the narrower the range. The simplest measure of dispersion is the **range**. The range is calculated by simply taking the difference between the maximum and minimum values in the data set. However, the range only provides information about the maximum and minimum values and does not say anything about the values in between[[22]](#footnote-22).

Formula:

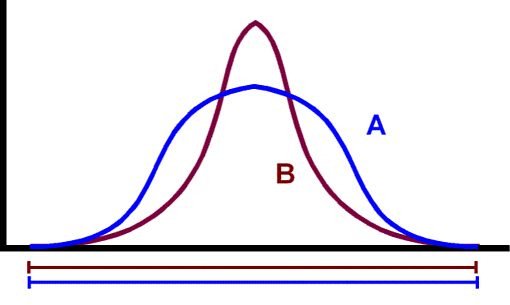
Range = Max value – Min value i.e. R = *xmax - xmin* or Highest score – lowest score



Distribution A has a larger range (and more variability) than Distribution B.

Because only the two extreme scores are used in computing the range, however, it is a crude measure.

For example:



The range of Distribution A and B is the same, although Distribution A has more variability.

For grouped frequency distributions of values in the data set, the range is the difference between the upper class limit of the last class and the lower class limit of the lowest class. In this case the range obtained may be higher than as compared to ungrouped data because of the fact that the class limits are extended slightly beyond the extreme values in the data set.

The coefficient of range is obtained using the following formula:

Coefficient of Range = H – L

H + L

The range is a crude measure of spread because:

1. Outliers distort it, giving the impression of a much wider spread of data between the majority of the data values than is really is the case
2. Range provides no information about the clustering of data values between the maximum and minimum data values, as it only uses two data values (i.e. *xmax* and *xmin*) to calculate it.

Range, therefore, is an unstable, volatile and unreliable measure of dispersion. However, it is always a useful first indication of the spread of data in a data set. It must be viewed with care and always examined along with other measures of dispersion.[[23]](#footnote-23)

**Application of Range**

1. Fluctuation in share prices: the range is useful in the study of small variations among values in a data set, such as variation in share prices and other commodities that are very sensitive to price changes from one period to another
2. Quality control: it is widely used in industrial quality control. Quality control is exercised by preparing suitable quality control charts. These charts are based on setting an upper control limit (range) and a lower control limit (range) within which produced items shall be accepted. The variation in the quality beyond these ranges requires necessary correction in the production process or system.
3. Weather forecast: the concept of range is used to determine the difference between maximum and minimum temperature or rainfall by meteorological departments to announce for the knowledge of the general public.[[24]](#footnote-24)

**Interquartile Range**

The interquartile range is the difference between the upper and the lower quartiles in a set of sample data values. It therefore, identifies the range (distance) of data values between the lower quartile and the upper quartile positions only.

Formula:

IQR = Q3 – Q1

Since the interval between Q1 and Q3 covers the middle 50% of data values, the interquartile range identifies the spread of the middle 50% of data values[[25]](#footnote-25).

It is the range of the middle 50% of a distribution. Because any outliers in our distribution must be on the ends of the distribution, the range as a measure of dispersion can be strongly influenced by outliers. One solution to this problem is to eliminate the ends of the distribution and measure the range of scores in the middle. Thus, with the interquartile range we will eliminate the bottom 25% and top 25% of the distribution, and then measure the distance between the extremes of the middle 50% of the distribution that remains.[[26]](#footnote-26)

Consider the following example:

|  |  |  |
| --- | --- | --- |
| Ranks | County | Expenditures (Kshs in 000) |
| 20  19  18  17  16  15  14  13  12  11  10  9  8  7  6  5  4  3  2  1 | Nairobi  Nyeri  Lodwar  Murang’a  Nyandarua  Mombasa  Kilifi  Marsabit  Isiolo  Bomet  Garissa  Kiambu  Laikipia  Kericho  Kisumu  Baringo  Kajiado  Meru  Uasin Gishu  Nakuru | 2,221  2,045  1,911  1,793  1,770  1,759  1,754  1,740  1,740  1,710  1,673  1,431  1,430  1,374  1,338  1,288  1,286  1,280  1,226  1,152 |

Note the scores have already been ordered from high to low. This makes the range easy to calculate, and it is a necessary step for finding the interquartile range. Of this 20 counties, Nairobi spent most of its per capita on public education (Kshs. 2,221,000) and Nakuru spent the least per capita (Kshs. 1,152,000) the range is therefore (2,221,000 – 1,152,000) or 1,069,000 (R= Kshs. 1,069,000)[[27]](#footnote-27).

To find Q we must locate the first and the third quartiles (Q1 and Q3). Q1 is determined by multiplying N by (0.25). Since (20 x 0.25) is 5, Q1 is the score associated with the 5th case, counting up from the lowest score. The fifth score is Baringo with a score of 1,288,000, thus Q1 = Kshs. 1,288,000. The case that lies at the third quartile (Q3) is given by multiplying N by (0.75), and (20 x 0.75) = 15th case. The 15th case again counting from the lowest is Mombasa with a score of 1,759,000 (Q3 = 1,759,000). Therefore:

IQR = Q3 – Q1

= 1,759,000 – 1,288,000

= 471,000

Half the distance between Q1 and Q3 is called the semi-interquartile range or the quartile deviation (QD).

QD = Q3 – Q1

2

The median is not necessarily midway between Q1 and Q3, although this will be so with symmetrical distribution. The median and quartiles divide the data into equal numbers of values but do not necessarily divide the data into equally wide intervals. In non-symmetrical distribution, the two quartiles Q1 and Q3 are at equal distance from the median, that is Median - Q1 = Q3 – Median. Thus Median ± quartile deviation covers exactly 50% of the observed values in the data set. A smaller value of quartile deviation indicates high uniformity or less variation among the middle 50% observed values around the median value. On the other hand, a higher value of quartile deviation indicates large variation among the middle 50% observed values.[[28]](#footnote-28)

**Coefficient of Quartile Deviation**

Since Quartile deviation is an absolute measure of variation, therefore its value gets affected by the size and number of observed values in the data set. Thus, the QD, of two or more than two sets of data may differ. Due to this, to compare the degree of variation in different data sets, we compute the relative measure corresponding to QD, called the coefficient of quartile deviation whose formula is:[[29]](#footnote-29)

Coefficient of Quartile Deviation (QD) = Q3 – Q1

Q3 + Q1

**Advantages of the Interquartile Range**

1. It removes some of the instability inherent in the range if outliers (extreme data values) are present
2. This modified range excludes these outliers and focuses on the spread of the middle 50% of the data values, thus it is more stable measure of dispersion than the range.

