UIDLC

PSY103
Quantitative Methods in PSYCHOLOGY

Course Manual

Benjamin Osayawe Ehigie, Ph.D

COURSE MANUAL

Quantitative Methods in Psychology

PSY103

University of Ibadan Distance Learning Centre
Ibadan Open and Distance Learning Course Series Development
Version ev1

Copyright

ISBN: 978-021-278-7

General Editor: Prof. Bayo Okunade

Page layout, instructional design & development by EDUTECHportal,
www.edutechportal.org

University of Ibadan Distance Learning Centre

University of Ibadan,
Nigeria

Telex: 31128NG

Tel: +234 (80775935727)
E-mail: ssu@dlc.ui.edu.ng
Website: www.dlc.ui.edu.ng

Vice-Chancellor's Message

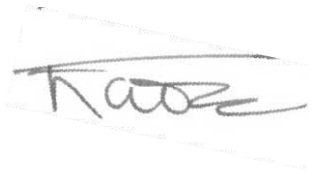
The Distance Learning Centre is building on a solid tradition of over two decades of service in the provision of External Studies Programme and now Distance Learning Education in Nigeria and beyond. The Distance Learning mode to which we are committed is providing access to many deserving Nigerians in having access to higher education especially those who by the nature of their engagement do not have the luxury of full time education. Recently, it is contributing in no small measure to providing places for teeming Nigerian youths who for one reason or the other could not get admission into the conventional universities.

These course materials have been written by writers specially trained in ODL course delivery. The writers have made great efforts to provide up to date information, knowledge and skills in the different disciplines and ensure that the materials are user-friendly.

In addition to provision of course materials in print and e-format, a lot of Information Technology input has also gone into the deployment of course materials. Most of them can be downloaded from the DLC website and are available in audio format which you can also download into your mobile phones, IPod, MP3 among other devices to allow you listen to the audio study sessions. Some of the study session materials have been scripted and are being broadcast on the university's Diamond Radio FM 101.1, while others have been delivered and captured in audio-visual format in a classroom environment for use by our students. Detailed information on availability and access is available on the website. We will continue in our efforts to provide and review course materials for our courses.

However, for you to take advantage of these formats, you will need to improve on your I.T. skills and develop requisite distance learning Culture. It is well known that, for efficient and effective provision of Distance learning education, availability of appropriate and relevant course materials is a *sine qua non*. So also, is the availability of multiple plat form for the convenience of our students. It is in fulfillment of this, that series of course materials are being written to enable our students study at their own pace and convenience.

It is our hope that you will put these course materials to the best use.



Prof. Isaac Adewole

Vice-Chancellor

Foreword

As part of its vision of providing education for "Liberty and Development" for Nigerians and the International Community, the University of Ibadan, Distance Learning Centre has recently embarked on a vigorous repositioning agenda which aimed at embracing a holistic and all encompassing approach to the delivery of its Open Distance Learning (ODL) programmes. Thus we are committed to global best practices in distance learning provision. Apart from providing an efficient administrative and academic support for our students, we are committed to providing educational resource materials for the use of our students. We are convinced that, without an up-to-date, learner-friendly and distance learning compliant course materials, there cannot be any basis to lay claim to being a provider of distance learning education. Indeed, availability of appropriate course materials in multiple formats is the hub of any distance learning provision worldwide.

In view of the above, we are vigorously pursuing as a matter of priority, the provision of credible, learner-friendly and interactive course materials for all our courses. We commissioned the authoring of, and review of course materials to teams of experts and their outputs were subjected to rigorous peer review to ensure standard. The approach not only emphasizes cognitive knowledge, but also skills and humane values which are at the core of education, even in an ICT age.

The development of the materials which is on-going also had input from experienced editors and illustrators who have ensured that they are accurate, current and learner-friendly. They are specially written with distance learners in mind. This is very important because, distance learning involves non-residential students who can often feel isolated from the community of learners.

It is important to note that, for a distance learner to excel there is the need to source and read relevant materials apart from this course material. Therefore, adequate supplementary reading materials as well as other information sources are suggested in the course materials.

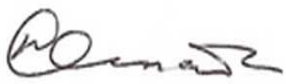
Apart from the responsibility for you to read this course material with others, you are also advised to seek assistance from your course facilitators especially academic advisors during your study even before the interactive session which is by design for revision. Your academic advisors will assist you using convenient technology including Google Hang Out, You Tube, Talk Fusion, etc. but you have to take advantage of these. It is also going to be of immense advantage if you complete assignments as at when due so as to have necessary feedbacks as a guide.

The implication of the above is that, a distance learner has a responsibility to develop requisite distance learning culture which includes diligent and disciplined self-study, seeking available administrative and academic support and acquisition of basic information technology skills. This is why you are encouraged to develop your computer skills by availing yourself the opportunity of training that the Centre's provide and put these into use.

In conclusion, it is envisaged that the course materials would also be useful for the regular students of tertiary institutions in Nigeria who are faced with a dearth of high quality textbooks. We are therefore, delighted to present these titles to both our distance learning students and the university's regular students. We are confident that the materials will be an invaluable resource to all.

We would like to thank all our authors, reviewers and production staff for the high quality of work.

Best wishes.



Professor Bayo Okunade

Director

Course Development Team

The University of Ibadan Distance Learning Centre wishes to thank those below for their contribution to this course manual:

Course Writer / Facilitator	Benjamin Osayawe Ehigie, Ph.D
Content Editor	Prof. Remi Raji-Oyelade
Production Editor	Dr. Gloria O. Adedoja
Learning Design & Technologist	Folajimi Olambo Fakoya
Managing Editor	Ogunmefun Oladele Abiodun
General Editor	Prof. Bayo Okunade

Contents

About this course manual

Quantitative Methods in Psychology PSY103 has been produced by University of Ibadan Distance Learning Centre. All Psychology course manuals produced by University of Ibadan Distance Learning Centre are structured in the same way, as outlined below.

How this course manual is structured

The course overview

The course overview gives you a general introduction to the course. Information contained in the course overview will help you determine:

- If the course is suitable for you.
- What you will already need to know.
- What you can expect from the course.
- How much time you will need to invest to complete the course.

The overview also provides guidance on:

- Study skills.
- Where to get help.
- Course assessments and assignments.
- Activity icons.
- Study sessions.

We strongly recommend that you read the overview *carefully* before starting your study.

The course content

The course is broken down into study sessions. Each study session comprises:

- An introduction to the study session content.
- Learning outcomes.
- Content of study sessions with learning activities.
- A study session summary.
- Assessments and/or assignment, as applicable.

Your comments

After completing this course, Quantitative Methods in Psychology, we would appreciate it if you would take a few moments to give us your feedback on any aspect of this course. Your feedback might include comments on:

- Course content and structure.
- Course reading materials and resources.
- Course assessments.
- Course assignments.
- Course duration.
- Course support (assigned tutors, technical help, etc).
- Your general experience with the course provision as a distance learning student.

You might forward your comments to coursereview@dlc.ui.edu.ng or post same on comment pad of your course website at UI Mobile Class. Your constructive feedback will help us to improve and enhance this course.

Course overview

Welcome to Quantitative Methods in Psychology PSY103

Psychology is a science and, therefore, adopts scientific procedures in its study, including the computation aspects. For every student desiring to study psychology, therefore, the knowledge of statistics is essential. This course - PSY103 - will expose you to the quantitative skills needed for scientific research.

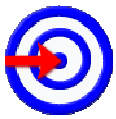
This course manual supplements and complements PSY103 UI Mobile Class Activities as an online course. The UI Mobile Class is a virtual platform that facilitates classroom interaction at a distance where you can discuss / interact with your tutor and peers while you are at home or office from your internet-enabled computer. You will also use this platform to submit your assignments, receive tutor feedback and course news with updates.

Quantitative Methods in Psychology PSY103—is this course for you?

PSY103 is a compulsory course for undergraduate students of psychology. The course attempts to equip you with quantitative skills needed for scientific research.

Course outcomes

Upon a successful completion of Quantitative Methods in Psychology PSY103 you will be able to:



Outcomes

- *explain* the concept of statistics.
- *justify* why statistics is essential in the study of psychology.
- *differentiate* among the types of measurements used for data collection in psychological research.
- *collect* and *summarize* data; and present facts in psychological research.
- *expose* data to various forms of descriptive statistical analysis, graphs and diagrammatic presentation.

Timeframe

How long?

This is a one semester course.

45 hours of formal study time is required.

Study skills

As an adult learner your approach to learning will be different to that from your school days: you will choose what you want to study, you will have professional and/or personal motivation for doing so and you will most likely be fitting your study activities around other professional or domestic responsibilities.

Essentially you will be taking control of your learning environment. As a consequence, you will need to consider performance issues related to time management, goal setting, stress management, etc. Perhaps you will also need to reacquaint yourself in areas such as essay planning, coping with exams and using the web as a learning resource.

Your most significant considerations will be *time* and *space* i.e. the time you dedicate to your learning and the environment in which you engage in that learning.

We recommend that you take time now—before starting your self-study—to familiarize yourself with these issues. There are a number of excellent web links & resources on the "Self-Study Skills" at your course website.

Need help?

Help

As earlier noted, this course manual complements and supplements PSY103at UI Mobile Class as an online course.

You may contact any of the following units for information, learning resources and library services.

Distance Learning Centre (DLC)
University of Ibadan, Nigeria
Tel: (+234) 08077593551 – 55
(Student Support Officers)
Email: ssu@dlc.ui.edu.ng

Head Office
Morohundiya Complex, Ibadan-Ilorin Expressway, Idi-Ose, Ibadan.

Information Centre
20 Awolowo Road, Bodija,
Ibadan.

Lagos Office
Speedwriting House, No. 16
Ajanaku Street, Off Salvation
Bus Stop, Awuse Estate, Opebi,
Ikeja, Lagos.

For technical issues (computer problems, web access, and etcetera), please send mail to webmaster@dlc.ui.edu.ng.

Academic Support

Help

A course facilitator is commissioned for this course. You have also been assigned an academic advisor to provide learning support. The contacts of your course facilitator and academic advisor for this course are available at onlineacademicsupport@dlc.ui.edu.ng

Activities

Activities

This manual features "Activities," which may present material that is NOT extensively covered in the Study Sessions. When completing these activities, you will demonstrate your understanding of basic material (by answering questions) before you learn more advanced concepts. You will be provided with answers to every activity question. Therefore, your emphasis when working the activities should be on understanding your answers. It is more important that you understand why every answer is correct.

Assessments

Assessments

There are three basic forms of assessment in this course: in-text questions (ITQs) and self assessment questions (SAQs), and tutor marked assessment (TMAs). This manual is essentially filled with ITQs and SAQs. Feedbacks to the ITQs are placed immediately after the questions, while the feedbacks to SAQs are at the back of manual. You will receive your TMAs as part of online class activities at the UI Mobile Class. Feedbacks to TMAs will be provided by your tutor in not more than 2 weeks expected duration.
Schedule dates for submitting assignments and engaging in course / class activities is available on the course website. Kindly visit your course website often for updates.

Bibliography

Readings

For those interested in learning more on this subject, we provide you with a list of additional resources at the end of this course manual; these may be books, articles or websites.

Getting around this course manual

Margin Icons

While working through this course manual you will notice the frequent use of margin icons. These icons serve to "signpost" a particular piece of text, a new task or change in activity; they have been included to help you to find your way around this course manual.

A complete icon set is shown below. We suggest that you familiarize yourself with the icons and their meaning before starting your study.

 Activity	 Assessment	 Assignment	 Case study
 Discussion	 Help	 Outcomes	 Reflection
 Study skills	 Summary	 Time	 Tip

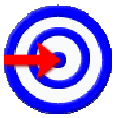
Study Session 1

The Meaning of Statistics

Introduction

a compulsory course for undergraduate students of psychology. The course

Statistics is perceived and interpreted differently by different professionals. This study session will therefore expose you to alternative definitions and explanations of the concept of statistics.



Learning Outcomes

When you have studied this session, you should be able to:

- *explain* the concept of statistics, in grammatical terms, and as used by scientific researchers. (SAQ 1.1)
- *explain* the two basic types of statistics. (SAQ1.2)

1.1 The Definition of Statistics

The word 'statistics' can be used in three different perspectives:

1. Layman's perspective
2. Research perspective
3. Grammar perspective

1.1.1 Layman's Perspective

As used in everyday language, statistics implies a collection of numerical data. This may be expressed in terms of the number of students in a school, the number of employees in a company, the number of children given birth to in a year, the number of accidents recorded in a year and so forth.

- **ITQ** In a lay-person's understanding, statistics means:
  - A. Addition, division, and subtraction of data.
  - B. A collection of numerical data.

Feedback on ITQs answers

- The correct answer was B
- If you have chosen A (addition, division and subtraction of data), then you have defined based on mathematical experience in primary school. In fact the question asks you to define statistics as a collection of numerical data.

1.1.2 Research Perspective

Statistics may also refer to the scientific methods for collecting, organizing, describing, analyzing, presenting, and interpreting data in line with the purpose for which it is required to serve. Valid conclusions and reasonable decisions are subsequently based on such analysis. Hence,

statistical results help researchers to make inferences about occurrences in the environment.

- **ITQ** From the empirical research's point of view, statistics means:

  A. Organization of data for presentation purposes.
  B. A collection of numerical data.
  C. Purposeful use of scientific methods to collect, organize, analyze and interpret data.

Feedback on ITQs answers

- If you have chosen A, then you have defined based on part of this definitive point of view.
- If you have chosen B, then you have defined based on a definitive perspective of layperson's view of statistics. In fact, the question asks you to define statistics as the scientific methods for collecting, organizing, describing, analyzing, presenting and interpreting data in line with the purpose for which it is required to serve.
- The answer is C.

1.1.3 Grammatical Perspective

Statistics may as well be used as a plural of statistic. Statistic is used as a summary measure of a set of data, after adequate analysis has been made. The mean or average, for instance, is a statistic when used as a summary measure of students' performance. When other summary measures like mode, median, standard deviation and variance are considered collectively, the word 'statistics' is most adequate.

- **ITQ** In grammatical terms, the word "statistic" implies:

  A. A fact from a study of data.
  B. A collection of numerical data.

Feedback on ITQs answers

- The correct answer was A. In grammar, statistic is the summary of measured data after adequate analysis has been made.
- Option B (a collection of numerical data) is a definition based on a lay-person's perspective of statistics. In fact the question asks you to define statistics as a summary of a measured set of data, after adequate analysis has been made.

1.2 Types of Statistics

In research parlance, there are basically two types of statistics: descriptive statistics and inferential statistics.

1.2.1 Descriptive Statistics

This type of statistics summarizes or describes the characteristics of a set of data in a clear and convenient way. Descriptive statistics mark the development of the discipline called statistics. They include all calculations, which reduce data to smaller sizes in which results of such

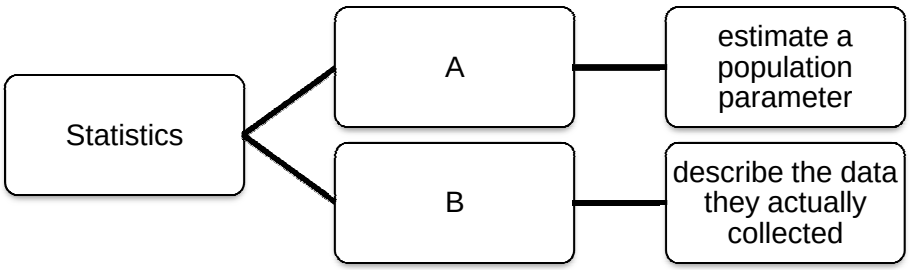
calculations could as well be presented in tabular form and diagrams or charts.

1.2.2 Inferential Statistics

Here, data are analyzed and interpreted in such a way that conclusion about the population is made, based on data received on the sample. Inferential statistical methods are devised to measure the degree of confidence on how close an estimated sample statistic is to the actual population parameter. This degree of confidence is expressed in terms of probability.

Descriptive statistics are considered as the preliminary stage for inferential statistics. Results of inferential statistics are what can be used in decision-making. Examples of descriptive statistics are the measures of central tendency (i.e. median, mode, and median), measures of variability (e.g. standard deviation, variance etc.), frequency distributions, and transformed scores. Inferential statistics include the t-test, analysis of variance (ANOVA), chi-square test, etc.

- **ITQ B** Look at the diagram below and fill the empty boxes A and with the most appropriate options.



Feedback on ITQs answers

- If you have chosen considered inferential statistics as option A then you are correct. Researchers use inferential statistics if they are using their results to estimate a population parameter
- Option B is in fact descriptive statistics. Researchers use descriptive statistics if they are using the results to describe the data they actually collected.

Study Session Summary

Summary

In this Study Session, the concept of statistics has been approached from three perspectives. First, from a lay-person's perspective that statistics is a collection of numerical data. Second, from a researcher's view that statistics is a scientific method of collecting, organizing, describing, analyzing, presenting and interpreting data. Third, in grammatical terms, statistics is a plural form of statistic.

You also learnt about types of statistics which are: descriptive and inferential statistics. The relationship between these types of statistics was also highlighted.

Assessment

Assessment

Now that you have completed this study session, you can assess how well you have achieved its Learning Outcomes by answering these questions. Write your answers in your Study Diary and discuss them with your Tutor at the next Study Support Meeting. You can check your answers with the Notes on the Self-Assessment Questions at the end of this Manual.

SAQ 1.1 (tests Learning Outcome 1.1)

> The definitive stands of statistics can be explained from three points of views (layperson's view, grammatical view and research view). What are the differences between these perspectives?

SAQ 1.2 (tests Learning Outcome 1.2)

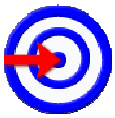
> A clinical psychologist conducted a study in which she gave some of her clients a new depression treatment. What form of statistics would she use if she wanted to describe the average depression score of only those clients who got the treatment?

Study Session 2

Why we Study Statistics

Introduction

This Study Session aims to provide explanations on why students of psychology and other related discipline require the knowledge of statistics.



Learning Outcomes

When you have studied this session, you should be able to:

- *justify* the inclusion of statistics in the study of psychology. (SAQ2.1)

Rationale for the Inclusion of Statistics into Psychology

Quantification translation to figures.

Statistics a scientific method of collecting, organizing, describing, analyzing, presenting and interpreting data.

In the present scientific age, almost all branches of knowledge; arts or the sciences is involved in one type of statistics or another e.g. **quantification**. People in the behavioural sciences (like psychology, sociology, economics, political science, and education) require the knowledge of **statistics** for the following reasons:

1. To understand professional literature like journals, magazines and other periodicals the knowledge of statistics is inevitable. Research reports published in scientific journals and books are based on statistical analyses. You will always find it difficult to understand many pertinent articles in these sources without the knowledge of statistics. Similarly, you will find it difficult to plan your own research to meet scientific standards without the knowledge of statistics.
2. For a scientific research to be taken serious, it must be statistical. Inferences are drawn about a population based on data obtained and analyzed from a sample. Thus, ignorance of statistics will limit your comprehension of inferences drawn in research reports, and you will not be skilled to make genuine inferences from your research.
3. As scientists, we require the knowledge of statistics to progress in any quantitative scientific pursuit. To carry out any significant behavioural science research, we have to engage in statistical analysis. In conducting successful research, we need to design the statistical analysis to test our hypotheses, prior to the collection of our data. By so doing, much information will be contained in our data and the objective of the research will be achieved. Even when a computer is to be used for analysis, a computer programmer needs to be guided on the adequate statistics needed to test hypotheses, which conform to the data-collected. When data are

wrongly collected, due to ignorance of the knowledge of statistics, the goal of the research will be defeated.

