Practical Approaches to Measurements, Sampling Techniques and Data Analysis

Felix Kutsanedzie
Sylvester Achio
Edmund Ameko

SciencePG
Science Publishing Group

Practical Approaches to Measurements, Sampling Techniques and Data Analysis

Felix Kutsanedzie

Sylvester Achio

Edmund Ameko

SciencePG
Science Publishing Group

Published by

Science Publishing Group

548 Fashion Avenue
New York, NY 10018, U.S.A.

http://www.sciencepublishinggroup.com

ISBN: 978-1-940366-58-6

Dedication

This book is dedicated to parents, relatives and all well wishers of the authors.

Preface

Measurements, sampling and data analysis remain very crucial if not the fulcrum of research activities. There is not a research study for which measurement and analysis is not done. For every research study undertaken, the researcher must have a concept for which he or she needs to find a way of measuring. However concepts are vague ideas and thus must be reduced to variables for them to be measured. It is thus only variables that can be measured for data to be collected as empirical evidence to support a research study being conducted.

Many novices find it difficult about how to reduce their concepts to variables to enable their measurability. This book contains several chapters on plethora of issues very vital to research that have being arranged in a systematic and logical way to aid the understanding of its readers. As if that is not enough, it provides hypothetical data and presents practical approaches for analyzing it with the appropriate statistical tools.

It contains chapters that explain how concepts can be reduced to variables and also help identify the levels of measurement each variable being studied in a research falls into. This is to help readers understand and use data analysis software applications that allow users to identify the levels of measurements of data to be analysed such as SPSS.

Sampling and sampling techniques in research is explained with a practical touch to enable the readers know where and when appropriately to use them to achieve the objectives for a research study. There are varied data collection methods and data collection instruments used in research. These methods and instruments are to be chosen based on the study being undertaken. The book gives a detailed explanation on how to choose the appropriate instrument for a particular study.

The book gives a detailed coverage to topics such as data analysis and tests; transformation of data from that which is not normal distributed to that which is normally distributed; tabular and graphical data summarizing tools; and the researcher's use of descriptive and inferential statistics for data analysis.

It is a complete book with several book chapters vital to understanding research through measurements, sampling and its techniques, data transformation and data analysis in a practical and comprehensive way. The book is intended to equipped lecturers, researchers, tertiary students and all those interested in conducting research studies. It would be of immense help to final year tertiary students conducting their research studies; and writing their research reports.

It is written in a way that can provide self-tutorials to readers and also spare them the ordeal of learning or equipping themselves with the requisite research skills through searching of different books.

The authors of the book acknowledge the support of every individual for their diverse roles played in getting the book published. We are indeed indebt to the Science Publishing Group for quality of their services and their unremitted efforts and drive towards achieving excellence and success.

Contents

Chapter 1

Data Collection Methods and Instrumentation

1 Data Collection Methods and Instrumentation

Felix Kutsanedzie[1]; Sylvester Achio[1]; Edmund Ameko[1]; Selassie Kwasi Diaba[3]; Ofori Victoria[2]

[1]Accra Polytechnic, Accra, Ghana
[2]Agricultural Engineering Department, KNUST, Ghana
[3]Anglican University College of Technology, Nkorazan, Sunyani–Ghana

Abstract

Most students and would-be researchers are not quite familiar with the various methods of data collections used in the design of various studies and the different types of instruments used in each case. They are therefore tempted to use any of the methods they are familiar with notwithstanding their appropriateness for the study intended to be embarked upon. This chapter explains the different data collection methods and the various instruments used in the collection of data as well as determine when a particular instrument must or should be used for the data collection.

Keywords

Instrumentation, Design, Instruments, Data Collection Methods

1.1 Data Collection Methods

There are four main data collection methods used in the research process. These include: observation, survey, contact and experimental methods. These are individually explained below:

1.1.1 Observation Method

The observation method basically involves collecting the needed data for the study via observation only. Since data is only collect through observation of the research in this case, there are some limitations to this method. It is difficult to collect data on people's feelings based on observation. One cannot be precise to conclude that someone is happy when he or she is laughing. It would be quite deceptive; also, the researcher if he or she is the one to do the observation is expected to be present at the place or venue the data is expected to be collected.

There are different types of data collection through observation:

Structured (plan) observation

This is used for descriptive research, where the researcher needs to provide a descriptive view of the situation. It is referred to as structured because the researcher plans about how to describe the situation.

Unstructured (unplanned) observation

This is used for exploratory research. With the unstructured observation, the researcher explores the situation at hand so has no thought out plan of anything in specific to collect data on via observation.

Participatory method

This is used when the researcher is also a participant in the process being observed. Taking for example, a student who wants to study the character of his class in an Institution. Once he or she is part of the class and participates in all activities of the class while taking data through observation, then the method being adopted by such a student is a participatory method of observation. In this case the research has first hand information and experience on what is on the ground.

Non participatory method

With this method, the researcher is not a participant but only present to take data on a particular area of study via observation.

Disguised method

The disguised method is used when the researcher does the observation while disguising him or herself. This method is often used for investigative research.

1.1.2 Survey Method

The survey method involves collecting data on facts or opinions from people through the use of instruments such as questionnaire, opinionnaire and interview guides. Survey method is classified into two: structured and unstructured surveys.

Structured survey method

This involves collection of data using questionnaires or opinionnaire. It is referred to as structured because of the instruments used in the data collection. This is because a questionnaire or opinionaire are set of structured questions

administered to targeted respondent who are required to give their responses for subsequent collection by the researcher.

Unstructured survey method

This method involves the use of an interview guide to collect data via interviewing of targeted respondent. It is referred to as an unstructured method because an interview guide does not contain a set of questions as a questionnaire or an opinionnaire. Instead, it has a list of items that serves as a guide for the interviewer to ask the necessary questions when he or she meets the targeted respondents (interviewees). The interviewer (the researcher) asks questions on item-by-item basis as indicated in the interview guide and can seek clarification on responses provided by the respondents (interviewees) until questions are exhausted on all items indicated in the guide.

1.2 Contact Method

The contact method is where the researcher deploys the instrument that would help collect data using means that can be used to reach the targeted clientele or the respondents. The contact methods include: mailing, telephoning, using face-to-face interview, focus group interview etc.

Mailing

This involves contacting the respondents via the mail where the appropriate data collection tool such as a questionnaire is sent to them for their responses. It has its own advantages and disadvantages. It reduces cost because the researcher does not have to travel to meet the respondents. However it is only respondents who are computer literate, and they are also likely to return their

responses at their own leisure and may not have clarification on some questions for appropriate responses to be provided.

Telephoning

This involves conducting an interview or collecting data via phoning the respondents. In this case the researcher uses the interview guide to interview the targeted respondents for their responses.

Face-to-face Interview

This is where the researcher collects data on the concepts (variable) understudy by conducting one-on-one interview with each respondent using a designed interview guide. With this method, data collection allows the researcher to make clarifications on questions to enable the respondents to give appropriate responses. However the researcher or the interviewer needs to be present to make the needed clarifications for the respondents.

Focus Group Interview

The focus group interview is where the researcher invites experts or the technical individuals from various sectors capable of providing responses to the subject to be discussed or interview to be conducted. Each invited individual or respondent represents a particular group or company. These selected experts are camped at a convenient venue where a forum is held and questions asked by the researcher for the respondents (the experts) to provide responses in an open discussion.

1.3 Experimental Method

This method involves the designing of experiment in either the laboratory or field and measuring data using the appropriate apparatii. The data collected in this case are mostly quantitative unlike human behaviour or attitude which is very difficult to quantify.

1.4 Data Collection Instruments

Instruments or tools for data collection in research are not necessarily the physical ones one can think of when measuring in the physical sciences such as the devices but can be designed forms that are administered to targeted clientele of a population on which a research is being conducted on. Basically the instruments for data collection are broadly grouped into enquiry forms, observation, interview, sociometry and psychological test.

Inquiry forms

These are designed forms which are used to collect data on enquiries made by the researcher. They include the use of questionnaire, checklist, score-card, schedules, rating scale, opinionnaire and attitude scale.

1.4.1 Questionnaire

A questionnaire as earlier explained is a set of questions administered to respondents to solicit their responses. A questionnaire is used for the collection of data on facts and opinions from the respondents.

Normally on questionnaires, before the required questions are asked, a brief on the purpose and the intent of the work or study is given for the perusal of the

respondent. This is preceded by the questions related to the respondent's background and other sets of questions relating to achieving the objectives of the study in a logical sequence.

Basically three different question types can be used in the design of a questionnaire: open ended questions, closed ended questions and the combination of the two. The open ended questions are used in soliciting the respondent opinions and hence not limited by the provision of options to be chosen from. The data obtained from these type of questions are mostly qualitative. Closed ended questions have options provided by the researcher to limit the respondents. This type of question gives rise to quantitative data. However in the design of a questionnaire, both types of questions are employed.

Types of questions to avoid in the design of questionnaires:

Double barreled questions

These are two-in-one questions or compound questions – a single question that demands two answers. The problem with the use of such questions for design of questionnaires is that there is the tendency of the respondent answering one part of the question and leaving the other part, thus instead of two responses, one is provided. Example: What is your occupation and rank? It is good to rather ask: what is your rank? What is your occupation? This will allow the responses to be provided for both questions.

Use of double negatives

When asking questions in a questionnaire, one needs to avoid the use of double negatives as much as possible because it makes understanding difficult and a bit complicated. The researcher should not subject his respondents to such an ordeal of thinking when responding to questions within the questionnaire. If

respondents do not understand the questions, they are likely to provide wrong or misleading responses. Example: Is stealing not uncommon in your area? Instead of asking samply: Is stealing common in your area?

Prestige bias questions

In the design of questionnaire, question that are asked bordering on respondents sensitivity such as age, educational status and health status are referred to as prestige bias questioned. For instance, people generally are not comfortable with asking them questions on how old they are. In asking them about them on their ages, it is likely that they may give responses on ages they are comfortable of associating with, thus falsifying the data which eventually affects the outcome of the study. The best approach to asking such questions is the use of closed ended questions where respondents are given the options in the form of age ranges to select from. The same applies to other sensitive issues such as health status, educational level etc.

Avoid asking leading questions

Normally in the designing of questionnaires, questions are expected to be posed to the respondents in a chronological order: first-question-first. This allows the respondents to answer and follow through easily. Whenever any question is asked that forces the respondent to answer it when he or she is not in the position to response, such a question can be referred to a leading question. For instance, asking a respondent how many times he travels overseas when one has not asked whether the respondent has even travel abroad before can be regarded as a leading question. The same can be said of asking the respondents how many times he or she washes his or her car when no question has established whether the respondent has a car or not. They are questions that force the respondents to give responses that are likely not to be true.

Long winded questions

The questions must be kept simple and short in order to avoid ambiguity or confusing the respondent respondents. Such long sentence questions create confusion.

Length of Questionnaire – Content

The questionnaire must not be overly long but rather must capture precisely questions that will exhaust the content for which it has been designed to ensure the needed data is provided for analysis. When a questionnaire is excessively long, it scares or puts off the respondents who oftentimes have to take the decision between trading of their time for responding to the questionnaire and business schedule. The consequent result is likely to be them not making the time to response to one's questionnaire (opportunity cost).

Checklist

This is an enquiry form which has sets of questions posed at the respondents to whom it is administered for which they are to consent to by checking boxes created against each question posed. It is usually used by producers or service providers to assess the clients' / customers' / consumers' needs or requirement for a product purchased or service provided. In order to develop a checklist and used it as a tool for data collection, one needs to know the various aspects of the product or service requirements to be met so as to check against the statements or questions.

Schedule

A schedule gives the tasks or activities planned for execution with their defined times. Schedule can be used as a tool for data collection by checking

whether activities meet their planned schedule. Thus the researcher would check the schedule against the real situation to generate his or her data.

Score-card

During data collections, sheets are designed on which scores or data are entered for each and individual subject been researched on. All the scores on the sheets for each individual subject under study are then collected on a score-card.

Rating scale

These are scales used in collecting data based on the researchers judgment of an object or a subject, character and situation. This judgment is expressed on a scale of values by the researcher and mostly used for quality assessment. For instance the rating scale indicates the various degrees or amount of different degrees of quality expressed on a linear scale. Thus the assessors have to choose a value that best suites the objects being assessed based on his or her judgment of them. For example the performance of student can be indicated on the scale below:

Excellent	Very good	Good	Credit	Pass	Fail
6	5	4	3	2	0

Rating can be achieved in three main ways: adopting paired comparison; ranking and rating

Attitudinal scale

This is one of the enquiry forms used for data collection based on the researchers judgment on the attitude of the subject under study.

1.4.2 Opinionnaire

This is also a set of question administered to respondents which is intended to collect data that borders on their opinion.

An opinionnaire has the same features as questionnaire but the difference here is in the questions contained in them. They are all intended at soliciting the opinions of respondents. The example below is an idea of how an opinnionaire is analysed:

Table 1.1 *Analysis of a Sample Opinionnaire.*

	SD	D	U	A	SA	
	-2	-1	0	1	2	Averages
clients preference for a service						1.024
surface	1	4	17	33	41	1.135417
air	2	2	14	42	36	1.125
land	1	6	19	54	16	0.8125
Producers preference for supplying the service						-0.333
air	0	0	1	1	6	1.625
surface	3	3	0	2	0	-0.875
land	4	4	0	0	0	-1.5

There is a scale for SD which stands for strongly disagree and assigned a score of (-2); D, disagree with a score of (-1); U, neutral with a score of (o); A, agree with a score of (1); and SA, strongly agree with a score of (2).

The various numbers under each of these classifications in the sample of the opinionnaire are the frequencies of the data collected and based on individuals' opinions or choices of each of the services rendered. The frequencies have been used for the calculation of the averages of the opinions of individuals on the likes and dislikes of the services being provided.

1.4.3 Observation

When observation method is used for data collection, the researcher can collect the data using a designed observation sheet or a checklist, scorecards etc. These are designed by the researcher based on the data type that needs to be collected in order to achieve the objectives of the study.

1.4.4 Interview

For data collection via interviews, the appropriate tool used is the interview guide. The interview guide differ from the questionnaire in that whereas the questionnaire contains the set of questions that the respondents are to respond to, the interview guide does not contain the list of questions but rather the outlines of the information on which an interview would be conducted. For instance on a interview guide can contain some outlines which are expected to cover the scope of the study that is being undertaken. Outlines on the interview guide can look like the following:

- Biodata information of interviewee
- Information on operational procedure of a machine
- Technical knowledge of interviewee on the machine
- Maintenance information of the machine

As can be seen, these are only outlines that would aid the interviewer to ask the interviewee the relevant questions based on the scope of study.

Bibliography

[1] Davies, W. M., Beaumont, T. J. (2007). *Research Proposals*, Teaching and Learning Unit, Faculty of Business and Economics, the University of Melbourone.

[2] Dawson, C. (2002). *Practical Research Methods*, New Delhi, UBS Publishers' Distributors.

[3] Kothari, C. R. (1985). *Research Methodology - Methods and Techniques*, New Delhi, Wiley Eastern Limited.

[4] Kumar, R. (2005). *Research Methodology-A Step-by-Step Guide for Beginners*, (2nd Ed.), Singapore, Pearson Education.

Chapter 2

Determination of Appropriate Sample Size for a Research

2 Determination of Appropriate Sample Size for a Research

Felix Kutsanedzie[1]; Sylvester Achio[1]; Edmund Ameko[1]

[1]Accra Polytechnic, Accra, Ghana

Abstract

In research studies, it is often difficult if not impossible to use the whole population for a study. This is however different when the study covers population that is relatively small that the whole population can be considered for the study. Otherwise it is expensive, sometimes the destructive nature of the study requires that only small fraction of the population referred to as a sample which is used to make inferences about the population understudy. Most often would–be researchers fall short of not considering the appropriate sample size for a study and thus the inferences made about the population are misconstrued. This paper addresses the challenge of giving a comprehensive understanding of how an appropriate size of sample can be taken from a population for a study. It explains the various calculations involved in sizing a sample of a given population size.

Keywords

Sample, Population, Representative, Size, Research

2.1 Introduction

Sample is defined as a fraction of a population taken that is considered for a study in order to make inferences about the population. For appropriate inferences to be made of the population using a selected sample requires that the sample in question should be representative in terms of its composition and size. In this regard the sampling technique to be used for selecting the sample as well as the way the sample size is to be selected from the population are of equal importance in constituting the representative sample. Sample size determination is the act of choosing the number of observations or replicates that should be included in a statistical sample. The sample size to use for a study depends on the data to be collected and the statistics that is required to be derived from it.

Sample sizes to be taken from a population depend on the expediency and availability of data. A sample size can result in wide confidence interval or risks of error in order to increase precision of the data to be collected but in other cases the accuracy for using larger sample sizes cannot be guaranteed because of systematic errors. All these notwithstanding, the law of large numbers and central limit theory underpin the use of large size samples for increasing statistical power of a selected sample.

In selecting the right sample size one needs to understand the concepts of confidence level, confidence interval as well as margin of error. Confidence level is expressed in percentage and it informs the percentage of confidence ascribed to obtaining the true mean of a selected sample being considered for a study. For instance confidence level of 95% or 0.95 means that there is a surety that 95% of the true mean of the sample would be obtained when the study is repeated for 100 times within an interval called the confidence interval. The

confidence interval is the range within which the true mean is expected to be located at the chosen confidence level.

2.2 Determination of Sample Size

In order to determine the sample size, parameters such as the margin of error and the confidence interval need to be considered. Usually when data is collected and the sample mean is calculated, it tends to differ from the population mean, and this difference between the two is termed the margin of error. The margin of error is simply the maximum difference between the observed sample mean ($\bar{x}$) and the true value of the population mean (μ). The margin of error is mathematically expressed as:

$$E = \frac{Z_{\frac{\alpha}{2}}}{(2\sqrt{n})}$$

To determine the margin of error, a confidence level must be chosen and this is often a value less than 100%, however mostly 99%, 95% and 90% is used. However, 99% is mostly used for medical experiments that require higher precision whereas 95% for other situations. Once the confidence level is chosen, say 95%, the α is derived by subtracting the chosen confidence level from 100% and expressing it in units:

$$\propto = 100\% - 95\% = 5\% = 1 - 0.95 = 0.05$$

From the normal distribution diagram, the total area under the curve is equal to 1; that under symmetric half of the curve is equal to 0.5; and the area of $Z_{\frac{\alpha}{2}}$ both to the left or right coloured red is $Z_{\frac{5}{2}} = 0.025$. The region of $Z_{\frac{\alpha}{2}}$ to the left and $Z_{\frac{\alpha}{2}}$ to the right of $Z = 0$ is therefore equal to:

$$0.5 - 0.025 = 0.475$$

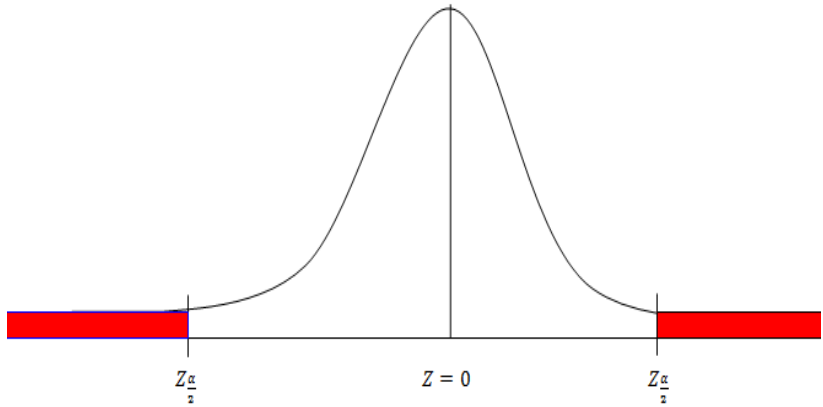


Figure 2.1 *Illustration* $\propto$ *on a normal curve.*

The critical Z value corresponding the area 0.475 under the normal distribution table is 1.9 located vertically plus 0.06 located horizontally (1.9 + 0.06 = 1.96) as show in the table below:

Table 2.1 *Areas for a Standard Normal Distribution.*

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857

Thus taking a sample size of 100, this can be substituted into the equation to obtain the margin of error as below:

$$E = \frac{1.96}{(2\sqrt{100})}$$

$$E = \frac{1.96}{(2)(10)}$$

$$E = \frac{1.96}{20}$$

$$E = 0.098 = 0.1 \; or \; 10\%$$

$$E = \pm 0.1 \; or \; \pm 10$$

$$Z_{\frac{\alpha}{2}} = 1.96, n = 100$$

However, if the margin of error is estimated at the beginning of the study as 5%, this same formula can be used to calculate the appropriate sample size to work with by making the n the subject of the equation as below:

$$n = (\frac{Z_{\frac{\propto}{2}}}{2E})^2 = (\frac{1.96}{2(0.05)})^2 = (\frac{1.96}{0.1})^2 = (19.6)^2 = 384.16 \cong 385$$

The formula $n = (\frac{Z_{\frac{\propto}{2}} \; x \; \partial}{2E})^2$, can be used when the population standard deviation is known. This might be obtained from other studies or a pilot test conducted or else might not be available at all.