NB: The *semi-Interquartile range* (or *SIR*) is defined as the difference of the first and third quartiles divided by two

* 1. The first quartile is the 25th percentile
  2. The third quartile is the 75th percentile

**SIR or QD = (Q3 - Q1) / 2**

**When to use the SIR**

1. The SIR is often used with skewed data as it is insensitive to the extreme scores

**Mean Absolute Deviation**

Since two measures of variation, range and quartile deviation, do not show how values in a data set are scattered about a central value or dispersed themselves throughout the range, therefore it is quite reasonable to measure the variation as a degree (amount) to which values within a data set deviate from either the mean or median.[[30]](#footnote-30)

The mean of deviation of individual values in data set from either actual mean is always zero so such as measure (zero) would be useless as an indicator of variation. This problem can be solved in two ways:

1. Ignore the signs of the deviations by taking their absolute value, or
2. Square the deviations because the square of a negative number is positive

Since the absolute difference between a value *x*i of an observation from A.M, is always a positive number, whether it is less or more than the A.M., therefore we take the absolute value of each such deviation from the A.M. (or median). Taking the average of these deviations from the A.M., we get a measure of variation called the mean absolute deviation (MAD)[[31]](#footnote-31).

Another method is to calculate the average difference between each data point and the mean value, and divide by the number of points to calculate the **average deviation** (mean deviation). However, performing this calculation will result in an average deviation of zero since the values above the mean will cancel the values below the mean. If this method is used, the absolute value of the difference is taken so that only positive values are obtained, and the result sometimes is called the *mean absolute deviation*. The average deviation is not very difficult to calculate, and it is intuitively appealing. However, the mathematics are very complex when using it in subsequent statistical analysis. Because of this complexity, the average deviation is not a very commonly used measure of dispersion.

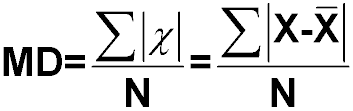
**Methods for Calculation of Mean Deviation:**

**Case I:** **For Ungrouped Data.**

 In this case the mean deviation is given by the formula

**Mean Deviation = M.D. = Σ|x - A|/n = Σ|d|/n,**

 where‘d’ stands for the deviation from the mean or median and **|d| is always positive** whether d itself is positive or negative and n is the total number of items.



**Case II: For Grouped data.**

 Let x1, x2, x3, …, xn  occur with frequencies f1, f2, f3, ,fn respectively and let Σf = n and M can be either Mean or Median, then the mean deviation is given by the formula.

**Mean Deviation = Σf|x-M|/Σf = Σf|d|/n**

 Where d = |x – M| and Σf = n.

 Coefficient of Mean Deviation = Mean Deviation / Median or = Mean Deviation / Mean (In case the deviations are taken from mean)

**Advantages of MAD**

1. The calculation of MAD is based on all observations in the distribution and shows the dispersion of values around the measure of central tendency
2. The value of MAD is easy to compute and therefore makes it popular among those users who are not even familiar with statistical methods
3. While calculating MAD, equal weightage is given to each observed value and thus it indicates how far each observation lies from either mean or median
4. Average deviation from the mean is always zero in any data set. The MAD avoids this problem by using absolute values to eliminate the negative signs[[32]](#footnote-32)

**Disadvantages of MAD**

1. The algebraic signs are ignored while calculating MAD. If the signs are not ignored, then the sum of deviations taken from the arithmetic mean will be zero and close to zero when deviations are taken from the median
2. The value of MAD is considered the best when deviations are taken from the median. However, median does not provide a satisfactory result in case of high degree variability in a data set. Moreover, the sum of deviations from the mean (ignoring the signs) is greater than the sum of deviations from median (ignoring the signs). In such a situation computations of MAD by taking deviations from mean is also not desirable.
3. The MAD is generally unwieldy in mathematical decisions[[33]](#footnote-33)

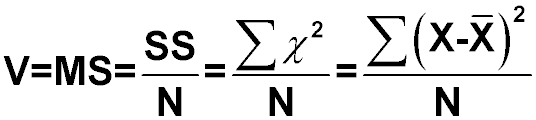
**Degrees of freedom**

The "n-1" term in the above expression represents the ***degrees of freedom*** (df). Loosely interpreted, the term "degrees of freedom" indicates how much freedom or independence there is within a group of numbers. For example, if you were to sum four numbers to get a total, you have the freedom to select any numbers you like. However, if the sum of the four numbers is stipulated to be 92, the choice of the first 3 numbers is fairly free (as long as they are low numbers), but the last choice is restricted by the condition that the sum must equal 92. For example, if the first three numbers chosen at random are 28, 18, and 36, these numbers add up to 82, which is 10 short of the goal. For the last number there is no freedom of choice. The number 10 must be selected to make the sum come out to 92. Therefore, the degrees of freedom have been limited by 1 and only n-1 degrees of freedom remain. In the SD formula, the degrees of freedom are n minus 1 because the mean of the data has already been calculated (which imposes one condition or restriction on the data set).

**Variance**

Variance is a measure of average squared deviation about the arithmetic mean. It is expressed in square units. Consequently, its meaning – in a particular sense – is obscure. To provide meaning, the dispersion measure should be expressed in the original units of measure of the random variable. This is provided by a measure of dispersion called the standard deviation[[34]](#footnote-34).

Another solution to the problem of [the deviations summing to zero](http://www.uwsp.edu/psych/stat/5/CT-Var.htm#sumdev) is to square the deviations. That is:



Thus another name for the Variance is the ***Mean of the Squared Deviations About the Mean*** (or more simply, the ***Mean of Squares (MS)***). The problem with the MS is that its units are squared and thus represent space, rather than a distance on the X axis like the other measures of variability

*Variance* is defined as the average of the square deviations:



**What Does the Variance Formula Mean?**

* First, it says to subtract the mean from each of the scores
* This difference is called a *deviate* or a *deviation score*
* The deviate tells us how far a given score is from the typical, or average, score
* Thus, the deviate is a measure of dispersion for a given score

**Why can’t we simply take the average of the deviates? That is, why isn’t variance defined as:**



**The above formula is wrong**

* One of the definitions of the *mean* was that it always made the sum of the scores minus the mean equal to 0
* Thus, the average of the deviates must be 0 since the sum of the deviates must equal 0
* To avoid this problem, statisticians square the deviate score prior to averaging them
  + Squaring the deviate score makes all the squared scores positive
* Variance is the mean of the squared deviation scores
* The larger the variance is, the more the scores deviate, on average, away from the mean
* The smaller the variance is, the less the scores deviate, on average, from the mean

**Variance of a Sample**

Because the sample mean is not a perfect estimate of the population mean, the formula for the variance of a sample is slightly different from the formula for the variance of a population:



Where s2 is the sample variance, X is a score, X is the sample mean, and N is the number of scores

**Example-** The table shows marks (out of 10) obtained by 20 people in a test

|  |  |
| --- | --- |
| Mark (x) | Frequency (f) |
| 1 | 0 |
| 2 | 1 |
| 3 | 1 |
| 4 | 3 |
| 5 | 2 |
| 6 | 5 |
| 7 | 5 |
| 8 | 2 |
| 9 | 0 |
| 10 | 1 |

Work out the variance of this data. In such questions, it is often easiest to set your working out in a table:

|  |  |
| --- | --- |
| Fx | fx2 |
| 0 | 0 |
| 2 | 4 |
| 3 | 9 |
| 12 | 48 |
| 10 | 50 |
| 30 | 180 |
| 35 | 245 |
| 16 | 128 |
| 0 | 0 |
| 10 | 100 |

Sf = 20   
Sfx = 118   
Sfx2 = 764   
variance =  Sfx2  - ( Sfx )2  =  764  -  (118)2   
     20       ( 20 )2   
                  Sf       (  Sf  )2  =  38.2 - 34.81 = 3.39

Standard deviation (denoted by *s* or *s.d.*) is the root mean square of the deviations from the arithmetic mean. The standard deviation indicates the average distance of the observations from the mean of the data set.