4. People acquainted with statistics accept that research, especially in the behavioural sciences, is mainly educated guesses. In other words, there are no perfect rights or wrong answers in any research, and no single research has a conclusive statement on any issue. The knowledge of statistics thus helps us to evaluate the strengths and weaknesses inherent in any research, in terms of the techniques adopted by the researcher in collecting information and drawing inferences.
5. The knowledge of statistics also helps us to be aware of our limitations as human beings in conducting our personal research. Sampling, research designs and statistical analysis are devices employed to check these inherent defects, and we are thereby warned of the dangers of invalid, biased or over-generalized conclusions. Thus, with the knowledge of statistics, we are aided to make less subjective conclusions.
6. The knowledge of statistics is indispensable to the absorption of most written information around us, for example in newspapers. In actual fact, data become more meaningful when treated statistically. For instance, comparisons and relationships are easily made when figures or numerals are used, and they are much clearer than verbal descriptions.

Study Session Summary

Summary

In this Study session you learned that statistics enables us to progress in scientific pursuit because data become more meaningful when treated statistically. The knowledge of statistics helps us to be aware of our limitations as human beings in conducting our personal research. The knowledge of statistics also helps us to evaluate the strengths and weaknesses inherent in any research.

Assessment

Assessment

Now that you have completed this study session, you can assess how well you have achieved its Learning Outcomes by answering these questions. Write your answers in your Study Diary and discuss them with your Tutor at the next Study Support Meeting. You can check your answers with the Notes on the Self-Assessment Questions at the end of this Manual

SAQ 2.1 (tests Learning Outcome 2.1)

Identify at least four reasons why it is necessary for you to study statistics?

Discussion

So far, which of the provided points in SAQ2.1 will you consider as most cogent and why? Or do you have a better reason or different perspective?

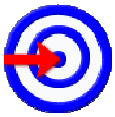
Post your response on UI Mobile Class forum for discussion.

Study Session 3

Common Terms used in Statistics

Introduction

The teaching of statistics involves the understanding of concepts that are used in scientific research. This study session will therefore focus on scientific concepts that are used in statistics.



Learning Outcomes

When you have studied this session, you should be able to:

- *explain* the meanings of concepts such as variables, data, population, parameter, and sample. (SAQ3.1)

3.1 Variable

Variables are any measurable feature, which has the potential of taking different values when it is quantified. It includes such human features like weight, sex, intelligence quotient (I.Q), etc. These human characteristics can assume different values when different individuals are evaluated, or when the same individuals are evaluated on different occasions.

3.1.1 Classification of Variables

Variables in research are classified into three:

Independent Variables (IV)

Independent variables are those that are easily manipulated by researchers so as to determine their effects on another set of variables called the dependent variables. In fact, the independent variables are the variables of interest in a research.

Dependent Variables (DV)

Dependent variables change in values as a result of changes in the independent variables. It is the values obtained on the dependent variables that constitute the data needed for statistical analysis in a research.

- **ITQ** The IV (independent variable) in a study is the

  A. variable expected to change the outcome variable.
  B. a outcome variable.

Feedback on ITQs answers

- A is correct.
- If you have chosen B, you are wrong. DV (dependent variable) is the outcome variable that is used to compare the effects of the different IV levels.

Extraneous Variables

Extraneous variables can be likened to weeds in a farmland. An okra vegetable found in a maize farm is regarded as a weed even though it is edible; because it is not the originally planted crop.

These are also referred to as secondary, nuisance, or concomitant variables. They are variables that have similar effect on the dependent variables like the independent variables, but they are not the independent variables of concern.

Extraneous variables have to be checked in every research to prevent their contaminating the dependent variable and, thereby, providing alternative explanations for any change observed. An example is a research designed to examine the effect of noise on academic performance of students. The IV is noise while the DV is academic performance. A possible extraneous variable is the students' level of intelligence, class of study, age, and the like.

3.1.2 Measurement of Variables

Variables can be measured in two forms: discrete and continuous measurements.

A discrete measurement of variables

This can take up only finite values of raw scores with no fraction or decimals. It could be countable data, such as the number of students who registered for a course. Data generated from the conversion of qualitative data to quantitative data contain discrete values. Discrete values are in the form of 1, 20, 70, 200 etc.; these are always whole numbers.

A continuous measurement of variables

This form of measurement can take up any values within a given range. Thus, they accommodate fractions or decimals. For instance, between the range of 1 and 2, we may have 1.33, 1.52, 1.75 to 2. Data on continuous measurement normally come from interval or ratio scales. These include measurements like height, weight, money etc., which could be expressed in decimals or fractions like 5.2 metres, 70.3 kilograms or N7.40.

- **ITQ** The DV (dependent variable) in a study is the

  A. variable expected to change the outcome variable.
  B. a outcome variable.

Feedback on ITQs answers

- A is not correct. In fact a variable that is expected to change the outcome variable is the manipulated variable which is otherwise known as IV (independent variable).
- B is correct.

3.2 Data

Data information in raw or unorganized form (such as numbers or symbols) that refers to conditions or objects.

Collections of numbers or measurements used to represent or quantify observation during research before transformation to other statistical forms is referred to as raw **data.** For instance, students' scores in an examination can be used to represent academic performance, and could be taken as data for the research on effect of noise on academic

performance.

3.2.1 Types of Data

Data can be grouped into two: quantitative or numerical data and qualitative or non-numerical data.

Quantitative data arise from measurable characteristics of objects such as height, weight, length, width, IQ scores, age, anxiety level, income and the like. On the other hand, **qualitative data** arise from non-numerical characteristics of objects, such as sex, religion, occupation, marital status, colour and some Yes or No responses.

The numerical characteristics of objects are referred to as variables, while the non-numerical characteristics are called attributes. However, non-numerical data could be converted to numerical data by counting the number of cases falling in a given attribute. This may vary from one to infinity. By so doing, such values can also be called variables, just like the numerical data. An example is sex as an attribute, which may be expressed in terms of the number of males and the number of females. Sex is thereby converted to a numerical value called variable.

3.3 Population

This is also referred to as universe. It is ideally defined as a large group of people, animals, objects, responses, or measurements that are alike in at least one respect. Thus, we have a population of all Yoruba students in the University of Ibadan, a population of all 'white rats' of a given genetic strain, the population of undergraduate students in the University of Nigeria, and the population of all federal government workers in Ibadan.

3.3.1 Parameter

This refers to the describable features of a population. They are the features or characteristics that a collection of people or objects possess that allow them to be called a population. For example, a population of undergraduate students has the feature of grouping all undergraduates, and thereby excluding the postgraduate students.

3.3.2 Sample

Representative Sample A subset of a statistical population that accurately reflects the members of the entire population.

This is a subset of a population, usually drawn from the population randomly or otherwise, for the purpose of generalization to the population. In the behavioural sciences and education, it is not often easy to study the entire population of interest. Consequently only a part of the population is taken for study and this part is representative of the entire members in the population under study. The part selected for a study is the sample. When the sample is selected in scientific ways, all the characteristics of the population are believed to be reflected in the chosen sample. Such sample is then referred to as a **representative sample** of the population. In a class of 60 students, 10 of them could be selected as a sample if the selection is done randomly. We will discuss the concept of sampling in the next Study Session.

Activity 3.1

Allow 15 minutes

So far, we have discussed some statistical terminologies. Now work through the following activity, referring to this Study Session. To complete the activity, you need a small collection of different types of books in your office or home (about 20 will do)

1. What data do you want to collect?
2. If you collected data initially as raw data, how would you record this data?
3. Make a tally of your data – use the table below. Add lines as necessary.

Major Colour of Front cover	Tally	Frequency

4. On the table above, as indicated in the third column, record the frequencies of the colours (answer in same table).
5. Is it necessary to group your data? Explain your answer.
6. Can you find the range for this data? Explain your answer.
7. What form of data collection did you use?
8. Did you use random selection to find your data?
9. What is the population of the data that you have collected?
10. Did you need to choose information from a sample when you collected your data?

NOTE

You may refer to Statistics (pages 1-3), an additional reading material in the bibliographic section for further help. This material is free and available for downloading at your course website (UI Mobile Class).

Activity 3.2

Allow 5 minutes

Once you have collected and recorded data in a table like you have in questions 3 and 4 of the activity above, you have **a frequency table** of your data. If you look at the frequency table, you can start to answer descriptive and interpretive questions about your data such as the questions in the following activity.

1. Which colour was the most common in your collection of books?
2. Which colour was the least common in your collection of books?
3. Do you think that another collection of books would have the same numbers of different colours as you have found?

Study Session Summary

Summary

In this Study Session, we defined and provided explanations on some statistical concepts. Anything that can change in value or form is referred to as a variable; it could exist as independent, dependent or extraneous variable. Measurement of variables could be expressed in discrete or whole number form, as well as in continuous form or measures with decimals or fractions. A collection of objects with similar features is called a population. The features which the members of a population possess that characterize their being called a population are referred to as parameter. A subset that is drawn from a population is a sample. Numbers or a measurement used to represent or quantify observations during research is called data, which could be qualitative or quantitative.

Assessment

Assessment

Now that you have completed this study session, you can assess how well you have achieved its Learning Outcomes by answering these questions. Write your answers in your Study Diary and discuss them with your Tutor at the next Study Support Meeting. You can check your answers with the Notes on the Self-Assessment Questions at the end of this Manual

SAQ 3.1 (tests Learning Outcome 3.1)

Carefully study the scenario in the short passage below, and fill in the blank spaces.

Mr. Biodun, a researcher studying depression gave a new treatment to a sample of 100 people with depression out of a____A____, a group of all things that share a set of characteristics. In this case, the “things” are people, and the characteristic they all share is depression.

Mr. Biodun wants to know what the mean depression score for the population would be if all people with depression were treated with the new depression treatment. In other words, he wants to know the____B___, the value that would be obtained if the entire population

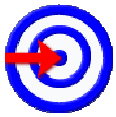
were actually studied. Of course, Mr. Biodun don't have the resources to study every person with depression in the world, so he must instead study a____C___, a subset of the population that is intended to represent the population. In most cases, the best way to get a sample that accurately represents the population is by taking a ___D___ from the population, when each individual in the population has the same chance of being selected for the sample.

So, he then uses the sample statistic value as an estimate of the population parameter value. When researchers use a sample statistic to infer the value of a population parameter it is called ___E___.

Study Session 4

Sampling

Introduction



Learning Outcomes

This Study Session will expose you to the meaning of sampling, and the various techniques of conducting sampling, specifically in social science research setting.

When you have studied this session, you should be able to:

- *use* sampling technique. (SAQ 4.1)
- *differentiate* between probability and non-probability sampling techniques. (SAQ 4.1 and 4.2)

4.1 What is Sampling?

Sampling is the process of taking or selecting any portion of a population or universe, scientifically, as representative of that population or universe. A representative sample is one which has all the characteristics of the population under study. A selection technique in which every member of the population has equal chance of being selected is called **random sampling.** On the other hand, a selection technique in which the members of the population do not have equal chances of being selected is called, **biased sampling**; each member has a higher or lower probability of being selected than others.

4.1.1 Important Concepts in Sampling

Sample Frame: By sampling frame, we refer to a list of items in the population from which the sample is drawn. For instance, the names of all undergraduate students in the University of Ibadan could stand as a sampling frame if the population is defined as undergraduate students of the university. To select the sample randomly, we need a complete sampling frame.

Sample Size: It is the total number of items in the population that is actually selected for a research. The sample size in the example given may be 200 undergraduate students of the University of Ibadan. The sample size desirable for any given research is determined by the sampling frame, the reliability of results expected, and costs of managing the sample. Generally, however, the greater the sample size, the higher its reliability and accuracy, but the higher its costs. This is because a larger sample size increases the representativeness of the population characteristics, but it could be more difficult to manage, in terms of conducting the research.

4.2 Sampling Techniques

These are the methods used to select a sample for a research. There are two main techniques for selecting samples; these are probability and non-probability sampling.

Probability Sampling: In this case, some forms of random sampling are used in one or more of their stages.

Non-probability Sampling: Non-probability techniques do not use random sampling.

4.2.1 Probability Sampling Techniques

A. Simple Random Sampling

By this technique, all members of interest have equal probabilities or chances of being selected. A common method adopted is the use of table of random numbers. The table of random numbers is generated by the computer and contains decimal digits of 0 to 9. In usage, the researcher should first prepare the sampling frame by listing all members of the population in alphabetical order, and number them serially. The sample size is decided before entering the table of random numbers.

To start the selection, the researcher enters the table at any point, without bias. To ensure objectivity, the researcher may close his or her eyes and just touch any number with the tip of a pencil. The random number digits can be read sideways (horizontally), or up and down (vertically) from this starting point. The number of random digits to select depends on the members of the population in the sampling frame. If they are between 1 and 99 then two-digit random numbers are selected at a time, from the table. The serial numbering of the members of the population in the sampling frame would range from 01 to 99, to reflect what would be observed on the table of random numbers. A frame having members up to 999 requires the use of three-digit random numbers. Thus, the serial numbers should range from 001 to 999.

A selected random number is taken as the serial number of the corresponding member of the sample frame in the population, and by inference that member is included in the sample as a participant. This exercise continues until the total number of participants desired for the research is exhausted. When a number greater than any in the frame is selected, it is discarded and the process continues until the sample size required is obtained.

Another method of simple random sampling is **balloting**. By this method, members of the population are also arranged alphabetically and numbered serially. Numbers are then written on pieces of papers to cover all the numbers in the population. Each number that is written on a slip of paper is folded, and kept in a container. After all the papers are folded and placed in the container, the papers are shuffled after which selection of papers from the container is made, to cover the sample size. Each selected number is taken as the serial number of the corresponding member of the population; the member is thus included in the sample.

B. Stratified Random Sampling

In stratified random sampling, the population is divided into subgroups, called strata. Each stratum is defined by unique characteristics of each subgroup in the population. In a population we may have our strata as males and females, blacks and whites, junior employees and senior employees, etc. Within each stratum, simple random sampling is used to select participants who will be included in the sample, hence, it is called stratified random sampling.

This sampling technique is used mostly when a researcher notices that some peculiar features of the population have to be adequately represented in the sample, which may not be achieved if the simple random sampling technique only is used. The simple random sampling may not reflect the heterogeneity of some populations. By dividing a population into distinct subgroups or strata, with each stratum differing from others, internal homogeneity within each stratum is ensured. Employing the simple random selection technique in each stratum presents a sample which reflects the heterogeneity of the population. Such sample is, therefore, a better representative of the population.

To improve on the heterogeneity of the population while adopting the stratified random sampling, the researcher may decide to select equal number of participants from each stratum or the selection may reflect the ratio in which strata exist in the population. For instance, regardless of the number of males and females in the sample frame, a researcher may decide to select just 20 males and 20 females as the sample. On the other hand, if he/she had 300 males and 100 females in the sample frame, this gives a ratio of 3:1. In selecting participants for a sample size of 40, using the simple random sampling in each stratum, he/she could select ¾ x 40 i.e. 30 males and ¼ x 40 i.e. 10 females. The 40 participants needed will, therefore, mean 40 males from the 300 available and 10 females from the 100 available.

However, the stratified random sampling is not adequate where a sampling frame does not exist or cannot be easily compiled. For instance, having a sampling frame for the population of Nigerian youths may be rather difficult. Also when a sample frame exists but members of the population are so scattered that one may not be able to clearly spell out a stratum as distinct from another, then the stratified random sampling is not adequate.

C. Cluster Sampling

Cluster sampling is best used when stratified sampling cannot be used. It is a multi-stage sampling technique that requires successive random sampling of the target sampling units or sets and subsets. It can be a two-stage, three-stage, or more exercises. For instance, states in a country can be randomly selected; the towns, streets, family, individual household, and individuals in a household may be subsequently selected.

D. Systematic Sampling

Systematic sampling entails the random selection of the first person and successive persons at a particular interval, called the nth interval. It starts by the alphabetical arrangement of all members of the population and

serial numbering of these. By this, the total numbering of the sample frame is determined. The researcher then decides on the sample size for his/her research. To determine the nth term for random selection, divide the population size by the sample size desired. Thus, if the population size is 100 and the sample size desired is 10, then the nth term becomes 100/10 i.e. 10. In selecting participants for the sample, every 10^{th} person in the sample frame is selected. Thus, the 10^{th}, 20^{th}, 30^{th}, 40^{th}, 50^{th}, - - - 100^{th} members in the frame are selected to make the sample. This will finally give us a total of 10 participants as the sample size.

4.2.2 Non-probability Sampling

A. Accidental Sampling

It is the weakest form of sampling but the most frequently used in survey research. One simply takes available samples at hand, rather than embark on the laborious process of obtaining a probability sample. A good example is the distribution of research questionnaires indiscriminately to people as they are found. Also are interviews of persons on the street by news reporters.

This technique is used when both time and money are limited. It is quick and easy to conduct accidental sampling. However, there is limitation of generalizing the result of the sample to the entire population; the results and interpretation are limited to the sample.

B. Quota Sampling

Quota sampling is the most systematic and scientific of the three non-probability sampling techniques. It involves dividing the population into a number of segments, just like the stratified random sampling. However, unlike the stratified random sampling, in this case, the researcher arbitrarily selects a quota sample from each segment. Thus, selection is not random so quota sampling does not ensure the selection of a representative sample. Despite this shortcoming, quota sampling remains an attractive alternative to stratified sampling.

C. Purposive Sampling

This is characterized by the use of judgment and a deliberate effort to constitute a representative sample by including presumably typical areas or groups in the sample. Thus, the degree of representativeness of the sample depends on the good judgment and luck of the researcher. However, results from such sample may not be accurate reflection of the population characteristics.

- **ITQ** Which of the following sampling techniques is more objective than the others?

  A. Cluster sampling
  B. Simple random sampling
  C. Purposive sampling
  D. Accidental sampling

Feedback on ITQs answers

- The correct answer is A because cluster sampling involves random sampling within specialized group clusters.
- If you chose B, you would be wrong because the simple random sampling may not reflect the heterogeneity of some populations.
- If you chose C, or D, you would be wrong. Purposive, and accidental sampling techniques are non-probability sampling, and do not allow for scientific objectivity. They do not provide all members of the sample frame equal probability or chance of being included in the sample size.

Study Session Summary

Summary

In this Study session you learned that sampling is the process of taking a representation from a population as research participants. Random sampling was explained as a case where every member of a population has equal chances of being included in the sample. This was differentiated from biased sampling where members of the population do not have equal chances of being selected. A sample frame was differentiated from sample size; the former being the total units in the population, while the latter is the number of participants selected in a study. Two broad sampling techniques were differentiated; probability and non-probability sampling techniques. While the former adopts some forms of random selection, the latter does not. Types of probability sampling techniques presented include simple random, stratified, systematic, and cluster sampling; while non-probability techniques include accidental, quota, and purposive sampling.

Assessment

Assessment

Now that you have completed this study session, you can assess how well you have achieved its Learning Outcomes by answering these questions. Write your answers in your Study Diary and discuss them with your Tutor at the next Study Support Meeting. You can check your answers with the Notes on the Self-Assessment Questions at the end of this Manual

SAQ 4.1 (tests Learning Outcome 4.1)

A researcher desires to select 200 participants for a research to examine the relationship between sex and academic performance at the University of Ibadan. She however discovers that the ratio of males : females in the institution's sample frame is 3:2. What number of males : females should be selected from this population for the study?

A. 150:50
B. 80:120
C. 120 : 80
D. 50:150

SAQ 4.2 (tests Learning Outcome 4.1 and 4.2)

Assuming that a population size is known to be 600 and the above researcher wishes to systematically sample 20 participants from that population. Determine the nth term of the random selection for the sample.