For an identically and independently distributed data with a population variance of σ^2, the formula below is used in estimating the sample size required:

$$n = \frac{16\sigma^2}{W^2}$$

n=sample size, σ^2 =population variance, W=width of confidence interval (Wald's unit)

For estimating a required sample size for survey data collection, the formula below can be used:

$$n = \frac{4}{W^2} = \frac{1}{B^2}$$

n=sample size, W= width of confidence interval, B=error bound on estimate (usually given as $\pm B$ and in precentage)

Also the Mead's resource equation is used for estimating the required size of laboratory animals to be used in an experiment as well as other types of resources with regards to experimental work. Its used for estimating the sample size that might be as accurate as the other methods but is very helpful where the standard deviations or differences between values of groups being considered for an experiment are difficult or hard to estimate. Mead's resource equation is given as:

$$E = N - B - T$$

E=degree of freedom of the error component which should range between 10 and 20, N=the total number of individuals or units in the study minus 1 or df of sample size, T=the number of treatment groups including the control or the number of questions asked minus 1 or df of treatment, B=the blocking component minus 1 or df of block, however when there is no stratification B=0.

Let assume an experiment needs to be conducted where forty (40) animals of eight (8) animals each put in five (5) treatment groups. This can be worked out as follows:

$$E = (40 - 1) - 0 - (5 - 1)$$

$$E = 39 - 0 - 4$$

$$E = 35$$

The value 35 exceeds the expected range of E which is between 10 and 20. It presupposes that eight (8) is not the right sample size. Assuming 2 animals per group of 20 is considered, and then E becomes:

$$E = (40 - 1) - 0 - (20 - 1)$$

$$E = 39 - 19$$

$$E = 10$$

Therefore the required sample size can be 20.

2.3 Determination of Sample Size with Levels of Significance

Another way of determining the appropriate size of sample to be selected from a known population size for a research study is to use the levels of significance as per the formula:

$$n = \frac{N}{1 + N(level\ of\ sgf.)^2}$$

where n=sample size, N=population size, level of sgf.=level of significance in units (5%=0.05, 1%=0.01)

Thus using the formula above, and considering a population size of 200 and a level of significance of 5%, the sample size to be used can be calculated as follows:

$$n = \frac{20}{1 + 20(0.05)^2}$$

$$n = \frac{20}{1 + 20(0.0025)}$$

$$n = \frac{20}{1 + 0.5} = \frac{20}{1.5} = 19.04 \cong 19$$

When above formula is used a table below can be obtained for the various populations sizes and their respective sample sizes

Table 2.2 *Population Sizes and their calculated Sample Sizes.*

Population Size	Sample Size
20	19
40	36
60	52
80	66
100	80
150	108
200	132

Bibliography

[1] Ary, D., Jacobs, L. C., Razavieh, A. (1996). *Introduction to research in education*. Fort Worth, TX: Harcourt Brace College Publishers.

[2] Browner, W. S., Newman, T. B. (1998). Sample size and power based on the population attributable fraction. Am J Public Health, 79(9): 1289-94.

[3] Castelloe, J. (2000), "Sample Size Computations and Power Analysis with the SAS System," Paper 265-25 in *Proceedings of the Twenty-Fifth Annual SAS User's Group International Conference*, Cary, NC: SAS Institute, Inc.

[4] Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), New York: Academic Press, New York.

[5] Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.

[6] Desu, M. M., and Raghavarao, D. (1990). *Sample Size Methodology*, Boston: Academic Press.

[7] Donald, M. N. (1967). Implications of non-response for the interpretation of mail questionnaire data. *Public Opinion Quarterly, 24*(1), 99-114.

[8] Fink, A. (1995). *The survey handbook*. Thousand Oaks, CA: Sage Publications.

[9] Freiman, J. A., Chalmers, T. C., Smith, H. Jr., Kuebler, R. R. (1978). The importance of beta, the type II error and sample size in the design and

interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med., 299(13): 690-4.

[10] Hagbert, E. C. (1968). Validity of questionnaire data: Reported and observed attendance in an adult education program. *Public Opinion Quarterly*, *25*: 453-456.

[11] Hair, J., Anderson, R., Tatham, R., Black, W. (1995). *Multivariate data analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

[12] Halinski, R. S. & Feldt, L. S. (1970). The selection of variables in multiple regression analyses. *Journal of Educational Measurement*, *7*(3), 151-158.

[13] Holton, E. H., Burnett, M. B. (1997). Qualitative research methods. In R. A. Swanson, & E. F.

[14] Houston, W. J. (1983). The analysis of errors in orthodontic measurements. Am J Orthod., 83(5): 382-90.

Chapter 3

Graphical and Tabular Statistical Data Summarizing Tools

3 Graphical and Tabular Statistical Data Summarizing Tools

Felix Kutsanedzie[1]; Sylvester Achio[1];

Selassie Kwasi Diaba[2]

[1]Accra Polytechnic, Accra, Ghana
[2]Anglican University College of Technology, Nkorazan, Sunyani–Ghana

Abstract

Data collected in every experiment or survey that has been conducted must be summarized using the appropriate statistical tool for the purpose of making a good presentation of the findings by the researcher. This when rightly done would ensure that the right interpretation is given to the data for the information to be communicated easily by the researcher to his targeted audience in the scientific community. This chapter seeks to explain the various tools used in summarizing qualitative and quantitative data collected during an experiment or a survey study.

Keywords

Tables, Qualitative, Quantitative, Data, Graphs

3.1 Concepts

Concepts are abstractions which are mental representations of real and agreed upon phenomena based on experience. Examples of mental representations of real phenomena include buildings, cars, animals; and those of agreed-upon phenomena – marriage, justice, trust, faith etc.

3.2 Constructs

Constructs are also abstractions which are theoretical creations of observations which cannot be seen directly or indirectly. Examples include intelligence, leisure, taste, etc. The constructs are almost the same as concepts which are agreed-upon phenomena.

Since concepts and constructs are abstractions or ideas that are within the mind of the researcher, they remain nebulous or vague to others until they are reduced to measurable units referred to as variables before others can appreciate and understand them the same way as the researcher.

3.3 Measurement and Variables

Measurement is a symbol assigned to a phenomenon or variable to give / denote / represent / describe it with a value. Thus the measurable label of a construct or a concept or phenomenon is referred to as a variable. When a variable is measured, its value is obtained. In terms of measurements, all things that can be measured are classified into three:

- Direct observables: all items that can be seen – length, volume, weight etc.

- Indirect observables: data from questionnaires - age, gender, income, marital status, religion etc.

- Constructs: variables that based on theoretical creations which cannot be seen directly or indirectly – leisure, pleasure, taste, love, etc.

To enable variables to be measured, there must be *theoretical definitions of measurement of the variable,* i.e. measurement of the variable as defined in books or in literature; and *operational definition of measurement of the variable*, i.e. how the researcher intents to measure the variable based on the work being carried out.

It should also be noted that concepts or constructs (phenomena) must be free to vary if they are to be measured as *variables* or else are regarded as *constants.*

There are four ways by which all variables can be measured. These include *"metering" (using devices), counting, ordering (ranking or referencing) and classifying (grouping without ranking).*

With all research studies excluding the descriptive ones, there are always two categories of variables that must be identified – *independent variables (IV)* and the *dependent variables (DV)*. Once these two variables are established in a research, one can posit a cause-effect relationship. The cause – effect relationship implies that a change in the dependent variable causes a change in the independent variables. However if all the variables in the research are all independent variables, one cannot posit this cause-effect relationship in the study, but the other variables outside the ambit of the research may depend on the independent variables. Independent variables are classified into Active and Attribute independent variables. The Active Independent variables are those whose levels are controlled or manipulated by the researcher whereas the Attribute Independent variables are those whose levels are controlled by the

subjects involved in the research (natural characteristics such as sex, age, race etc.). There are also Intervening variables (extraneous variables) which involve uncontrolled variables that may account for changes in the dependent Variables (DV). Controlled variables are variables that are controlled and statistically measured and accounted for in order not to affect the DV.

3.4 Levels of Measurement

Having understood what variables are and the different types, it becomes necessary to know the levels of measurement that can be applied to them to enable the use of the appropriate statistical tools for their measurements. There are basically four levels of measurement into which every variable falls. These include nominal (classificatory), ordinal (ranking), interval, and ratio which together are represented by the acronym NOIR.

Nominal Level (classificatory scale)

This level of measurement applies to all variables that are mutually exclusive and collectively exhaustive – which implies that it puts all items into one and only one group at a time and once an item is captured in a group it is excluded from belonging to another group at the same time and thus the chance of being found in a different group at that same time exhausted. It is primarily used for classifying hence the name classificatory scale). Assuming gender is being considered as a variable, all items are to be considered or grouped into either a male or female and not both (hermaphrodite) at the same time. Again considering a variable like religion – Christianity, Islam, pagan; race – African, Asian, American etc.; marital status – single, married; body colour- dark, white, brown, yellow etc. It should however be noted the nominal level of measurement is used in only classifying and not ranking and they make no mathematical sense or cannot be

subjected to any mathematical operation to make sense. All variables that this level of measurement applies to are considered qualitative except when there is the issue of coding where they are given codes in order to be counted.

Ordinal (Ranking Scale) Level of Measurement

The ordinal levels of measurement of variables are similar to those of the nominal or classificatory scale but the distinctive feature that distinguishes the two is the fact that the ordinal level involves ranking or ordering of items under consideration. Example is grading of students into: first class, second class, and third class. It thus not only put students into classes but also ranks them. More examples can be given as taste of a sumptuous meal – palatable, more palatable, most palatable, classifying types of shoes in terms of their quality – good, better, best. Range of Students Marks such as 31 – 40, 41 – 50, 51 – 60. The ordinal level of measurement also makes no mathematical sense or cannot be subjected to any mathematical operations but rather used in ordering items. Where ordinal levels of measurement are applied to a variable, the data collected are mostly qualitative but can be quantitative when values of variables are coded.

Interval Level of Measurement

This is applied to variables that are ordered based on equal intervals. They make mathematical sense or can be subjected to mathematical operations such as addition and subtraction. It has no absolute zero value. Which means that it has no true zero value and any zero value assign to a variable based on the interval level of measurement is arbitrary. For example, the fact that a Celsius thermometer is used to measure the temperature of a body to be 0℃ does not imply that there body has no heat in it but rather the instrument used for measuring the temperature could not pick up the latent heat (hidden) of the body. Therefore should a different thermometer scale such as the Kelvin or the thermodynamic

scale be used the value would not be recorded. It thus implies that the zero (0) value recorded on the Celsius scale is arbitrary or not the true zero. However every unit on the on Celsius thermometer is equal in value and the differences between temperatures are equal and mathematically meaningful irrespective of the scale used - (20 ℃ - 15 ℃) = (293.15K – 288.15K) = 20. Thus examples of variables that the interval level of measurement can be subjected to include temperature, intelligence quotient (IQ) etc. Interval measurements can assume negavtive values.

Ratio Level of Measurement

This is applied to all variables that can be subjected to all mathematical operations such as addition, subtraction, multiplication, division etc. to make sense. They all have true zero values or true absolute values. It has equal units and can also be ordered. Examples include weight, height, amount of money, distance, volume etc.

3.5 Problems Associated with Data Collected

When variables are measured or responses are obtained – that is to say data is collected. It must be noted that there are problems associated with data collected in research studies with regards to their reliability and validity. The data collected in a study can be classified into qualitative and quantitative. However since qualitative data is mostly classificatory, validity and reliability is less of concern to qualitative data. The same cannot be said of quantitative data.

Validity of data refers to the extent or the degree to which an instrument used in measuring or collecting data on a concept reduced to variable would actually measure to exactness. It actually refers to the accuracy of the measurement. Thus validity of data depends on whether one is measuring the intended concept rightly

or using the right instrument. If the right instrument is not being used, then the data can not be validated.

Reliability of data refers to the measure of variations in data collected on a particular concept (reduced to variable) over a repeated number of trials. It is thus synonymous to consistency of the collected data. Reliability can be systematic – a change in a subject value overtime or repeated trials or random – which results from sampling error.

3.6 Measure of Reliability of Collected Data

Retest Correlation

To measure reliability of the data collected depends on the data type. The retest correlation involves using the data collected over repeated trials for the same variable (concept) and then finding the correlation between the first collected data and the subsequent. If the correlation value is 1, it means it is a perfect correlation hence the more consistent the data obtained, if the correlation reduces towards zero, it means the data is less consistent.

Kappa coefficient

It is used to measure the reliability of nominal variables over a repeated collection of data. It values range from – 1 to 1.

For the measurement of the reliability of variables of average scores of two or more items on multi questionnaire or inventory, and the alpha reliability is used.

Measurement error

During a study or an experiment, errors emanate in measurement due to sampling, subjects or experimenter or researcher effects as well as on occasions

where there is reduced reliability and validity. Errors due to sampling arise when the right sampling techniques are not deployed. There are also errors due to the subjects or experimental materials and researcher or experimenter effects which results in bias. Last but not the least; errors arise when there is reduced reliability and validity.

3.7 Tools for Summarizing Data Collected

Table 3.1 *Data Summarizing Tools.*

			Frequency distribution
			Relative frequency distribution
			Percent frequency distribution
		Tabular Method	*Crosstabulation*
	Qualitative Data		
		Graphical Method	*Bar graph*
			Pie chart
Data			
			Frequency distribution
			Relative frequency distribution
			Cumulative frequency distribution
			Crosstabulation
			Cumulative Relative frequency distribution
			Stem-and-leaf display
		Tabular Method	
	Quantitative Data		
		Graphical Method	*Dot plot*
			Histogram
			Ogive
			Scatter
			Diagram

Qualitative Data Summarizing Tools

The data collected from research is basically classified into twofold: qualitative and quantitative data. For both types of data, the researcher must be able to identify the type of data collected – whether qualitative or quantitative before the right statistical tools can be deployed for summarizing and analyzing them.

3.8 Tabular Method

Frequency Distribution

The frequency table is used for summarizing both qualitative and quantitative data. It is a descriptive statistical tool used in counting the number of times a particular data (responses) is obtained for a measured variable. Applications like Excel and SPSS can be used to find the frequency distribution of data collected in order to summarize it. Taking for example data collected on the colours of different cars observed on a road during a defined period as follows: green, green, red, green, yellow, white, black, white, yellow and red. The frequency distribution table for the data above is shown below:

Table 3.2 *Frequency Table.*

Colour of Cars	Frequency(f)
Yellow	2
Red	2
White	2
Green	3
Black	1
Total	$\sum f = 10$

3.9 Relative Frequency Distribution

The relative frequency distribution table is a type of frequency distribution which is expressed as unity or where the total frequency is equal to 1. Thus it is the ratio of the frequency of each data categories to the total frequency.

Taking the same data on the colour of cars, the relative frequency table for the data is drawn below:

Table 3.3 *Relative Frequency Table.*

Colour of Cars	Frequency(f)	Relative frequency (Rfq)
Yellow	2	$^{2}/_{10} = {}^{1}/_{5} = 0.2$
Red	2	$^{2}/_{10} = {}^{1}/_{5} = 0.2$
White	2	$^{2}/_{10} = {}^{1}/_{5} = 0.2$
Green	3	$^{3}/_{10} = 0.3$
Black	1	$^{1}/_{10} = 0.1$
Total	∑f =10	∑Rfq = 1

Percent Frequency Distribution

This is a frequency distribution table in which the frequencies of responses or data collected are expressed as percentage. Using the same data on the colours of cars, the percentage frequency table is shown below:

Table 3.4 *Percentage Frequency Table.*

Colour of Cars	Frequency(f)	Percent frequency (%fq)
Yellow	2	$^{2}/_{10}\,x\,100 = {}^{1}/_{5} = 20$
Red	2	$^{2}/_{10}\,x\,100 = {}^{1}/_{5} = 20$
White	2	$^{2}/_{10}\,x100 = {}^{1}/_{5} = 20$
Green	3	$^{3}/_{10}\,x\,100 = 30$
Black	1	$^{1}/_{10}\,x\,100 = 10$
Total	∑f =10	∑%fq = 100

Crosstabulation

Cross tabulation is a tabular summary of data on two variables. The data on one variable is put in the column and other in a row. It helps to describe the relationship between the variables. Assuming one needs to draw a crosstabulation on Mechanical Engineering students in Accra Polytechnic with the following data: first year class – 70 good, 30 poor; second year class – 50 good, 40 poor; and third year class – 10 good, 90 poor. A cross tabulation of this data can be done as shown below:

Table 3.5 *Crosstabulation of Data on two variables.*

	Mechanical Engineering Students		
Performance Rating	**Year 1**	**Year 2**	**Year 3**
Good	70	50	10
Poor	30	40	90
Total	100	90	100

3.10 Qualitative Data

Graphical Method

Bar Graph

For summarizing of qualitative data, the bar graph or chart is used instead of the histogram. This is because the frequency of the data is represented on the vertical axis and on the horizontal axis the label of data items whose frequencies are indicated on the vertical axis. The items on the horizontal axis are descriptive in nature because the axis gives no indication of measurement in terms of counting or quantifying but rather classifying. This implies that the tool is used purely or purposely for describing a data set and not to establish quantity. It is distinguished from histogram by the space left between the various drawn

bars. Hence it is not appropriate tool for summarizing quantitative data. A bar chart of the data on colour of cars is represented below:

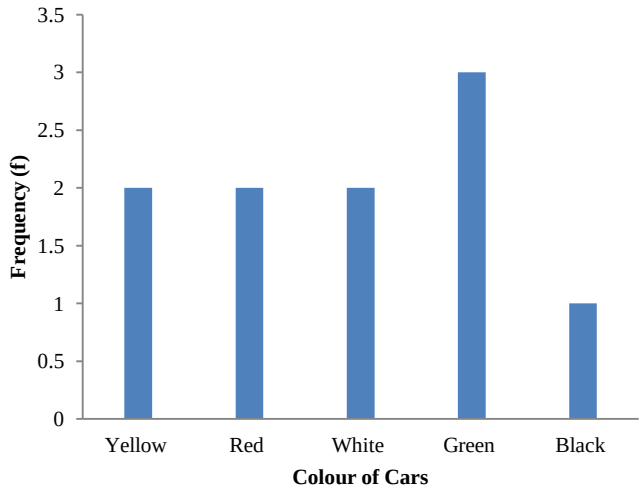


Figure 3.1 *Colour of Cars Observed on a Road within a Period.*

Pie Chart

The pie chart is used for summarizing qualitative. It involves the frequency of the data in a circle - distributing the data according the total angle distribution within a circle which is 360°. It can be expressed in the degrees or degrees subsequently converted into percentage as seen in the related table. The data on the colour of cars can be represented with a pie chart as shown below:

Table 3.6 *Table on the conversion of frequencies into degrees for drawing of a pie chart.*