Image13

To get a better estimate of the standard deviation of the population (denoted by s), standard deviation is often computed with *n*–1 instead of *n* in the denominator. However, for large values of *n* (*n ³* 30) there is practically no difference between the two definitions. Variance is the square of the standard deviation.

**Standard Deviation**

Standard deviation expresses dispersion – as computed by the variance – in the original units of the random variable. A standard deviation is found by taking the square root of the variance i.e.

Standard Deviation = √Variance

S = √S2

σ identifies a population standard deviation, while *s* describes a sample of standard deviation.

Formula:

Thus the mathematical formula for a standard deviation of a sample data is the square root of the variance formula, as follows

*X)*

*s* = ∑*xi* -

n – 1

* When the deviate scores are squared in variance, their unit of measure is squared as well
  + E.g. If people’s weights are measured in pounds, then the variance of the weights would be expressed in pounds2 (or squared pounds)
* Since squared units of measure are often awkward to deal with, the square root of variance is often used instead[[35]](#footnote-35)
  + The standard deviation is the square root of variance
* Standard deviation = √variance
* Variance = standard deviation2

**Computing the Standard Deviation**

*X)*

|  |  |  |
| --- | --- | --- |
| Scores (*x*i) | Deviations | Deviations Squared (*x*i - ) |
| 10 | (10 – 30) = -20 | (-20)2 = 400 |
| 20 | (20 – 30) = -10 | (-10)2 = 100 |
| 30 | (30 – 30) = 0 | (-0)2 = 0 |
| 40 | (40 – 30) = 10 | (10)2 = 100 |
| 50 | (50 – 30) = 20  *X )X)X)* | (20)2 = 400  *X)* |
| ∑ *x*i = 150 | ∑(*x*i - ) = 0 | ∑(*x*i - ) 2 = 1,000 |

*s* = ∑(*xi* -

*X)*

N

*s* = 1,000

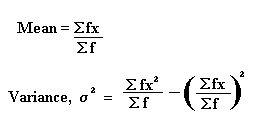
5

*s* = √200

*s* = 14.14

**Computational Formula**

There are many ways of writing the formula for the standard deviation. The one below is for a basic list of numbers. The formula for the variance when the data is grouped is as follows. The standard deviation can be found by taking the square root of this value.



When calculating variance, it is often easier to use a computational formula which is algebraically equivalent to the definitional formula:

Where σ2 is the population variance, X is a score, μ is the population mean, and N is the number of scores



**Advantages of Standard Deviation**

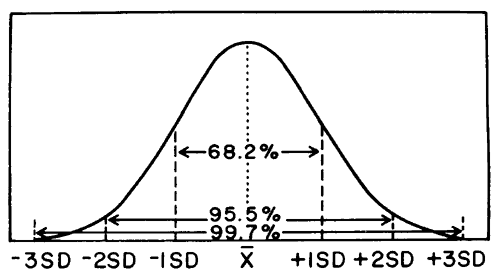
1. The value of standard deviation is based on every value in the data set. It is the only measure of variation capable of algebraic treatment and less affected by fluctuations of sampling as compared to other measures of variation
2. It is possible to calculate the combined standard deviation of two or more sets of data
3. Standard deviation has a definite relationship with the area under symmetric curve of a frequency distribution. Due to this reason standard deviation is called a standard measure of variation
4. Standard deviation is useful in further statistical investigations. For example, standard deviation plays a vital role in comparing skewness, correlation, and so on, and also widely used in sampling theory

**Disadvantages**

1. As compared to other measures of variation, calculations of standard deviation are difficult
2. While calculating standard deviation, more weight is given to extreme values and less to those near mean. Since for calculating SD, the deviations from the mean are squared, therefore large deviations are more proportionately more than small deviations. For example the deviations 2 and 10 are in the ratio of 1:5 but their squares of 4 and 100 are in the ratio of 1:25.[[36]](#footnote-36)

**Normal or Gaussian distribution**

**Percentage(%) of data values within 1, 2, and 3 standard deviations of the mean**



Traditionally, after the discussion of the mean, standard deviation, degrees of freedom, and variance, the next step was to describe the normal distribution (a frequency polygon) in terms of the standard deviation "gates." The figure here is a representation of the frequency distribution of a large set of laboratory values obtained by measuring a single control material. This distribution shows the shape of a normal curve. Note that a "gate" consisting of ±1SD accounts for 68% of the distribution or 68% of the area under the curve, ±2SD accounts for 95% and ±3SD accounts for >99%. At ±2SD, 95% of the distribution is inside the "gates," 2.5% of the distribution is in the lower or left tail, and the same amount (2.5%) is present in the upper tail. Some authors call this polygon an error curve to illustrate that small errors from the mean occur more frequently than large ones. Other authors refer to this curve as a probability distribution.

**Coefficient of variation**

The standard deviation is a measure of absolute variability, which means that the spread of data variables can only be measured in relation to its own arithmetic mean. On its own the standard deviation conveys no meaningful information.

To illustrate:

An *s* = 10 can be large if the arithmetic mean of the data values is 15, but an *s* = 10 can be small if the arithmetic mean is 250.

Formula:

To interpret the magnitude of a standard deviation, the standard deviation must be compared to its own mean. This is expressed in the coefficient of variation:

Coefficient of Variation (CV) = Standard Deviation %

Mean

**CV% = (SD/Xbar)100**

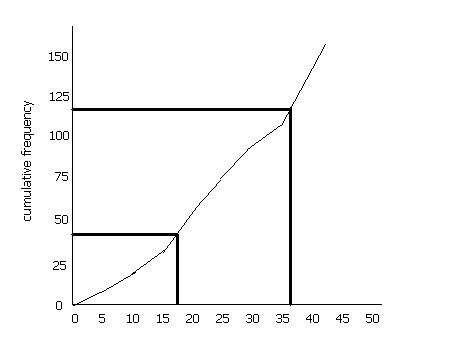
The coefficient of variation is interpreted as a percentage. The smaller the CV, the more concentrated the data values are about their mean; conversely, a larger CV implies that the data values are more widely spread about their mean. The lower limit of a CV is zero, but there is no upper limit.[[37]](#footnote-37)

**Quartile, Deciles and percentiles**

If we divide a cumulative frequency curve into quarters, the value at the lower quarter is referred to as the lower quartile, the value at the middle gives the median and the value at the upper-quarter is the upper quartile.