A. 20
B. 30
C. 600
D. 400

Assignment

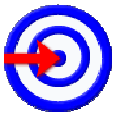
Find the first and last individuals that would be included in the sample in SAQ 4.2. Post your answer on Study Session 4 Assignment Page at the UI Mobile Class.

Study Session 5

Scales / Levels of Measurement

Introduction

Measurement scale is very pivotal in determining what statistical analysis would be appropriate during an analysis. This Study Session is therefore designed to acquaint you on various forms in which data can be expressed in research.



Learning Outcomes

When you have studied this session, you should be able to:

- *define* , *use* correctly and *differentiate* all of the key words printed in **bold**. (SAQ 5.1)
  - **Nominal scale**
  - **Ordinal scale**
  - **Interval scale**
  - **Ratio scale**

To put data of a variable into statistical use, they are recorded in some systematic ways. The procedure of describing the variables is called **scaling**, and the descriptions which emerge are called **scales**. The scales on which variables are measured have different properties.

Scales of measurement are also referred to as levels of measurement. This is because the numbers used to represent some variables have more meaning than those used to represent other variables.

Four scales are conventionally used in converting observations into numerical data. These are nominal, ordinal, interval and ratio scales.

5.1 Nominal Scale

This is the simplest and weakest level of measurement. By this method, observations of variables are classified into defined groups or categories, and each group or category is given a name or label for identification purposes. As a rule, each observation must fall under, at least, one category, and must not be considered under more than one category. All observations that have the features of a defined category are grouped into that category and given the appropriate label for the category. Observations that are not of the same features are grouped into different categories and given different labels. Labels used to describe the categories could be names like males or females for sex variables. They could as well be numerical numbers like '1' for males and '2' for females.

Numbers are, however, used as labels merely to classify observations into different categories, or used to describe different categories, without implying that one category is numerically related to any other, or

numerically higher than the other. Thus, the fact that females are represented with '2' in the example above and males with '1', does not imply that females are superior to males; it could be numbered the other way round. Another example is the numbers used to represent footballers on the football field, without any number indicating level of performance. In fact, such numbers used in nominal scaling do not imply order of importance. Invariably, this scale does not permit any arithmetic computation. Examples of variables in the nominal scaling form are sex, religion, ethnic group, vehicle number, hair colour, name of organization, marital status, state of origin, etc.

5.2 The Ordinal Scale

While nominal scales only classify, ordinal scales classify and order the classes in order of superiority. Ordinal scales, thus, indicate the order of observations on a particular attribute but do not show their separations on the scale. In other words, ordinal scales permit discussion of "more than" or "less than", but the extent to which an observation is more than, or less than another is not evident from the scale. In this case too, numbers could be used in ordering or ranking observations on the degree of importance or superiority, but the numbers do have some quantitative meanings, unlike the nominal scales where numbers are mere labels. However, we cannot say anything about the relative differences between pairs ranking.

An example of ordinal scale is the positioning of students on their examination performance. Numbers could be used to indicate the relative performance of students in terms of 1^{st}, 2^{nd}, 3^{rd} and so forth. It is an ordinal scale because the student occupying the first position performed better than the others, the second student performed better than others, except the first student. However, the numbers do have quantitative meanings. But we cannot say ideally that the actual difference in performance between the 1^{st} and the 2^{nd} students is equivalent to the difference in performance between the 2^{nd} and 3^{rd} student. Although ordinal scales permit the use of numbers with quantitative meaning, the numerals employed are non-quantitative in that arithmetic computations of addition, subtraction, multiplication and division are not computable, just like the nominal scales. Variables expressed in ordinal terms include job status, class levels at school, level of education, etc.

- **ITQ** Religious grouping may be referred to which type of measurement.

  A. Nominal scale.
  B. Ordinal scale.

Feedback on ITQs answers

- The correct answer is A, because religious grouping is just a descriptive category.
- If you chose B, you would be wrong because ordinal scales are not mere categories, they also show relative size.

5.3 Interval Scale

This is the highest scale or level of measurement usually attained in the behavioural sciences. Interval scales have all the qualities of ordinal scales and an added quality. Not only do they represent the ordering of observations on the characteristic being measured, they also represent the relative separation of items in the scale. Equal intervals between any two pairs of observations are equal. For instance, the distance or interval between 9 and 10 in the interval scale is the same as that between 112 and 113 or between any two adjacent observations.

With an interval scale, we can say precisely how much bigger, smaller, or better an observation or event is, compared with another. Thus, with nominal scale, we can say that today is hot while yesterday was cold. With ordinal scale, we can say today is colder than yesterday. In terms of interval scale, we can say that today is 15^{o} colder than yesterday. Interval scales permit the use of numerical values, which are quantitative. Thus, the operations of arithmetic computations such as adding, subtracting, multiplying and dividing are allowed. We can therefore say that the average of 9 and 10 is 9.5 without violating the properties of the attribute being measured. To achieve this, measurements must be based on some agreed units such as degrees, seconds, number of responses and so forth. It is the unit that permits us to equate intervals between points on the scale. Examples of variables expressed in interval measurement scale are intelligence quotient (IQ), anxiety, personality, aggression, etc

5.4 Ratio Scale

This scale has all the qualities of the interval scale with some additional qualities. That is, only ratio scale permits the making of statements that relate to ratio of numbers in the scale. For instance, 6cm is to 3cm as 2cm is to 1cm so they both have ratio 2:1. We can then meaningfully say that one is twice as good as the other. Ratio scale is mostly adequate for the measurement of physical objects or variables like height, weight, length, and so forth. Variables measured in the behavioural sciences are mostly in interval form. So if a student had 80% in a course and another 40% in the same course, we may not be right to say that the former is twice as intelligent as the latter on the course. This is because performance in examination in this form is measured in interval scale. But we can convincingly say that a father is twice as tall as his son because height is in ratio scale.

The ratio scale also employs a true or absolute zero mark, whereas the interval scale permits an arbitrary zero mark. With ratio scale a length of 0cm or time of 0.00 seconds implies absence of the variables being measured; in other words, zero is real. But with interval scale, if a student scores 0% in statistics, for instance, it does not necessarily imply that the student does not have any knowledge of statistics. In statistics, however, there is no clear-cut difference in the treatment of the interval and the ratio scales.


Tip

All ratio scales are interval scales; all interval scales are ordinal scales, and all ordinal scales are nominal scales. But all nominal scales are not ordinal scales; all ordinal scales are not interval scales, and all interval scales are not ratio scales.

- **ITQ** A measuring scale that permits you to determine whether data is greater than another but does not mean that does imply that an individual can score nothing may be referred to as which one of the following scale types?

  A. Nominal scale.
  B. Ordinal scale.
  C. Interval scale.
  D. Ratio scale.

Feedback on ITQs answers

- The correct answer is C because interval scales tells you the extent of difference between two individuals; and does not confer absolute zero on the individual that scores zero.
- Option A is a mere categorization scale. Although option B can enable you to determine qualities of superiority, it is also a mere categorisation scale. Option D provides an absolute zero

Study Session Summary


Summary

In this Study session you learned that the process of describing variables for statistical use is called scaling. When variables are classified into defined groups or categories and each group or category is given a name or label for identification purposes, it is called a nominal scale. Ordinal scales exist where observations on a particular attribute are expressed in order of importance. Measures of behaviour are expressed in interval forms, while physical measures are in ratio forms; both allow for equal distances between any two pairs of observations

Assessment


Assessment

Now that you have completed this study session, you can assess how well you have achieved its Learning Outcomes by answering these questions.

SAQ 5.1 (tests Learning Outcome 5.1)

Fill the empty columns in the table below with appropriate terms

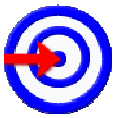
Scale of Measurement	Uses of the Scale	Example
		Pets: 5 dogs, 12 cats, 7 fish, 2 hamsters
	RANK some things as having more of something than others (but NOT QUANTIFY how much of it they have)	
Interval		
	QUANTIFY how much of something there is and a score of zero means the absence of the thing being measured	

Study Session 6

Frequency Distribution

Introduction

In trying to understand, interpret, predict, and modify behaviour, psychologists engage in various forms of research that involve data collection. Such data represent observable behaviours that are recorded in numerical forms. This study session will expose you to how to summarize data into manageable forms, using frequency distributions.



Learning Outcomes

When you have studied this session, you should be able to:

- *Demonstrate* how data can be put in a simple frequency distribution pattern. (SAQ6.1)
- *construct* class intervals, and *enhance* the use of tally marks with subsequent frequency values. (SAQ6.2)
- *determine* exact limits and mid-point values in grouped frequency distribution. (SAQ6.2)

6.1 Regular Frequency Distribution

Often times when research works are conducted by economists, educationists, psychologists, political scientists or sociologists, the dependent variable of interest is measured and represented with numerical numbers. Each participant's score on the dependent variable is thus recorded, leaving the research with a mass of unordered data for the whole research participants. These data are of little or no meaning until they have been organized and classified in a systematic way that will allow the introduction of statistical analysis like the average, standard deviation, and variance.

Suppose that a training manager wants to evaluate how well his trainees have performed at the completion of a training exercise in accounting. He gave his trainees a test and their scores fall between 0 and 10. Look at the scores the group of 50 trainees in Table 6.1.

Table 6.1

Number of Problems solved correctly by 50 Trainees

5	8	8	7	10	8	5	4	6	9
8	7	3	9	5	3	1	5	6	6
6	4	4	9	9	8	6	7	5	5
9	7	8	5	6	2	5	5	4	3
8	6	5	7	5	5	4	4	10	2

Frequency Distribution An organized tabulation/graphical representation of the number of occurrence in each category on the scale of measurement.

As you can observe, the data in Table 6.1 are unorganized or unordered and are, therefore, confusing and difficult to comprehend. In statistical terms, they are called raw or unclassified data. Items or raw data become more comprehensible when classified. The resulting table obtained, after classifying raw data, is called a **frequency distribution**. When the number of different scores in the raw data, is not too large, the regular frequency distributions are an excellent way of summarizing a set of data. The scores in Table 1 range between 1 and 10, and it gives the number of different scores as 11. This is not too large, so regular frequency distribution can be used to summarize the raw data, as shown in Table 6.2.

Table 6.2

Regular Frequency Distribution for Data in Table 6.1

Scores (X)	Tally Marks	Frequency (f)
10	||	2
9	𝍸	5
8	𝍸 ||	7
7	𝍸	5
6	𝍸 ||	7
5	𝍸 𝍸 ||	12
4	𝍸 |	6
3	|||	3
2	||	2
1	|	1
0		0
		Σ f = N = 50

6.1.1 Steps in Constructing a Regular Frequency Distribution Tally= a current amount

1 List every raw score value in the first column of a table and denote this by the symbol [X], with the highest score at the top and the lowest score at the bottom. Thus, 10 is at the top while 0 is at the bottom.

2 Run through the raw data either horizontally or vertically, recording each observed score against identical score in the first column with a tally mark. Show the tally marks in the second column. After every four tallies, the fifth tally is used to cross the previous four (it is a rule). It is aimed at easing the recording of frequencies.

3 When the tally is completed, the tally marks for each score in the first column are added up to find the *frequency* (f) and recorded in

the third column. The frequency of each score shows the number of times a given score was obtained. You can easily read from Table 6.2 that two students had a score of 10, five students had a score of 9, seven students had a score of 8, etc. Cross check with Table 6.1

4 Add up the frequencies and place the total at the bottom of the frequency column. The Greek sign Σ (Sigma) means the "sum of" and Σf means "the sum of frequencies" is used to show this. It means the total number of scores in the sample. Σf and N must be equal, they are therefore used to check error in tallying. Thus, if Σf and N are not equal, the tallying exercise should be repeated. Even when they are equal, it is still advisable to repeat the exercise to check for error of placing one or more tally marks wrongly. If the first tallying was done vertically, the second may be done horizontally to check for such error.

Activity 6.1

Allow 15 minutes

You have learnt that a frequency distribution is a convenient way to organize a set of data. You have also been exposed to the steps of constructing a frequency distribution table.

You will now prepare a frequency distribution for "the heights of 24 students" under this guided activity.

The heights in inches of 24 students are as follows:

66, 68, 65, 67, 64, 68, 64, 66, 64, 69, 69, 64, 67, 63, 63, 68, 67, 65, 69, 65, 67, 66, 69, 67

Prepare a frequency distribution

Solution: Step 1: Make three columns.

Step 2: List the heights (in order of size) in the first column, the tally in the second column, and the frequency in the third column.

Heights	Tally	Frequency
63	||||	2
64	||	4
65	|||	Fill the remaining columns accordingly
66		
67		
68		
69		

Post your findings on Activity 6.1 Page at the UI Mobile Class.

6.2 Grouped Frequency Distributions

When data obtained in a research are relatively more, the regular frequency distribution might not be adequate in summarizing the data; rather, the grouped frequency distribution becomes necessary. This calls for the grouping of raw data into class intervals and the number of scores in each class interval is then recorded as frequency. An interval in which scores are grouped is called a *class interval.*

Suppose a production manager is interested in the production efficiency of employees in his department. He selected 50 members of his department and recorded their performances in terms of number of cartons packaged in one hour, as presented in Table 6.3.

Table 6.3

Performance Data of Employees in a Packaging Firm

63	53	62	40	54	58	61	68	50	67
48	46	74	26	24	19	23	27	32	62
36	20	68	45	35	43	41	45	33	30
47	39	36	51	22	66	40	70	20	34
72	55	60	57	56	75	44	61	50	58

You can see that the scores in Table 6.3 range between 19 and 75. Therefore, the grouped frequency distribution should be constructed to summarize the data as shown in Table 6.4.

Table 6.4

Grouped Frequency Distribution of Performance for Employees in a Packaging Firm

Class Intervals	Tally Marks	Frequency (f)
73-77	||	2
68-72	|||	3
63-67	||||	4
58-62	𝍸	5
53-57	𝍸	5
48-52	𝍸 ||	7
43-47	𝍸 |	6
38-42	||||	4
33-37	||||	4
28-32	|||	3
23-27	||||	4
18-22	|||	3
		$\Sigma f = N = 50$

6.2.1 Steps in Constructing a Grouped Frequency Distribution

1. **Calculate the range of the raw scores**. This is obtained by subtracting the lowest value from the highest value. In Table 6.3 the highest value is 75, while the lowest is 19. Therefore, the range is 75 - 19, which gives 56.
2. **Determine the number of class intervals and the *class interval size* or *class width* of each interval.** Points to note here are:
   a. The number of class intervals must not be less than 10 or above 20. It is, however, better to have intervals between 10 and 15 for convenience.
   b. Certain interval sizes or width are preferred; these are 2,3,5 or a multiple of 5 like 10, 15, 20 etc.
   c. All intervals should be the same size. This means that, the range must be divided into a number of equal intervals.

This formula could be used

Where C.I. = Class interval
H = Highest score
L = Lowest Score
I = Class Interval size chosen
H –L = Range

For the case under study we have:

The higher the interval size chosen, the fewer the number of class intervals. The lower the interval size chosen, the larger the number of class intervals that will emerge.

3. **Arranged classes in a decreasing order of sequence with the lowest measure placed at the bottom of the table and the highest measure at the top, according to accepted rule.** The first class interval from below is therefore 18-22; this covers the performance scores (units packaged) of 18, 20, 21 and 22, which makes the class interval of 5. The second interval of 23-27 covers the scores of 23, 24, 25, 26 and 27. The lowest score in each interval is called the *lower limit* while the highest score is called the *upper limit.* Thus, 18, 23,28, etc are lower limits while 22, 27, 32 etc are respectively upper limits in the example in Table 4.

As guidelines note the following:

a. Ensure that the lowest score in the raw data is neither the lower limit nor the upper limit of the lowest class interval, or else the number of class intervals emerging may exceed or be less than that expected from your calculations.

a. Classes must be mutually exclusive; that is they must not over lap. This means that the upper limit of a class interval must not be the same as the lower limit of the next following interval. For instance, 18-22, 22-26, 26-30 are over-lapping class intervals.

b. Classes must also be collectively exhaustive, that is, should cover all observations. To ensure this, the lowest class interval should be able to accommodate the lowest raw score, while the top-most class interval should be able to accommodate the highest score. Thus, the class interval 18-22 accommodates 19, which is the lowest score, while 73-77 accommodates 75 the highest score, in our example.

c. Make the lowest score in each following interval a multiple of the interval size. Thus, we have 18, 23, 28 etc. with 5 as interval size.

d. Gaps between classes must be avoided as much as possible.

4. Run through the raw data either horizontally or vertically and score each observation with a tally mark against its class until all observations are covered.

5. The tally marks in each class interval are added up to find the frequency for the interval. The frequency opposite a given class interval indicates the number of cases with scores in that interval. Thus, grouped frequency distributions lose information, since the exact value of each score is not provided.

6. Add the frequencies and place the total at the bottom of the frequency column.

7. Always give your tables suitable headings.

6.3 The Exact Limits and Mid-Points of Class Intervals

In grouped frequency distributions, class intervals do not show how values in continuous measurement forms should be presented. This is because class intervals make provisions only for discrete measures. This Study Session would show how to correct this anomaly using class intervals, exact limits of class intervals are as well used in grouped frequency distribution. Within a class interval, a set of values are grouped together, thereby making it all scores in a class interval to lose identity. To represent all the scores in a class interval, therefore, the mid-point value is determined.

6.3.1 Exact Limits and Mid-Points

The limits (lower and upper) of the intervals in Table 6.4 do not show exactly where each interval begins and ends. However, in statistical situations, it is very important to think in terms of exact limits. The

interval 18-22 contains the scores 18, 19, 20, 21 and 22. A measure of 18, in essence, means anything from 17.5 to 18.5, which could be 17.5, 17.6, 17.7 …… 18.5. Thus, an interval of 18-22 has exact limits of 17.5 to 22.5. An interval of 23-27 has exact limits of 22.5 to 27.5. Exact limits are, therefore, established for each interval by subtracting .5 from the lower limit and adding .5 to the upper limit. These are then referred to as *lower exact limits* and *upper exact limits* respectively. This principle holds, no matter the size of the interval. Exact limits are also called *real limits*. So, we can refer to lower real limit and upper real limit of class intervals.

It will easily be observed that each interval begins exactly where the one below it ends, which is as it should be. The interval with its exact limits should read 17.5 up to 22.5, which means it includes 22, 22.1, 22.2, 22.3, 22.4 but not 22.5. This is the interpretation, which is different from how it is read ordinarily.

Table 6.5 **Exact Limits and Mid-Points of the Distribution in Table 6.4**

Intervals	Exact Limits	Mid-Points (x)	Frequencies (f)
73-77	72.5 - 77.5	75	2
68-72	67.5 - 72.5	70	3
63-67	62.5 - 67.5	65	4
58-62	57.5 - 62.5	60	5
53-57	52.5 - 57.5	55	5
48-52	47.5 - 52.5	50	7
43-47	42.5 - 47.5	45	6
38-42	37.5 - 42.5	40	4
33-37	32.5 - 37.5	35	4
28-32	27.5 - 32.5	30	3
23-27	22.5 - 27.5	25	4
18-22	17.5 - 22.5	20	3

The mid-point of a class interval is the most central score in the interval. This is calculated easily by adding the lower and upper limits of each interval and dividing the sum by 2. The exact limits could as well be added and divided by 2; they give the same mid-point. The mid-point is represented by small letter (x).