Colour of Cars	Frequency	Expressing data in degrees and percentage	
Yellow	2	$\frac{2}{10}\ x\ 360° = 72°,$	$\frac{72°}{360°}\ x\ 100 = 20\%$
Red	2	$\frac{2}{10}\ x\ 360° = 72°,$	$\frac{72°}{360°}\ x\ 100 = 20\%$
White	2	$\frac{2}{10}\ x\ 360° = 72°,$	$\frac{72°}{360°}\ x\ 100 = 20\%$
Green	3	$\frac{3}{10}\ x\ 360° = 108°,$	$\frac{108°}{360°}\ x\ 100 = 30\%$
Black	1	$\frac{1}{10}\ x\ 360° = 36°,$	$\frac{36°}{360°}\ x\ 100 = 10\%$
Total	10	360°	100%

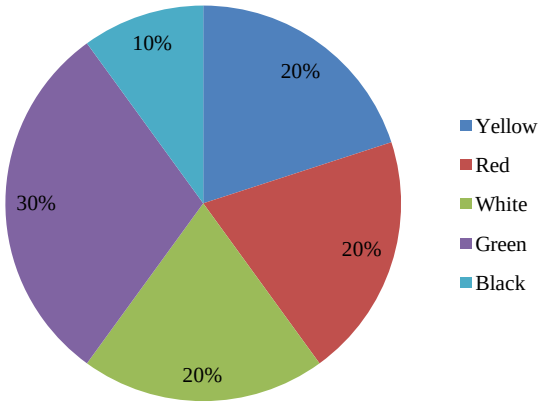


Figure 3.2 *Pie Chart.*

Quantitative Data Summarizing Tools

Tabular Tools

Frequency Distribution

With summarizing quantitative data, the frequency distribution is one of the tools that can be used. Taking for example count of cars that have covered measured distances within a specific time. The frequency distributions – the

frequency tables such as relative frequency table and the percentage frequency table can be drawn as follows:

Table 3.7.1 *Frequency distribution table.*

Distances Covered	Number of cars (frequency)
10	2
20	2
30	2
40	3
50	1
	$\sum f = 10$

Table 3.7.2 *Relative frequency distribution table.*

Distances Covered	Number of cars (frequency)	Relative frequency
10	2	$\frac{2}{10} = 0.2$
20	2	$\frac{2}{10} = 0.2$
30	2	$\frac{2}{10} = 0.2$
40	3	$\frac{3}{10} = 0.3$
50	1	$\frac{1}{10} = 0.1$
	$\sum f = 10$	

Table 3.7.3 *Percent frequency distribution table.*

Distances Covered	Number of cars (frequency)	Percent frequency
10	2	$\frac{2}{10} x\ 100 = 20$
20	2	$\frac{2}{10} x\ 100 = 20$
30	2	$\frac{2}{10} x\ 100 = 20$
40	3	$\frac{3}{10} x\ 100 = 30$
50	1	$\frac{1}{10} x\ 100 = 10$
	$\sum f = 10$	

Cumulative frequency distributions

The cumulative frequency distribution such as cumulative relative frequency and cumulative percentage frequency are used in summarizing quantitative data. This is because cumulative frequency is found by adding the frequencies preceding classes or group to those succeeding classes or groups. Since it is only quantitative data that can be added to make mathematical sense, it presupposes that cumulative frequency table must be a quantitative tool. The cumulative frequency distribution tables of the data above are drawn as follows:

Table 3.8 *Cumulative frequency distribution table.*

Distance	Frequency	Cumulative frequency	Cumulative relative frequency	Cumulative Percentage frequency
10	2	2	0.2	20
20	2	4	0.4	40
30	2	6	0.6	60
40	3	9	0.9	90
50	1	10	1.0	100
	∑f = 10	∑Cf = 10	∑Rf = 1	∑%f = 100

Cross tabulation

Cross tabulation can also be used to summarize quantitative data. Taking for instance the data on two car brands, X and Y used to cover a distance with their respective fuel consumptions, a cross table can be drawn below:

Table 3.8.1 *Crosstabulation table on cars and the distances covered.*

Distance Covered / Km	Car Brands and their Fuel Consumption in Km/hr	
	X	Y
10	5	6
20	8	9
30	10	15
40	15	20
50	18	30

3.11 Graphical Tools

3.11.1 Stem and Leaf Display

This is a quantitative data summarizing tool where the data or values of the collected variables are arranged or displayed in a pattern mimicking a stem and it leaves to enable easy summarization of the data for analysis. The values or numbers are separated into stems and leaves, where the 'stems' are first identified for each value, and then the 'leaves' of the value identified and placed close to their respective identified 'stems'. For example, taking the following numbers as values of variables within a data set obtained for a study: 52, 54, 53, 61, 65, 93, 98, 102, 44, 48, 31. The stem and leaf display of this data set can be represented as below:

Stem	Leaf		
3	1		
4	4	8	
5	2	3	4
6	1	5	
9	8		
10	2		

Figure 3.3 *Stem and leaf display.*

3.11.2 Dot Plot

This is one of the graphical tools used in summarizing quantitative data. It is a graph where dots are used to represent values on top of a single graded or calibrated horizontal line or axis. Taking for instance the data set: 5, 6, 6, 7, 8, 10, 10, 10, 12, 13, 15, 15, 6, 6, 7, 1, 1, 1; the dot plot for such a data set is drawn as fig 3.4.

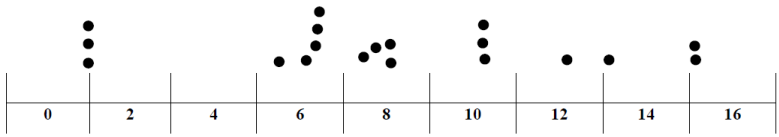


Figure 3.4 *Dot Plot.*

3.11.3 Histogram

The histogram is a tool used for summarizing quantitative data. It looks different from the bar chart in that there are no spaces between its bars and its horizontal and vertical axis are numerical whereas the y –axis of the bar chart always represent the frequency of a data which is numerical but the horizontal axis whether it is numerical or not is only used descriptively. For example data collected on class scores such as: 20, 25, 20, 20, 30, 10, 20, 30, 40, 50, 30 can be represented using the histogram as below:

Table 3.8.2 *Summary of data for drawing of histogram.*

Class Scores	Frequency
10	1
20	4
30	2
40	1
50	1
	$\sum f = 9$

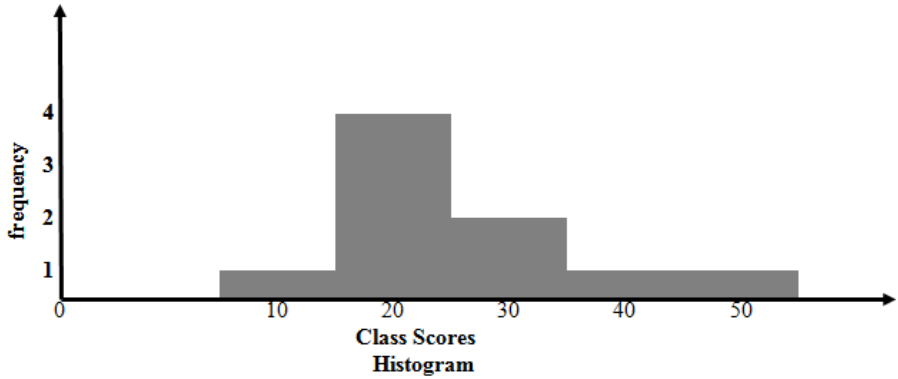


Figure 3.5 *Stem and leaf display.*

3.11.4 Ogive

Ogive is also a tool used for summarizing quantitative data. It is a graph obtained by plotting the cumulative frequency of a data set against its raw scores or marks or values of the data set. The cumulative frequency is always on the y-axis as it depends on the scores or the values of the data set. Taking the data used in the plot of the histogram, the ogive can be drawn as below:

Table 3.8.3 *Table for summarizing scores used for drawing the ogive.*

Class Scores	Frequency
10	1
20	4
30	2
40	1
50	1
	$\sum f = 9$

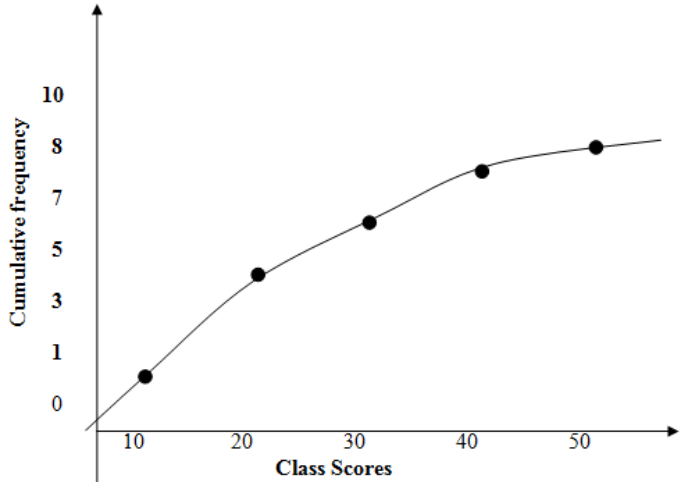


Figure 3.6 *Ogive.*

3.11.5 Scatter Diagram

The scatter diagram is used as a graphical tool for summarizing quantitative data, aiding in finding the relationship between two or more variables. With the scatter diagram there is no line made through the plots. Thus the coordinates are plotted and the pattern used to determine the relationships between the plotted variables on which data have been collected in a study.

Taking for example the data below:

Table 3.8.4 *Data for drawing the relationship between X and Y Coordinates.*

X	Y
1	5
2	10
3	15
4	20
5	25

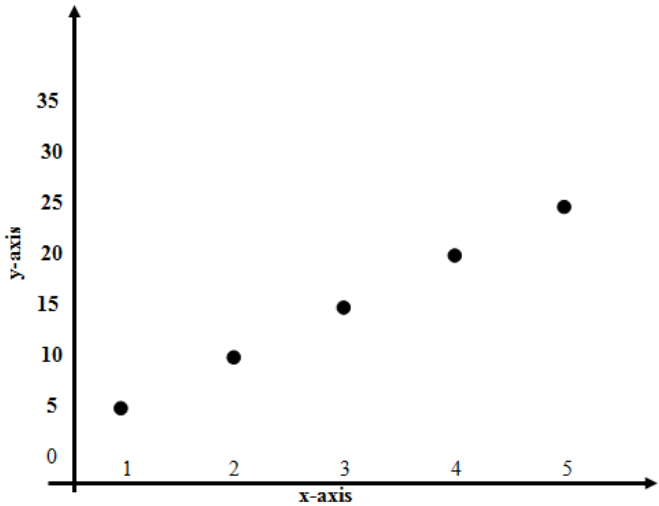


Figure 3.7 *Scatter Diagram.*

Various relationships can be established with scatter diagrams. Below are some of the identified relationships that can be established using the scatter diagram for various collected data sets

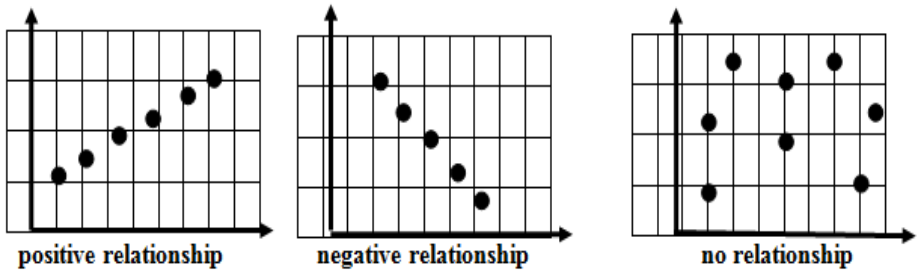


Figure 3.8 *Box and Whisker.*

3.11.6 Box and Whisker

The box and whisker plot is used in summarizing quantitative data. it looks like a number line with a box in the middle divided into three sections, representing the first quartile, second quartile and the third quartile starting from left to right. Also attached to the box are whisker, the tail one to the left

indicating the minimum value and frontal one to the right the maximum value. Taking for instance the data below: 2, 2, 4, 4, 1, 3, 2, 4, 5, 6. The box and whisker can be drawn as follows:

The values are arranged first in ascending order shown below:

1,2,2,2,3,4,4,4,5,6,6,6

$$The\ position\ of\ the\ First\ Quartile\ Mark\ = (\frac{1}{4}\ x\ 12)^{rd} = 3^{rd}$$

Therefore looking at the numbers arranged in an ascending order, the 2 is the mark that occupies the third position.

Therefore the first Quartile Mark = 2

$$The\ position\ of\ the\ Second\ Quartile\ Mark\ = (\frac{2}{4}\ x\ 12)^{th} = 6^{th}$$

Hence from the number arranged in an ascending order the 6^{th} position is occupied by the numbers 4 and 4 counting from both sides. Therefore to find the 6^{th} position. The average of both numbers must be used.

$$Thus\ average\ of\ the\ numbers\ = (\frac{4+4}{2}) = 4$$

Hence the Second Quartile Mark = 4

$$The\ position\ of\ the\ Third\ Quartile\ Mark\ = (\frac{3}{4}\ x\ 12)^{th} = 9^{th}$$

Thus the 9^{th} position is occupied by 5, hence the Third Quartile Mark = 5

From the given data set, the minimum value = 1; and the maximum value = 6

This computed information from the data set can be used to draw the box and whisker plot as shown Figure 3.9:

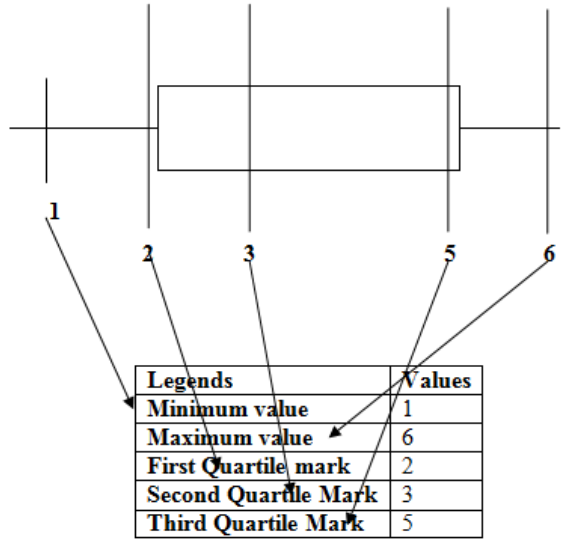


Figure 3.9 *Stem and leaf display.*

Bibliography

[1] Fetterman, D. M. (1989). *Ethnography: Step by Step.* Applied Social Research Methods Series, Vol. 17. Newbury Park, CA: Sage.

[2] Guba, E. G., and Lincoln, Y. S. (1981). *Effective Evaluation*. San Francisco, CA: Jossey-Bass.

[3] Lincoln, Y. S., and Guba, E. G. (1985). *Naturalistic Inquiry*. Beverly Hills, CA: Sage.

[4] Patton, M. Q. (1990). *Qualitative Evaluation and Research Method*, 2nd Ed. Newbury Park, CA: Sage.

[5] Yin, R. K. (1989). *Case Study Research: Design and Method*. Newbury Park, CA: Sage.

[6] Sofaer, S. (1999) Qualitative research methods: what are they and why use them? Special Supplement on Qualitative Methods in Health Services Research, December 1999, Part II. *Health Serv. Res.* 34: 1101-1118.

[7] Shortell SM. The emergence of qualitative methods in health services research. Special Supplement on Qualitative Methods in Health Services Research, December 1999, Part II. *Health Serv. Res.* 34: 1083-1090.

[8] Edmunds, H. (2000). *The Focus Group Research Handbook.* Lincolnwood, IL: NTC Publishing Group.

[9] Templeton, J. F. Focus Group: *A Strategic Guide to Organizing Conducting and Analyzing the Focus Group Interview*. Chicago, IL: The McGraw-Hill Companies 1996.

[10] Krueger, R. A., Casey, M. A. (2000). *Focus Groups: A Practical Guide for Applied Research.* Thousand Oaks, CA: Sage Publications.

[11] Morgan, D. L. (1998). *Focus Groups as Qualitative Research* (2nd edition). Thousand Oaks, CA: Sage Publications.

[12] Strauss, A., Corbin, J. (1998). *Basics of Qualitative Research*. (2nd edition). Thousand Oaks, CA: Sage Publications.

Chapter 4

Sampling and Sampling Techniques

4 Sampling and Sampling Techniques

Felix Kutsanedzie[1]; Sylvester Achio[1];

Ofori Victoria[2]; Goddey Paul[1]

[1]Accra Polytechnic, Accra, Ghana
[2]Agricultural Engineering Department, KNUST, Ghana

Abstract

Sampling is required when a researcher has to do a research on a large population for the purpose of inferring about the population from the sample. However sampling might not be required in all researches especially when the population is small and can easily be handled. In order to take samples from population for a study, sampling techniques are used. The concept of sampling and the applications of the various known sampling techniques are not well understood. This chapter explains with examples the various sampling techniques, and when and how to use them in sampling.

Keywords

Samples, Sampling, Techniques, Population, Research

4.1 Introduction

There are several reasons for which sampling is done. However the paramount of them is to use it to estimate or predict the character of a population. The cost of studying every subject or object is expensive and takes time. Sampling thus reduces these by taking sizeable fraction or portion or part of the population that is most representative of it to conclude on the behaviour and character of it. It is very important to note that some experiments are destructive and if all subjects or objects of the population are to be used for the study in such cases, it would thus point to the whole population being destroyed. In order to avoid these mass destruction, sampling becomes a necessity. Also, it is good to add here and now that sampling is not required in all researches. For instance where the size of the population is too small such that it can easily be handled, there is no need for sampling. Taking for example, where a researcher has chief executive officers of ten financial institutions in a particular country as the targeted clientele for a research, the researcher can study the whole population since it is very small.

4.2 Definition of Sampling, Sampling Techniques, Samples and Population

Sampling can be defined as the act of selecting part or fraction of a population which is most representative of it in order to use it to make inferences about the population. The technique adopted in taking the sample from a population is referred as the *sampling technique.*

A sample thus is defined as the part or fraction of a population selected in order to make inferences about the population while a *population* is the entirety of all the objects or subjects being researched on.

4.3 Sampling Techniques

There are several sampling techniques known and in taking or selecting samples from a population for a study. These are broadly classified into two – probability and non probability sampling techniques. However the type to be used in the selection of a sample depends entirely on the purpose of the study. But generally the non probability types are prone to bias compared to the probability ones as they depend on the whims and caprices of the researcher.

4.4 Probability Sampling Techniques

These types of sampling are based on probability and thus it presupposes that all subjects or objects within the population under consideration are given equal chances of being selected as part of the sample. It thus avoid or reduces bias on the part of the researcher in adding subjects or objects that he or she wished or willed to be selected to constitute a sample. There are different types of the probability sampling techniques which are used in varying situations for the selection of samples from different populations. However it behoves on researcher to know the population under consideration and the appropriate sampling technique to use and how to use them in every situation encountered when conducting a research study. The probability sampling techniques are explained in detailed with applicable examples as follows:

Simple Random Sampling Techniques

This is the commonest and simplest type of probability sampling upon which other types are based. Since it is a probability sampling techniques all the subjects or objects of the population are given equal chances of being selected. In this type of sampling, once the sample sizes are determined from the

population using the appropriate formula as below, then the techniques is used in selecting this size of object from the population:

$$n = \frac{N}{1 + N(level\ of\ sgf.)^2}$$

To select the sample from the population using the simple random sampling, practically all members or objects of the population are assigned numbers on pieces of papers. This pieces of papers on which the numbers are written are shuffled and the required number is picked one after the other from all the papers until the total sample size is obtained. Objects or subjects of the population tagged with or assigned such numbers automatically become objects or subjects of the population selected to constitute the sample for a particular study.

Instead of writing numbers on pieces of papers and assigning it to the members of the population which are then shuffled and picked randomly to select the sample, alternatively a random number table can be used.

Systematic Sampling Technique

It is an example of probability sampling used for taking samples during batch production. This sampling technique requires that with a batch of items being manufactured every nth item is selected or sampled to be tested for quality assurance before the product is passed to be sold out to the ultimate users.