A set of numbers may be as follows: 8, 14, 15, 16, 17, 18, 19, 50. The mean of these numbers is 19.625. However, the extremes in this set (8 and 50) distort the range. The inter-quartile range is a method of measuring the spread of the numbers by finding the middle 50% of the values. It is useful since it ignores the extreme values. It is a method of measuring the spread of the data.   
  
The lower quartile is (n+1)/4 th value (n is the cumulative frequency, i.e. 157 in this case) and the upper quartile is the 3(n+1)/4 the value. The difference between these two is the inter-quartile range (IQR).

In the above example, the upper quartile is the 118.5th value and the lower quartile is the 39.5th value. If we draw a cumulative frequency curve, we see that the lower quartile, therefore, is about 17 and the upper quartile is about 37. Therefore the IQR is 20 (bear in mind that this is a rough sketch- if you plot the values on graph paper you will get a more accurate value).

****

**Quartile Deviation:**

 Quartile deviation = Q3 - Q1 / 2

 Coefficient of Quartile Deviation = Q3 - Q1 / Q3 - Q1

The method by which the median is determined can further be extended to divide the distribution into more than two parts. The distribution can be divided into 4, 10, or 100 parts. The value of the items which divides the distribution into four equal parts is called quartiles. On the other hand the value of the item which divides the distribution into ten equal parts is called the deciles, similarly the value of the item dividing the distribution into hundred equal parts is known as percentiles

The 2nd quartile, the 5th decile and the 50th percentile is the median.

Calculating quartiles, deciles and percentiles is the same as median in both ungrouped and grouped data

**Ungrouped data**

e.g. Q1 position = n+1/4 and Q3 position = 3(n+1)/4 etc

6th decile position = 6(n+1)/10 etc

25th percentile position = 25(n+1)/100 etc

**Grouped data**

Examples for computation of quartiles/deciles & percentiles

|  |  |  |
| --- | --- | --- |
| **Age** | **F** | **Cf** |
| 15-19 | 17 | 17 |
| 20-24 | 55 | 72 |
| 25-29 | 45 | 117 |
| 30-34 | 31 | 148 |
| 35-39 | 18 | 166 |
| 40-44 | 11 | 177 |
| 45-59 | 9 | 186 |
| 50-54 | 14 | 200 |
| **Total** | **200** |  |

1. Q3 = 35+ [{3(200/4)}-148]/18\*5 Q3 position = 3(cf/4)
2. 4th Decile = 25+ [{4(200/10)}-72]/45\*5 4th decile position = 4(cf/10)
3. 25th Percentile = 20+ [{25(200/100)}-17]/55\*5 25th Percentile position = 3(cf/4)

**Inter–decile range = Q9 – Q1**

**When to use the range**

* The range is used when
  + you have ordinal data or
  + you are presenting your results to people with little or no knowledge of statistics
* The range is rarely used in scientific work as it is fairly insensitive
  + It depends on only two scores in the set of data, XL and XS
  + Two very different sets of data can have the same range:  
    1 1 1 1 9 vs 1 3 5 7 9

**Summary**

Estimation is the goal of inferential statistics. We use sample values to estimate population values. The symbols are as follows:

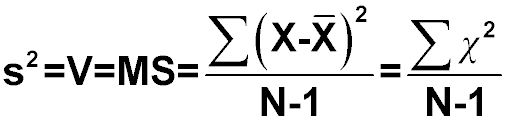
|  |  |  |
| --- | --- | --- |
| **Measure** | **Sample** | **Population** |
| **Mean** | **Eq-14** | **Eq-02** |
| **Variance** | **s2** | **2** |
| **Standard Deviation** | **S** | **** |

It is important that the sample values (estimators) be unbiased. An ***unbiased estimator*** of a parameter is one whose average over all possible random samples of a given size equals the value of the parameter.

Eq-14

While is an unbiased estimator of ****, **s2** is not an unbiased estimator of **2**.

In order to make it an unbiased estimator, we use N-1 in the denominator of the formula rather than just N. Thus:



Note that this is a *defining* formula and, as we will see below, is not the best choice when actually doing the calculations.

Variance and standard deviation of a [population](http://simon.cs.vt.edu/SoSci/converted/glossary.html#population) are designated by  sigma_sqand  sigma, respectively. Variance and standard deviation of a [sample](http://simon.cs.vt.edu/SoSci/converted/glossary.html#sample) are designated by ***s2*** and ***s***, respectively.

|  |  |  |
| --- | --- | --- |
|  | **Variance** | **Standard Deviation** |
| **Population** | pop_var | pop_stdv |
| **Sample** | samp_var | samp_stdv |

|  |  |  |
| --- | --- | --- |
| **Column A** | **Column B** | **Column C** |
| **X value** | **X value-Xbar** | **(X-Xbar)2** |
| 90 | 90 - 87.7 = 2.30 | (2.30)2 = 5.29 |
| 91 | 91 - 87.7 = 3.30 | (3.30)2 = 10.89 |
| 89 | 89 - 87.7 = 1.30 | (1.30)2 = 1.69 |
| 84 | 84 - 87.7 = -3.70 | (-3.70)2 = 13.69 |
| 88 | 88 - 87.7 = 0.30 | (0.30)2 = 0.09 |
| 93 | 93 - 87.7 = 5.30 | (5.30)2 = 28.09 |
| 80 | 80 - 87.7 = -7.70 | (-7.70)2 = 59.29 |
| 90 | 90 - 87.7 = 2.30 | (2.30)2 = 5.29 |
| 85 | 85 - 87.7 = -2.70 | (-2.70)2 = 7.29 |
| 87 | 87 - 87.7 = -0.70 | (-0.70)2 = 0.49 |
| ls14f1X = 877 | ls14f1(X-Xbar) = 0 | ls14f1(X-Xbar)² = 132.10 |

In these equations,  mu is the population mean,  capX_bar is the sample mean, ***N*** is the total number of scores in the population, and ***n*** is the number of scores in the sample.

**Examples**

The dispersion of values about the mean is predictable and can be characterized mathematically through a series of manipulations, as illustrated below, where the individual x-values are shown in column A.