Study Session Summary

Summary

In this Study Session, we have demonstrated how raw data can be presented in regular frequency distribution form. This involves the use of tally marks that are later converted / added up to find the frequency. You have also learnt that the grouped frequency distribution is used when there are more than 15 values of scores. It calls for the grouping of raw data into class intervals and the number of scores in each class interval is recorded as frequency.

You also learned that exact limits are calculated by subtracting .5 from the lower limit to obtain the lower exact limit and adding .5 to the upper limit to have the upper exact limit. The mid-point value is obtained by dividing the result of the summation of the lower and upper limits by 2.

Assessment

Assessment

Now that you have completed this study session, you can assess how well you have achieved its Learning Outcomes by answering these questions.

SAQ 6.1 (tests Learning Outcome 6.1)

A security unit of a psychological centre measured the speed of 25 cars that entered into its premises. The resulting speeds were:

29, 23, 30, 30, 27, 24, 30, 25, 23, 28, 25, 24, 28, 30, 23, 30, 27, 25, 29, 24, 23, 26, 30, 28, 25

Prepare a frequency distribution for this data.

Assignment

Assignment

Below is a set of intelligence quotient (IQ) scores of 50 undergraduate students.

IQ Scores of 50 Undergraduate Students

109	99	101	126	94	106	97	110	115	106
104	112	84	116	105	122	110	92	112	99
125	107	95	103	122	102	119	120	103	103
108	98	118	114	114	109	109	118	105	111
107	111	127	122	88	121	112	128	117	107

Task

1. Using a class interval size of 5, and with the lower limit of the lowest class interval as 80, develop a grouped frequency distribution for the IQ Scores.
2. How many class intervals emerged?
3. What is the range of the raw scores?

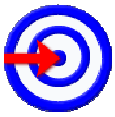
Send in your answers to your tutor by submitting at Study Session 6 Assignment Panel at UI Mobile Class.

Study Session 7

Graphic Representation of Frequency Distribution

Introduction

So far, Study Sessions 6 and 7 have shown how data could be collected and organized statistically. As earlier explained, statistics also involve the presentation and interpretation of data. This Study Session will show how facts of frequency distribution can be presented clearly in some graphic ways. Such presentations help to interpret the statistical data, by translating numerical information, which are sometimes difficult to comprehend into pictorial forms that are more readily understandable. Graphic presentations, thus give better picture of a frequency distribution by setting out its general contour and showing more precisely the number of cases in each interval.



Learning Outcomes

When you have studied this session, you should be able to:

- *present* frequency distributions in polygon, histogram, bar chart, and relative bar chart forms. (SAQ 7.1)

7.1 The Frequency Polygon

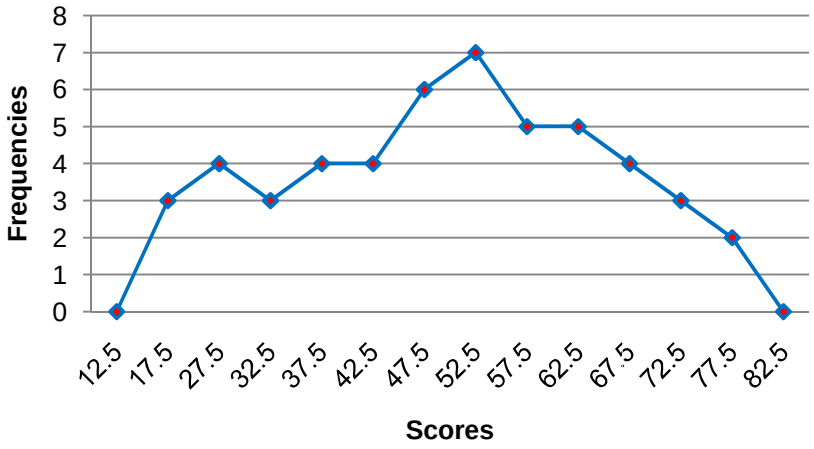
Polygon means *many-sided figure*. Simple frequency polygons are particularly adequate for continuous data where scores can emerge between score values. Generally, scores are always treated as continuous, whether or not fractions occur. The steps involved in drawing a frequency polygon are as follow:

1. Note the size of the range of the distribution and rule a base line of suitable length to cover the number of class intervals. This line is referred to as the x-axis or abscissa.
2. Clearly mark the exact limits of the class intervals on the X-axis. However, in drawing the diagram, always allow two extra intervals, one at each end of the axis so as to allow you to bring down the diagram to the base line. Using the frequency distribution in Table 4, we have 12 intervals; adding two extra intervals make it 14 intervals. The base line is therefore divided into 14 equally spaced intervals, beginning with 12.5 and ending with 82.5 as shown in Figure 9.1. The intervals 12.5 – 17.5 and 77.6 – 82.5 are the two intervals added.
3. At the left-hand side of the base line, draw a perpendicular line, called *Y-axis or ordinate*, and mark off on this axis equally spaced units to represent the frequencies. Note that the height of

the diagram should not exceed the width. The values on the Y-axis start with 0.

4. At the mid-point of each interval on the X-axis, go up on the Y-axis until a position is reached corresponding to the frequency for that interval and place a dot here.
5. Connect all dots with straight lines to give the frequency polygon.
6. The polygon looks hanging. To bring the polygon to the base line, extend the straight lines to the mid-points of the two extra intervals. Thus, the frequency to be recorded at the mid-points of the added intervals is a zero. The resulting frequency polygon provides a pictorial illustration of the frequency distribution.

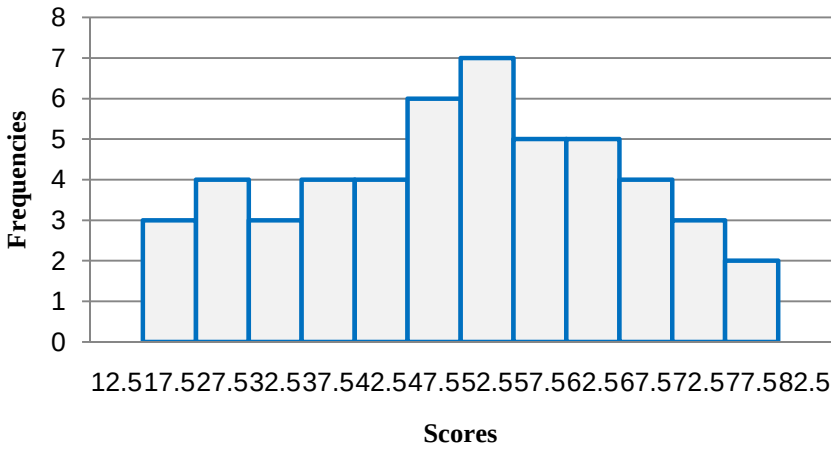
Fig 7.1 **Frequency Polygon of Employee Performance in Table 6.5 (Exact Limits and Mid-Points of the Distribution in Table 6.4)**



8.2 The Histogram

The histogram graph is sometimes called a column diagram. It looks like a bar chart except that there are no spaces between the "bars" of the histogram, as can be seen in Figure 8.2

Fig 7.2 **A Histogram plotted from the Distribution of Employees Performance in Table 6.5 (Exact Limits and Mid-Points of the Distribution in Table 6.4)**



The procedure in marking out the histogram is identical to that involved in plotting the frequency polygon. However, there are some exceptions.

1. There may be no need to include the two extra intervals as in frequency polygon.
2. After the points representing the frequencies are located as before at the mid-points of the intervals, these points are not joined by straight lines. Rather, a horizontal line is drawn through each point and extended to the exact limits of each interval, after which two vertical lines are drawn to the lower exact limit and upper exact limit respectively, on the X-axis. These then complete the rectangles or columns; thus, the histogram is called a column diagram.

A Comparison of the Frequency Polygon and the Histogram

The frequency polygon and the histogram show basically the same thing. Both enable us to see graphically how scores are distributed, whether symmetrically or asymmetrically (piled up at the low or high end). Both enable us also to see at a glance the number of cases falling in each interval, but the histogram is more exact in this than the polygon. This is because in the histogram, each measurement occupies exactly the same area size, and each rectangle is directly proportional to the number of measures within that interval.

The frequency polygon is often preferred to the histogram when there is need to gives a clearer picture of the contour or shape of the distribution. It is also easy to compare two frequency distributions with the frequency polygon.

7.3 The Bar Chart

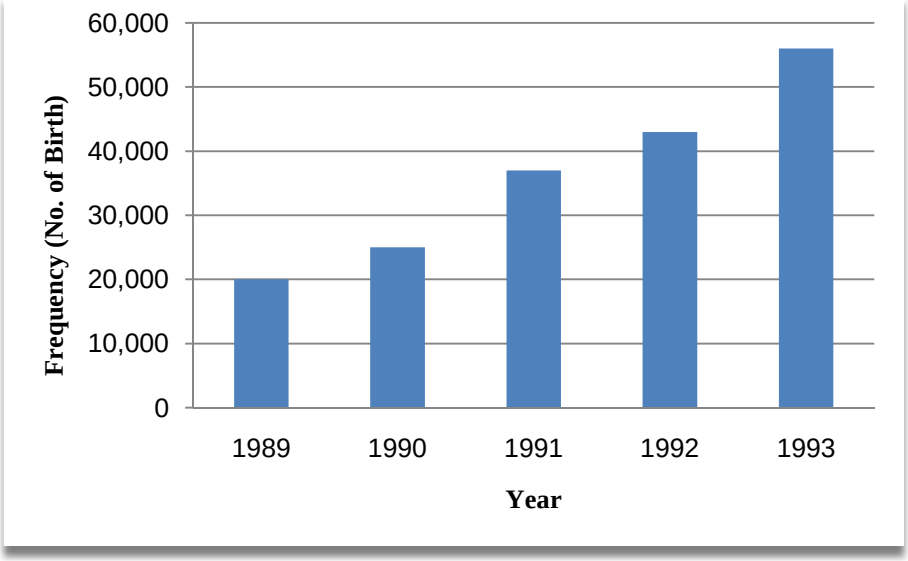
The bar chart is like the histogram in shape but with spaces between the "bars". This is because bar charts are used in displaying frequency distribution of qualitative or attribute data. Such data have distinct points that cannot be merged, as the case with histogram. Assume the birth and the death rates in a country for five consecutive years are shown on a table as we have it in Table 7.1:

Table 7.1 **Birth and Death Rates in a Country for Five Years**

Year	Birth	Death
1989	20,000	10,000
1990	25,000	15,000
1991	37,000	20,000
1992	43,000	32,000
1993	56,000	45,000

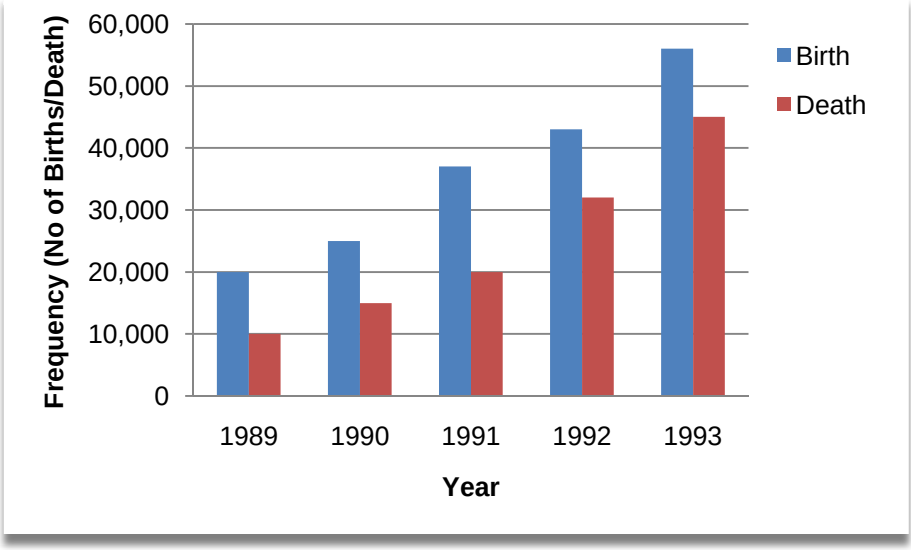
To represent the birth rate graphically on a bar chart, the attribute (year) is shown on the x-axis, while the frequency (number of births) is plotted on the y-axis as shown in Figure 7.3

Fig 7.3 **Bar Chart on the Number of Births**



Bar charts can also be used for graphical comparisons between two or more classes of attribute data. Look at Fig 7.4 for graphical comparison between the birth and the death rates in Table 7.1.

Fig 7.4 **Comparative Bar Chart on the Number of Births and Deaths**



- **ITQ** What type of graph is used for discrete data or qualitative data?

  A. bar graph.
  B. histogram.

Feedback on ITQs answers

- The correct answer is A. Option B (histogram) is used when the data are continuous.

Study Session Summary

Summary

In this Study Session, you have seen that there are three basic types of graphs that we use for most data: (1) frequency polygon, (2) bar graphs, and (3) histograms. The simple frequency polygons are adequate for continuous data where scores can emerge between score values. The names of the last two are a bit misleading because both are created using bars. The only difference between a bar graph and a histogram is that in a bar graph the bars do not touch while the bars do touch in a histogram. In general, bar graphs are used when the data are discrete or qualitative. The space between the bars of a bar graph emphasize that there are no possible values between any two categories. For example, when graphing the number of children in a family, a bar graph is appropriate because there is no possible value between any two categories (e.g., 1 and 2 children). When the data are continuous, we use a histogram. The bars touch in a histogram to indicate that there are possible values between any two categories. For example, if we were graphing time to complete a test, the bars would touch to indicate that there are possible values between any two times (e.g., 27 and 28 minutes).

Assessment

Self Assessment

Now that you have completed this study session, you can assess how well you have achieved its Learning Outcomes by answering these questions.

SAQ 7.1 (tests Learning Outcome 7.1)

Draw a bar graph of frequency distribution car speed in SAQ6.1

Assignment

Now that you have completed this study session, attempt the following assignment (a tutor marked assessment).
The table below shows the population of undergraduate students in the Faculty of the Social Sciences at the University of Ibadan for the 2005/2006 academic session.

Departments	**Levels of Study**			
	1	2	3	4
Psychology	100	150	180	60
Sociology	120	180	90	50
Political Science	80	110	160	100

Geography	110	165	200	140
Economics	90	140	170	200

Task

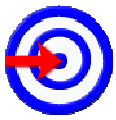
1. Present the data on the table in a suitable graph.
2. Graphically present the statistics of the students in the faculty in terms of their departments.
3. Graphically present the statistics of the student population in the faculty in terms of their levels of study.

Study Session 8

Diagrammatic Representation of Frequency Distribution

Introduction

Statistics involve, among others, the presentation and interpretation of data in some diagrammatic ways. This Study Session will discuss the representation of frequency distribution using pie chart, pictogram and stem-and-leaf display.



Learning Outcomes

When you have studied this session, you should be able to:

- *present* raw data in the form of pie chart, pictogram, and stem-and-leaf display. (SAQ 8.1)

8.1 The Pie Chart

It is a circular diagram in which classes or attributes are represented by sectors. Each sector represents a class or attribute at a proportion equal to the percentage or relative frequency of the class or attribute. The relative frequency of an attribute is the ratio of the frequency of that attribute to the total number of occurrences of observations.

Using the example of the number of employees in the various departments of a company (see Table 8.1), the steps involved in drawing a pie chart are provided after the table.

Table 8.1 **Distribution of Employees in a Company**

Department	Student Population
Marketing	153
Accounts	60
Production	102
Personnel	85
Works	98
Total	**498**

1. Find the relative frequency of percentage of the employees in each department using the formula:

   ——————————————————— ——

   For marketing, R.F is —— ——

   For Accounts, it is —— ——

   For Production, it is —— ——

   For Personnel, it is —— ——

   For Works, it is —— ——

Look at Table 8.2.

Table 8.2 **Relative Frequency for a Pie Chart Distribution of the Employees in a Company in Table 8.1**

Department	**Student Population**	**Relative Frequencies**
Marketing	153	30.7%
Accounts	60	12.0%
Production	102	20.5%
Personnel	85	17.1%
Works	98	19.7%
Total	**498**	**100%**

Note that the sum of relative frequency must be 100.

2. Calculate the angle of the sector for each department, using the formula

   Angle of sector —— ——

   For Marketing it is —— ——

   For Accounts, it is —— ——

   For Production, is —— ——

For Personnel, it is — —

For Works, it is — —

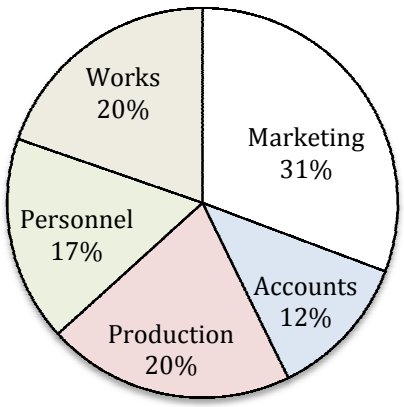
Note that the angles must sum up to 360°. See Table 8.3

Table 8.3 **Pie Chart Distribution of the Employees in a Company in Table 8.1**

Department	Student Population	Relative Frequencies	Angle of Sector
Marketing	153	30.7%	110.60°
Accounts	60	12.0%	43.37°
Production	102	20.5%	73.74°
Personnel	85	17.1%	61.45°
Works	98	19.7%	70.84°
Total	**498**	**100%**	**360°**

3. Draw the Pie Chart: This can be done neatly with a pair of compass and a protector. Using a pair of compass, measure any length and draw a circle, showing the circumference. From the center of the circle, draw a straight line to the circumstance (i.e. radius). At the radius, use your protector to measure the various degrees calculated for each department, and show the corresponding percentage in the division. The resultant diagram is a pie chart which you can see in Fig 8.1.

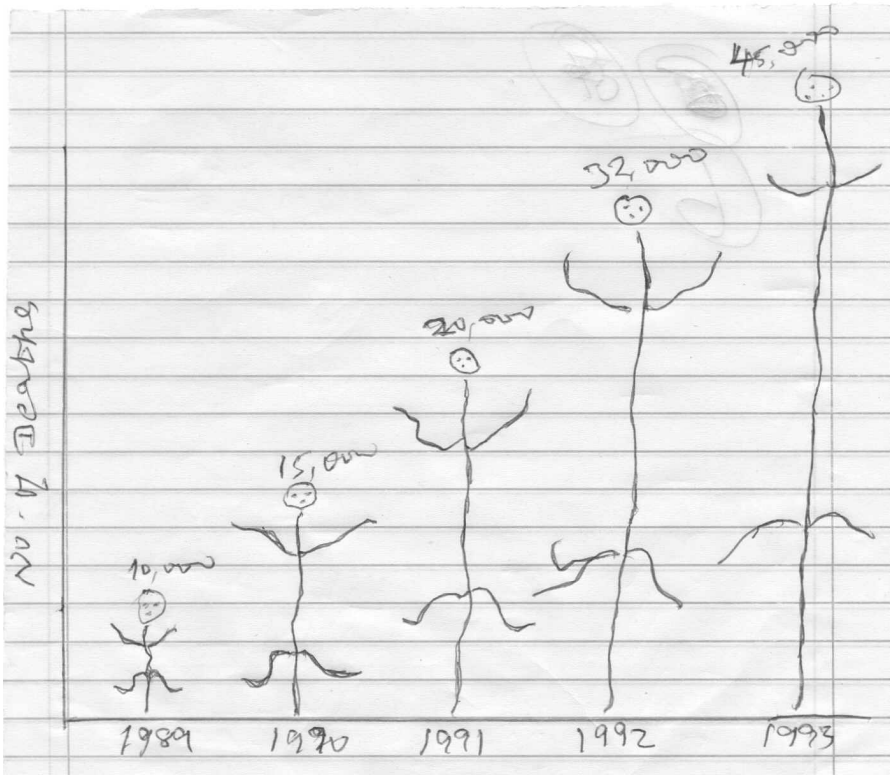
Fig 8.1 **Pie Chart of Employees in a Company in Table 8.3**



8.2 The Pictogram

The pictogram is also called the pictograph or ideograph. It is a graphic way of presenting data with emphasis on the aesthetic aspect of the display and not accuracy. The symbol or picture chosen for any set of data must have relationship or resemblance with the concept being portrayed. For example, Look at Fig 8.2, it shows the pictogram of deaths recorded in Table 7.1

Fig 8.2 **Pictogram of the Total Number of Deaths in a Country**



8.3 The Stem-and-Leaf Display

This is a simple, useful but uncommon technique for summarizing a set of data by combining the features of the frequency distribution and the histogram. The class intervals form the "stems" while specific values of the raw data within each interval form the "leaves". An example using scores in Table 6.3 and illustrated in Table 8.2 is shown in Table 8.4.