Assuming a car manufacturing company is doing a batch production of 1000 cars per day, and has as their corporate policy to sample 20 cars out of those produced in a batch for quality testing and check. Knowing the population size as 1000 and the sample size as 20, the researcher can use the formula below to select the 20 cars for the sample from the car population of 1000:

$$i^{th} = \frac{N}{n}$$

$$i^{th} = the\ position\ of\ item\ selected$$

$$N = Population$$

$$n = sample\ size$$

$$i^{th} = \frac{1000}{20} = 50$$

Since the i^{th} is 50, it suggest moving within the chain of production, every 50th car in the chain is to be sampled or selected as part of the sample. Thus the 50th, 100th, 150th, 200th ….etc are to be selected until all the 20 cars are selected to constitute the sample.

Taking for instance within a brewery company, 20 cartons of beer is produced and are in a production line as shown within the table with the number of each carton indicated as a right superscript of the cartons represented as letters. Assuming 10 of the cartons are to be selected for the sample, it means that by application of the formula, i.e. $i^{th} = \frac{20}{10} = 2^{nd}$.

Table 4.1 *Labeled and Numbered Cartons in the Production Line.*

^{1}[A]	^{2}[**B**]	^{3}[C]	^{4}[**D**]	^{5}[E]	^{6}[**F**]	^{7}[G]	^{8}[**H**]	^{9}[I]	^{10}[**J**]
^{11}[K]	^{12}[**L**]	^{13}[M]	^{14}[**N**]	^{15}[O]	^{16}[**P**]	^{17}[Q]	^{18}[**R**]	^{19}[S]	^{20}[**T**]

Thus every second carton in the line of production has to be sampled until the 10 cartons are obtained to constitute the sample. Thus from the above calculation, the cartons that need to be selected have been indicated in table 4.2.

Table 4.2 *Selected Cartons in the Production Line.*

^{1}[A]	^{2}[B]	^{3}[C]	^{4}[D]	^{5}[E]	^{6}[F]	^{7}[G]	^{8}[H]	^{9}[I]	^{10}[J]
^{11}[K]	^{12}[L]	^{13}[M]	^{14}[N]	^{15}[O]	^{16}[P]	^{17}[Q]	^{18}[R]	^{19}[S]	^{20}[T]

From the table, [B], [D], [F], [H], [J], [K], [K], [L], [N], [P], [R] and [T] were the cartons selected.

Stratified Sampling Technique

This technique is also a probability technique adopted when the population under consideration is made up of different groups for which if a representative sample should be obtained, all the consisting groups are to be considered for the sampling. Strata refers to groups, hence stratified sampling implies sampling based on groups.

A hypothetical case for which this kind of sampling technique can be deployed is when the behaviour of first degree mechanical degree students within a university are to be studied. If the programme duration is four years, it means the students can be put into groups such as first year, second year, third year and fourth year. Therefore the population can be grouped into four based on the levels of the students and then a number sampled from each of the groups and bulked together to form the representative sample.

Assuming the population of the mechanical students in the university is 400, comprising 100 from each level. The sample size for the population is determined by the appropriate formula and then the number divided into four. Taking for instance that a sample of 100 is to be selected. The 100 would be divided by 4 to get 25, meaning 25 students are to be selected from each of the four groups based on the levels. In order to select the 25 from the hundred students constituting each of the groups, a simple random sampling technique is used. Once the researcher succeeds in selecting the 25 students from each group, all 25 students selected from each group are put together to constitute the sample for the population understudy.

Table 4.3 *Table on how Stratified Sampling is done.*

Description of Population	**Mechanical Engineering Students**			
Size	**400**			
Stratification of Population (putting Members of Population into strata	Strata (Level 1)	Strata (Level 2)	Strata (Level 3)	Strata (Level 4)
Population size of each strata	100	100	100	100
Sample size from each strata using simple random sampling technique	25	25	25	25
Bulking of the sample from each strata to form the representative sample for the taking from the population	=25+25+25+25 100			

The table above illustrates how stratified sampling is done.

Cluster Sampling Technique

It is a probability sampling where samples are selected based on classification into various areas. It is almost handled the same way as stratified sampling but the difference here is that with the latter the population is put into different groups prior to sampling while the former the population is classified into areas prior to sampling.

Taking a hypothetical case where a researcher wants to study the garages located in Accra, because there are many garages scattered over the land of Accra, the researcher would have to demarcate Accra into zones or areas. Thus Accra could be divided into four zones – Northern, Southern, Western and Eastern. Once the population of the garages in Accra is known, the appropriate formula as known can be used to determine the sample size which must be divided into four or based on the number of areas from which samples are to be drawn.

If the number of samples to be taken from each area is established, simple random sampling techniques then is used to sample from each area sample numbers established and then bulked together to form the representative. Thus

assuming the population of garages in Accra is 200 with 50 in each zone and a sample size of 80 is to be taken. The table below shows how cluster sampling can be used to achieve that:

Table 4.4 *Table on how Cluster Sampling is done.*

Description of Population	**Garages in Accra**			
Size	**200**			
Putting Garages of Population into clusters (Areas)	Area 1 (North)	Area 2 (South)	Area 3 (East)	Area 4 (West)
Population size of each Area	50	50	50	50
Sample size from each area using simple random sampling technique	20	20	20	20
Bulking of the sample from each strata to form the representative sample for the taking from the population	=20+20+20+20 80			

Multistage sampling technique

The multistage sampling is applied when the population is large, varied and scattered such that it becomes practically impossible to get a representative sample with sampling only once. This means for a representative sample of the population to be obtained several stages of sampling would be necessary, and due to this the name multistage sampling technique was derived.

Assuming mechanical engineering students in Ghana is a target population that a researcher needs to sample from, the size of the population is first established and the sample size determined by the appropriate formula as given earlier. Since there might be so many universities and polytechnics training engineering students, the first stage of sampling probably might be to sample the identified training institutions by location or areas which results in cluster sampling; the next stage can be sampling within the samples obtained via clustered sampling to obtain samples that consist of students put into year groups or streams in the various institutions i.e. first year, second year, ……..,

final year via stratified sampling, and then the third stage sampled via the use of simple random sampling to obtain the representative sample for the targeted population.

4.5 Non Probability Sampling Techniques

Sampling techniques under this category are all deterministic and thus does not follow any probability. The choice of the samples depend mainly on the purpose for which the researcher is conducting the study; the convenience of the researcher; his judgment, the quota of the sampling of the researcher's choice; etc.

Purpose Sampling

With purposive sampling, the researcher selects the sample from the population based on the purpose for which the study is being conducted. This is non probability because members of the population are not given equal chances of being selected to constitute the sample.

Assuming the researcher's target population is the most intelligent students within Accra Polytechnic, he can choose to sample using the cumulative grade points of the students, and thus selecting those with the highest grade point. Another researcher can say some programmes are more difficult than others and so select students with the highest grade point from such programmes to constitute his or her sample instead of using the whole polytechnic. This way of sampling introduces biases into the research but nevertheless allows the researcher to sample based on the purpose of his or her study.

Convenience sampling

This is also a non probability sampling in which the researcher selects sample from a population based on his or her convenience.

Taking for instance where a researcher stays in Accra and wants to conduct a research on the programmes in the Polytechnics in Ghana. In this case the population under consideration are the Polytechnics in Ghana and since the researcher stays in Accra, the researcher can choose Accra Polytechnic to be part of the sample based on his convenience in terms of its proximity to him or her.

Judgmental Sampling

In this sampling technique the sampling is done based on the judgment of the researcher and therefore not based on probability. The sample is predetermined on the judgment of the researcher.

Assuming a lecturer has to study about the performances of students that are expected to be sent on scholarship to pursue a new programme. Based on the lecturer's knowledge and judgment about the students, he or she can sample students from the population of the students.

Quota Sampling

Quota sampling is non probability sampling. It is where the researcher takes samples from a population based on the researcher's quota. With this sampling technique, unlike stratified sampling if the researcher is to study Mechanical students within Accra Polytechnic, the researcher can choose to take whatever quota from year one, year two and year three without any formula or can decide on even the quota to be used as the sample size.

Voluntary response sampling

This is a technique that involves respondent volunteering to be considered to constitute the sample within a particular targeted population. It is non probability because the members of the sample are selected based on them volunteering.

Assuming a researcher wants to study the environmental conditions prevailing in a particular locality, because of the import of the outcome of the research, members of the locality can willingly volunteer to be part of the sample to be taken from the population to be interviewed on the situation prevailing or issues at stake.

Bibliography

[1] Armitage, P., Berry, G. (1994). *Statistical Methods in Medical Research.* Oxford: Blackwell Scientific Publications.

[2] Bosch, O., Jonge de, E. (2008). "Visualising official statistics", in *Statistical Journal of the IAOS*. Retrieved on 4th June 2015 from available at: http://iospress.metapress.Com/content/v0376364l348/?p=fc2e171758ee4053a01be16bbbae10eb&pi=0.

[3] Few, S. (2004), *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, Oakland CA, Analytics Press.

[4] Fisher, R. A., Yates, F. (1974). *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed. London: Longman.

[5] Harris, R. L. (2000). *Information Graphics*, New York and Oxford, Oxford University Press.

[6] Miller, J. E. (2004). *The Chicago Guide to Writing About Numbers*, Chicago, University of Chicago Press.

[7] Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *Am Statistician*, 55(3): 187-93.

[8] MacFarlane, T. V. (2003). Sample size determination for research projects. J Orthod., 30(2): 99-100.

[9] Phillips, C. (2002). Sample size and power. What is enough? Semin Orthod., 8: 67-76.

[10] Thornley, B., Adams, C. (1998). Content and quality of 2000 controlled trials in schizophrenia over 50 years. *BMJ*, 31(317): 1181-4.

[11] Torgerson, D. J., Miles, J. N. (2007). Simple sample size calculation. *J Eval Clin Pract.*, 13(6): 952-3.

Chapter 5

Use of Random Number Table for Sample Selection

5 Use of Random Number Table for Sample Selection

Felix Kutsanedzie[1]; Sylvester Achio[1]; Edmund Ameko[1]; Mensah Edith[1]

[1]Accra Polytechnic, Accra, Ghana

Abstract

The use of the random number table is unknown by many individuals engaged in research. However selection of elements or members of a sample remain very crucial to the data to be collected. Bias in research is expected to be reduced to the barest minimum if it cannot be outrightly eliminated. Due to this random numbers, for that matter the random table is used to help or aid the selection of members of sample for a given population. The way and manner the table is used in achieving this has been demonstrated in this chapter. Though the tables exist, appropriate Excel Application functions can be used to generate specific random numbers that can be assigned to subjects to help with sample selection. The chapter gives an in-depth knowledge on how the table and the Excel Application can be used to generated random numbers for the purposes of sampling from a given population.

Keywords

Random, Number, Sample, Probability, Population

5.1 Introduction

A random number is that which is based on probability distribution rather than any deterministic approaches or basis. Thus it has no definite plan or formula for its determination. It thus presupposes that random numbers occurrence is based on chances and therefore no manipulation can be done by the researcher in making them appear for selection. This thus satisfies one of the basic characteristics of research that it must be *unbiased.* It is to this end that the use of random number table and it application to the selection of sample from a given population is crucial to the data to be collected and analysed to provide an *empirical evidence* – another characteristics of research.

Use of the random number table

The random number table is a table with various generated numbers which can be used in sampling of objects from a given population given the required sample size. Assuming the sample size is determined as nineteen (19) from a population size of twenty (20), one can assign numbers from 1 to 20 to the objects or subjects within the population. Once this is done, the researcher can open a page of the random number table and close his or her eyes to point any number in the table.

Table 5.1 *Random Number Tables.*

91005	44463	22662	65905
13000	15389	20013	67770
04444	75941	65905	13557
67909	61149	11268	18850
09999	41417	55100	69440
47489	28357	10806	14570
52000	16783	19015	66164
78888	30950	94820	29881
24414	52995	44157	28660

Assuming the researcher opens to the page with the random numbers above, he or she would close the eyes and point to anyone of them. Taking for instance that the researcher points to 65905 as marked red in the table; because the possible numbers to be selected based on the population size under consideration are expected to be from 1 to 20. Therefore only the first two digits of the numbers are to be considered from the table. In applying this principle, the number marked red would be ignored because the two first digits - 65 would not make it fall under the population size of 1 - 20. The researcher can move through the table and then select random numbers that qualify to be selected as indicated in the table with bold fonts and underlined. The researcher is expected to go through the random table till all the first nineteen 19 random numbers are obtained. Once the first nineteen (19) random numbers are obtained, it presupposes that the objects or subjects within the population assigned the numbers obtained from the random table have been selected to constitute the sample.

Use of Excel Application to generate Random Numbers

The Excel application can be used to generate random numbers that can be assigned to objects of a population for the purposes of selecting samples. In order to use the random number, one needs to use the appropriate commands to generate the numbers before they are randomly assigned to objects or subjects within the population.

For one to generate the random numbers for using EXCEL Application the following commands or functions can be applied:

Type in a cell: = RAND ( )

This function when typed into the cell and the enter button pressed would generate random numbers with decimals. One can choose to ignore the decimals

and used the numbers or integers. It implies that the researcher is to select the numbers that appear and fall within the domain of the population until the sample size is attained. The subsequent numbers that appear after the sample size is reached or attained can be ignored.

Using the function above would generate the following random numbers when the cells are dragged over a range:

Table 5.2 *Random Numbers Generated.*

0.84119	0.41447	0.96359	0.79984
0.09976	0.29623	0.84550	0.26919
0.64255	0.45713	0.49517	0.58345
0.30392	0.58731	0.73212	0.81216
0.22872	0.49228	0.39392	0.71131
0.56884	0.21029	0.40610	0.01592
0.80367	0.01590	0.66877	0.26521
0.82651	0.76603	0.06450	0.01026
0.67440	0.14069	0.62185	0.20182
0.03624	0.75226	0.28048	0.12603
0.73470	0.21909	0.61547	0.84535

Thus ignoring the decimal points in Table 5.3 would change to:

Table 5.3 *Discarding Decimal Points from Random Numbers.*

84119	41447	96359	79984
09976	29623	84550	26919
64255	45713	49517	58345
30392	58731	73212	81216
22872	49228	39392	71131
56884	21029	40610	01592
80367	01590	66877	26521
82651	76603	06450	01026
67440	14069	62185	20182
03624	75226	28048	12603
73470	21909	61547	84535

Therefore assuming the researcher is selecting a sample size of 10 from a population size that is from one (1) to twenty (20), the subjects assigned the numbers underline in the table from the first to the tenth would be selected to constitute the sample size.

However in order to eliminate the decimal using a function and the number of digits being considered, the following functions are appropriate:

Type into the cell: = INT (RAND()*(10-1) + 1)

When this command is type in a cell and the enter button is pressed, it would churn out an integer and when the cell is drag over a range, it would give random integers ranging from 1and 10. Thus if a sample size is to be selected from a given population, the first random numbers that appear till the sample size is attained or reached are selected. Sample of numbers generated by applying above function is given below:

Table 5.4 *Generating Single Digits Random Numbers.*

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

It should however be noted that the same number cannot be assigned more than once to a subject or an object within a population and also the numbers are assigned prior to using the random numbers to select them. Therefore each and every member of the population has equal chance of being selected and can be selected only once.

Type into the cell: = INT (RAND()*(100-1) +2)

This function when applied in the same way would generate random integers ranging from 1 and 100, eliminating the decimals. Applying this function the following random numbers are churned out with the dragging of the cells:

Table 5.5 *Generating Double Digits Random Numbers.*

16	56	64	54	50
65	51	94	55	22
40	57	92	19	19
80	32	98	60	66
18	29	65	10	77
18	4	35	73	92

Type into the cell: = INT (RAND()*(1000-1) +3)

This function when applied would generate random integers ranging from 1 and 1000.

Table 5.6 *Generating Triple Digits Random Numbers.*

675	536	481	545
507	13	526	404
902	693	889	826
397	917	737	687
166	819	368	207
578	642	190	474
620	396	198	910

Type into the cell: = INT (RAND()*(10000-1)+4)

This function when applied would generate random integers ranging from 1 and 1000.

Table 5.7 *Generating Four Digits Random Numbers.*

6147	9287	8660	401	671
6182	6098	8102	5868	8379
1995	4846	4875	32	8884
1662	7512	1272	3770	2442
1886	8554	9858	8320	8859
6274	8052	5219	4765	6159

5.2 Random Numbers Generation with Population Sizes

Type into the cell: = INT (RAND()*200)

Once the population size is known, it can be used in generating the random numbers such that the random numbers are selected starting from the first till the total number of the sample is reached or attained. Thus assuming the population size is 200, when the function above is entered into a cell and dragged over a range the following random numbers are generated and can be selected from the beginning until the sample size is attained:

Table 5.8 *Generating Random Numbers from using a Sample Size.*

128	175	10	35	7
146	85	113	120	55
46	35	156	63	77
143	52	23	154	59
15	166	182	58	35
159	149	163	159	47
119	117	181	133	105
167	27	189	67	29

5.3 Assigning of Random Numbers to Subjects or Objects within a Population

In order to select a sample from a given population, the sample size determined determined with an appropriate formula demonstrated in the earlier chapter. The subjects or objects are then assigned a number by the researcher. Once this is done, the random numbers can be obtained and those that appear from the first until the number expected in the sample size is attained. Once the random numbers are selected it means subjects or objects within the population with numbers assigned corresponding with the random numbers selected would constitute the sample.

Assuming the researcher is to sample from a given population size of 20 and using the appropriate formula, the sample size must be obtained first.

Considering the members or subjects of the population to be the first 20 English alphabets, the research can assign each of them the numbers from 1 to 20, starting from A to T as shown in the table below:

Table 5.9.1 *Table Illustration of Members of Sample Selected from a given Population.*

[1] [A]	[2] [B]	[3] [C]	[4] [D]	[5] [E]	[6] [F]	[7] [G]	[8] [H]	[9] [I]	[10] [J]
[11] [K]	[12] [L]	[13] [M]	[14] [N]	[15] [O]	[16] [P]	[17] [Q]	[18] [R]	[19] [S]	[20] [T]

From the table the letters represents the subjects or objects within the population from which the sample must be taken. The sample size of 19 was determined using the formula below:

$$n = \frac{N}{1 + N(level\ of\ sgf.)^2}$$

$$n = \frac{20}{1 + 20(0.05)^2}$$

$$n = \frac{20}{1 + 20(0.0025)}$$

$$n = \frac{20}{1 + (0.05)}$$

$$n = \frac{20}{(1 + 0.05)} = \frac{20}{1.05} = 19.04 \cong 19$$

Using the random number table, the researcher can open and point to any number at all on it as a starting point for selection. Now the eyes can be opened to check whether the number pointed falls or does not fall within the population assigned numbers. The researcher can select the number if it falls within the assigned numbers of the population or otherwise ignore it and move to the next number until the total number for the sample is selected.

When the Excel Application is to be used instead of the random table, the function: = INT(RAND()*(20) can be typed into an EXCEL cell with the enter button pressed and then cell dragged until the sample size is obtained. When this is done random numbers as found below would be generated:

Table 5.9.2 *Generated Random Numbers to be used for Sample Selection.*

4	7	1	19	5
8	11	4	12	14
7	3	4	11	1
18	14	12	12	2
7	12	6	1	14
17	17	9	1	5
5	18	0	16	16
10	9	3	11	18
0	15	4	8	18
19	13	12	13	14
7	16	12	0	3
1	5	19	3	6

Form these random numbers the numbers underlined are those that have been selected. It should be noted that each of the random numbers can be selected once and the rest of the numbers that are same ignored once the first of its kind is selected. The random numbers selected are then corresponded to the numbers assigned the subjects or objects within the population and then practically selected to constitute the sample to be study as per the given population understudy. Thus the objects or subjects selected within the population are indicated in the table in bold font. Within the table only subject represented with letter [J] that had not been selected.