* The first mathematical manipulation is to sum (summation symbol) the individual points and calculate the mean or average, which is 877 divided by 10, or 87.7 in this example.
* The second manipulation is to subtract the mean value from each control value, as shown in column B. This term, shown as X value - Xbar, is called the difference score. As can be seen here, individual difference scores can be positive or negative and the sum of the difference scores is always zero.
* The third manipulation is to square the difference score to make all the terms positive, as shown in Column C.
* Next the squared difference scores are summed.
* Finally, the predictable dispersion or standard deviation (SD or s) can be calculated as follows:

ls14f2  
= [132.10/(10-1)]1/2 = 3.83

**Various formulas**

**a) s.d =** √{∑d2/n-1 ungrouped data **or** s.d = √{∑fd2/∑f-1 grouped data where d = x,- actual mean

**b) s.d =** √{∑fx2/(∑f-1)} – {∑fx/(∑f-1)}2

**Assumed mean method**

**c) s.d =** √{∑fd’2/(∑f-1)} – {∑fd’/(∑f-1)}2 where d’ = x,- assumed mean

**Arbitrary method**

**d) s.d = {**√[∑fu’2/(∑f-1)] – [∑fu’/(∑f-1)]2 }i where u’ = (x,- assumed mean)/i

**Deducing the formula for standard deviation**

s.d = √{∑fd2/∑f = √{∑f(x-**Eq-14**)2/∑f

s,d2  = ∑f(x2- 2x**Eq-14** + **Eq-14**2) Expanding the formula given that (a-b)2 = a2 – 2ab+ b2

∑f

s,d2  = ∑fx2- 2**Eq-14**∑fx + **Eq-14**2∑f

∑f ∑f ∑f

s,d2  = ∑fx2- 2**Eq-14**2 + **Eq-14**2

∑f

s,d2  = ∑fx2- **Eq-14**2 but **Eq-14 =** ∑fx

∑f ∑f

s,d = √∑fx2 - (∑fx/∑f )2

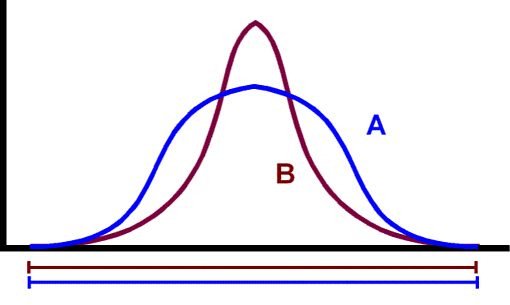
∑f

**Properties of the Variance & Standard Deviation**:

* + Are always positive (or zero).
  + Equal zero when all scores are identical (i.e., there is no variability).
  + Like the mean, they are [sensitive to all scores](http://www.uwsp.edu/psych/stat/5/CT-Var.htm#sens).

**Overall Example**

Let's reconsider an example from above of two distributions (A & B):



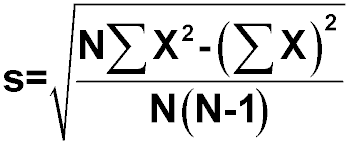
Consider a possibility for the scores that go with these distributions:

|  |  |  |
| --- | --- | --- |
| **Distribution** | **A** | **B** |
| **Data** | 150 | 150 |
| 145 | 110 |
| 100 | 100 |
| 100 | 100 |
| 55 | 90 |
| 50 | 50 |
|  | **600** | **600** |
| **N** | **6** | **6** |
| **Eq-01** | **100** | **100** |
| **Range** | **150-50=100** | **150-50=100** |

Notice that the central tendency and range of the two distributions are the same. That is, the mean, median, and mode all equal 100 for both distributions and the range is 100 for both distributions. However, while Distributions A and B have the same measures of central tendency and the same range, they differ in their variability. Distribution A has more of it. Let us prove this by computing the standard deviation in each case. First, for Distribution A:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **Eq-25** | **** | **2** |
| 150 | 100 | 50 | 2500 |
| 145 | 100 | 45 | 2025 |
| 100 | 100 | 0 | 0 |
| 100 | 100 | 0 | 0 |
| 55 | 100 | -45 | 2025 |
| 50 | 100 | -50 | 2500 |
|  | **600** |  | **0** | **9050** |
| **N** | **6** |  |  |  |

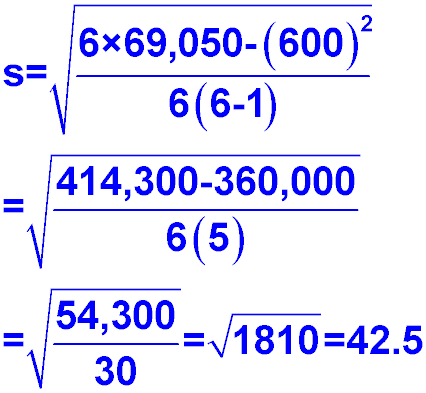
Plugging the appropriate values into the *defining* *formula* gives:



Doing the computations for Distribution A in this manner gives:

|  |  |  |
| --- | --- | --- |
|  | **A** | **X2** |
| 150 | 22500 |
| 145 | 21025 |
| 100 | 10000 |
| 100 | 10000 |
| 55 | 3025 |
| 50 | 2500 |
|  | **600** | **69050** |
| **N** | **6** |  |

Then, plugging in the appropriate values into the computational formula gives:

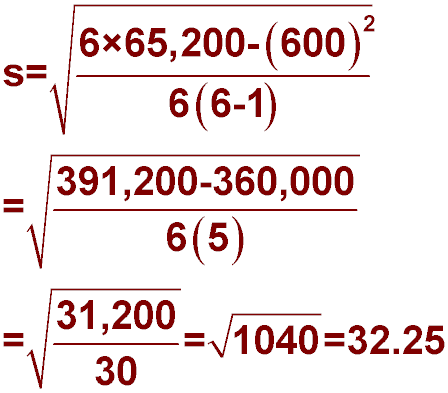


Note that the defining and computational formulas give the same result, but the computational formula is easier to work with (and potentially more accurate due to less rounding error).

Doing the same calculations for Distribution B yields:

|  |  |  |
| --- | --- | --- |
|  | **B** | **X2** |
| 150 | 22500 |
| 110 | 12100 |
| 100 | 10000 |
| 100 | 10000 |
| 90 | 8100 |
| 50 | 2500 |
|  | **600** | **65200** |
| **N** | **6** |  |

Then, plugging in the appropriate values into the computational formula gives:



Thus, Distribution **A** clearly has more variability than Distribution **B**.

# UNIT V (CHS 1.5) Standard Hospital Statistics and Formulae

1. Definition, and calculation of standard administrative statistics formulae
2. Importance of hospital and administrative statistics formulae
3. Types of hospital and administrative statistics
4. Occupied bed days
5. Available bed days
6. Vacant bed days
7. Excess bed days
8. Admissions, discharges and deaths including well people
9. Average occupancy
10. Average length of stay
11. Turn over interval
12. Turn over per bed
13. Average available beds
14. Crude hospital death rate
15. Average discharge rate
16. Average admission rate
17. Specific death rate
18. Average out-patient attendances
19. MCH/FP average attendance
20. Family planning dropout rate