Table 8.4 **Stem-and-Leaf Display for data in Table 6.3**

Stems (Intervals)	Leaves (Observations)	Frequency (f)
73-77	4 5	2
68-72	2 8 8	3
63-67	3 6 6 7	4
58-62	1 2 2 8 8	5
53-57	3 4 5 6 7	5
48-52	0 0 0 1 1 8 9	7
43-47	3 4 5 5 6 7	8
38-42	0 0 1 9	4
33-37	4 5 6 6	4
28-32	0 2 2	3
23-27	3 4 6 7	4
18-22	0 0 9	3

Each leaf is made up of the number of scores falling within a given interval with each score represented mainly by the last digit of the score, if they are two-digit scores. The scores in each leaf are then ordered from low to high. With the class interval of 23-27 in Table 9.1 for instance, reading the raw data in Table 6.3 horizontally, the scores falling in this interval are 26, 24, 23 and 27. The last digits for these are 6, 4, 3 and 7 respectively; these form the leaf for the interval 23-27. When ordered from low to high, the leaf stands as 3, 4, 6 and 7. Do the same for other intervals.

The stem-and-leaf display shows the same graphic representation as the histogram because each observation takes up one space. Also, it is very easy to recall each score in the raw data from the display. For instance, it can be read from the display that the scores within class interval 18-22 are 20, 22 and 19 since the leaf is made up of 0, 2 and 9. The 0 is for 20, 2 is for 22, and 9 is for 19 (remember the interval is 18-22). For data with three or more digits, you may choose to increase, correspondingly, the number of digits to use for the leaves.

Study Session Summary

Summary

In this Study Session, you have learnt that the pie chart is a circular diagram in which classes or attributes are represented by sectors. A pictogram is a graphic way of presenting data with emphasis on the aesthetic aspect of the display and not accuracy. The stem-and-leaf display is a technique for summarizing a set of data by combining the features of the frequency distribution and the histogram.

Assessment

Self Assessment

Now that you have completed this study session, you can assess how well you have achieved its Learning Outcomes by answering these questions. Write your answers in your Study Diary and discuss them with your Tutor at the next Study Support Meeting. You can check your answers with the Notes on the Self-Assessment Questions at the end of this Manual

SAQ 8.1 (tests Learning Outcome 8.1)

The table below shows the monthly spending of Mrs. Halima

ITEM	AMOUNT SPENT (₦:K)
Food	7000
Transportation	1250
Accommodation	6000
Savings	1500
Miscellaneous	1000.00

Represent the data in a pie chart

Assignment

In an election that was conducted in a Local Council Ward, among the 2, 500 votes recorded, 500 voted for party A; 230 voted for party B; 800 voted for party C; 180 voted for party D, and the others voted for party E. Present the result of the election in an appropriate diagram.

List any three variables that could be presented in a pictographic form. Demonstrate how each could be presented using self-generated data.

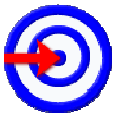
Post your answer to your tutor on Study Session 8 Assignment page at UI Mobile Class.

Study Session 9

Cumulative Distributions

Introduction

In the simple frequency distribution, you can easily determine the number of cases falling within a class interval by adding up frequencies in that interval. But when additional information is required, like the number of cases occurring above or below a given interval, then cumulative frequency distribution becomes necessary. This Study Session will expose you to how to use the cumulative frequency distribution which is an alternative method that is used when information is required in form of "more than" or "less than".



Learning Outcomes

When you have studied this session, you should be able to:

- *infer* from frequency distributions, the number of cases falling at, and above or less than, a particular point.
- *plot* graphs showing "more than" and "less than" cumulative frequency distributions.

9.1 Constructing Cumulative Distribution

Let us take a look at the information in Table 6.3. Can you state the number of employees whose performance scores are 38 and above or those who are 62 and below. You most likely observe that this cannot be easily decided. To achieve such a task, the cumulative frequency distribution is required, by cumulating the frequencies to see how many employees are "more than" or "less than" a certain level.

9.1.1 Constructing a "More Than" Cumulative Distribution

The cumulating starts from the highest class interval and descends. The highest class interval in Table 11.1 is 73-77 with a frequency of 2. This explains that two employees scored 73 and above, thus 2 appears in the column of "more than" cumulative for the interval 73-77. For the interval 68-72, the frequency is 3 and it is interpreted that three employees scored between 68 and 72. The number of employees whose scores are 68 and above will then be 2+3=5, so 5 appears as the "more than" cumulative frequency for the interval 68-72. For the interval 63-67, it is 2+3+4=9.

The successive addition continues until the class intervals are exhausted. It should be noted that the last cumulative frequency must be equal to Σf or N, which is the total number of observations; otherwise, a mistake must have been done. In our example, Σf or N = 50, this tallies with what is on the Table 9.1. It implies that 50 employees scored 18 or more.

Table 9.1 **"More than" and "Less than" Cumulative Distribution of Performance for Employees in a Packaging Firm**

Class Intervals	Frequency (f)	More than Cumulative	Less than Cumulative
73-77	2	2	50
68-72	3	5	48
63-67	4	9	45
58-62	5	14	41
53-57	5	19	36
48-52	7	26	31
43-47	6	32	24
38-42	4	36	18
33-37	4	40	14
28-32	3	43	10
23-27	4	47	7
18-22	3	50	3
	Σ f = N = 50		

9.1.2 Constructing a "Less Than" Cumulative Frequency Distribution

It is very much similar to the "more than" cumulative frequency distribution, except that we now begin to cumulate the frequencies from the bottom class interval. As shown In Table 9.1, there are three employees whose performance scores are between 18 and 22 or less. 3 therefore appears at the bottom of the "less than" cumulative frequency column for the interval 18-22. The following interval shows that 4 employees had performance scores range between 23-27. The total number of employees whose performance scores are 27 or less is calculated as 3+4=7. Thus, the "less than" cumulative frequency at the interval 23-27 is 7. This procedure also continues until all the intervals are exhausted. Again, the "less than" cumulative frequency for the top most intervals must be equal to Σf or N. In this case, it is 50, and it is interpreted that 50 employees have performance scores of 77 or less.

9.2 Graphical Presentations of Cumulative Distribution

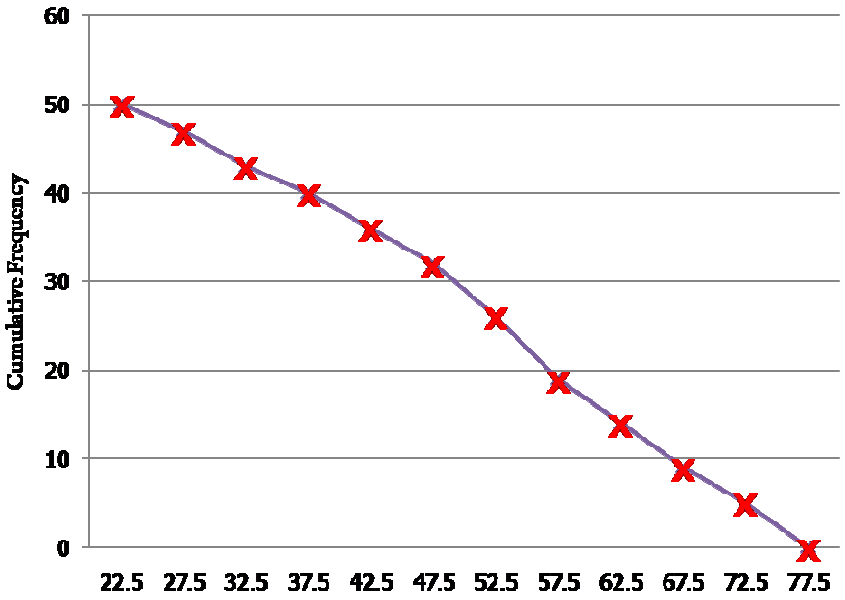
When distributions are placed in cumulative forms, they could also be presented graphically. The procedure involved in constructing the

cumulative distributions graphically is what this section is designed to achieve.

9.2.1 Graphical Presentation of a "More Than" Cumulative Distribution.

Cumulative frequencies are plotted on the Y-axis, while the corresponding performance scores are on the X-axis. The method of plotting is quite similar to that of plotting the simple frequency distribution. However, while the frequency polygon type of graph is adopted, the histogram type of graph is not applicable here. When plotting the frequency polygon, the frequency in each interval is taken at the mid-point of the interval but in the "more than" cumulative frequency curve, each cumulative frequency is plotted at the lower exact limit of the interval. For the "more than" cumulative frequency curve of the data in Table 9.1, take a look at Figure 9.1. Also two extra intervals are not added in this case.

Fig 9.1 **"More than" Cumulative Distribution Curve of Employees' Performance**



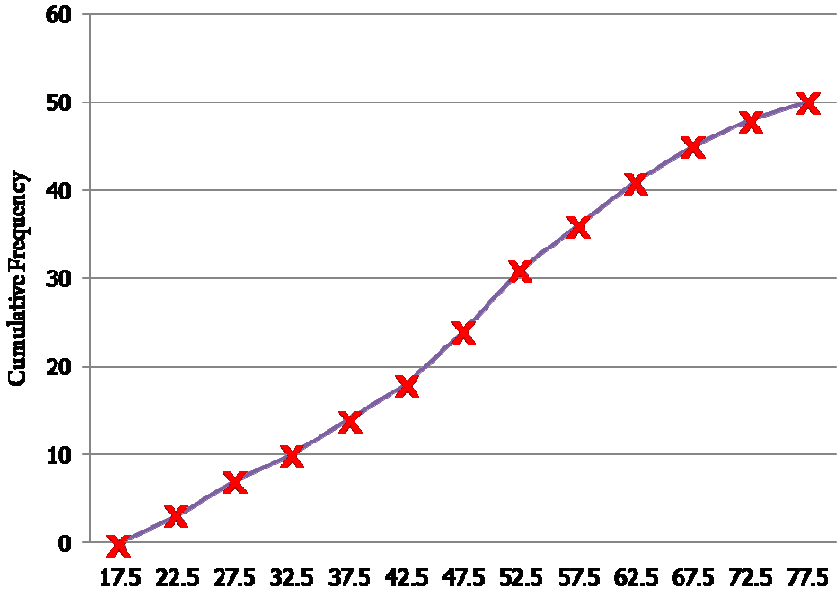
When all the cumulative frequencies have been plotted, the points are joined by straight lines. To bring the curve down to the base line, the cumulative frequency of 0 is plotted at the upper exact limit of the interval 73-77, which is 77.5. This means that there are no employees with performance scores of "more than" 77.5.

9.2.2 Graphical Presentation of "Less Than" Cumulative Distribution.

In the "less than" cumulative frequency, the graph is plotted between the cumulated frequencies and the upper exact limit of the interval. The graph is brought down to the base by placing the cumulative frequency of 0 at the lower exact limit of the bottom interval. Thus, there is no employee who performed less than 17.5.

From the graph in figure 9.2, you can now easily read the number of employees who performed less than any given score. So also, you can read the number of employees who performed more than any given score, with the "more than" cumulative distribution.

Fig 9.2 **"Less than" Cumulative Distribution Curve of Employees Performance Score.**



Study Session Summary

Summary

In this Study session you learned "more than" cumulative distribution shows how many cases fall at, and above a given point, while the "less than" distribution shows how many cases fall at, and below a given point. We also noted that the method of plotting cumulative distribution graphs is quite similar to that of plotting the simple frequency distribution.
However, when plotting the "more than" cumulative frequency curve, each cumulative frequency is plotted at the lower exact limit of the interval. To bring the curve down to the base line, the cumulative frequency of 0 is plotted at the upper exact limit of the highest class interval. In the "less than" cumulative frequency, the graph is plotted between the cumulated frequencies and the upper exact limit of the interval. The graph is brought down to the base by placing the cumulative frequency of 0 at the lower exact limit of the bottom interval.

Assessment

Self Assessment

Now that you have completed this study session, you can assess how well you have achieved its Learning Outcomes by answering these questions. Write your answers in your Study Diary and discuss them with your Tutor at the next Study Support Meeting. You can check your answers with the Notes on the Self-Assessment Questions at the end of this Manual

SAQ 9.1 (tests Learning Outcome 9.1)

Cumulative frequency distribution is that it helps to calculate the number of people falling at a particular score (a) True (b) False.

Assignment

With the skills acquired in this Study Session, answer the following questions using Table 6.5 as a guide:

1. How many employees had score of 43 and above?
2. How many had score of 32 and below?
3. How many had scores between 32 and 43?

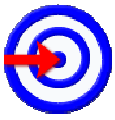
Post your answers on Study Session 9 assignment page at UI Mobile Class.

Study Session 10

Measures of Central Tendency: Mean

Introduction

The frequency distribution with its graphic representations primarily shows the organization of data into a manageable form. They are the preliminary steps toward quantitative treatment of data. The method of statistically treating the distribution gives numerical descriptions which are the basic features of the distribution. These further enable concise and definite comparison of the features of one distribution with others. This study session will attempt to expose you to measures of central tendency, which is one of the types of numerical descriptions of frequency distributions.



Learning Outcomes

When you have studied this session, you should be able to:

- *compute* and *interpret* the arithmetic mean or average score for both ungrouped and grouped data.
- *compute* geometric and harmonic means.

A **measure of central tendency** is a location or representative point on a scale, a central point where data are concentrated, centered, packed or clustered with other scores scattering on either side. Such measure serves the following purposes:

1. It is a measure that represents all measures in a sample, which could be utilized in comparing one distribution with another.
2. A sample is a representative of a particular population. A measure of central tendency, therefore, describes indirectly, with some amount of accuracy, the population from which the sample has been drawn. Thus, we can estimate a measure of central tendency for a population from the knowledge of the sample measure of central tendency. Generalizations and predictions are therefore made beyond that of the sample to the population. This makes scientific research in the behavioural sciences and education possible.

There are basically three measures of central tendency; these are: the mean, the median and the mode.

Three types of mean are used in statistics. These are:

1. the arithmetic mean;
2. the geometric mean; and
3. the harmonic mean

The arithmetic mean is the most frequently used, while the other two have a more restricted use.

10.1 The Arithmetic Mean

10.1.1 Ungrouped Data

Another name for arithmetic mean is average. To calculate the arithmetic mean for ungrouped data, add up all the measurements and divide by the number of measurements added. The formula for the calculation is given as:

$$\bar{X} = \frac{\Sigma X}{N}$$

Where $\bar{X}$ = Arithmetic mean

Σ = "the sum of"

X = an individual score

ΣX = sum of X scores

N = the number of measurements

Given that the scores (in percent) obtained by a student in seven subjects are 43, 59, 61, 73, 55, 64, and 49. The average performance of the students on the seven subjects is calculated thus:

$$\bar{X} = \frac{\Sigma X}{N}$$

ΣX = 404

N = 7

10.1.2 Regular Frequency Distribution

With scores in the form of a regular frequency distribution, the mean is computed, using the formula:

$$\bar{X} = \frac{\Sigma fx}{N}$$

Where $\bar{X}$ and Σ remain the same

Fx = *x* score multiplied by the frequency of that score

Σfx = sum of fx values

N = total number of scores i.e. Σf

Assume that a company recorded the number of accidents experienced per worker in a year, as shown in Table 8.1 in a regular frequency distribution form; calculate the average number of accidents recorded in the company for the year.

Table 10.1 **Regular Frequency Distribution of Accidents recorded in a Company**

No. of Accidents (X)	No. of Employees (f)	Fx
9	2	18
8	2	16
7	6	42
6	5	30
5	3	15
4	4	16
3	5	15
2	4	8
1	0	0
	Σ f = N = 31	Σfx = 160

Σ fx = 18 + 16 + 42 + - - - - - + 0 = 160

N = total number of employees = 31

The average number of accidents recorded in the company for the year is therefore 5.16, which may be approximated to 5 since number of accidents can only be expressed in a discrete measurement form.

10.1.3 Grouped Frequency Distribution with Class Intervals

The formula used for this frequency distribution is the same as that used for the regular frequency distribution. In fact, both distributions are grouped frequency distributions. While the former is without class intervals, the latter is with class intervals. For the grouped frequency distribution with class intervals:

Where and Σ remain the same

X = mid-point of class interval

F = number of cases in an interval

Fx = multiply each *x* score by corresponding *f* value

Σfx = sum of fx values

N = total number of scores i.e. Σf

Look at Table 10.2. Two employees scored between 90 and 99. Since the individual scores of the employees is lost in the course of grouping, we assume that two employees scored 94.5; this being the mid-point for the interval 90-99. This, however, is not a true representation of all intervals. Nonetheless, the error in any interval is small and in the final calculation of the mean, most of the small errors tend to cancel each other out, bringing out a final result that is essentially correct.

Table 10.2 **Grouped Frequency Distribution of Performance Scores for 100 Employees**

Class Interval	Mid-points	F	FX
90-99	94.5	2	189
80-89	84.5	4	338
70-79	74.5	3	223.5
60-69	64.5	15	967.5
50-59	54.5	24	1308
40-49	44.5	21	934.5
30-39	34.5	14	483
20-29	24.5	7	171.5
10-19	14.5	5	72.5
0-9	4.5	5	22.5
		N=100	Σfx = 4710

10.2 The Geometric Mean

The geometric mean is used as summary statistics in special situations which call for:

averaging of ratios, and
averaging rates of change

Generally, the geometric mean of n numbers is obtained by multiplying the n numbers together, and finding the nth root of the product. Thus, if we have n observations X_1, X_2, X_3, - - -, Xn then its geometric mean (G.M.) is:

If we have two numbers, the geometric mean of the two numbers is found by multiplying the two numbers together and finding the square root of

the product. For instance, if we have two numbers 7 and 5, the geometric mean of 7 and 5 is

$$\sqrt{7 \times 5}$$

If there are three numbers, the geometric mean is found by multiplying the three numbers together and finding the cube root of the product. The geometric mean of 6, 9 and 8 is

$$\sqrt[3]{6 \times 9 \times 8}$$

When the numbers whose geometric mean is to be calculated are many, then this method of stepwise multiplication and finding the root would be problematic, especially when they are more than three. This calls for the use of logarithms.

The formula for geometric mean is:

$$\text{Log G.M.} = \frac{\Sigma \log X}{n}$$

Where log X = logarithm of each observation and n remains the same

$\frac{\Sigma \log X}{n}$ = mean of the logarithms of the observations

The G.M. is then given as the antilog of log G.M. Thus, the geometric mean (G.M.) of a number of observations is the antilogarithm of the mean of the logarithm of the observations.