Table 5.9.3 *Members Selected to Constitute Sample.*

[1] [A]	[2] [B]	[3] [C]	[4] [D]	[5] [E]	[6] [F]	[7] [G]	[8] [H]	[9] [I]	[10] [J]
[11] [K]	[12] [L]	[13] [M]	[14] [N]	[15] [O]	[16] [P]	[17] [Q]	[18] [R]	[19] [S]	[20] [T]

Bibliography

[1] Altman, D. G. (1991). *Practical Statistics for Medical Research.* London: Chapman & Hall.

[2] Armitage, P., Berry, G. (1994). *Statistical Methods in Medical Research.* Oxford: Blackwell Scientific Publications.

[3] Campbell, M. J., Machin, D. (1993). *Medical Statistics: A Commonsense Approach.*2nd ed. Chichester: John Wiley.

[4] Fisher, R. A., Yates, F. (1974). *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed. London: Longman.

[5] Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *Am Statistician,* 55(3): 187-93.

[6] MacFarlane, T. V. (2003). Sample size determination for research projects. *J Orthod.,* 30(2): 99-100.

[7] Phillips, C. (2002). Sample size and power. What is enough? Semin Orthod., 8: 67-76.

[8] Strike, P. W. (1991). *Measurement and control, Statistical Methods in Laboratory Medicine.* Oxford: Butterworth-Heinemann.

Chapter 6

Research Errors Encountered in Data Handling

6 Research Errors Encountered in Data Handling

Felix Kutsanedzie[1]; Ofori Victoria[2]; Selassie Kwasi Diaba[3]; Anthony Deku[1]

[1]Accra Polytechnic, Accra, Ghana
[2]Agricultural Engineering Department, KNUST, Ghana
[3]Anglican University College of Technology, Nkorazan, Sunyani–Ghana

Abstract

The human beings are not perfect hence in every human engagements there are elements of errors that are encountered and must be handled in order that results are not drastically affected. Likewise in carrying out a research, there are so many errors that can be encountered in the taking of measurements which eventually affect the reliability and validity of data collected. Statistics or numbers are used for estimation and the extent to which they would perfectly represent the facts depends on plethora of factors such as the instruments, the influence of the researcher and the environment. This chapter concentrates on the various errors that are encountered by a researcher in the conduct of research and how to note them and the measures to adopt in reducing or treating them such that they do not easily influence the results of findings unduly. This thus prevents false interpretations for consumers of research information, and denting the reputation of the researcher in question.

Keywords

Errors, Scores, Estimation, Interval, Confidence Level

6.1 Introduction

Research is defined as a planned enquiry which utilizes the scientific method to develop and test hypotheses and also find solutions that are generally practicable in addressing the myriads of problems in a society. Therefore the two important purposes of research are to develop and test hypotheses, and to apply findings to practically solve problems.

The developing and testing hypotheses is meant to expand the frontiers of knowledge, thus allowing students, researchers, lecturers and other academics to tap from this knowledge pool. Also by applying research findings to solve problems make the world better for all. However, the knowledge contributed by a researcher should be proven to be working and can stand the test of time. Therefore whatever results of findings the researcher comes out with must be foolproof else it would lead to deceiving all those who tap from it as well as those who apply it, which consequentially can have devastating effects. This is why the data collected must be error tight or if not possible must cater for error management in order to serve the empirical characteristic of research.

6.2 Types of Scores

Usually in the conduct of research, data is collected and the outcome of the research is dependent on the data obtained. However in collecting data, errors can occur through measuring, recording, use of faulty instruments, researcher's influence, etc. Whenever an instrument is used in taking a measurement, the

value obtained is just an estimation of the variable being measured. Thus the taking of such a measurement once and then using it to conclude can result in grave errors being made, hence the need for replication and averaging of values of variables. The following scores in terms of measurements must be explained to enhance understanding of the errors to be considered in this chapter:

True score

The true score is referred to as the actual value of the measurement being taken and it is the value that is expected to be exact estimation of the variable. This cannot easily be obtained by taking only one measurement and it can be a value that lies within a range of measured values of variables. For instance if a researcher is taking the temperature of the body, how will he or she know by taking the reading once that, the value obtained must be the true score? What if the value is taking for the first time and it read 37 ℃, and then the second time 39 ℃, third time 35 ℃, fourth time 30 ℃ and finally the fifth time 40 ℃? Which will be the true score of the variable? This presupposes that the true value is hidden within 30 ℃ and 40 ℃ with the two numbers inclusive. Thus the reason for the instrument not giving the exact true value is due to error.

Obtained / achieved score

Obtained / achieved score refers to the value obtained when an instrument is used for the measurement of a variable. In this case once a thermometer is used to take the temperature of a body and it reads 37 ℃, this becomes obtained or achieved score. What this means is that for the obtained or achieved scores there are always the element of error because the possibility of another value of the same variable being obtained in the second instance of taking the measurement is ruled out which needs to be checked.

Error score

Now the error score is the score associated with error, and it is the error that is associated with the obtained score which prevents getting the true score. Assuming a weight of a body is taking on a scale and it records 60kg and the same body without being tampered with is weighed a minute later with the same scale and records 85kg. What could have happened with this drastic weigh changed in the same body having not been tampered with in anyway? This difference can be attributed to error.

Having explained these types of scores, it is now ripe to introduce the various types of estimation of values of variables recorded; how to account for the errors associated with the data obtained in the conduct of a research study; and the interpretations that can be given to the data for the consumption of the end-users of the findings.

6.3 Types of Estimations

There are three main types of estimations: point estimation; interval estimation, and confidence lnterval level estimation. Each of these estimations enable the people to know the errors associated with the data from which the findings emanate in order to know the confidence to repose in them.

Point Estimation

With regards to handling data by point estimation, the value of the variable recorded when measurement is taken is considered as the true score. It is assumed that the true score is equal to the obtained or achieved score, disregarding the possibility of error. Thus if a researcher handles data this way, it shows that the possibility of error occurring in the value has not been

accounted for and therefore obtained score is considered as exact, when in actual fact cannot be so. The findings from such researches must be applied cautiously knowing that error has not been catered for.

Data handled in this way when variables are measured once and the values taken as exact without replicated to establish variations demonstrate point estimation. For example using a thermometer to take the temperature of a body once by means of contact at one end of the body and then using the value obtained as the true score. This implies that the researcher is ignoring the fact that there can be researcher fault or calibrating problems of the instruments.

Interval Estimation

When interval estimation is used, the value obtained from the measurement of a variable is taken as the obtained or achieved scores and not the true score. It is assumed that the true score lies within a range of the obtained or achieved scores. It presupposes that in other to find the true score more than one measurement must be taken for errors to be accounted for, for the range within or the interval within which the true score lies to be established.

In interval estimation measurement of variables are replicated so that more than one obtained or achieved scores are obtained for the mean of the scores, variations and the standard error to be calculated to account for the error in the data. Assuming the interval estimation is to be used for estimating the temperature of a body say B. A thermometer is used to take the temperature of the body B at various points of contact with the body B as T_1, T_2, T_3, T_4, T_5 ….etc. T_1, T_2, T_3, T_4 and T_5 become the obtained or achieved score of the measurement. Let say their respective values are 20, 22, 21, 24 and 26. Once the obtained or achieved scores are identified, the researcher must proceed to find

the range within which the true score lies and then account for the error in the data collected as regard their measurement.

To find where possible, the true score lies, the mean of the obtained or achieved scores is calculated as follows:

$$\bar{x}\ (obtained\ scores) = \frac{T_1 + T_2 + T_3 + T_4 + T_5}{n}$$

$$\bar{x}\ (obtained\ scores) = \frac{20 + 22 + 21 + 24 + 26}{5}$$

$$\bar{x}\ (obtained\ scores) = \frac{113}{5} = 22.6$$

$\bar{x}$=mean of obtained or achieved scores, T_1 ……. T_5 represent the obtained scores, n= number of temperature readings taken.

Thus the mean of the obtained scores is 22.6. Now the standard error of the mean is computed to establish the range within which the true score lies from the mean.

Confidence Interval Level Estimation

The confidence interval level estimation is more like the interval estimation but unlike the latter, it takes into consideration the certainty or the confidence level ascribed to the true score of being found within the established interval. The interval estimation helps to establish the interval or the range within which the true scores lies with respect to the mean of the obtained or achieved scores but in this case the researcher cannot assign any certainty or probability of the true scores being found within the established interval. Like the interval obtained in the interval estimation example given earlier, what confidence level will the researcher have that the true score lies within such an interval? Could one be 100% or 99% or 20% sure that the true score lies within the interval or what? The confidence interval level estimation allows one to know to what level

of confidence to repose in the data collected and thus goes beyond the information that can be provided by interval estimation.

The confidence level interval estimation takes into consideration the assigning of probability or calculating the chances of the true scores being found within the established interval determined from the obtained or achieved scores. For instance when a lecturer examines students via examination over a total of 100 scores, the obtained scores are likely to range between 0 and 100 with both values inclusive. So the interval within which the true scores of the students would lie is from 0 to 100, and the probability of the true scores of the student being found in this range would be 100%, this percentage thus becomes the confidence level.

Likewise when a die is tossed the interval within which the obtained scores would lie are from 1 to 6, however if the true score is expected to be an even number, then the confidence level that can be assigned to the true scores of being found within the established interval is the probability of even number appearing when the die is tossed, which is $\frac{3}{6} = 0.5 = 50\%$. This implies that the research is 50% certain that the true score would be found within the established interval.

Bibliography

[1] Dong, Y., Peng, C. Y. J. (2013). *Principled missing data methods for researchers. Springer Plus*.

[2] Ercan, I., Ocakoglu, G., Sigirli, D, Ozkaya, G. (2012) Assessment of Submitted Manuscripts in Medical Sciences According to Statistical Errors. Turkiye Klinikleri *J Med Sci* 32(5): 1381-1387.

[3] Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. (1998). *Multivariate data analysis*. Prentice Hall, New Jersey..

[4] Little, R. J. (2002). Rubin DB. *Statistical analysis with missing data*. John Wiley and Sons, New York.

[5] Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc*. 83(404): 1198-202. .

[6] Little, R., Rubin D. B. (1987). *Statistical analysis with missing data*. John A. Wiley & Sons, Inc., New York. .

[7] Simpson, E. H. (1951). The interpretation of interaction in contingency tables. J Royal Stat Soc Series B (Methodological), 1951: p. 238-241.

[8] Steyerberg, E. W. (2009). *Clinical prediction models*. Springer, New York. .

[9] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3): 581-92.

[10] Vittinghoff, E., Shiboski, S., McCulloch, C. E. (2005). *Regression methods in biostatistics*. Springer, New York.

[11] McGuigan, S. (1995) The use of statistics in the British Journal of Psychiatry. *Br J Psychiatry* 167(5): 683-688.

[12] He, J., Jin, Z., Yu, D. (2009) Statistical reporting in Chinese biomedical journals. *Lancet* 373(9681): 2091-2093.

Chapter 7

Handling Research Data with Descriptive Statistics

7 Handling Research Data with Descriptive Statistics

Felix Kutsanedzie[1]; Sylvester Achio[1]; Appiah Lewis Gyekye[1]

[1]Accra Polytechnic, Accra, Ghana

Abstract

Statistics is a tool used in research analysis and its classified into different branches such as descriptive statistics and inferential statistics. Each of these branches are used to achieve different purposes in research in terms of summary, data analysis and interpretations. The data type collected in the conduct of research determines the type of statistical tool. It can either fall under descriptive or inferential statistics, which must be deployed in data summary, analysis and interpretations. Most researchers and students do not know the tools that fall under descriptive and how to deploy them to handle data garnered during research studies. This chapter therefore identified the various statistical tools under descriptive statistics that can be used to for data processing.

Keywords

Research, Descriptive Statistics, Data, Tools, Analysis

7.1 Introduction

Descriptive statistics is one of the braches of statistics which involves the use of statistical tools to analysis an entire population given that all data sets representing the population are known. For descriptive statistics to be used for data analysis it presupposes that all values of all variables of every subject or entire subject of a population data set is known. Once the dada of the whole population is known, descriptive statistical tools can be used to summarize, analyse, and determine the pattern of the data set representing the population in order to give a decisive interpretation to the data collected in the conduct of a research study. The first stage in data analysis is to describe or summarize data, and the whole analysis may involve calculating and interpreting descriptive statistics. It should be made known that statistics refers to an index used to measure the performance of a sample while parameter is used to refer to a population. Hence for a descriptive statistic to be used for data analysis, all the values of the variables within the sample must be known. Likewise for descriptive statistics to be performed on a population all values of it variables must be known.

Descriptive statistics looks at the measures of central location or tendency; measures of dispersion or variation; distribution shapes; and measures of position. For each research study in when the values of variables for a population is known, all the following statistical tools has the potential of being used for its summary, analysis and interpretation based on the objective to be achieved. Thus in the conduct of a research, the data conducted may require the use of descriptive statistics or inferential depending on the data set that is collected. These statistical descriptive tools and procedures are explained to the readers in a way to enable them know how these tools are used for data analysis. Although these are vital statistical tools, the essence is not to explain the

statistics or mathematics but to emphasize their use and application for data analysis and interpretation in research studies.

7.2 Measures of Central Tendency or Location

The measures of central tendency or location are used to determine the average scores of the sample or population with values of variables known. Usually when instruments are used for taking measurements; and the measurements of the same variable is repeated they may be often times differ. Sometimes these values might be at their extremes. Taking for instance a class score of students for a test out of ten (10) in mathematic are given as follows: 0, 1, 5, 9, 9, 0, 0, 1, 0, 9, 8, 7. The measures of central tendency are used to determine whether the scores fall closer to the central point or the average so that a single value or figure or number is used representatively to describe the data collected. Thus the average of the collected score can be computed and used to describe the data; the median value; and the mode can all be used. However the most appropriate one depending on the type of data, level of measurement as well as the purpose of the study. For instance if the level of measurement of the data is nominal, the mode is appropriate; if however it is ordinal, the median and if interval or ratio, and the mean.

Mode as Measure of Central Location or Tendency

The mode is one of the measures of central location which means it is used as a single number or value or value label to describe or summarize a given set of collected data for a study. It is the number or a value label with the highest occurring frequency within a data set or the sample or population. For instance taking into consideration the test score in mathematic case given earlier - 0, 1, 5, 9, 9, 0, 0, 1, 0, 9, 8 and 7, the most frequent occurring value is 0, and so the modal

value is 0, meaning that majority of the students scored zero. However this is one of the least used measures of central location and it is the most unstable measure. It is appropriate to use this measure for data with nominal level of measurement. Assuming a researcher collects data on the colour of shirts worn in a class as: red, red, green yellow, blue, red, yellow, yellow and pink, yellow. In this case the data type is qualitative and the level of measurement that can be applied is nominal. Hence the use of the mean and the median cannot be computed or determined for the data and used as a single value or value label to represent or summarize, or describe the data appropriately. The only option available to obtain a single value or value that can be used to represent such data is the use of the mode, i.e. to find the modal value label or colour. Since yellow is the most occurring shirt colour, it becomes the modal colour which is used to describe or represent the data.

The mode can be found for both group and ungroup data that are collected.

However, it should be noted that the appropriate level of measurement of data for which the mean can be appropriately used to summarize as a single number is ratio or interval level.

The mean score for grouped and ungrouped data can be calculated as follows:

Data collected from a research study or an experiment can either be grouped or ungrouped. It is referred to as an ungrouped not classified and grouped when it is put into categories.

Taking for instance the weight of students taken using a weighing scale and recorded as follows:

50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, 71. This dataset is ungrouped and therefore the determination of the measures of

central tendency would follow a different means of determination compared to the grouped data.

When the same dataset is put into categories with the following intervals: 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89, then the data set would be referred to as grouped.

To find the mode of an ungrouped data, it is simply finding the frequency of the individual values or scores within the dataset and then taking the one with the highest frequency as the mode or the modal value or item.

Using the given ungrouped dataset - 50, 51, 60, 51, 67, 51 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, 71; the frequencies of the individual scores are derived below and used for the modal value determination.

Table 7.1 *Frequency Table of an Ungrouped Dataset.*

Weights (kg)	**47**	**48**	**46**	**50**	**51**	**60**	**63**	**64**	**66**	**67**	**68**	**71**	**75**	**77**	**78**	**81**	**85**	**86**
Frequencies	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1

Thus the modal value or weight is 51kg because it is the weight measured with the highest frequency. It is a unimodal data, meaning it has one mode. The modal value 51kg indicates that the majority of the students weighed 51kg or the value that can be used to represent the weights of the students in the class is 51kg.

For the grouped data for the same dataset - 50, 51, 60, 51, 67, 51 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, 71 and 72, grouped into intervals such as 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89. The frequency table for the grouped data is shown in table 7.2

Table 7.2 *Frequency Table of a Grouped Dataset.*

Weight Categories of Students	Frequencies
40 – 49	3
50 – 59	4
60 – 69	6
70 – 79	5
80 – 89	3

From the frequency table developed, the modal class of student weights is 60 – 69, but to determine the modal weight of the modal class or interval, the formula below is used:

$$Mode = L + \left(\frac{D_1}{D_1 + D_2}\right) C$$

L=Lower class boundary of the modal class. It is found by substracting 0.5 from the lower limit of the modal class, thus the lower class boundary of modal class =60-0.5=59.5, D_1=the difference in frequences between the modal class and the class before it,

$$i.e. D_1 = 6 - 4 = 2$$

D_2=the difference in frequencies between the modal class and the class after it,

$$i.e. D_2 = 6 - 5 = 1$$

C=the modal class size=upper class limit-lower class limit+1

$$C = 69 - 60 + 1 = 9 + 1 = 10$$

$$Modal\ Weights\ of\ students = 59.5 + \left(\frac{2}{2+1}\right) x\ 10$$

$$Modal\ weight\ of\ students = 59.5 + \left(\frac{2}{3}\right) x\ 10 = 59.5 + 6.67 = 66.16 \cong 66$$

Thus the mode is the most unstable measure of central tendency compared to all the other measures and thus most appropriate for data that is nominal level of measurement has been applied.

Median as Measure of Central Location or Tendency

The median is also another measure of central tendency used as a single value or value to describe or summarize or represent a given data set. The median is regarded as the value or value label that occupies the median position of the data. Again with the example given in the case of the test score - 0, 1, 5, 9, 9, 0, 0, 1, 0, 9, 8 and 7. The median is the value that occupies the middle position when the data set is arranged in an ascending or descending other. So in this case the data arranged in ascending order becomes: 0, 0, 0, 0, 1, 1, 5, 7, 8, 9, 9, 9. Counting from both sides, two numbers occupy the middle position. Thus the $median\ value = \frac{1+5}{2} = 3$. Therefore for the median value to be found required that the average of the two vales are computed and the value used as the median value. It should be noted here that just like the mode, the appropriate level of measurement of the data that the median can be applied to is the ordinal level.

The median score can be found for grouped and ungrouped data

To determine the median for an ungrouped data such -50, 51, 60, 51, 67, 51 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, 71; the values within the dataset are arranged in ascending or descending order and the number or value that occupies the central position is taken as the median.

46, 47, 48, 50, 51, 51, 51, 60, 64, 65, 66, 67, 68, 71, 75, 77, 78, 81, 85, 86

To find the position of the median value, divide the total frequency of the weight by two

$$median\ Position = \frac{20}{2} = 10th$$

However, if two numbers or values compete for the 9th position, the average of the two numbers is computed and then used as the median value or weight in

this case. Thus for the data above, the 10th position is occupied by two numbers i.e. 64 and 65 respectively counting from the left and right, therefore the Median mark or value $= \frac{65+66}{2} = 65.5$.

However when an odd number ungrouped dataset is used such as 2, 3, 3, 4, 5, 6, 2; the values within the dataset are arranged in either ascending or descending order such as 2, 2, 3, 3, 4, 5, 6 and their total frequency determined with one (1) added for the computation of the position of the median value.

Thus the position of the median value $= \frac{7+1}{2} = \frac{8}{2} = 4^{th}$, therefore the number or value occupying the 4th position becomes the median value.