# HEALTH ADMINISTRATIVE STATISTICS (IN-PATIENT STATISTICS)

Definitions of various terms

1. **In-patient**- person/ patient who have occupied a hospital bed/admitted. He/she has undergone all the admission procedures. Any admitted person/patient must be issued with a hospital number/in-patient number/admission number.
2. **Daily Bed Return (DBR – MOH 340)** – A tool of collecting data pertaining to in-patient services (admissions & discharges). It is filled in duplicate by staffs/nurses of respective ward. Each ward retains its DBR and each day has a DBR. DBRs are filled from midnight when the day starts and collected in the morning (8.ooam) for processing.
3. **Bed complement or allocated/authorized/available/staffed beds** - # of beds that are **permanently allocated** to a unit for use by patient. The unit may be a hospital, ward. These are both beds and cots
4. **Bed state -** # of patients occupying a bed at any given time in a unit
5. **Bed turnover/turnover per bed/through put per bed** – average # of **patients** expected to be treated per bed
6. **Turnover interval** – Average # of **days** that bed lie vacant between two successive admissions/patients
7. **% occupancy** – is the % of the actual patient’s days to the maximum authorized/available patients’ days as determined by patients beds at any given period.
8. **Average Length of Stay (ALOS)** – average number of **days** a patient occupies a beds
9. **Average daily occupancy** - number of beds that are occupied in a day
10. **Occupied bed days (OBD) -** numberof patients remaining in the hospital/ward added together in a particular period.
11. **Amenity bed** – Beds provided in a single room or separate room in a ward for which can be available on extra fees.
12. **Day case / Out-patient** – is a person attending a health facility as a non-residential patient for specific test, treatment, investigation/procedure.
13. **Excess beds** – is the number of beds would be required at one moment to cover the **extra** patients in the unit/hospital
14. **Excess bed days** – is the total # of patients days in unit who were above the total maximum capacity of the unit/hospital in a specified period of time
15. **Well persons** – These are care takers in the ward
16. **Paroles**- Are patients who have been **temporarily** discharged from the ward but their beds preserved. It’s applicable in psychiatric hospitals.

## The Daily Bed Return

The DBR is a major tool of data collection used in health services for collecting data of the in-patient state in any health institution. Thus it is used when an institution has the patient's admission status. It is abbreviated DBR. The resultant data from this tool is referred to as hospital administrative statistics.

**The Contents of DBR**

The data is usually collected on daily basis hence the name daily bed return. The contents include

* Name if the institution
* Ward/service area i.e. medical ward, surgical ward
* Date which includes day, month and year
* Hospital numbers for patients admitted
* Names of Patients admitted
* Hospital numbers of patients discharged from ward/service unit
* Name of patient discharged for a ward/service area
* Description of the destination of the discharged patient i.e. death, transfer to or from another ward or hospital
* Particulars of patients transferred within the wards of the same hospital (Transfer in or transfer out)
* Previous daily bed state of the same ward or service area.
* The daily bed state of that particular day for the ward/service area
* Authority of completion by a ward staff - usually the nursing officer in-charge of the ward/service unit
* Record of summary of the data carried on the daily bed state usually by the health records and information staff

**Procedure of Using a Daily Bed Return**

It is the responsibility of HRI department to ensure that the DBR is available and issued to the point of use. It is the responsibility of the point of use staff to enter all patients admitted and discharged from those points on daily basis.

**Entry - DBR**

* Assumption of right and left division of the DBR
* All categories of admissions are entered on the left hand side of the DBR i.e. direct admission or transfers from other wards of the same hospital
* All categories of discharges are entered on the right hand side of the DBR
* Admission entries are made immediately since the patient should occupy hospital bed
* Discharge entries are made immediately to indicate that the patient is leaving the ward

**The Lay Out of the DBR**

**MINISTRY OF HEALTH**

**DAILY BED RETURN**

**Hospital …………………………………………….**

**Ward………………………………………………..**

 Admissions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Hospital Number | Name | Hospital Number | Name | Discharge to |
| 1  2  3  4 | 1  2  3  4 | 1  2  3  4 | 1  2  3  4 |  |

Inter-ward Transfers within the Hospital

 Admissions Discharges

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| From Ward | Hospital Number | Name | To Ward | Hospital Number | Name |
| 1  2  3  4 | 1  2  3  4 | 1  2  3  4 | 1  2  3  4 | 1  2  3  4 |  |

Previous Daily Return Numbers Today's Daily Return Numbers

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Beds | Cots | Total |  | Beds | Cots | Total |
| Patients  Well People  Vacant  Total |  |  |  | Patients  Well People  Vacant  Total |  |  |  |

Records Officer Use Only

Patients: ……………………. Well People …………………………………

Previously ………………….. Previously …………………………………..

Admissions …… + …………. Admissions ………………… + ……………

Total ……………… ………… Total …….….………………………………

Discharges …….. - …….……. Discharges ……..……….. - …….………….

Total ………………………… Total ..………………………………………

Records Copy Checked by ………………………………….

GPK (L)

Usually DBR is completed in duplicate such that the original copy is released to the HRI department and duplicate remains in the ward or source point. Before the original copy is released it should be completed upto the level of being signed by the nursing officer in-charge.

It is required that DBR be balanced at mid-night since it is the changeover of dates i.e. from one date to another. This is a completion of a 24 hour day.

**Balancing of the DBR in the Ward**

This is done by the nursing in the ward or the staff working in the service point. It is usually facilitated by:

* Physical count of patients remaining in the ward at particular time of balancing the DBR
* Entries made on admissions and discharges for the particular date in consideration
* Previous day's (date's) DBR. To balance the day's stay, obtain the previous day's stay, and add on all the day's admission and subtract all the day's discharges and deaths

**Balancing**

Previous state + admission - Discharges and deaths

Today's daily bed state = Previous bed state + admission - Discharges and deaths

Example One

8th Feb 2014

|  |  |  |  |
| --- | --- | --- | --- |
|  | Beds | Cots | Total |
| Patients | 46 | - | 46 |
| Well People | - | - | - |
| Vacant | - | - | - |
| Total | 46 | - | 46 |

On 9th Feb 2014 there are 10 Discharges

Today's Daily Bed State (TDBS) = Previous state + Admission - Discharges and Deaths

9th TDBS = 46 + 0 - 10

= 36

**Collection and Assembling of the DBR**

It is the responsibility of HRI officer to collect the DBR at the point of service. He/she should also check to ensure that the DBR is completed well accurately balanced so as to take the immediate collection. After the DBR is used to confirm (for checking) the availability of patients records those who have been discharged, the officer should collect one copy and take it for filing in the DBR file.