Table 11.3 will be used as an illustration of the use and computation of the geometric mean.

Table 10.3

Census Records of a Country for 1980 and 1990

States	*Census Records*	*% Change*	*LogX 1980*	*1990*
A	20,000	25,500	127.5	2.1055
B	7,250	11,320	156.1	2.1934
C	25,100	37,800	150.6	2.1778
D	18,275	25,410	139.0	2.1430
E	14,500	20,530	141.6	2.1511
F	9,473	17,840	188.3	2.2749
G	10,779	19,428	180.2	2.2558
H	15,000	22,320	148.8	2.1727
I	28,332	32,480	114.6	2.0592
J	16,100	24,150	150.0	2.1761

N=10	ΣLog X = 21.7094

Average increase percent in population is 48.2

The table shows the population size of 10 states in a country, for two census records in 1980 and 1990 respectively. The task is to calculate the rate of population growth in the country. The steps involved are:

1. Calculate the percentage change in population for each state by expressing the latter population size as a percentage of the earlier. That is, divide the value in column 3 by that in column 2 and multiply by 100. For state A, this gives 25,500/20,000, multiplying the result by 100 gives 127.5. Enter this in column 4. Do the same for each of the other states.
2. Find the log of each value in column 4 and enter this in column 5. For state A, the log of 127.5 is 2.1055.
3. Sum the value of the log in column 5 to give Σ log X. This gives 21.7094
4. To calculate the mean of the logs, divide the result in step 3 by 10, the number of items. This gives 2.1709; this is the log of the geometric mean.
5. Calculate the G.M. which is the antilog of the value in step 4. The antilog of 2.1719 is 148.2.
6. Interpretation: The problem is to find the average percentage change in population. The values in column 4 are not individual measurements but are ratios, and to average these ratios the geometric mean was required. The value obtained as the geometric mean is 148.2. Because the change in population size was expressed in percentages, the average change in population is determined by subtracting 100 from the geometric mean. This gives the direction of the change; whether increase or decrease. A positive result indicates an increase in change, while a negative result shows a decrease in change. In our example, 148.2 – 100 = 48.2. Thus, we conclude that there was 48.2 percent average increase in the population.

   i.e.

Tip

Geometric mean is useful when rates of increase of population, performance, production, sales, product prices, profits, etc are to be determined. It is, however, practically adapted to short series problems.

10.3 The Harmonic Mean

The harmonic mean is used in problems requiring the averaging of time rates. A rate is a ratio and as such it may be stated in either of two forms. If two measurements are taken, X and Y for instance, their ratios can be

expressed as X/Y, or Y/X. Take an example of the time taken by factory workers in a bottling company to assemble crates of drinks. In determining production rate in the company, two measurements would be required. These are the number of units produced within a given time, and the time taken to produce one unit. Using N to represent the number of units produced within one hour, and T the time to produce one unit, then the production rate may be taken in either of these two forms:

1. N/T units per hour
2. T/N hours per unit

The Table 10.4 shows the time taken by seven factory workers in a bottling company to assemble crates of drink, and it is used to explain the computation of the harmonic mean.

Table 10.4 **Production Rates of a Company**

Units per Hour	Time per Unit
A. 7 crates per hour	= 8.57 minutes per crates
B. 14 crates per hour	= 4.29 minutes per crates
C. 5 crates per hour	= 12 minutes per crates
D. 8 crates per hour	= 7.5 minutes per crates
E. 11 crates per hour	= 5.46 minutes per crates
F 12 crates per hour	= 5 minutes per crates
G. 9 crates per hour	= 6.67 minutes per crates
Total = 66 Crates per hour	= 49.49 minutes per crates
= 9.43 Crates per hour	= 7.07 minutes per crates

Therefore,

The production rate for each employee has been expressed in two forms:

1. the number of crates assembled per hour and
2. the time taken to assemble one crate.

The average number of crates assembled by the employees per hour is 66/7 = 9.43, in the first series. In the second series the average time (in minutes) taken to assemble one crate is 49.49/7 = 7.07 minutes. To express the unit of measurement in the first series as that in the second series, that is, to express the average crates per hour in the form of minutes per crate, observe the following calculations.

1 hr = 60 minutes

9.43 crates per hour implies

9.43 crates = 60 minutes

Therefore, 1 crate = 60/9.43 = 6.36 minutes

From the first series the average time to produce one crate is 6.36 minutes, while in the second series it is 7.07 minutes. This makes a difference of about 11 percent. This difference exists because the two series are not comparable at all, they have not been brought to the same basis for comparison. This is why the computation of the harmonic mean is needed.

If we wish to calculate the average time required in assembling one crate, and the data recorded are in the form of those in the first series (i.e. Units produced per Hour), then the harmonic mean is to be computed. If the data recorded is as that in the second series, then the arithmetic mean computed stands. On the other hand, if we are to compute the average number of crates assembled per minute and the information gathered are as those in the second series (i.e. Time per Unit), the harmonic mean would be needed to compute that, while the arithmetic mean is adequate if the data are in the form of the first series.

To calculate the harmonic mean, let us return to the case of calculating the average time required to assemble one crate when the data are as those in the first series. The formula for the harmonic mean is:

$$\frac{1}{H.M} = \frac{1}{N}\sum\frac{1}{X}$$

Where H.M. = harmonic mean

N = number of cases

X = an individual measurement

The steps followed in computing the harmonic mean are presented, using the data on Table 10.5.

1. Find the reciprocals of each of the numbers. The reciprocal of a number is the inverse of the number. The reciprocal of 7 is 1/7 = 0.1429
2. Find the arithmetic mean of the reciprocals. This you do by adding the reciprocals and divide by the number of cases whose reciprocals were added. This gives 1/N x 1/X in the formula. In this case it is 1/7 x 0.8246 = 0.1178
3. Calculate the harmonic mean (H.M.). H.M. is the reciprocal of the mean determined in step 2. Thus,

   H.M. = 1/0.1178 = 8.489 crates per hour

   The harmonic mean gives the average number of crates assembled per hour as 8.489, as against 9.43 initially obtained when the arithmetic mean was computed.

4. To calculate the time (in minutes) required to assemble one crate, divide 60 by the resulting step 3 (60 minutes make one hour). Thus, 60/8.489 = 7.07 minutes per crate.

Table 10.5

Computation of a Harmonic Mean of Rate of Assembling Crates in a Bottling Company

Worker	No. of Crates per hour (X)	Reciprocals of Rates (1/X)
A.	7	0.1429
B	14	0.0714
C	5	0.2000
D	8	0.1250
E	11	0.0909
F	12	0.0833
G	9	0.1111
N = 7		Σ 1/X = 0.8246

1/H.M = 1/7 x 0.8246 = 0.1178

H.M = 1/0.1178 = 8.489 crates per hour

60/8.489 = 7.07 minutes per crate

By the harmonic mean, 7.07 minutes is required to assemble one crate of mineral. This is equivalent to what is obtained by finding the arithmetic mean of the rates in the second form.

Study Session Summary

Summary

In this Study Session, we have differentiated between three types of mean; arithmetic, geometric, and harmonic. We stated that the arithmetic mean or average is calculated for ungrouped data by adding all the values together and dividing by the number of cases. For the grouped frequency data with class intervals, you have to multiply the mid-point value of each class interval with corresponding frequency value, and add all the multiples, and then divide the outcome by the total number of cases grouped. The geometric mean is calculated when an average rate of change is to be determined. Alternatively, the harmonic mean is computed when the average of measures expressed in ratios is to be determined.

Assessment

Assignment

1. In a population census conducted, statistics were taken of the number of families that have a particular family size in a community. This is reported on the table below:

Number of families	1	2	3	4	5	6	7	8
Size of family	10	6	5	8	3	2	7	4

What is the average family size in the community?

2. A consumer psychologist is interested in studying the inflation rate in a town. He obtained the prices of selected products for two different years. The data obtained are presented on the Table below:

Products	**Prices**	
	2003	**2005**
1. Bournvita	350.00	430.00
2. Sugar	120.00	180.00
3. Milk	280.00	360.00
4. Butter	150.00	200.00
5. Cornflakes	180.00	245.00
6. Bread	100.00	120.00
7. Milo	370.00	450.00
8. Lipton tea	60.00	75.00
9. Nescafe	80.00	95.00

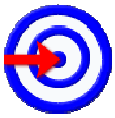
What is the average inflation rate?

Study Session 11

Measures of Central Tendency: Median

Introduction

This study session will expose you to how scores are arranged. The median is the mid-point of a set of scores. It is the point on a continuum of scores above which fall exactly one half of the cases and below which fall the other half. When scores are arranged in order of magnitude, either in ascending or descending order, the median has the same number of scores above and below it.



Learning Outcomes

When you have studied this session, you should be able to:

- *locate* the median position of a set of ungrouped data, as well as calculate the median value of grouped frequency distribution, with and without class intervals.
- *determine* the median value of a distribution, using the distribution curve, called OGIVE.

11.1 The Median of Ungrouped Data

The calculation of the median of ungrouped data takes two forms, as determined by the number of observations, N. When N is odd, the median is determined as the ½(N+1)th observation. When N is even, the median is the average of the two most central scores. For example, with scores like 9, 11, 8, 7, 15, 13, 8, 9, 9. N is 9, which is an odd number, therefore, the median will be calculated as Md = ½(9 + 1)th value, i.e. ½(10)th value = 5^{th} value. To decide on the 5^{th} value, we have to re-arrange the values in order of magnitude. This gives us: 7, 8, 8, 9, (9), 9, 11, 13, 15.

The 5^{th} value is 9, therefore, the median for the set of scores is 9. For another set of scores like 9, 11, 8, 7, 15, 13, 8, 9 the number of observations is 8, which is even. The median will be taken as the mean of the two middle values, that is, the mean of the 4^{th} and 5^{th} observations. Arranging the scores in order of magnitude, we have 7, 8, 8, (9, 9), 11, 13, 15. The 4^{th} and 5^{th} observations are 9 and 9 respectively. Md is ½ (9 + 9) = ½ (18) = 9.

11.2 The Median of Grouped Data

When the data are grouped in a regular frequency distribution or in a grouped frequency distribution with intervals, the calculation of the median takes a different pattern. The computational formula is given as:

Where LRL	=	lower real limit of median interval
N/2	=	one half of the total number of cases
SFB	=	sum of all the frequencies up to, but not including the median interval
F	=	the frequency within the median interval
I	=	class interval size

11.2.1 The Median of Grouped Data with Regular Frequency Distribution

The computational procedures shall be explained, using the data in Table 10.1, which is now shown in Table 11.1 for the median calculation.

Table 11.1 **Regular Frequency Distribution of Accidents recorded in a Company**

No. of Accidents (X)	No. of Employees (f)
9	2
8	2
7	6
6	5 5 cases down here
5	3 **Median interval**
4	4 13 cases up here
3	5
2	4
1	0
	Σ f = N = 31

1. Determine the median interval. The task of calculating the median is to determine the point on the scale on either side of which lie half the cases. As the total number of cases, N, in this example is 31, then 15.5 cases will lie on either side of the median, i.e. N/2. However, there is a problem locating this in a frequency distribution where the individual identity of the items has been lost while grouping.

   From the bottom of the frequency column, add up the frequencies until the interval having the 15.5^{th} cases is reached. This is called the median interval or critical interval. In the table, the median interval is 5. The sum of the frequencies before the median

interval (i.e. "less than" cumulative frequency) is 13, and the frequency of the median interval is 3. Therefore, the 15.5th case must lie within that interval; thus, it is regarded as the median interval.

2. Determine the lower real limit (LRL) of the median interval, which is 4.5
3. Calculate half the total number of cases N/2 = 31/2 = 15.5
4. Sum the frequencies below the median interval, SFB = 13.
5. Determine the frequency within the median interval, f = 3
6. Determine the class width (I) = 1
7. Calculate the median (Md) = 5.33.

11.2.2 The Median of Grouped Data with Grouped Frequency

The steps involved in the computation are similar to those of the regular frequency distribution. Using the data in Table 10.2, we will present the calculation of the median. The median interval is the class interval, having the N/2, i.e 100/2th case. This makes the 50th case the median point. The corresponding interval, 40-49, is the median interval. Follow the steps outlined before:

LRL =39.5

N =100

SFB = 31

F =21

I =10

Theoretical Explanations

In Table 11.2, the median interval is the interval 40 – 49 since the 50th case resides within the interval. Remember, the 50th case is the median point for the distribution as N/2 i.e. 100/2 is 50. Thirty-one cases are counted up to the top of the interval 30-39 (less than cumulative frequency). The next interval 40-49, which is the median interval contains 21 cases, suggesting that the 50th case lies somewhere within this interval because adding 21 to 31 gives 52; this is greater than the 50th case desired. The problem now is how to determine where the 50th case lies within the median interval?

For the sake of interpolation, we assume that the 21 cases to the median interval are evenly distributed over the whole range of the interval with exact limits as 39.5 and 49.5. Actually, we require 19 cases of the 21 cases in the median interval to be added to the sum of the preceding cases (i.e. 31) to make the desired 50 cases. Thus, we need a fraction of 19/21 of the median interval. Since the class width of the median interval is 10, it implies we desire 19/21 of 10, which is 9.05 units of the interval to correspond with the 50th case desired. Adding the 9.05 to the exact lower limit of the interval (39.5), we now have 39.5 + 9.05, which is 48.6 as the median value. So 48.6 corresponds with the 50th case in the distribution. You can as well observe that 48.6 falls within the median interval.

Table 12.2 **Calculation of the Median of Performance Scores for 100 Employees**

Class Interval	F	Cf
90-99	2	100
80-89	4	98
70-79	3	94
60-69	15	91
50-59	24	76 48 cases down here
40-49	21	52 **Median Interval**
30-39	14	31 31 cases up here
20-29	7	17
10-19	5	10
0-9	5	5
	N=100	

As the median is the dividing point of a distribution into two halves, it could as well be determined from the top-end of a frequency distribution, that is, by working down from the top. Starting at the top of the frequency column ("more than" cumulative frequency), 48 cases are count down to the interval 50-59. We need 2 more cases to make up the 50^{th} case desired. This implies going 2/21 of the way down into the median interval. Again, the class width is 10, therefore we need to go down 2/21 of 10, which is 0.95. Since we are working from the top intervals, we now subtract 0.95 from the exact upper limit of the median interval (49.5). We now have 49.5 - 0.95, which is 48.6 as the median value. This is exactly what we obtained before when calculating from the bottom interval.

Thus, working from the top class intervals, the formula for median becomes

Where URL = upper real limit of median interval

N	= total number of cases
N/2	= one half of the total number of cases
SFA	= sum of all the frequencies above the median interval but not including the median interval
F	= frequency within the median interval
I	= class interval size

For the data in Table 11.2

URL = 49.5

N = 100

SFA = 48

F = 21

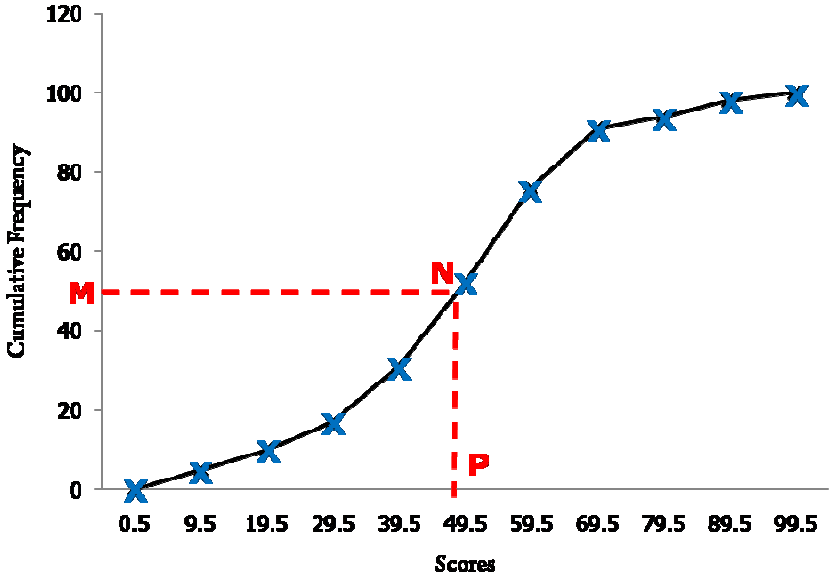
I =10

11.3 Determining the Median using the Cumulative Frequency (Or OGIVE Curve)

To find the median of grouped data, the cumulative frequency distribution curve can as well be used. From the curve, we can easily read off the median, having determined the median point as N/2 on the Y-axis for the cumulative frequency.

Figure 11.1 is the OGIVE of data in Table 11.2, employing the "less than" cumulative frequency distribution. The median is the dividing point of the desired point, that is N/2, which is 100/2 = 50. Read off 50 on the Y-axis to meet the curve at N. From N, draw another dotted line perpendicular to the Y-axis to meet the X-axis at P. The corresponding value of P on the X-axis is the median of the data.

Fig 11.1 **OGIVE showing Median Scores**



Study Session Summary

Summary

In this Study session you learned that that the median is the dividing point of a distribution into two halves. In ungrouped data, it is determined by arranging the data in ascending or descending order and then locating the most central value. For grouped data with class intervals, it involves a longer process.

Assessment

Assignment

Now that you have completed this study session, attempt the following assignment (a tutor marked assessment).

1. Ten employees in an organization reported the years they have spent on the job as follows:

   14, 20, 16, 14, 12, 10, 18, 16, 11, and 12

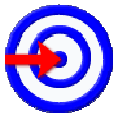
   What is the median year of the employees on the job?

Study Session 12

Measures of Central Tendency: The Mode

Introduction

This Study Session aims to provide explanations on the mode, which is the most frequent score in a distribution of scores. The mode is the value with the highest frequency.



Learning Outcomes

When you have studied this session, you should be able to:

- *compute* the modal value for ungrouped and grouped frequency distributions with and without class intervals.

Given a set of scores like these in Table 12.1, the mode is 5. The highest frequency is 15, while the corresponding score is 5, indicating that the score 5 occurred 15 times in the set of data. Thus, 5 is the mode, being the most frequent score.

Table 12.1

Determining the Mode of a Distribution

Score	0	1	2	4	5	7	8	9
Frequency	7	3	11	7	15	6	4	5

12.1 Mode of Grouped Data with Intervals

When data are grouped in the class interval form, the crude mode or interpolated mode may be determined. The interpolated mode is, however, more accurate than the crude mode.

12.1.1 The Crude Mode

The crude mode is simply the mid-point value of the class interval having the highest frequency. In Table 11.2, for instance, the class interval with the highest frequency is 50-59. The mid-point is 54.5 therefore the crude mode of the distribution is 54.5. In some cases a distribution may have two class intervals with the same frequency as the highest. Care must be taken in deciding what the mode is.

The following rules should be noted:

1. If there are more than one class intervals separating the two class intervals, like distribution A in Table 12.2, then we say that the distribution has two modes or it is bimodal. The modes will be the two mid-points of the respective class intervals. Thus, we have 9.5 and 3.5 as the modes for distribution A in Table 12.2.

2. If the two intervals are separated by only one intervening interval, like distribution B in Table 12.2, then it is possible that the distribution is actually unimodal. It is more certain to be unimodal when the intervening interval has a relatively high frequency. In such a case, the crude mode is not adequate as it will not be possible to decide what the crude mode is.
3. When the two class intervals with the highest frequency occupy adjacent positions like distribution C in Table 12.2, then the crude mode is decided as the dividing point between the two intervals. The dividing point is simply the upper real limit of the lower class interval, which is equivalent to the lower real limit of the higher of the two class intervals. For distribution C in Table 12.2, the crude mode is 6.5. Such distribution is invariably unimodal.