Median value therefore =4

Using the dataset: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71 grouped into intervals such as 40 – 49, 50 -59, 60 – 69, 70 – 79, 80 – 89. The frequency table is developed and the median value calculated as follows:

Table 7.3 *Table Showing the Median Mark for a grouped Dataset.*

Weight Categories of Students	Frequencies	Cumulative frequency
40 – 49	3	3
50 – 59	4	7
60 – 69	6	13
70 – 79	5	18
80 – 89	3	20
	∑f = 20	

To compute the median value the total frequency of the dataset is determined and the median class is obtained by dividing the total frequency by 2, and tracing the class within which it falls.

$$Median\ Class\ position\ = \frac{20}{2} = 10th$$

Using the cumulative frequency of the dataset, the 10th position can be located within 60-69, thus the median class becomes 60 – 69.

Once the median class has been identified, the formula given below is used to compute for the median mark or value:

$$Median\ value\ = L + \left(\frac{\frac{\sum f}{2} - f_c}{f_m}\right)C$$

L=Lower class boundary of the median class. It is found by substracting 0.5 from the lower limit of median class, thus the lower class boundary of median class =60-0.5=59.5, f_c=cumulative frequency before the median class or sum of frequencies before the median class, f_c=3+4 = 7, f_m=frequency of the median class, f=Total frequency of the dataset, C=the modal class size=upper class limit-lower class limit+1

$$C = 69 - 60 + 1 = 9 + 1 = 10$$

$$Median\ value\ = 59.5 + \left(\frac{\frac{20}{2} - 7}{6}\right)x10$$

$$Median\ value\ = 59.5 + \left(\frac{20 - 7}{6}\right)x\ 10$$

$$Median\ value\ = 59.5 + \left(\frac{13}{6}\right)x\ 10 = 59.5 + \frac{130}{6} = 59.5 + 21.67$$

$$Median\ value\ = 59.5 + 21.67 = 81.17$$

Mean as Measure of Central Location or Tendency

The mean also referred to as average of a data set is considered as the most stable and most used measure of central location as it is used as a single number

or value to describe or summarize or represent the data set as a point of convergence where the true score lies. It is found to be the sum of all the scores and dividing it by the total number of occurrences. Thus using the test score - 0, 0, 0, 0, 1, 1, 5, 7, 8, 9, 9, 9;

$$\text{Thus the } mean\ value = \frac{0+0+0+0+1+1+5+7+8+9+9+9}{2} = 4.08$$

Taking an ungrouped dataset such as 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71; the mean or average of the dataset can be determined by the formula

$$Mean\ (\bar{x}) = \frac{\sum x_i}{n}$$

$$Mean\ (\bar{x}) = \frac{50 + 51 + 51 + 51 + \cdots 71}{20}$$

$$Mean\ (\bar{x}) = \frac{1290}{20} = 64.5$$

For group data of the same dataset: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71 grouped into intervals such as 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89.

A table is developed and the mean calculated as follows:

Table 7.4 *Table Showing Mean determined for a Grouped Dataset.*

Weight Class of Students	Frequencies (f)	Midpoint (x)	Fx
40 – 49	3	44.5	133.5
50 – 59	4	54.5	218
60 – 69	6	64.5	387
70 – 79	5	74.5	372.5
80 – 89	3	84.5	253.5
	∑f = 20		∑fx = 1364.5

The midpoint is found for each class by summing upper class limits and lower class limits for each class and dividing by 2. i.e. for class 40 – 49, the midpoint = (40 +49)/2 = 44.5

$$Mean\ (\bar{x}) = \frac{\sum fx}{n} = \frac{\sum fx}{\sum f} = \frac{1364.5}{20} = 68.23$$

7.3 Measures of Variation or Dispersion

These measures are single numbers or indices which are used to describe or summarize the variations in dataset collected in the conduct of a research study or experiment. Just as the measures of central location are single numbers or values or indices used to describe the points at which the dataset converges to a central point, the measures of dispersion or variation rather looks at the spread of values within the dataset collected from a research study. The measures of dispersion include the *range* of the dataset, *interquartile range or quartile deviation, variance, and standard deviation.*

Range as Measure of Variation

Range is the difference between the highest and lowest value recorded in a particular dataset collected from a research study. It gives an indication of the spread within the data though it is not a stable measure of dispersion.

For the dataset: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71; the range for this dataset is 86 - 46 = 40, and indication that the spread between or the variation between the lowest value and the highest value within the dataset is 40 units. Though it is not the best type of measure in terms of variations in a data collected, the lower the range the closer the differences or less variations in the data set. The level of measurement of data the range can be applied to is the ordinal level.

Quartile Deviation or semi-interquartile range

This is found by finding the difference between the third (3^{rd} or the 75^{th} percentile) or upper quartile and the first quartile (1^{st} or the 25^{th} percentile) or lower quartile and then dividing by 2. Before the quartile deviation or semi-interquartile range, the first and third quartile must be computed for the dataset.

To calculate the quartile deviation or semi-interquartile range for an ungrouped data such as:

50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71

The data is re-arranged in either ascending or descending order and the following formula are used to determine the positions of the first and the third quartile:

Position of the First (1^{st} or the 25^{th} Percentile) Mark $= \frac{1}{4}$ x N + 1

Position of the Second (2^{nd} or the 50^{th} Percentile) Mark $= \frac{2}{4}$ x N + 1 $= \frac{1}{2}$ x N + 1

Position of the Third (3^{rd} or the 75^{th} Percentile) Mark $= \frac{3}{4}$ x N + 1

Where N = size of the data in terms of number = 20.

The data re-arranged in ascending order becomes: 46, 47, 48, 50, 51, 51, 51, 60, 64, 65, 66, 67, 68, 71, 75, 77, 78, 81, 85, 86.

Therefore,

Position of the First (1st or the 25th Percentile) Mark $= \frac{1}{4}$ x $20 + 1 = 5 + 1 = 6^{th}$.

Thus the Mark that occupied the 6th position within the re-arranged data becomes the First Quartile Mark as show below:

The First Quartile Mark (Q1) = 51

Position of the Second (2nd or the 50th Percentile) Mark $= \frac{1}{2}$x $20 + 1 = 10 + 1 = 11^{th}$

46, 47, 48, 50, 51, 51, 51, 60, 64, 65, 66, 67, 68, 71, 75, 77, 78, 81, 85, 86

Thus the Second Quartile Mark (Q2) = 66

Position of the Third (3rd or the 75th Percentile) Mark $= \frac{3}{4}$ x $20 + 1 = 15 + 1 = 16^{th}$

46, 47, 48, 50, 51, 51, 51, 60, 64, 65, 66, 67, 68, 71, 75, 77, 78, 81, 85, 86

Thus the Third Quartile Mark (Q3) = 77

Quartile Deviation or semi-interquartile range mark$= \frac{Q_3 - Q_1}{2} = \frac{77-51}{2} = \frac{26}{2} = 13$

Therefore when the quartile deviation is small there is less variations within the dataset and vice versa.

For data such as: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71 grouped into intervals such as 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89.

The quartile deviation can be computed as follows:

First, a frequency table such as done below is constructed for the dataset

Table 7.5 *Frequency Table of an Ungrouped Dataset used for the Determination of the Quartile Deviation for a Grouped Dataset.*

Weight Categories of Students	Frequencies	Cumulative frequency
40 – 49	3	3
50 – 59	4	7
60 – 69	6	13
70 – 79	5	18
80 – 89	3	20
	∑f = 20	

Use the following formula to identify the first, second and third quartile classes:

Position of the First (1st or the 25th Percentile) Class= $\frac{1}{4}$ x N + 1

Position of the Second (2nd or the 50th Percentile) Class= $\frac{2}{4}$ x N + 1 = $\frac{1}{2}$ x N + 1

Position of the Third (3rd or the 75th Percentile) Class= $\frac{3}{4}$ x N + 1

Position of the First (1st or the 25th Percentile) Class= $\frac{1}{4}$ x 20 + 1 = 5 + 1 = 6^{th}

Position of the Second (2nd or the 50th Percentile) Class= $\frac{1}{2}$ x 20 + 1 = 10 + 1 = 11^{th}

Position of the Third (3rd or the 75th Percentile) Class=$\frac{3}{4}$ x $20 + 1 = 15 + 1 = 16^{th}$

Therefore the 6th item in the data is supposed to be located within the First Quartile and using the frequency the 6th item would be located within the class 50 – 59, hence the First Quartile Mark can be located within 50 – 60; the Second Quartile Class is 60 – 69 since data item with position 11 falls within it; the Third Quartile Class is 70 – 79, since data item with the position 16 falls within it. Thus one can then proceed to determine the First, Second and Third Quartile Marks with the following respective formula:

$$Q_1 = L_1 + \left(\frac{\frac{N}{4} - f_c}{f_1}\right)c$$

Q_1=First quartile mark, N=Total frequency of data or data size=20, f_c=cumulative frequency before the first quartile class or total frequency before first quartile class=3, f_1=frequency of the first quartile class=4, L_1=lower class boundary of the first quartile class=49.5, c=first quartile class size=10

$$Q_1 = 49.5 + \left(\frac{\frac{20}{4} - 3}{4}\right) \text{x } 10 = 49.5 + \left(\frac{5-3}{4}\right) \text{x } 10 = 49.5 + (0.5)(10) = 54.5$$

$$Q_2 = L_2 + \left(\frac{\frac{N}{2} - f_c}{f_2}\right)c$$

Q_2=Second quartile mark, N=Total frequency of data or data size=20, f_c=cumulative frequency before the second quartile class or total frequency before second quartile class=7, f_1=frequency of the second quartile class=6, L_1=lower class boundary of the second quartile class=59.5, c=second quartile class size=10

$$Q_2 = 59.5 + \left(\frac{\frac{20}{2} - 7}{6}\right) \text{x } 10 = 59.5 + \left(\frac{10 - 7}{6}\right) \text{x } 10 = 59.5 + \left(\frac{3}{6}\right)(10) = 59.5 + 5$$
$$= 64.5$$

$$Q_3 = L_3 + \left(\frac{\frac{3N}{4} - f_c}{f_3}\right) c$$

Q_=Third quartile mark, N=Total frequency of data or data size=20, f_c=cumulative frequency before the third quartile class or total frequency before third quartile class=13, f_1=frequency of the third quartile class=5, L_1=lower class boundary of the third quartile class=69.5, c=third quartile class size=10

$$Q_2 = 69.5 + \left(\frac{\frac{3(20)}{4} - 13}{5}\right) \text{x } 10 = 69.5 + \left(\frac{15 - 13}{5}\right) \text{x } 10 = 69.5 + \left(\frac{2}{5}\right)(10)$$
$$= 69.5 + 4 = 73.5$$

Hence, $the\ Quartile\ Deviation\ or\ Semi - interquartile\ mark = \frac{Q_3 - Q_1}{2}$.

The Quartile Deviation or Semi-interquartile mark= $\frac{73.5 - 54.5}{2} = \frac{19}{2} = 9.5$.

Hence the value 9.5 is the measure of the variations within the given dataset.

Mean Deviation as a Measure of Variation or Dispersion

This is also a measure of variation, it is found by finding the summation of the differences between the mean of a dataset and the individual data values and dividing by the size of the dataset or the number of individual data it consist of. It thus suggests that the lower the mean deviation, the smaller the variation in the dataset and vice versa. Taking for instance the ungrouped data: 50, 51, 60,

51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71. The mean deviation is computed using the formula below:

$$Mean\ Deviation = \frac{\sum(x_i - \bar{x})}{n}$$

Table 7.6 *Calculation of the Mean Deviation for an Ungrouped Dataset.*

Weights (x) kg	Mean ($\bar{x}$)	x-$\bar{x}$
47	64.5	-17.5
48	64.5	-16.5
46	64.5	-18.5
50	64.5	-14.5
51	64.5	-13.5
60	64.5	-4.5
63	64.5	-1.5
64	64.5	-0.5
66	64.5	1.5
67	64.5	2.5
68	64.5	3.5
71	64.5	6.5
75	64.5	10.5
77	64.5	12.5
78	64.5	13.5
81	64.5	16.5
85	64.5	20.5
51	64.5	-13.5
51	64.5	-13.5
86	64.5	21.5
		$\sum$(x-$\bar{x}$) = -5

$$Mean\ Deviation = \frac{-5}{20} = -0.25$$

For the data: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71 grouped into intervals such as 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89. The mean deviation is computed by developing a table as shown in table 7.7.

Table 7.7 *Calculation of the Mean Deviation for a Grouped Dataset.*

Weights Classes	Midpoints (x)	Freq(f)	mean ($\bar{x}$)	x-$\bar{x}$	f(x − $\bar{x}$)
40 – 49	44.5	1	68.23	-23.73	-23.73
50 – 59	54.5	3	68.23	-13.73	-41.19
60 – 69	64.5	1	68.23	-3.73	-3.73
70 - 79	74.5	1	68.23	6.27	6.27
80 - 89	84.5	1	68.23	16.27	16.27
				∑(x − $\bar{x}$) =-18.65	∑f(x − $\bar{x}$) =-46.82

$$Mean\ Deviation = \frac{\sum f(x_i - \bar{x})}{n} = \frac{-46.82}{20} = -2.34$$

Variance and Standard Deviation as a Measure of Variation or Dispersion

It is a measure of the spread within a dataset. To calculate the variance for a data set, the mean for the collected data is computed, the differences in the mean and each data item is then calculated and each squared and then sum, and finally the outcome divided by the number or size of the dataset to obtain the variance. When the variance is small it means the variation within the dataset is also small and vice versa. Usually when sum of the differences between the mean and each individual value in the data is zero; the square of the differences in the mean and each individual value or item in the data set is rather determined. The square root of the variance is referred to as standard deviation. The standard deviation is therefore the most stable and the most used index of variability in a dataset. The appropriate level of measurement of data the standard deviation is applied to is the interval and ratio levels. The formula for the computation of variance and standard deviation are given as follows:

For a given ungrouped data such as: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71. The mean is first determined as follows:

$$Mean\ (\bar{x}) = \frac{\sum x_i}{n}$$

$$Mean\ (\bar{x}) = \frac{50 + 51 + 51 + 51 + \cdots 71}{20}$$

$$Mean\ (\bar{x}) = \frac{1290}{20} = 64.5$$

$$Variance\ (v) = \frac{\sum(\mathrm{x} - \bar{x})^2}{n}$$

$$Standard\ deviation\ (s) = \sqrt{\frac{\sum(\mathrm{x} - \bar{x})^2}{n}}$$

After which the table below can be prepared and used for the mean differences and their respective squares

Table 7.8.1 *Calculation of the Variance and Standard Deviation for an Ungrouped Dataset.*

Weights (x) kg	Mean ($\bar{x}$)	x-$\bar{x}$	(x-$\bar{x}$)^2
47	64.5	-17.5	306.25
48	64.5	-16.5	272.25
46	64.5	-18.5	342.25
50	64.5	-14.5	210.25
51	64.5	-13.5	182.25
60	64.5	-4.5	20.25
63	64.5	-1.5	2.25
64	64.5	-0.5	0.25
66	64.5	1.5	2.25
67	64.5	2.5	6.25
68	64.5	3.5	12.25
71	64.5	6.5	42.25
75	64.5	10.5	110.25
77	64.5	12.5	156.25
78	64.5	13.5	182.25
81	64.5	16.5	272.25
85	64.5	20.5	420.25
51	64.5	-13.5	182.25
51	64.5	-13.5	182.25
86	64.5	21.5	462.25
		$\sum$(x-$\bar{x}$) = -5	$\sum$(x-$\bar{x}$)^2 =3367

$$Varaince = \frac{\sum(\mathrm{x} - \bar{x})^2}{n} = \frac{3367}{20} = 168.35$$

$$Standard\ deviation\ (s) = \sqrt{\frac{\sum(\mathrm{x} - \bar{x})^2}{n}}$$

$$Standard\ deviation\ (s) = \sqrt{\frac{3367}{20}} = \sqrt{168.35} = 12.97$$

$$Therefore\ s = \sqrt{v}$$

For the data: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71 grouped into intervals such as 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89. The variance and standard deviation is computed by developing a table as below:

Table 7.8.2 *Calculation of the Variance and Standard Deviation for a Grouped Dataset.*

Weights Classes	Midpoints (x)	Freq (f)	mean ($\bar{x}$)	x-$\bar{x}$	(x-$\bar{x}$)^2	f(x-$\bar{x}$)^2
40 – 49	44.5	1	68.23	-23.73	563.11	563.11
50 – 59	54.5	3	68.23	-13.73	188.51	565.53
60 – 69	64.5	1	68.23	-3.73	13.91	13.91
70 - 79	74.5	1	68.23	6.27	39.31	39.31
80 - 89	84.5	1	68.23	16.27	264.71	264.71
				$\sum(\mathrm{x} - \bar{x}) = -18.65$	$\sum(\mathrm{x} - \bar{x})^2 = 1069.57$	$\sum \mathrm{f}(\mathrm{x} - \bar{x})^2 = 1446.57$

For group data variance and standard deviation, the formula below are used:

$$Varaince = \frac{\sum f(\mathrm{x} - \bar{x})^2}{n}$$

$$Variance = \frac{1446.57}{20} = 72.33$$

$$Standard\ deviation\ (s) = \sqrt{\frac{\sum f(\mathrm{x} - \bar{x})^2}{n}}$$

$$Standard\ deviation\ (s) = \sqrt{\frac{1446.57}{20}} = \sqrt{72.33} = 8.50$$

For data to be considered as relatively normal each individual value of the dataset should fall within the range of $\bar{x} \pm 3$SD (3- Standard deviation from the mean of the dataset).

7.4 Measures of Position

It indicates the relative position of an individual score to the scores of others within the dataset or measured on the same variable. It is used when one wants to measure the performance of an individual among his or her colleagues measured on the same variable. The measures of relative position include percentile and the standard scores.

Percentile rank as a Measure of Position

Percentile rank is an indication of whether a percentage score falls on or below a given score. Oftentimes it is deceptive to think a student performed badly just looking at his or her scores in given subjects. For instance taking a student who scored 45 marks and 65 marks out of 100 respectively in Science and English, it would appear as if the student performed better in English in his or her class compared to science but the story might actually be different when subjected to statistically analysis. When the student's performance in English is compared to the scores or performance in the same subject, it might be realized that though the student scored 65 marks in English, his or her score could be the lowest in the class while the 45 marks scored in science could be the highest score in the class. In order to analyse this situation so that the relative position of the student in each of these subject among the classmates is established, the

percentile ranking is used. Thus when the percentile ranking is computed for the scores of the students in English and the student's score of 65 marks for instance falls on or corresponds to 15^{th} percentile, it means that 15 percent of the scores of the students in the class scored lower than 65 marks. If thus the score of 45 marks obtained by the student correspond to the 90^{th} percentile, it thus means 90 percent of the class scored below the student's score of 45 marks.

The percentile can be computed for ungrouped and grouped data as below:

Ungrouped data

To determine the percentile for an ungrouped data of marks such as: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71.

In order to compute any percentile for an ungrouped data, the scores must be arranged in the ascending order.

This formula determines the position of the mark that corresponds to a particular percentile.

For the position of the 1^{st} percentile score $= \frac{1}{100}\ x\ N + 1$

P^{th}=the percentile position, N=total frequency or size of data

Thus to calculate the position of the 40^{th} percentile score;

$$P^{th} = \frac{40}{100}\ X\ 20 + 1 = 9^{th}$$

It means when the scores are arranged in an ascending order, the score that occupies the 9^{th} position is the score corresponding to the 40^{th} percentile.

46, 47, 48, 50, 51, 51, 51, 60, 64, 65, 66, 67, 68, 71, 75, 77, 78, 81, 85, 86.

The 40th percentile mark thus correspond to 64 mark.

Grouped Data

For the data such as: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71 grouped into intervals such as 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89. First, a frequency table such as done below is constructed for the dataset.