**Balancing of the DBR at the HRI Department**

Check for the accuracy in completion of the DBR at the individual entries and the daily state balance

Balance at the position of the "Records use only"

**Example**

Patients Well People

Previous 28 Previous

Admissions + 0 Admissions +

Total 28 Total

Discharges - 2 Discharges -

Total 26 Total

**The Daily Bed State Summary**

The DBSS should be maintained monthly i.e each month on its one page with entries comprising of its respective data. In making the daily bed state summary the balancing principle should be observed:

Total Daily Bed State = Previous state + admission - Discharges and Deaths

**Hospital Name**

**Daily Bed State Summary Sheet**

**Male Ward Month January**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dates | Admission | T/In | T/Out | Discharge | Death | WPD | IPD |
| 1 | - | - | - | - | 2 | - |  |
| 2 | 12 | - | - | 2 |  |  |  |
| 3 | 2 | - | 2 | 6 | - | - |  |
| 4 | - | - | 1 | 4 | - | - |  |
| 5 | - | - | - | 4 | - | - |  |
| 6 | - | - | - | - | 1 | - |  |
| 7 | 1 | 4 | - | 2 | 1 | - |  |
| 8 | - | - | - | - | 2 | - |  |
| 9 | - | - | - | - | - | - |  |
| 10 | 30 | - | - | 2 | - | - |  |
| 11 | 4 | - | 3 | 4 | 2 | - |  |
| 12 | - | - | - | 4 | - | - |  |
| 13 | - | - | - | 4 | 1 | - |  |
| 14 | 1 | - | - | 9 | 1 | - |  |
| 15 | 1 | - | - | 2 | 1 | - |  |
| 16 | - | - | - | - | 1 | - |  |
| 17 | 1 | - | - | 5 | 1 |  |  |
| 18 | 31 | - | - | 5 | 3 | - |  |
| 19 | 1 | - | 4 | 2 | - | - |  |
| 20 | - | 1 | - | 4 | 1 | - |  |
| 21 | - | - | - | 2 | 1 | - |  |
| 22 | - | - | - | - | - | - |  |
| 23 | - | - | - | - | 1 | - |  |
| 24 | - | - | - | 5 | 1 | - |  |
| 25 | 3 | - | - | 4 | 1 | - |  |
| 26 | 42 | - | - | 2 | - | - |  |
| 27 |  | - | 3 | 2 | - | - |  |
| 28 | - | - | - | 15 | 1 | - |  |
| 29 | - | - | - | 1 | - | - |  |
| 30 | - | - | - | - | - | - |  |
| 31 | - | 2 | - | 5 | - | - |  |
| **Total** |  |  |  |  |  |  |  |

Once the daily summary is completed we end up with the monthly in-patient state summary

This includes

* Total Monthly admissions for the ward or service unit
* Total monthly transfer ins
* Total monthly transfer outs
* Total monthly discharges
* Total monthly deaths
* Total monthly well people days
* Total monthly in-patients days

All these totals are in accordance to ward or service unit. When the monthly summary state is obtained an annual summary is obtained to indicate these specific monthly summaries. The layout for the annual state summary would appear as follows:

**Hospital Name**

**Monthly Bed State Summary**

**Male Ward Year 2013**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Month | Admission | T/In | T/Out | Discharge | Death | WPD | IPD |
| January |  |  |  |  |  |  |  |
| February |  |  |  |  |  |  |  |
| March |  |  |  |  |  |  |  |
| April |  |  |  |  |  |  |  |
| May |  |  |  |  |  |  |  |
| June |  |  |  |  |  |  |  |
| July |  |  |  |  |  |  |  |
| August |  |  |  |  |  |  |  |
| September |  |  |  |  |  |  |  |
| October |  |  |  |  |  |  |  |
| November |  |  |  |  |  |  |  |
| December |  |  |  |  |  |  |  |
| Total |  |  |  |  |  |  |  |

When the monthly bed state summary is prepared we end up with the annual bed state for the particular ward/service unit. This gives us annual totals of admissions, transfer in, transfer out, discharges, deaths, well people days, in-patient days.

There are also provisions for the monthly and annual bed state summary to be maintained on service/ward basis. The layout for such summary would be as follows:

* Monthly bed state service unit/ward summary

**Hospital Name**

**Monthly Ward/Service Unit Bed State Summary Sheet**

**Month……………………. Year………………….**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Ward | Admission | T/In | T/Out | Discharge | Death | WPD | IPD |
| 1 – Surgical |  |  |  |  |  |  |  |
| 2 – Maternity |  |  |  |  |  |  |  |
| 3 – Palliative |  |  |  |  |  |  |  |
| 4 – Male |  |  |  |  |  |  |  |
| 5 – Female |  |  |  |  |  |  |  |
| 6 – Pediatric |  |  |  |  |  |  |  |
| 7 – Medical |  |  |  |  |  |  |  |
| 8 – Amenity |  |  |  |  |  |  |  |
| Total |  |  |  |  |  |  |  |

* Annual bed state service unit/ward summary

**Hospital Name**

**Annual Ward/Service Unit Bed State Summary**

**Month……………………. Year………………….**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Ward | Admission | T/In | T/Out | Discharge | Death | WPD | IPD |
| 1 – Surgical |  |  |  |  |  |  |  |
| 2 – Maternity |  |  |  |  |  |  |  |
| 3 – Palliative |  |  |  |  |  |  |  |
| 4 – Male |  |  |  |  |  |  |  |
| 5 – Female |  |  |  |  |  |  |  |
| 6 – Pediatric |  |  |  |  |  |  |  |
| 7 – Medical |  |  |  |  |  |  |  |
| 8 – Amenity |  |  |  |  |  |  |  |
| Total |  |  |  |  |  |  |  |

**ANALYTICAL FORMULAE**

Available Bed Days (ABD) = Authorized/Available beds X Period (in days)

**Example 1**

A hospital had 92 beds in 1985. Calculate ABD

ABD = AB x Period (days)

= 92 x 365 = 33580

Average Available beds = ABD

Period (days)

Occupied bed Days (OBD) = ABD – VBD

**Or** OBD = % occ. x ABD

**Or** OBD = ALOS x (D + D)

**Example 2**

A hospital with 170 beds had 89% bed occupancy in 1985

OBD = 89 x (170 x 365) = 55225

100

**Or** OBD = ABD – VBD

= (170 x 365) – 11 x ABD

100

= 62050 – 6825.5 = 55225

Percentage occupancy (% Occ.) = OBD x 100

ABD

= 55225 x 100 = 89%

62050

Average # of occupied bed/ average # of patients (OB) = OBD

Period (days)

= 55225 = 152 beds

365

VBD = ABD – OBD = 62050 – 55225 = 6825

Or VBD = % un-Occ. x ABD = 11 x 62050 = 6825

100

Or VBD = TOI x (D + D)

**Other formulae**

ALOS = OBD

(D + D)

Turnover per bed/discharge per bed (TOB) = D + D

AB

Therefore AB = D + D

TOB

Excess bed days = OBD - ABD

|  |  |  |  |
| --- | --- | --- | --- |
| **Ward** | **# of (D + D)** | **# of IPD** | **# of AB** |
| **1** | 250 | 91 | 25 |
| **2** | 120 | 9900 | 30 |
| **3** | 150 | 4050 | 40 |
| **4** | 300 | 980 | 27 |
| **5** | 450 | 18000 | 50 |
| **6** | 100 | 2200 | 60 |
| **7** | 210 | 980 | 25 |
| **Total** | ***1580*** | ***36201*** | ***257*** |

**TASK/ASSIGNMENT**

The following data was recorded in Msambweni district hospital which had seven wards during the year 1982

**Calculate the following**

1. Overall hospital % occupancy
2. Overall hospital average length of stay
3. Overall hospital bed occupancy
4. Overall turnover interval
5. Overall turnover per bed
6. Excess patients days in ward 6
7. Excess patients days in ward 7
8. Extra beds needed in ward 7 to cover the extra patients sleeping on the floor during the year
9. The overall overage available beds for the hospital during the year