Table 12.2 **Illustrations of when the Crude Mode is Adequate**

Class Interval	Distributions (Frequency)		
	A	B	C
11 – 12	2	2	2
9 – 10	7	7	6
7 – 8	6	6	7
5 – 6	2	7	7
3 – 4	7	2	2
1 – 2	3	3	3
	Bimodal	**Unimodal**	

12.1.2 The Interpolated Mode

With the limitation in the use of the crude mode, the interpolated mode is favoured. The crude mode is ideal for small sample of less than 100. When the sample is large, and the distribution is skewed, the mid-point value of crude mode is not sufficiently accurate. This calls for an interpolation within the modal interval to obtain a more accurate estimate. The formula for calculating the interpolated mode is given as:

Where LRL = lower real limit

d_1 = difference between the frequency of the modal interval and frequency of the preceding interval

d_2 = difference between the frequency of the modal interval and the frequency of the next following interval

I = class interval size

An example of calculating the interpolated mode will be applied to the data in Table 11.2. Some extracts from the Table are shown below:

60 – 69	15
50 – 59	24 Modal Interval
40 – 49	21

The modal interval is the interval with the highest frequency, that is, 50 – 59. The preceding interval is 40 – 49, the following interval is 60 – 69.

LRL = 49.5

$d_1 = 24 - 21 = 3$

$d_2 = 24 - 15 = 9$

I = 10

$$= 49.5 + 3/12 \times 10$$

$$= 49.5 + 2.5 = 52$$

The crude mode is 54.5 while the interpolated mode is 52, which is a "pull away" towards the preceding interval. The reason is that the preceding interval has a frequency of 21, while the following interval has a frequency of 15; hence, the pull towards the preceding interval. If the frequency of the following interval was large, then the pull would have been directed towards the following interval, making the interpolated mode above 54.5 but less than 59. Thus, the interpolation of the mode within the modal class interval improves the estimate of the mode by allowing the adjoining frequencies to add their weight in arriving at a final estimate.

12.2 Determining the Mode using the Histogram

The mode of grouped data can easily be read off from histogram. An illustration is made from the data in Table 11.2, and the histogram is displayed in Figure 12.1.

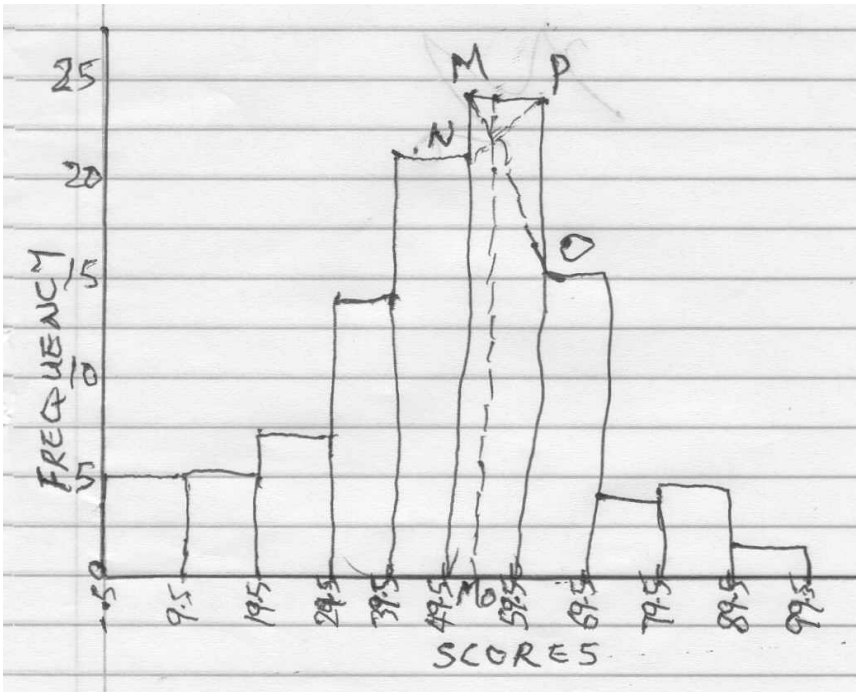


Fig 12.1 **Showing how to determine Mode using the Histogram**

To determine the mode from a histogram, draw two diagonals to join the corners of the modal class, labelled as Mo and Np respectively in Figure 12.1. At the point where the diagonals intersect, draw a perpendicular line through to meet the x-axis at Mo. The value of Mo on the x-axis is read as the mode, which is 52.5 in this case.

12.3 The Characteristics and Use of Measures of Central Tendency

The mean, median, and mode, already referred to as the measures of central tendency are commonly used in statistical computations. There are, however, situations in which one is more adequate than the others. Knowledge of the characteristics and uses of each of the measures will invariably expose when to use any of the measures.

The Mean

1. The arithmetic mean is often used when the distribution is reasonably symmetrical. This is because the mean takes into account all measures in a distribution. The extremely small values and the extremely large ones, as well as those near the centre of the distribution, influence the mean. When, however, there are extreme values on one side of the distribution that are not balanced by extreme values on the other side, the mean is greatly influenced. This results in a misleading picture of the location of the measure of central tendency. Consequently, the mean is not adequate for a distribution that is highly skewed.
2. For a normal distribution, the sum of differences (or deviations) of all scores from the mean is zero. That is, X – Mean = 0. This implies that the mean balances the sums of the positive and negative deviations. This will be expanded later.

3. The mean is the most reliable and stable of the three measures of central tendency. This is because when different samples are drawn from the same population and their means computed on the same measure, there will be little or no difference in the means. Also, errors of measurement tend to neutralize one another around the arithmetic mean. If any error exists in the use of a mean, the error is considerably smaller than the error of a single measure in the distribution.
4. The mean is useful in further statistical computation, while other measures of central tendency are usually not. For the characteristics of the mean, it is used in many of the procedures of inferential statistics.

The Median

1. The median is a positional measure of central tendency. Its position is determined directly by the number of values in a distribution, not the magnitude of the values. The magnitude of values only indirectly affects the median by determining the serial ordering before decision on what the median is.
2. Though all measures contribute to the calculation of the median, just like the mean, the median is less affected by the extreme of the distribution than is the arithmetic mean. For example, the median of the scores 3, 5, 7, 9, and 9 is 7 while the arithmetic mean is 6.6. If the extreme values are changed and we now have 4, 6, 7, 12, and 13, the median is till 7 but the arithmetic mean is now 8.4. If the scores are 4, 6, 7, 9, 9, 12, and 13 then the median becomes 9. Thus, the median is mainly influenced by the number, rather than the size of the extreme values in a distribution.
3. The median is an easy measure of central tendency to compute.

The Mode

1. It is the quickest estimate of central tendency available.
2. It is the most typical measure in that it tells the most occurring in a set of objects.

Study Session Summary

Summary

In this Study Session, we have discussed that the mode is the most occurring value in a set of values. We have also looked at the importance of the different measures of central tendency.

Assessment

Assignment

1. What is the modal family size of the data in Assignment 10.1.
2. Calculate the mode of the values in Assignment 11.1.

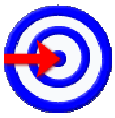
Forward your findings to your tutor by submitting answers at Assignment 12 page on UI Mobile Class.

Study Session 13

Measures of Variability

Introduction

A measure of central tendency tells much about a distribution but does not give a complete picture of the distribution. When two distributions are to be compared, if it is based solely on averages, quite wrong conclusions may be drawn. Other descriptive statistical measurements need to accompany averages to amplify the description of a set of data, hence this Study Session will focus on measures of variability are essential.



Learning Outcomes

When you have studied this session, you should be able to:

- *differentiate* between measures of central tendency and measures of dispersion or variability.
- *compute* and *interpret* measures of dispersion like range, mean absolute deviation, standard deviation, and variance by inference. (SAQ 13.1)

A measure of variability tells how spread out or scattered scores in a distribution are. The measures of variability are needful when comparing two distributions, for instance, two distributions of statistical data may be symmetrical and may have the same means, medians and modes; yet they may differ in the distribution of the individual scores about the measure of central tendency. The measures of variability, thus, explain and illustrate the methods of describing the dispersion and skewness of distributions by the use of single numbers. For example, consider the three distributions below:

A	B	C
8	10	20
8	9	9
8	8	8
8	7	3
3	3	3
3	3	2
3	3	1
3	2	1
3	2	0

The arithmetic mean of each of the three distributions is the same, i.e. 5, while the median and mode is 3 and it is the same for the three

distributions. If we are to compare the three distributions on the basis of the measures of central tendency, then we may conclude that the distributions are the same. This may be rather erroneous because a close observation of the scores in each distribution shows that they are dispersed differently. For instance, the lowest score in distribution A is 3 while the largest is 8 and the scores do not vary much from one another. For distribution B, the lowest is 2 while the largest is 10 and the scores are more spread out than distribution A. Distribution C has 0 as the lowest score and 20 as the largest. The scores in distribution C are more dispersed than the other two distributions.

A measure of dispersion or variability is needed for a more accurate comparison to be made among the distributions. Thus, the central locations of two distributions may be the same but the variability may differ. Also, the variability may be the same but the central locations may differ. Locations and variability are actually independent. While variability refers to the distance between each score and any other and is measured with the measures of variability, the measures of central tendency show locations in a set of data. A measure of variability show, numerically, the degree to which scores spread around the average. It is a measure which tells whether scores cluster closely around the mean or whether they scatter widely.

Measures of variability are useful in many areas of life, although it is not frequently reported because it is not familiar to many as the measures of central tendency. A measure of variability tells us how representative the average is. If the numerical value of the measure is small, it means that the individual scores are close to the average. If the variation score is large, the mean can be used with less assurance because the scores are rather far from the mean.

13.1 The Range

The range is the simplest and quickest method of summarising the variability of a distribution. It is simply defined as the difference between the highest value, H, and the lowest value, L. This can be symbolized as:

$$R = H - L$$

Where R = range

H = highest value

L = lowest value

For grouped data, the range is calculated by subtracting the lower limit of the bottom interval from the upper limit of the top interval. Thus,

$$R = Ult - Llb$$

Where R = range

Ult = upper limit of top interval

llb = lower limit of bottom interval

The range is, however, the most unreliable of the measures of variability. This is because it is determined by only two measures, with all the other individual values in between having no effect in the calculation. There is

often a gap between those extreme values used in determining the range and the next highest or lowest value in a distribution. If these extreme cases are absent from the distribution then there would be a significant difference in the calculated range. For instance, consider these two distributions.

Distribution A: 20 16 16 15 15 15 15 14 14 10

Distribution B: 20 19 19 18 17 16 15 14 11 10

The range for the two distributions is 20 - 10, which is 10. An observation of the values in distribution A shows that the scores are relatively clustered near the middle of the distribution while those of distribution B are more spread out. The range is therefore a weak measure of variability of distribution A. For instance, when the two extreme scores are eliminated from the two distributions, the range for distribution A becomes 16 - 14, which is 2; this is rather far from the earlier range of 10. For distribution B, the range becomes 19 - 11, which is 8; it is still close to the initial range of 10. Since this type of change occurs frequently, the range is considered as a crude measure of variability, just like the mode, and it is avoided as much as possible in the behavioural sciences.

Generally, the range is more unreliable with a small sample size than a large sample size. This is because with a small sample size the extreme values may be marked by one case. But when the sample size is large, there may be more than one case of each of the extreme values. Thus, the probability of the range being altered as a result of missing scores is higher for a smaller sample than for a larger sample.

13.2 The Median Absolute Deviation (MAD)

The mean deviation is also referred to as the mean variation. It is a measure of dispersion which considers the deviation of each measure in a distribution from a measure of central tendency. The deviation from the mean, as it is always called, is obtained by subtracting each value in the distribution (X) from the mean value (X), of the distribution. If a particular measure is greater than the average, it will have a positive deviation, with a positive sign (+), and it is said to have a positive deviation. If the score is less than the mean value, it has a negative sign (-), and it is said to have a negative deviation. A measure that is identical with the measure has a zero deviation. If all the deviations from the mean were calculated for all the measures in a distribution, and the average of the sum computed, the resultant value is the mean of deviations from the mean. If the measures in the distribution were widely dispersed or scattered, then their deviations would be relatively large and this would be reflected by the mean of the deviations. If the measures were clustered or concentrated about the mean of the distribution, then the mean of the deviations would be small. Thus, the mean deviation is an excellent measure of dispersion.

In calculating the mean deviation, however, the signs of the deviations of each measure from the mean are usually ignored. This is because in a perfectly symmetrical distribution, the algebraic sum of deviations from the median or mode would be zero, and such is the case in any distribution; the difference between the positive and negative deviations

(i.e. algebraic sum of deviations from the mean) must always be zero. Recall that it is the mean of this sum of deviations that is calculated to obtain the mean deviation.

Consequently, the deviation of a measure from the mean is meaningless in this wise, rather it is the absolute deviation from the mean that is meaningful. By absolute deviation from the mean, the signs of the deviations are ignored, while the magnitude of the deviation, that is, the actual value of deviations from the mean are emphasized. This is referred to as mean absolute deviation (MAD). Thus, all deviations are treated as positive. Some examples are now considered from ungrouped and grouped data.

Table 13.1 shows the raw scores of trainees on a 20-point test. The interpretation of the mean absolute deviation measure of variability is that, on the average, the individual scores in the distribution vary from the mean score of 16.2 by approximately 1.6. Thus, scores within the limits of 1 MAD below the mean and 1 MAD above it will range between 14.6 and 17.8, while +- 2 MAD range between 13 and 19.4. If you subtract 1.6 from 16.2 it gives 14.6, while if you add 1.6 to 16.2 you will have 17.8 (i.e. for + - 1 MAD). For + - 2 MAD multiply 1.6 by 2, then subtract and add respectively to 16.2 to have 13 and 19.4 respectively.

Table 13.1

Calculation of Median Absolute Deviation and Standard Deviation from Ungrouped Data

Col. 1 Score (X)	Col. 2 Deviation ( )	Col. 3 Absolute Deviation ( )	Col. 4 ( )²	Col. 5 ²
17	0.8	0.8	0.64	289
11	-5.2	5.2	27.04	121
16	-0.2	0.2	0.04	256
17	0.8	0.8	0.64	289
15	-1.2	1.2	1.44	225
18	1.8	1.8	3.24	324
20	3.8	3.8	14.44	400
15	-1.2	1.2	1.44	225
16	-0.2	0.2	0.04	256
17	0.8	0.8	0.64	289
Σ X = 162 **N = 10** **= 16.2**	**Σ( ) =0**	**Σ( ) = 16.0**	**Σ( )² = 49.60**	**ΣX^2= 2674**

Where **MAD** = Mean Absolute Deviation

= Individual Score

= an average of the individual scores

= the absolute deviation of each score

13.3 The Standard Deviation (S.D.)

This is statistically the most reliable measure of dispersion, denoted by S for sample and δ for population. It varies less than any other measure of variability, from sample to sample; the samples being chosen at random from the same population. The standard deviation and variance are the most frequently used measures of dispersion or variability because they have certain properties that make them preferred to the other measures. Variability, in actual sense, refers to the extent to which scores differ from one to another. In statistical terms, it means the difference between each score and every other score in a distribution. Since the mean is the most frequently used measure of central tendency, and best represents all scores in a distribution, it is more reasonable to define the deviation of a single score from the other scores as its distance from the mean. This is symbolized by:

Deviation score =

Where $\boldsymbol{X}$ = individual score in the distribution

$\boldsymbol{X}$ = the average of the distribution

When the deviation of a single score from each of the other scores in a distribution is determined; that is, by computing the difference between the score and each of the other scores, and the average of these deviation scores is calculated, it will not be much different (if at all different) from the deviation of the score from the mean of the distribution

(i.e **)**. Thus, the deviation or distance of a single score from the mean is a representation of all possible deviations or distances of the score from each of the other scores in the distribution.

When a score is extremely different from the mean, the numerical deviation score for that score will be large. But scores that are close to the mean will have small numerical deviation scores. The word "numerical" is emphasized because it is actually the magnitude of the deviation that is of interest. The deviation scores will normally have some positive and some negative values which, when the deviation scores are added (i. e. $\boldsymbol{\Sigma(X - X)}$, it is always equal to zero. Thus, if we are to calculate the average variability by averaging the deviation scores, that is, by summing the deviation scores and dividing by N, i.e.

Then we will always have a zero score, which will make a nonsense of the whole exercise. To overcome this problem, we have to take the absolute value of each deviation by ignoring the sign, and then utilizing

its numerical value for the computation of the mean absolute deviation (MAD). Though the MAD is a descriptive measure of variability, but it is usually rejected for its inadequacy in further statistical analysis. This is because absolute values which are incorporated in MAD are unsuitable for use in inferential statistical analysis.

The measure which takes care of this problem without using absolute values, and is useful for further statistical analysis is that which requires the squaring of each of the deviations before taking the average of the squared deviations. By squaring the deviation, the negative signs are taken care of and all the squared deviations become positive. The sum of the squared deviations from the mean, that is, $\Sigma(X - X)^2$ is called the sum of squares and is symbolized by SS. The average of the sum of squares is a measure of variability, called variance, which is symbolized by $\boldsymbol{\delta^2}$.

This is, however, the formula for the variance of a population. When the data of a sample are to be used in estimating the variance of the population from which the sample was drawn, then the sample variance is computed with some slight changes in the formula:

Whereas the population variance is represented with a square of the Greek letter sigma ($\boldsymbol{\delta^2}$), the sample variance is represented by ; although both are often used interchangeably. For the population variance, the sum of squared deviations is divided by N, while that of the sample is divided by **N – 1**. This is to enable to be unbiased estimate of the population variance. **N – 1** is the number of deviation about the mean that are free to vary. This number is called the degree of freedom. For reasonably large samples, however, the actual value of variance is affected a little, whether we divide by **N – 1** or **N**.

In each of the formulas above, the steps involved are:

1. Subtract the mean from each score.
2. Square each result to eliminate the minus signs
3. Sum the squares
4. Divide the sum by **N** if population variance, or **N – 1** if sample variance.

The computation of S.D. differs from the M.D. in two ways. First, in computing S.D. the deviation of each measure from the mean is squared. Second, the deviations are always taken from the arithmetic mean, whereas that of the M.D. may be taken from the median. Also, having squared the deviation these squares are then summed, and the sum is divided by **N**. Then the square root is taken, to give S.D.

Be informed that variance is a square of the original units in which the mean deviations are based; this is aimed at eliminating the negative signs that would have yielded a total of zero if the deviations from the mean were added. The average of these squared deviations will certainly not produce an average variability that is the most ideal measure of variability. To accomplish this, we have to take the positive square root

of the variance, bearing in mind that the variance is the average of squared deviations. This yields a measure of variability called, the standard deviation, symbolized by δ for population standard deviation or S for sample standard deviation. The formulas are:

13.3.1 Calculation of S.D. from Ungrouped Data

Data from Table 14.1

- This formula is used for *manual calculation* and is called the definition formula

- This formula is used with a *scientific calculator*, and is called the computational formula

Population

- This formula is used for *manual calculation*, called definition formula

- This formula is used with a *scientific calculator*, called computational formula.

The average deviation of scores from the mean is 2.35 and 2.23 for sample and population respectively, in the distribution.

13.3.2 Calculation of the MAD and S.D. from Grouped Data

Where M.D., f, , and N remain the same

X = the mid-point of each class interval

These formulae are called definition formula. They are used for manual calculation.

$$\frac{\sum f(X-\bar{X})^2}{N} \qquad \frac{\sum f(X-\bar{X})^2}{N-1}$$

$$\sqrt{\frac{\sum f(X-\bar{X})^2}{N}} \qquad \sqrt{\frac{\sum f(X-\bar{X})^2}{N-1}}$$

These formulae are called computational formula. They are used with a scientific calculator.