Table 7.9 *Calculation of the Percentiles for a Grouped Dataset.*

Weight Categories of Students	Frequencies	Cumulative frequency
40 – 49	3	3
50 – 59	4	7
60 – 69	6	13
70 – 79	5	18
80 – 89	3	20
	∑f = 20	

To determine the percentile in the case of grouped data, the percentile class is first identified by using the formula:

The first (1st) percentile class is given by

$$P^{th} = \frac{1}{100} X N + 1$$

$$P^{th} = \frac{1}{100} X 20 + 1 = 1.2 \cong 1^{st}$$

Thus the class in which the frequency of 1.2 falls becomes the first (1st) percentile class. Thus the percentile class is 40 – 49.

Therefore to calculate the first (1st) percentile mark

$$P_1 = L_1 + \left(\frac{\frac{N}{100} - f_c}{f_1} \right) c$$

P_1=1st percentile mark, N=Total frequency of data or data size=20, f_c=cumulative frequency before the percentile class or total frequency before the percentile class=0, f_1=frequency of the first (1st) percentile class=0, L_1=lower class boundary of the first percentile class= 39.5, c=first percentile class size=10.

$$P_1 = 39.5 + \left(\frac{\frac{20}{100} - 0}{4} \right) x\ 10 = 39.5 + \left(\frac{0.2 - 0}{4} \right) x\ 10 = \ 39.5 + (0.05)(10) = 40$$

Thus the first percentile corresponding to the score 40 marks and thus the lowest mark of the class.

Standard Scores as a Measure of Position

Standard score is a derived score that expresses how far a given raw score is from a referenced point. Some examples of standard scores that are known include z-score and t-score. They are used in converting averages or means in terms of standard deviation units. Sometimes it becomes difficult to have an appropriate scale on which various averages of different tests on performance of individual can be determined. For instance determining the averages of scores in terms of students performance in English and in mathematics on one scale, the standard scores becomes a means of handling that. It should be noted that standard scores are only accurate to a degree where the scores are normally distributed. Thus when one wants to convert raw scores that are not normally distributed, there is a need to transform such scores to what is referred to as normalized scores before their conversion to standard scores.

Z-scores

It is a basic standard score which expresses the mean score in terms of standard deviation units. When scores are transformed into z-score the new mean of the distribution becomes 0 and the standard deviation is 1. On the z-score the mean of the data becomes 0 and a score that is 1 standard deviation above the mean it correspond to z-score of +1; and if a score is 1 standard deviation below the mean it corresponds to z-score of -1. It allows different sets of test scores to be compared on the same scale.

Assuming a student score 40 marks in mathematics and 60 marks in English, from just this statement any one would be tempted to conclude that the student is better in English compared to mathematics. However if it is added to the statement that the average or mean class score in mathematics is 20 and English is 70, then the whole story takes a different dimension in that it indicates that though the student scored a low mark of 40 in mathematics, he had above the average score in his or her class in the same subject; and the contrary can be said of the student's score in English. It thus presupposes that the student performed better in Mathematics as compared to English in the class. Again if the idea of standard deviation of the class scored in both subject is given as 20, then more detailed information on the students score can be obtained. The given information indicates that the student's performance in mathematics would be (20 + 20 = 40).

$$Z = \frac{x - \bar{x}}{SD} = \frac{40 - 20}{20} = \frac{20}{20} = +1$$

z=z-score, x=raw score, SD=standard deviation of scores in the subject, $\bar{x}$=mean of class scores in subject.

Thus the student's performance in mathematics on the z-score would be z-score of +1, meaning his or her performance in mathematics is 1 SD (Standard Deviation) above the mean. In the case of the subject English, (70 - 10 = 60)

$$Z = \frac{x - \bar{x}}{SD} = \frac{60 - 70}{20} = \frac{-10}{20} = -0.5$$

Thus the student's performance in English on the z-score would be z-score of - 0.5; this means the student's performance is 0.5 SD (Standard Deviation) below the mean.

T- Scores

The t-scores is also handled like the z-scores but in the case of t-scores, the size of the data is less, normally it is less than 30, above that size it is treated as z-score.

7.5 Distribution of Shapes

This refers to the shapes assumed by a dataset. An example is skewness of dataset. A dataset is considered normal when the mean, median and mode are equal. The distribution of a normal dataset is given by the normal curve, which is symmetrical in nature or shape, meaning it can be divided into equal halves. However if a dataset or distribution of a dataset is not normal, then it is said to be skewed either to one side or the other, i.e. left or right. When a dataset is skewed, it means most of it elements or data values are more packed on one side than the other. The distribution is termed positively or right skewed if the extremes scores are the upper end of the distribution, that is when the mean is greater than the median and the mode (mean > median > mode). It is negatively or left skewed when the extreme scores are the lower end of the distribution, that is the mean is less than the median and the mode (mean < median < mode).

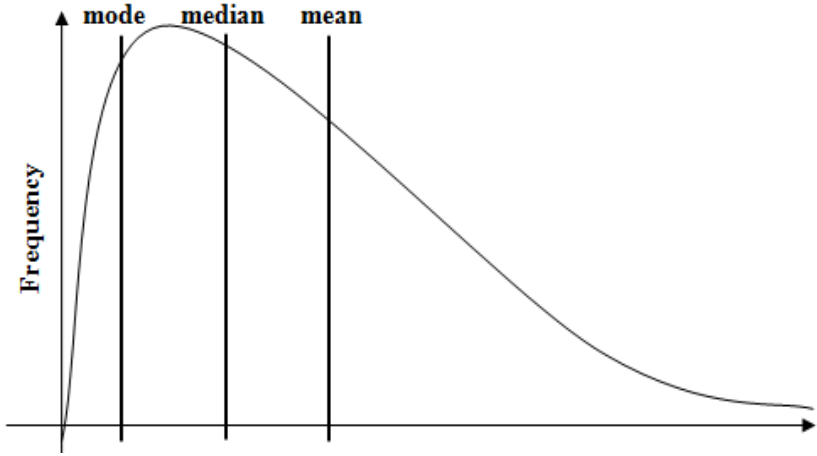


Figure 7.1 *Right skewed or positively data.*

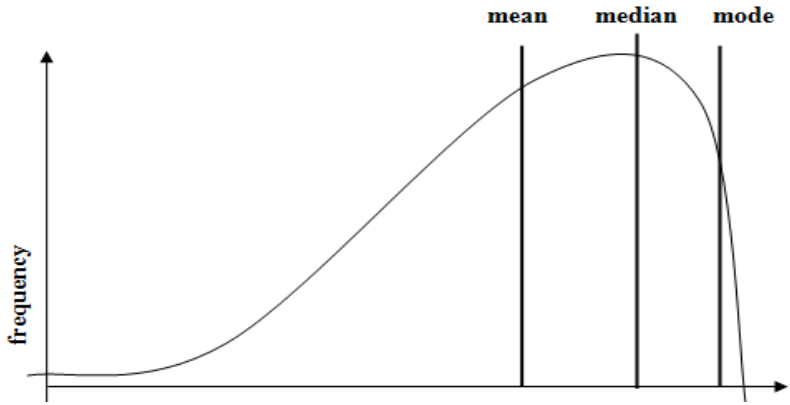


Figure 7.2 *Left skewed or negatively data.*

Table 7.9.1 *STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.*

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

Table 7.9.2 *STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the RIGHT of the Z score.*

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997

Bibliography

[1] Altman, D. G., Boca, R. (1999). *Practical statistics for medical research.* London, New York, Washington D. C.: Chapman & Hall/CRC.

[2] Armitage, P., Berry, G. (1994). *Statistical Methods in Medical Research.* Oxford: Blackwell Scientific Publications.

[3] Bruning, J. L., and Kintz, B. L. (1977). *Computational Handbook of Statistics.* 2nd ed. Glenview, Illinois, Scott, Foresman.

[4] Greenfield, M. L. V. H., Kuhn, J. E., Wojtys, E. M. A. (1997) statistics primer: descriptive measures for continuous data. *Am J Sports Med.*, 25: 720-723.

[5] He, J., Jin, Z., Yu, D. (2009) Statistical reporting in Chinese biomedical journals. *Lancet* 373(9681): 2091-2093.

[6] McHugh, M. L. (2003). Descriptive statistics, part I: level of measurement. *JSPN*, 8: 35-37.

[7] Overholser, B. R., Sowinski, K. M. (2007). Biostatistics primer: part I. *Nutr Clin Pract.* 22: 629-635.

[8] Strasak, A. M., Zaman, Q., Pfeiffer, K. P., Gobel, G., Ulmer, H. (2007). Statistical errors in medical research - a review of common pitfalls. *Swiss Med Wkly 137* (3-4): 44-49.

[9] Strike, P. W. (1991). *Measurement and control, Statistical Methods in Laboratory Medicine.* Oxford: Butterworth-Heinemann.

[10] Wallis, W. A., Roberts, H. V. (1956). *Statistics: A New Approach.* Glencoe, Illinois, Free Press.

Chapter 8

Handling Research Data with Inferential Statistics

8 Handling Research Data with Inferential Statistics

Felix Kutsanedzie[1]; Sylvester Achio[1]; Ofori Victoria[2]

[1]Accra Polytechnic, Accra, Ghana
[2]Agricultural Engineering Department, KNUST, Ghana

Abstract

Inferential statistics just like descriptive statistics has various tools under it which have their peculiar uses as regard data processing. Inferential statistical tools are used for making inferences from samples about the population from which they are selected. It is the field of statistics that has statistical tools for estimating, predicting and making decisions about population and testing hypotheses. The analytical tools within this field of statistics are not well understood and also applied by some researchers, would-be researchers as well as students. This chapter explains the various statistical tools thoroughly.

Keywords

Inferential, Research, Estimation, Sample, Population, Analysis

8.1 Introduction

With descriptive statistics it is required that all the variables considered are known before a dataset can be described with a descriptive tool. However in the case of inferential statistics, not all the variables are known. Inferential statistics involves taking a representative sample from a population to make inferences from it about the sample because it would be difficult to study all members or elements of a population. Inferential statistics is a statistical method used to test hypotheses that relate to the relationship between two variables. Descriptive statistics can be used to describe the relationships based on the data pattern or trend but inferential statistics provide a more rigorous prove as to whether there exist a relationship between two variables. Inferential statistics are used in the case of hypothesis testing and tests such as z-test, t-test, Analysis of Variance (ANOVA) and chi-square test.

8.2 T-test

It is used to test the difference between two groups that are continuous variables. For instance, testing the differences between the weights of individuals based on their ages i.e. testing whether there is weight decrease or increase as an individual ages. The equation for the t-test to use depends on whether the researcher is doing an independent samples t-test – comparing two different groups; dependent samples t-test (paired t-test) – comparing two same groups on two different periods of time or different groups matched on an important variable. There is also the one sample t-test, which is used when the researcher wants to test the group scores of a population with a known mean. It must also be stated that the equation used would also vary when doing the independent sample t-test based on the whether the two groups have the same

sample size or not. The t-test is used to handle a sample size of 30 and below which is usually referred to as small sample size. The t-test is preferred when the sample size is 30 and the standard deviation of the population is unknown.

The equation for a single (one) sample t-test is given as:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where $\bar{x}$ *=sample mean of the group,* μ *=population mean of the group, s=sample standard deviation, n=sample size, t=the t statistics.*

Let us take for instance a case where a researcher collects the test scores obtained by 25 students out of 100 marks for a class size of 30 students summarized below:

Table 8.1.1 *Scores of Students in a Class.*

61.5	**54.3**	**1**	**28**	**2**
15.1	28	33	60	40
12.5	69	19	31	58
50	25	91	67	75
1.5	15	40.4	33	81

Assuming the population mean mark for the class is 35, the t-test can be used to test where there sample mean is different from the population mean. In another scenario, a given mean score of 55, one can test whether there is a difference between this mean score and the sample mean. The t-test is used to test whether the mean of a sample differ from a known or a given mean.

From the data given the mean of the sample can be calculated using the formula:

$$\bar{x} = \frac{\sum x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n}$$

$$\bar{x} = \frac{\sum 61.5 + 54.3 + 1 + \ldots\ldots + 81}{25} = 39.65$$

$$where\ \bar{x} = sample\ mean$$

The sample standard deviation is given by the formula:

$$S = \sqrt{\frac{\sum(\mathrm{x} - \bar{x})^2}{n}}$$

$$S = \sqrt{\frac{\sum(\mathrm{x} - \bar{x})^2}{n}} = \sqrt{\frac{\sum(\mathrm{x}_1 - \bar{x})^2 + (\mathrm{x}_2 - \bar{x})^2 + (\mathrm{x}_3 - \bar{x})^2 + \cdots + (\mathrm{x}_n - \bar{x})^2}{n}}$$

$$S = \sqrt{\frac{\sum(61.5 - 39.65)^2 + (54.3 - 39.65)^2 + (1 - 39.65)^2 + \cdots + (81 - 39.65)^2}{25}}$$

$$= 25.94$$

where n=sample size s=sample standard deviation.

Therefore

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{39.65 - 35}{\frac{25.94}{\sqrt{25}}}$$

$$t_{cal} = \frac{39.65 - 35}{\frac{25.94}{\sqrt{25}}} = \frac{4.65}{\frac{25.94}{5}} = \frac{4.65}{5.19} = 0.90$$

Thus t-calculated $(t_{cal}) = 0.90$

$$where\ \mu = population\ mean$$

Now the degree of freedom (df) of the sample which is given by the total number of sample minus one (1) i.e. $df = n - 1 = 25 - 1 = 24$. Once this is done, the degree freedom (df) obtained (24) can be used with the selected level of significance chosen either 5% (0.05) or 1% (0.01) to read the t-critical value

or tabulated value. The t-calculated and t-critical values are then compared and the decisions taken as follow:

if $t_{cal} > t_{critical\ or\ tabulated}$ $(at\ df=24, \alpha=0.05\ or\ 0.01$.

Then the sample mean is statistically significant from the hypothesized mean or assumed population mean.

However if $t_{cal} < t_{critical\ or\ tabulated}$ $(at\ df=24, \alpha=0.05\ or\ 0.01$.

Then the sample mean is not statistically significant from the hypothesized mean or assumed population mean.

A demonstration of how to read the t-critical value from the t-test table in figures 8.1.1 and 8.1.2

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05
df							
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052

Figure 8.1.1 *T-table for 0.05 Level of Significance.*

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02
df								
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479

Figure 8.1.2 *T-table for 0.01 Level of Significance.*

Decision:

$$t_{cal}(1.71) > t_{critical\ or\ tabulated\ at\ df=24,\alpha=0.05}(0.90)$$

$$t_{cal}(2.49) > t_{critical\ or\ tabulated\ at\ df=24,\alpha=0.05}(0.90)$$

Conclusion:

Since t_{cal} *is greater than* $t_{crit\ at\ df=24}$ *at both 5% and 1% levels of significance, then the sample mean is significantly different from the population mean or the hypothesize mark.*

Using the independent t-test

The independent t-test is used to test whether the means of two independent groups are the same or differ from each other. For this test to be used the two

groups must be independent from each other. The formula for the computation of the t-calculated value is given by the formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{ss_1 + ss_2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sum x_1^2 - \frac{\sum (x_1)^2}{n_1} + \sum x_2^2 - \frac{\sum (x_2)^2}{n_2}}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where SS_1=Sum of Squares of Group one, SS_2=Sum of Squares of Group two, where n_1 and n_2 are sample sizes of group one and two respectively.

It should be noted that the sample sizes may be equal or unequal depending on the data being collected.

Now let us assume a researcher want to test whether the sample means of the performance of two groups of students pursuing two different programmes based on the data summarized below:

1. *Equal Sample Size*

Table 8.2.1 *Data to Illustrate Equal Sample Size.*

Sample No.	A	B
1	25	34
2	45	59
3	63	73
3	21	43
4	19	17
5	78	47

For the table above the sample sizes for the two independent groups (A and B) are equal. Each group has a sample size of 6.

In order to compute the t-calculated value, the sample means and the sum of squares for each of group is determined.

The means of each group is calculated as follows:

$$\bar{x}_A = \frac{\sum x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n_A}$$

$$\bar{x}_A = \frac{\sum 25 + 45 + 63 + \ldots\ldots + 78}{6} = 41.83$$

$$\bar{x}_B = \frac{\sum x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n_B}$$

$$\bar{x}_B = \frac{\sum 34 + 59 + 73 + \ldots\ldots + 47}{6} = 45.5$$

The Sum of Square for each of the groups is also determined as follows:

$$SS_A = \sum x_A^2 - \frac{\sum (x_A)^2}{n_A}$$

$$SS_A = \sum (25^2 + 45^2 + 63^2 + \ldots + 78^2) - \frac{\sum (25 + 45 + 63 + \cdots + 78)^2}{6}$$

$$SS_A = 13505 - 10500.17 = 3004.83$$

$$SS_B = \sum x_B^2 - \frac{\sum (x_B)^2}{n_B}$$

$$SS_B = \sum (34^2 + 59^2 + 73^2 + \ldots + 47^2) - \frac{\sum (34 + 59 + 73 + \cdots + 47)^2}{6}$$

$$SS_B = 14313 - 12421.5 = 1891.5$$

$$t_{cal} = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\left(\frac{\sum x_A^2 - \frac{\sum (x_A)^2}{n_A} + \sum x_B^2 - \frac{\sum (x_B)^2}{n_B}}{n_A + n_B - 2}\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

$$t_{cal} = \frac{41.83 - 45.5}{\sqrt{\left(\frac{3004.83 + 1891.5}{6 + 6 - 2}\right)\left(\frac{1}{6} + \frac{1}{6}\right)}}$$

$$t_{cal} = \frac{-3.67}{\sqrt{\left(\frac{(4896.33)}{10}\right)\left(\frac{2}{6}\right)}} = \frac{-3.67}{\sqrt{489.97}} = \frac{-3.67}{22.14} = -0.17$$

The degree of freedom $(df) = (n + n - 2) = (6 + 6 - 2) = 10.$

The t-critical value can now be read from the t-test table and then compared to the t-calculated to take the decision.

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05
df							
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201

Figure 8.2.1 *T-table for 0.05 Level of Significance.*

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02
df								
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718

Figure 8.2.2 *T-table for 0.01 Level of Significance.*

Decision:

$$t_{cal}(2.23) > t_{critical\ or\ tabulated\ at\ df=10,\alpha=0.05}(-0.17)$$

$$t_{cal}(2.76) > t_{critical\ or\ tabulated\ at\ df=10,\alpha=0.01}(-0.17)$$

Conclusion:

Since t_{cal} *is greater than* $t_{crit\ at\ df=10}$ *at both 5% and 1% levels of significance, then the sample mean of the two independent group is significantly different.*

2. Unequal Sample Size

Table 8.3.1 *Data to Illustrate Equal Sample Size.*

Sample No.	A	B
1	25	34
2	45	59
3	63	73
4	21	43
5	19	
6	78	

In the case given above, the two independent groups have unequal sample sizes: Group has a sample size of 6 and B, a sample size of 4. This case can be handled in a similar way as done below: The means of each group is calculated as follows:

$$\bar{x}_A = \frac{\sum x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n_A}$$

$$\bar{x}_A = \frac{\sum 25 + 45 + 63 + \ldots\ldots + 78}{6} = 41.83$$

$$\bar{x}_B = \frac{\sum x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n_B}$$

$$\bar{x}_B = \frac{\sum 34 + 59 + 73 + 43}{4} = 52.25$$

The Sum of Square for each of the groups is also determined as follows:

$$SS_A = \sum x_A^2 - \frac{\sum (x_A)^2}{n_A}$$

$$SS_A = \sum(25^2 + 45^2 + 63^2 + \ldots + 78^2) - \frac{\sum(25 + 45 + 63 + \cdots + 78)^2}{6}$$

$$SS_A = 13505 - 10500.17 = 3004.83$$

$$SS_B = \sum x_B^2 - \frac{\sum(x_B)^2}{n_B}$$

$$SS_B = \sum(34^2 + 59^2 + 73^2 + 43^2) - \frac{\sum(34 + 59 + 73 + 43)^2}{4}$$

$$SS_B = 11815 - 10920.25 = 894.75$$

$$t_{cal} = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\left(\frac{\sum x_A^2 - \frac{\sum(x_A)^2}{n_A} + \sum x_B^2 - \frac{\sum(x_B)^2}{n_B}}{n_A + n_B - 2}\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

$$t_{cal} = \frac{41.83 - 52.25}{\sqrt{\left(\frac{3004.83 + 894.75)}{6 + 4 - 2}\right)\left(\frac{1}{6} + \frac{1}{4}\right)}}$$

$$t_{cal} = \frac{-3.67}{\sqrt{\left(\frac{(3899.58)}{8}\right)\left(\frac{5}{12}\right)}} = \frac{-3.67}{\sqrt{487.86}} = \frac{-3.67}{22.09} = -0.17$$

The degree of freedom$(df) = n + n - 2 = (6 + 4 - 2) = 8$

Now the researcher can take the decision by reading t-critical value from the t-test table and then compared to the t-calculated value.