# SEMESTER IV

# UNIT VI (CHS 1.6) Vital and Health Statistics

1. Definition and description of various vital and health statistics rates
2. Importance of vital statistics
3. Types of vital statistics
   1. Morbidity rate
   2. Fertility rate
   3. Crude death rate
   4. Infant mortality rate
   5. Crude death rate
   6. Incidence and prevalence rates
   7. Hospital activity analysis
4. Formulae for calculating vital statistics
5. Revision of all the units

# References

1. Koch G. (2008). *Basic Allied Health Statistics and Analysis 3rd Edition*. Cengage Learning.
2. Jackson S. L. (2010). *Statistics: Plain and Simplified, 2nd Edition*. Cengege Learning.
3. Narayanan N. E. (2011). *Statistics*. PHI Learning Private.
4. Kumar R. (2011). *Research Methodology: A Step-by-Step Guide for Beginners 3rd Edition*. SAGE.
5. Healey F.J. (2010). *The Essentials of Statistics: A Tool for Social Research, 2nd Edition*. Wadsworth Cengage Learning.
6. Dawidowicz P. (2010). *Literature Reviews Made Easy: A Quick Guide to Success*. Information Age.
7. Wegner T. (2007). *Applied Business Statistics: Methods and Excel-Based Applications*. Juta & Co., Cape Town.
8. Lacey M. F., Matheson L., & Jesson J. K. (2011). *Doing Your Literature Review: Traditional and Systematic Techniques*. SAGE.
9. Hart C. (2005). *Doing a Literature Review: Releasing the Social Science Research Imagination*. SAGE.
10. Sharma J.K. (2007). *Business Statistics*. Dorling Kindersley, India.
11. Dutta D. (2006). *Statistics and Operations Research: A Unified Approach*. Laxmi.
12. Gupta S.P. (2012). *Statistical Methods, 2nd Edition*. Sultan Chad and Sons.
13. Harper W.M. (1991). *Statistics, 6th Edition*. Prentice Hall.
14. Osborn C.E. (2006). *Statistical Application for Health Information Management, 2nd Edition*. Jones and Barlett, UK.
15. Anderson D., Sweeney D., and Williams T. (2011). *Statistics for Business and Economics, 11th Edition*. Cengage Learning.

1. Dutta D. (2006). *Statistics and Operations Research: A Unified Approach* [↑](#footnote-ref-1)
2. Koch G. (2008). *Basic Allied Health Statistics and Analysis 3rd Edition* [↑](#footnote-ref-2)
3. Narayanan N. E. (2011). *Statistics* [↑](#footnote-ref-3)
4. Jackson S. L. (2010). *Statistics: Plain and Simplified, 2nd Edition* [↑](#footnote-ref-4)
5. Narayanan N. E. (2011). *Statistics* [↑](#footnote-ref-5)
6. Jackson S. L. (2010). *Statistics: Plain and Simplified, 2nd Edition* [↑](#footnote-ref-6)
7. Koch G. (2008). *Basic Allied Health Statistics and Analysis 3rd Edition* [↑](#footnote-ref-7)
8. Jackson S. L. (2010). *Statistics: Plain and Simplified, 2nd Edition* [↑](#footnote-ref-8)
9. Koch G. (2008). *Basic Allied Health Statistics and Analysis 3rd Edition* [↑](#footnote-ref-9)
10. Kumar R. (2011). *Research Methodology: A Step-by-Step Guide for Beginners 3rd Edition* [↑](#footnote-ref-10)
11. Kumar R. (2011). *Research Methodology: A Step-by-Step Guide for Beginners 3rd Edition* [↑](#footnote-ref-11)
12. Dawidowicz P. (2010). *Literature Reviews Made Easy: A Quick Guide to Success* [↑](#footnote-ref-12)
13. Gupta S.P. (2012). *Statistical Methods, 2nd Edition* [↑](#footnote-ref-13)
14. Harper W.M. (1991). *Statistics, 6th Edition* [↑](#footnote-ref-14)
15. Koch G. (2008). *Basic Allied Health Statistics and Analysis 3rd Edition* [↑](#footnote-ref-15)
16. Osborn C.E. (2006). *Statistical Application for Health Information Management, 2nd Edition* [↑](#footnote-ref-16)
17. Osborn C.E. (2006). *Statistical Application for Health Information Management, 2nd Edition* [↑](#footnote-ref-17)
18. Wegner T. (2007). *Applied Business Statistics: Methods and Excel-Based Applications* [↑](#footnote-ref-18)
19. Healey F.J. (2010). *The Essentials of Statistics: A Tool for Social Research, 2nd Edition* [↑](#footnote-ref-19)
20. Sharma J.K. (2007). *Business Statistics* [↑](#footnote-ref-20)
21. Anderson D., Sweeney D., and Williams T. (2011). *Statistics for Business and Economics, 11th Edition* [↑](#footnote-ref-21)
22. Wegner T. (2007). *Applied Business Statistics: Methods and Excel-Based Applications* [↑](#footnote-ref-22)
23. Anderson D., Sweeney D., and Williams T. (2011). *Statistics for Business and Economics, 11th Edition* [↑](#footnote-ref-23)
24. Anderson D., Sweeney D., and Williams T. (2011). *Statistics for Business and Economics, 11th Edition* [↑](#footnote-ref-24)
25. Wegner T. (2007). *Applied Business Statistics: Methods and Excel-Based Applications* [↑](#footnote-ref-25)
26. Healey F.J. (2010). *The Essentials of Statistics: A Tool for Social Research, 2nd Edition* [↑](#footnote-ref-26)
27. Healey F.J. (2010). *The Essentials of Statistics: A Tool for Social Research, 2nd Edition* [↑](#footnote-ref-27)
28. Sharma J.K. (2007). *Business Statistics* [↑](#footnote-ref-28)
29. Anderson D., Sweeney D., and Williams T. (2011). *Statistics for Business and Economics, 11th Edition* [↑](#footnote-ref-29)
30. Anderson D., Sweeney D., and Williams T. (2011). *Statistics for Business and Economics, 11th Edition* [↑](#footnote-ref-30)
31. Sharma J.K. (2007). *Business Statistics* [↑](#footnote-ref-31)
32. Sharma J.K. (2007). *Business Statistics* [↑](#footnote-ref-32)
33. Sharma J.K. (2007). *Business Statistics* [↑](#footnote-ref-33)
34. Wegner T. (2007). *Applied Business Statistics: Methods and Excel-Based Applications* [↑](#footnote-ref-34)
35. Wegner T. (2007). *Applied Business Statistics: Methods and Excel-Based Applications* [↑](#footnote-ref-35)
36. Sharma J.K. (2007). *Business Statistics* [↑](#footnote-ref-36)
37. Anderson D., Sweeney D., and Williams T. (2011). *Statistics for Business and Economics, 11th Edition* [↑](#footnote-ref-37)