Table 13.2

Computation of M.D. and S.D. from a Distribution of English Exam Marks (Mean = 59.5)

Col.1 Scores	Col.2 Mid-Point (X)	Col.3 f	Col. 4 fX	Col.5 X^2	Col.6 fX^2	Col.7 X –	Col.8 $(X-)^2$	Col.9 f(X-)	Col.10 $f(X-)^2$
79-83	81	1	81	6561	6561	21.5	462.25	21.5	462.25
74-78	76	3	228	5776	17328	16.5	372.25	49.5	816.75
69-73	71	3	213	5041	15123	11.5	132.25	34.5	396.75
64-68	66	4	264	4356	17424	6.5	42.25	26.0	169.00
59-63	61	7	427	3721	26047	1.5	2.25	10.5	15.75
54-58	56	4	224	3136	12544	3.5	12.25	14.0	49.00
49-53	51	1	51	2601	2601	8.5	72.25	8.5	72.25
44-48	46	4	184	2116	8464	13.5	182.25	54.0	729.00
39-43	41	1	41	1681	1681	18.5	342.25	18.5	342.25
34-38	36	2	72	1296	2592	23.5	552.25	47.0	1104.50

Study Session Summary

Summary

In this Study Session, you discovered that the measures of variability tell us the degree to which scores vary from one another, and from the mean. The measures include the range, mean absolute deviation, standard deviation, and variance.

The range is the difference between the highest value and the lowest value in a distribution. The mean absolute deviation is the deviation of each measure in a distribution from a measure of central tendency. The standard deviation is the positive square root of the average sum of square deviations from the mean, while the square of this value is the variance. Thus, the average of the sum of squares is variance.

Assessment

Assessment

SAQ 13.1

Given the following frequency distribution, find the standard deviation of the data.

X	**6**	**7**	**8**	**9**
F	**2**	**3**	**3**	**2**

Assignment

1. The ages of a sample of students is given as:
   17, 12, 20, 22, 19, 19, 23, 24, 18, 19, 20 and 21.

   Calculate the range, mean absolute deviation, standard deviation, and variance of these ages and interpret each result.

2. The data below is the accident record of a sample of factory drivers in one year. It shows the number of accidents each driver had in the year.

Number of Accidents	9	8	7	6	5	4	3	2
Drivers involved	2	2	6	5	3	4	5	4

   Calculate and interpret the dispersions in the accident data.

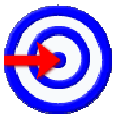
Send your findings to your tutor for feedback at Study Session 13 Assignment Page at UI Mobile Class

Study Session 14

Transformed Scores

Introduction

If you are informed that you have obtained 75% in PSY 103 examination, you will definitely need additional information in order to determine how well you performed, compared with other students who took the examination. If the examination was generally easy for most students and there were many high scores, then your score of 75% may be an average or even below average score. If, on the other hand, the examination was a little bit difficult for the whole class, then your score may be among the highest or even the highest. One way of getting this additional information is to transform the original score to show at a glance how well you performed in comparison to other students in the course. Thus, this Study Session addresses the different ways of transforming scores to allow for relative comparison of an individual's scores across different tasks, as well as relative comparisons of individuals' scores on the same task.



Learning Outcomes

When you have studied this session, you should be able to:

- *compute* transformed scores.
- *interpret* percentile, standard score (Z score), and T score.

14.1 Percentiles

The percentile rank of a score is a single number which gives the percentage of cases in the specific reference group, scoring at or below that score. If, for example, your raw score in an examination is 75%, and it corresponds to a percentile rank of 65, it means that 65% of the class obtained equal or lower scores than you did, while 35% of the class received higher scores. A percentile is, thus, the score at or below which a given percent of the case lie. The percentile shows directly how an individual score compares to the scores of a specific group. It is important to keep in mind that a percentile rank cannot be correctly interpreted, unless the reference group is taken into consideration.

14.1.1 Computational Procedure: Computing the Corresponding Percentile Rank from a given a Raw Score

To illustrate the computation of percentile ranks, reference will be made to Table 14.1. Our task is to find the percentile rank corresponding to the raw score of 43.

Table 14.1

Hypothetical Examination Scores for Students in PSY 103 Test: Transferring a raw score of 43 to a percentile rank

Class Interval	Frequency	Cumulative Frequency (CF)	
48 – 50	1	87	
45 – 47	5	86	6
42 - 44	6	81	Critical Interval
39 – 41	8	75	
36 – 38	9	67	
33 – 35	14	58	
30 – 32	12	44	
27 – 29	9	32	75
24 – 26	7	23	
21 – 23	5	16	
18 – 20	5	11	
15 – 17	6	6	

	F	Percent (= f/N)
All higher intervals	6	6/87 = 6.9% (H %)
Critical interval (42 – 44)	6	6/87 = 6.9% (I %)
All lower intervals	75	75/87 = 86.2% (L %)

Step 1: Locate the class interval wherein the raw score falls; call this the "critical interval".

Step 2: Group the frequencies (F) into three categories, embracing those with scores higher than the critical interval, those corresponding to all scores in the interval, and those corresponding to all scores lower than the critical interval. These are 6, 6, and 75 respectively.

Step 3: Each frequency is then converted to a percent by dividing by N; in this case 87. We will denote the percent of people scoring in interval higher than the critical interval by H% (for higher), the percent of people scoring in the critical interval by I% (for in), and the percent of people scoring lower than the critical interval by L% (for lower)

Step 4: It is clear at a glance that the score of 43 is better than at least 86.2% of the scores, i.e., those below the critical interval. Thus, the percentile rank must be at least 86.2%. It is also clear that 6.9% of the scores are better than 43, i.e. the ones above the critical interval. However, it is not evident whether the raw score of 43 is higher than all the scores that fall in the critical interval, or it is lower than, since the interval carries 6.9% of the entire scores. The solution is to look at the

score in comparison to the size of the interval; the higher the score in relation to the critical interval the more people in that interval is assumed to have been outscored.

Step 5: To determine the standing of score 43 in the critical interval, first ascertain the lower real limit of the interval, which is 41.5. Subtract this from the raw score of 43 (43 – 41.5) = 1.5. Since the interval size is 3, this distance is expressed as a fraction, and is equal to 1.5 points/3 points, = 0.5 of the interval. Thus, score 43 is 0.5 of the 6.9% of students in the critical interval. This gives a value of 3.45, which should be added to 86.2% (i.e. L %). The final result is 89.65%; implying that 89.65% of those that took the examination had 43 or less, which is the percentile rank. The percentile rank is determined from the formula:

L% and I% remain the same

Score = raw score in question

LRL = Lower real limit of critical interval

H = interval size

The percentile rank is therefore 89.65%, indicating that approximately 90% of the class received equal or lower scores and only about 10% received higher scores.

14.1.2 Computational Procedure: Computing the Corresponding Raw Score from a Given Percentile

This is a reverse of the former, requiring finding the raw score that corresponds to a specified percentile value. For example, suppose the Study Sessionr wishes to give a special teaching to the bottom 32% of the class; what raw score should be used as the cutting line? In this case the percentile (32%) is specified and the raw score is needed. The steps to adopt are as follows:

Step 1: Convert the percentile to a case number by multiplying the percentile by N. (i.e. .32 x 87 = 27.84). Thus, the score that corresponds to the individual whose rank is 27.84 from the bottom of the class (i.e. the person scoring at the 32nd percentile) is the cutting line needed.

Step 2: Find the interval in which the case number computed in step 1 falls. This is easily accomplished by starting at the bottom of the cumulative frequency distribution and proceeding upward until you find the first value equal to, or greater than the critical case (27.8); the corresponding interval is the critical interval. This is illustrated with data on Table 14.2.

Table 14.2

Hypothetical Examination Scores for Students in Psy 103 Test: Finding the raw score corresponding to the 32nd percentile

Class Interval	Frequency	Cumulative Frequency (CF)	
48 – 50	1	87	
45 – 47	5	86	
42 - 44	6	81	
39 – 41	8	75	
36 – 38	9	67	
33 – 35	14	58	
30 – 32	12	44	
27 – 29	9	32	first of Z = 27.84
24 – 26	7	23	
21 – 23	5	16	
18 – 20	5	11	
15 – 17	6	6	

pN = .32 x 87 = 27.84

Step 3: The 27.8th case must have a score of at least 26.5; the lower real limit of the interval in which it appears. However, the critical value covers three score points (26.5 – 29.5); what point value corresponds to the 32nd percentile?

There are 23 cases below the critical interval, so that the 27.84th case needs 27.84 – 23, i.e. 4.84 cases up in the interval. The total number of cases in the interval is 9, so this distance expressed as a fraction is 4.84 cases /9 cases = .54. Therefore, in addition to the lower real limit of 26.5, .54 of the three points included in the critical interval must be added in order to determine the point corresponding to the 27.84th case. The cutting score is therefore equal to 26.5 + (.54 x 3) = 26.5 + 1.62 = 28.12.

A general formula to obtain this is:

Where

Score p = Score corresponding to the pth percentile

LRL – lower real limit of critical interval

P – specified percentile

N – total number of cases

SFB – sum of frequencies below critical interval

F – frequency within critical interval

H – interval size

Thus, people with scores 28 or less need special teaching, while students with scores of 29 and above do not need.

14.2 "Z and T" Scores

These scores help us to derive a transformed score that shows at a glance the relationship of an original raw score to the mean, using the standard deviation of the reference group as the unit of measurement. Suppose a student obtains a raw score of 70, 65, and 55 in tests on PSY 101, PSY 102, and PSY 103 respectively, it might seem as though the student's best score is in PSY 101 and the poorest is in PSY 103. However, the raw scores of 70, 65, and 55 cannot be compared directly because they come from different distributions with different means and different standard deviations. Thus, the units of measurement are not the same from test to test. You can overcome this by transforming the scores on each test to a common scale with a specified mean and standard deviation. This new scale would then enable the transformed scores of different tests to be compared directly.

14.2.1 Standard Scores (Z Scores)

A procedure is to convert the original scores to new scores called Z scores or standard scores, with a mean of 0 and a standard deviation of 1. Standard scores have two major advantages. Since the mean is zero, it can be said at a glance whether a given score is above or below average; an above average score is positive and a below average score is negative. Also, since the standard deviation is 1, the numerical size of a standard score indicates how many standard deviations above or below average the score is; a score of one standard deviation above average (i.e. a standard score of +1) would demarcate approximately the top 16% in a normal distribution, while a score of two standard deviations above average (a standard score of +2) would demarcate approximately $2^{1/2}$% in a normal distribution.

You can use this formula to convert a set of scores to standard scores:

Where **Z** = standard score

X = raw data

= mean score

= standard deviation

Converting each of the original test scores to Z scores yield:

	Psy101	*Psy102*	*Psy103*
X	75	65	55
	80	60	50
	10	15	5
Z	_____	_____	_____

The standard scores show at a glance that the student was half standard deviation below the mean in PSY 101, less than half standard deviation above the mean in PSY 102, and one standard deviation above the mean in PSY 103. When raw scores are transformed into Z scores, the shape of the distribution remains the same. The Z scores give an accurate picture of the standing of each score relative to the reference group, no matter where the original scores are measured from or what scale is used. Standard scores are used extensively in the behavioural sciences.

14.2.2 T Scores

Standard scores are a little bit difficult to explain to one not well versed in statistics. Since behavioural scientists try to report test scores to people who are not statistically sophisticated, several alternatives to Z scores have been developed. One of such alternative, called T scores is defined as a set of scores with a mean of 50 and a standard deviation of 10. The T scores are obtained from this formula:

$$T = 10Z + 50$$

For example, the score of 55 in the example is converted to a Z score of +1.00 by using the Z formula. Then T is equal to (10) (+1.00) + 50 = 60.

Since the mean of T scores is 50 it can still be seen at a glance whether a score is above average (it will be greater than 50) or below average (it will be less than 50). Also one can tell how many standard deviations above or below average a score is. For example, a score of 40 is exactly one standard deviation below average (i.e. a Z score of -1.00) since the standard deviation of T scores is 10. But the T score of 60 confirms the Z score of +1.00 which is interpreted as one standard deviation above the mean.

Study Session Summary

Summary

In this Study session you learned that transformed scores allow for relative comparison of an individual's scores across different tasks, as well as relative comparisons of individuals' scores on the same task. You also learnt that the percentile rank of a score is a single number which gives the percent of cases in the specific reference group, scoring at or below that score. Z score or standard score is a set of scores with a mean of 0 and a standard deviation of 1. T score is a set of scores with a mean of 50 and a standard deviation of 10.

Assessment

Assignment

1. A psychologist administered three psychological tests on 10 job applicants and had the following scores:

Applicant	Tests		
	A	B	C
1	12	20	9
2	17	27	7
3	11	13	6
4	18	22	4
5	10	14	6
6	16	21	7
7	14	19	5
8	15	23	9
9	16	19	5
10	13	14	3

In what test did each applicant performed best?

References

Guilford, J.P. and Fruchter, B. (1986). Fundamental statistics in Psychology and Education. London: MacGraw-Hill Book Company.

Howitt, D. and Cramer, D. (1977). An Introduction to Statistics in Psychology: A Complete Guide for Students. London: Prentice Hall.

Welkowitz, J., Ewen, R.B. and Cohen, J. (2002). Introductory Statistics for the Behavioural Sciences. USA: John Wiley & Sons, Inc.

Feedback on SAQs

SAQ 1.1 A layperson's perspective views statistics "as collection of data", while empirical researcher's perspective differs on the bases of "planning and reporting data in line with its specific objective at hand". In grammatical terms however, statistics is "a summary provided after adequate analysis has been made on collected data".

SAQ 1.2 She would be using descriptive statistics.

SAQ 2.1 Your answers probably include the following:

1. Statistics will enable you to understand professional literature like journals
2. Knowledge of statistics will maximize your comprehension of inferences drawn in research reports, and you will be skilled to make genuine inferences from your research.
3. As scientists, it will help you progress in any quantitative scientific pursuit.
4. Statistics will help you to evaluate the strengths and weaknesses inherent in any research, in terms of the techniques adopted by you, the researcher in collecting information and drawing inferences.
5. The knowledge of statistics will help you to make less subjective conclusions.
6. It will enable you to easily make comparisons and relationships via the use of figures which is much clearer than verbal descriptions.

SAQ 3.1 Valid options to the exercise are as follows:

A. population
B. population parameter
C. sample / representative sample
D. random sample
E. inferential statistics

SAQ 4.1 The correct answer is C because the ratio of 3:2 means that if you divide the population into groups of five members (3+2), each group would be made up of 3 males and 2 females. Therefore, to draw a proportionate representative sample of each of the sexes from this population, the desired sample size of 200 similarly needs to be divided into groups of fives (200/5). This gives you 40 groups, each of which is made up of 3 males and 2 females similar to the population from which they have been drawn. Thus, 3 males x 40 groups = 120 males, and 2 females x 40 groups = 80 females.

If you chose A, you would be wrong because 3+2 is not equal to 4. (150:50 is a ratio of 3:1).

If you chose B, you would be wrong. Although 3+2 = 5, you may have wrongly swapped the relative proportions of male : female ratios

If you chose D, you would be wrong. Again, 3+2 is not equal to 4. (50:150 is a ratio of 1:3).

SAQ 4.2 The correct answer is B because the nth term is obtained by dividing the sample frame or population (600) by the desired sample size (200), giving 30.

If you chose A, you would be wrong because you may have assumed that the nth term was the sample size.

If you chose C, you would be wrong. You may have taken the sample frame or population as the nth term.

If you chose D, you would be wrong. You may have subtracted the sample size from the sample frame or population.

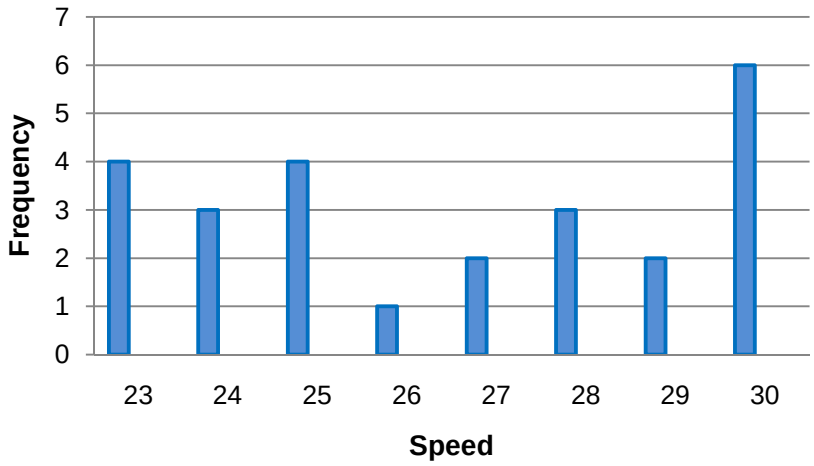
SAQ 5.1 We do not know what you have provided in your table, but it may be filled like as shown below:

Scale of Measurement	Uses of the Scale	Example
Nominal	COUNT the number of things within different categories	Pets: 3 dogs, 12 goats
Ordinal	RANK some things as having more of something than others (but NOT QUANTIFY how much of it they have)	Speed (measured by place of finish in a race): 1st, 2nd, 3rd, etc.
Interval	QUANTIFY how much of something there is but a score of zero does not mean the absence of the thing being measured	Temperature: –2° F, 98° F, 57° F; 0° F is not the absence of heat
Ratio	QUANTIFY how much of something there is and a score of zero means the absence of the thing being measured	Number of text messages sent in a day: 0, 30, 15, etc.

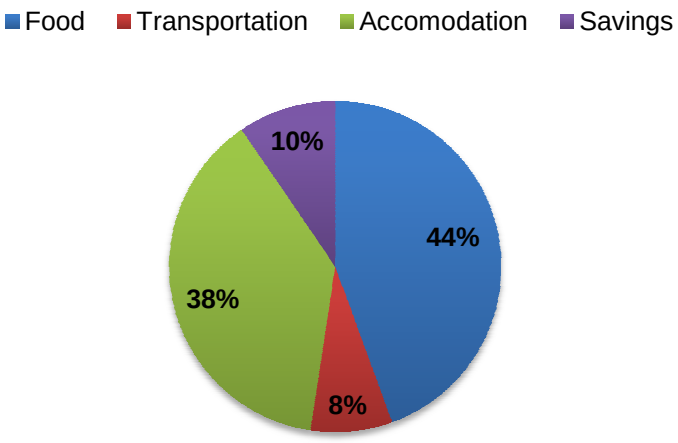
SAQ 6.1 A Frequency Distribution Table

Speeds	Tally	Frequency
23	////	4
24	///	3
25	////	4
26	/	1
27	//	2
28	///	3
29	//	2
30	~~////~~ /	6

SAQ 7.1 Frequency Distribution Bar Graph (Car Speed)



SAQ 8.1 Pie chart showing the monthly spending of Mrs Halima



SAQ 9.1 The answer is (b). The cumulative frequency distribution enables the calculation of the number of cases or people falling at, below, or above given scores in the distribution.

SAQ 13.1 Variance formula is:

$$\frac{\overline{\quad}}{\overline{\quad}} \qquad \frac{\overline{\quad}}{\overline{\quad}}$$

$$\frac{\overline{\quad}}{\quad} \qquad \overline{\quad}$$

And the standard deviation is:

$$\overline{\quad}$$