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781

Figure 8.3.1 *T-table for 0.05 Level of Significance.*

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02
df								
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718

Figure 8.3.2 *T-table for 0.01 Level of Significance.*

Decision:

$$t_{cal}(2.31) > t_{critical\ or\ tabulated\ at\ df=8,\alpha=0.05}(-0.17)$$

$$t_{cal}(2.90) > t_{critical\ or\ tabulated\ at\ df=8,\alpha=0.01}(-0.17)$$

Conclusion:

Since t_{cal} *is greater than* $t_{crit\ at\ df=10}$ *at both 5% and 1% levels of significance, then the sample mean of the two independent group is significantly different.*

Z-test

The Z-test also functions like the t-test but the difference here is that when the sample size exceeds 30 and the standard deviation of the population is known, the

data is considered as a large sample and thus the Z-test becomes the appropriate test. There are several Z-tests just like the t-tests. These include: one sample Z-test for proportions; two sample z-Test for Proportions; one sample Z-test; two sample Z-test with equal variance; two sample Z-test with unequal variance.

One sample Z-test for proportions

This is used to test whether the proportions of a sample are the same or different. Let us say there is a claim made by buyers that 5 out of 10 cars manufactured by a company are faulty. To test this claim by the buyers assuming a random sample of 35 cars manufactured by the company out of 60 were faulty; the one sample test for proportions can be used.

To test this claim, the one sample test for proportions is used by following the procedure below:

State the hypothesis

$$P = \frac{S}{n} = \frac{9}{10} = 0.9$$

$$\hat{P} = \frac{s_T}{n_T} = \frac{35}{60} = 0.58$$

where p=proportion n=number of proportion, s_T=Scores or counts out of random sample, nT=Total random sample number, s=scores or counts out of number of proportion.

$$H_o: P = 0.5$$

$$H_1: P \neq 0.5$$

Choose the level of significance

Let us use $\alpha = 5$ *or* 0.05.

State the decision rule:

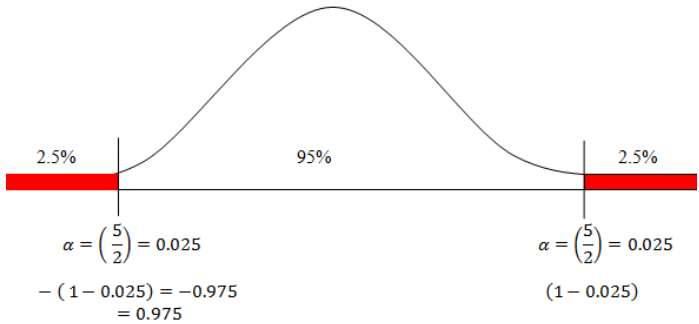


Figure 8.4.1 *Illustration how the chosen* $\boldsymbol{\alpha}$ *should be handled.*

Since the hypothesis is non directional i.e. (the equal sign has been used). It means it is a two tailed tests, therefore we use $\alpha = \frac{5}{2} = 0.025$, in the right and left directions, hence to find the Z-value in the left direction look for the Z-value for $(1 - 0.025) = 0.975$ in both directions. The Z- value for 0.0975 in the table is -1.96 and 1.96.

$$Reject\ H_0\ if\ Z < -1.96\ or\ Z > 1.96$$

$$Fail\ to\ reject\ H_0\ if\ -1.96 \leq Z \leq 1.96$$

Calculate the test statistics (Z- calculated)

$$Z = \frac{\hat{P} - P}{\sqrt{\frac{P(1-P)}{n_T}}}$$

$$Z = \frac{0.58 - 0.90}{\sqrt{\frac{0.90(1-0.90)}{100}}}$$

$$Z = \frac{-0.32}{\sqrt{\frac{0.90(0.1)}{100}}} = \frac{-0.32}{\sqrt{0.0009}} = \frac{-0.32}{0.03} = -10.67$$

State the results

$$Z = -10.67$$

$$Reject\ H_0\ because\ Z < -1.96$$

Make the conclusion

It can therefore be concluded there that the claim made by the buyers that 5 out of 10 manufactured cars are faulty is true.

Two sample Z-test for proportions

The two sample proportion test is used to test whether proportions of two samples are the same or different. If there is a claim that drug A is more efficacious than drug B, and a random sample of 40 out of 100 patients and 56 out of 100 patients recovered from the same disease as a result of administering drugs A and B respectively to them.

State the hypothesis

$$P_1 = \frac{s_1}{n_1} = \frac{44}{100} = 0.40$$

$$P_2 = \frac{s_2}{n_2} = \frac{56}{100} = 0.56$$

$$\hat{P} = P_1 + P_2 = \frac{40}{100} + \frac{56}{100} = 0.40 + 0.56 = 0.96$$

where P_1=proportion 1, P_2=proportion 2, n_1=number of proportion 1, n_2=number of proportion 2, s_1=scores or counts out for proportion 1, s_2=scores or counts out for proportion 2

$$H_o: P_1 = P_2$$

$$H_1: P_1 \neq P_2$$

Choose the level of significance

Let us choose $\propto = 0.05\ 0r\ 5\%$

State the decision rule

$$Reject\ H_0\ if\ Z < -1.96\ or\ Z > 1.96$$

$$Fail\ to\ reject\ H_0\ if\ -1.96 \leq Z \leq 1.96$$

Calculate the test statistics (Z- calculated)

$$Z = \frac{P_1 - P_2}{\sqrt{\hat{P}(1-\hat{P})}\ x\ \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$Z = \frac{0.40 - 0.56}{\sqrt{0.96(1-0.96)}\ x\ \sqrt{\frac{1}{100} + \frac{1}{100}}}$$

$$Z = \frac{-0.16}{\sqrt{0.96(0.04)}\ x\ \sqrt{\frac{2}{200}}}$$

$$Z = \frac{-0.16}{\sqrt{0.04}\ x\ \sqrt{0.01}} = \frac{-0.16}{0.2\ x\ 0.1} = -8.0$$

State the results

$$Z = -8.0$$

$$Reject\ H_0\ because\ Z < -1.96$$

Make the conclusion

The null hypothesis therefore must be rejected on the grounds that the data does not support it hence the alternate hypothesis is true i.e. drug A is more efficacious compared to drug B.

One sample Z-test

It is used for testing the claims when data is collected on two samples with equal variance. For instance, let us consider a case where the weights of mean in a country are normally distributed with a mean of 67kg and standard deviation of 6. Data on the weight of a sample size of 40 men within a city in that country was collected, and the mean weight determined to be 65Kg. One can then test whether the mean weight of men in the city is higher than that of the country. To test this, the following procedures are followed:

State the hypothesis

$$H_o: \bar{x} = \mu$$

$$H_1: \bar{x}_1 \neq \mu$$

Choose the level of significance

$$Let \ \propto = 0.05 \ or \ 5\%$$

State the decision rule

$$Reject \ H_0 \ if \ Z < -1.96 \ or \ Z > 1.96$$

$$Fail \ to \ reject \ H_0 \ if \ -1.96 \leq Z \leq 1.96$$

Calculate the test statistics (Z-calculated)

$$Z = \frac{\bar{x} - \mu}{\sqrt{\frac{\delta}{n}}}$$

$$\bar{x} = 65 \ \mu = 67 = 70 \ \delta = 6 \ n = 40$$

$$Z = \frac{65 - 67}{\sqrt{\frac{6}{40}}} = \frac{-2}{\sqrt{0.15}} = \frac{-2}{1.08} = -1.85$$

State the results

$$Z = -1.85$$

$$Fail \ to \ reject \ H_0 \ because \ -1.96 \leq Z \leq 1.96$$

Make the conclusion

The weight of the men in the city does not differ significantly from the weights of men in the country.

Two sample z-test with equal variance

It is used for testing the claims when data is collected on two samples with equal variance. For instance, let us consider a case where the weights of 60 men each from two different cities A and B in a country are normally distributed with means 65kg and 70kg respectively with equal standard deviation of 6. A researcher can test whether the weights of men in the two cities differ.

The case above is handled as follows:

State the hypothesis

$$H_o: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Choose the level of significance

$$Let \ \propto = 0.05 \ or \ 5\%$$

State the decision rule

$$Reject \ H_0 \ if \ Z < -1.96 \ or \ Z > 1.96$$

$$Fail \ to \ reject \ H_0 \ if \ -1.96 \leq Z \leq 1.96$$

Calculate the test statistics (Z-calculated)

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\delta\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\bar{x}_1 = 65 \ \bar{x}_2 = 70 \ \delta = 6 \ n_1 = 60 \ n_1 = 60$$

$$Z = \frac{65 - 70}{6\sqrt{\frac{1}{60} + \frac{1}{60}}}$$

$$Z = \frac{-5}{6\sqrt{\frac{2}{60}}} = \frac{-5}{6\sqrt{0.03}} = \frac{-5}{6\ x\ 0.18} = \frac{-5}{1.08} = -4.63$$

State the results

$$Z = -4.63$$

$$Reject\ H_0\ because\ Z < -1.96$$

Make the conclusion

The weights of the men in the two cities are significantly different.

Two sample z-test with unequal variance.

Assuming a drink was manufactured and there is a claim that the drink increases weight when taken. Let us again assume that a sample size of 80 men of equal weights were selected in the country and 40 of them were given the drink and the other forty not. After six months weights of the men given the drink and those not given were determined and the following results obtained: Those given the drinks had a mean weight of 45kg with a standard deviation of 4 while those not given had a mean weight of 35kg with a standard deviation of 5. One can test whether the drink caused any significant change in the weights of men given the drink and those not. This can be tested below.

State the hypothesis

$$H_o: \bar{x}_1 = \bar{x}_2$$

$$H_o: \bar{x}_1 \neq \bar{x}_2$$

Choose the level of significance

State the decision rule

$$Reject\ H_0\ if\ Z < -1.96\ or\ Z > 1.96$$

$$Fail\ to\ reject\ H_0\ if\ -1.96 \leq Z \leq 1.96$$

Calculate the test statistics (Z- calculated)

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}}}$$

$$\bar{x}_1 = 45\ \bar{x}_2 = 35\ \delta_1^2 = 4\ \delta_2^2 = 5\ n_1 = 40\ n_1 = 40$$

$$Z = \frac{45 - 35}{\sqrt{\frac{4}{40} + \frac{5}{40}}}$$

$$Z = \frac{10}{\sqrt{\frac{9}{40}}} = \frac{10}{0.47} = 21.08$$

State the results

$$Z = 21.08$$

$$Reject\ the\ null\ hypothesis\ (H_0)\ because\ Z > 1.9$$

Make the conclusion

It can therefore be concluded that weights of men given the drink are significantly different from those not given. Hence the drink might have caused this difference in weight.

Chi-square Test

The chi-square test is used to test whether a relationship exist or not between two categorical variables; and also establish whether a number of outcomes are

occurring with equal frequencies or not; or conforming to a known distribution or not.

Basically the chi-square is used for the following: test for hypothesized ratios; test for homogeneity of data collected from experimental trials that can be repeated; test for the independence of two groups attribute data – in other words it implies testing for whether they have a relationship.

The formula below is used in computing the chi-square (x^2) calculated value:

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$x^2 = Chi - square\ value, O = observed\ value\ E = expected\ value$$

Testing whether two outcomes occur with equal frequencies or not

Taking for instance a scenario where an individual is blindfolded and allowed to pick randomly each time a medal among five different medals placed on a table for 100 times, the data below is obtained:

Table 8.4.1 *Data used for Illustration of Chi-square Test.*

Medal type	Observed frequency selected medals
Diamond	25
Gold	30
Bronze	10
Silver	20
Aluminium	15

Now with this data, a researcher can test whether the frequency of the selection of medals are equal or not or follow a particular distribution using a chi-square test. In order to use a chi-square test, two particular set of data is required: the

observed values and expected values. The observed values are those collected from the experiment or the study. The expected values are based on calculations.

For instance in the case at hand, it can be said that once the section was done 100 times and each of the medals given equal chances of being selected, it presupposes that the expected frequencies of each medal of being selected is 100. Thus a table below can be developed for the data:

Table 8.4.2 *Determining the Expected frequency from a given Data for calculating the Chi-square Test.*

Medal type	Observed frequency selected medals	Expected frequency of selected medals
Diamond	25	100
Gold	30	100
Bronze	10	100
Silver	20	100
Aluminium	15	100

The chi-square test for the data is computed as follows:

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$x^2 = \sum_{i=1}^{5} \frac{(O_d - E_d)^2}{E_d} + \frac{(O_g - E_g)^2}{E_g} + \frac{(O_b - E_b)^2}{E_b} + \frac{(O_s - E_s)^2}{E_s} + \frac{(O_a - E_a)^2}{E_a}$$

$$x^2 = \sum_{i=1}^{5} \frac{(25 - 100)^2}{100} + \frac{(30 - 100)^2}{100} + \frac{(10 - 100)^2}{100} + \frac{(20 - 100)^2}{100} + \frac{(15 - 100)^2}{100}$$

$$x^2 = \sum_{i=1}^{5} \frac{(-75)^2}{100} + \frac{(-70)^2}{100} + \frac{(-90)^2}{100} + \frac{(-80)^2}{100} + \frac{(-85)^2}{100}$$

$$x^2 = \sum_{i=1}^{5} \frac{5625}{100} + \frac{4900}{100} + \frac{8100}{100} + \frac{6400}{100} + \frac{7225}{100}$$

$$x^2 = \sum_{i=1}^{5} 56.25 + 49 + 81 + 64 + 72.25 = 322.5$$

Thus $x^2_{calculated} = 322.5$

$$df = n - 1 = 5 - 1 = 4$$

Therefore if $x^2_{calculated} > x^2_{critical}$ $(df = 4, \alpha = 0.05\ 0r\ 0.01)$, then the frequencies of selecting the medals are significantly different.

However if $x^2_{calculated} < x^2_{critical}$ $(df = 4, \alpha = 0.05\ or\ 0.01)$, then the frequencies of selecting the medals are not significantly different or equal.

Before a decision can be taken, the chi-square critical value must be read from the figure as show below:

The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it from one, and then look it up (ie: 0.05 on the left is 0.95 on the right)

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548

Figure 8.4.2 *Chi-square Test Table.*

From the table

$$\mathrm{x}^2_{\mathrm{critical}}\ (\mathrm{df} = 4, \alpha = 0.05) = 9.488;$$

$$x^2_{critical}\ (df = 4, \alpha = 0.01) = 13.277$$

Decision:

$$x^2_{calculated} (> x^2_{critical}\ at\ df = 4, [\alpha = 0.01\ (13.28)] and\ [\alpha = 0.05\ (9.49)]$$

Conclusion:

The frequencies of selecting the medals are significantly different or are not equal.

Test for hypothesized ratios.

This test is used when there is an existing hypothesized ratio that is known to be working and a researcher has carried out a study relating to the ratio, he or she then can use chi-square to test whether the data collected in the study confirms or not the hypothesized ratio. Assuming there is known hypothesized ratio from a study that indicates that the number of birth per animal is in the ratio of 1:2 for male and female offspring respectively. If a researcher conducts an experiment on the births per animal over a period and recorded 92 and 212 male and female progenies respectively, then the chi-square test can be used to test this data collected against the hypothesized ratio.

In order to compute the chi-square value for this data, the expected values for each of the observation must be calculated.

Since the total birth per animal over the period is 92 + 212 = 304, the expected values for each of the birth per gender for the animal over the period can be calculated by using the hypothesized ratio i.e. 1:2. Thus the expected values for the birth per gender for the animal over the period are calculated as follow:

$$E_m = \frac{Ratio\ value\ for\ Male}{Total\ Ratio}\ x\ Total\ Observed\ Value$$

$$E_m = \frac{1}{3}\ x\ 304 = 101.33$$

$$E_m = \frac{Ratio\ value\ for\ Female}{Total\ Ratio}\ x\ Total\ Observed\ Value$$

$$E_f = \frac{2}{3}\ x\ 304 = 202.67$$

Table 8.5.1 *Data used as Illustration for testing of hypothesized ratio using Chi-square Test.*

	Birth per animal	
Gender	**Observed Values**	**Expected Values**
Male	92	101.33
Female	212	202.67
Total	304	304

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$x^2 = \sum_{i=1}^{2} \frac{(O_m - E_m)^2}{E_m} + \frac{(O_f - E_f)^2}{E_f}$$

$$x^2 = \sum_{i=1}^{5} \frac{(92 - 101.33)^2}{101.33} + \frac{(212 - 202.67)^2}{202.67}$$

$$x^2 = \sum_{i=1}^{5} \frac{87.05}{101.33} + \frac{87.05}{202.67} = 0.86 + 0.43 = 1.29$$

$$x^2 = 1.29$$

$$df = n - 1 \ = 2 - 1 = 1$$

Therefore if $x^2_{calculated} > x^2_{critical}$ $(df = 1, \alpha = 0.05 \; or \; 0.01)$ then the birth per gender animal is significantly different from the hypothesized ratio, that is to say the data collected does not confirm the hypothesized ratio.

However if $x^2_{calculated} < x^2_{critical}$ $(df = 1, \alpha = 0.05 \; or \; 0.01)$, then the birth per gender animal is not significantly different from the hypothesized ratio, that is to say that the data collected confirms the hypothesized ratio.

The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it from one, and then look it up (ie: 0.05 on the left is 0.95 on the right)

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	*3.841*	5.024	*6.635*	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757

Figure 8.5.1 *Chi-square Test Table.*

The

$$x^2_{critical}(\,at\ df = 1, \alpha = 0.05\,) = 3.841;$$

$$x^2_{critical}(\,at\ df = 1, \alpha = 0.01\,) = 6.635$$

$$However\ if\ x^2_{calculated} < x^2_{critical}\ (\,df = 1, \alpha = 0.05\ or\ 0.01),$$

Decision:

$$\mathrm{x}^2_{\mathrm{calculated}} = 1.29 > x^2_{\mathrm{critical}}(\,\mathrm{at\ df} = 1, \alpha = 0.05\,) = 3.841$$

$$\mathrm{x}^2_{\mathrm{calculated}} = 1.29 > x^2_{\mathrm{critical}}(\,\mathrm{at\ df} = 1, \alpha = 0.01\,) = 6.635$$

Conclusion:

The birth per gender animal is significantly different from the hypothesized ratio, that is to say the data collected does not confirm the hypothesized ratio.

Test for homogeneity of data collected from experimental trials that can be repeated.

Once the researcher is able to confirm that the collected data from a study follows the given hypothesized ratio and the study can be repeated, the researcher can test the homogeneity of the data collected on repeated trials for

the study by the chi-square test. Thus the homogeneity of the data below can be tested by Chi-square test.

Table 8.5.2 *Data for Illustration of the Test of Homogeneity using Chi-square Test.*

Births per animal per time	Male	Female	Row Total
1st	12	26	38
2nd	9	16	25
3rd	15	27	42
4th	17	33	50
Column Total	53	102	155
Grand Total			155

In this situation, the researcher can compute the expected values by using the formula below:

$$E_V = \frac{R_T \, x \, C_T}{G_T}$$

E_V=Expected value, R_T=Row Total, C_T=Column Total, G_T=Grand Total.

Thus expected values for the 1st Trial:

Table 8.5.3 *Determining the Expected Values for the Data used for Test of Homogeneity.*

Births per animal per time	Male		Female		Row Total
1st	12	$E_V = \frac{38 \, x \, 53}{155}$	26	$E_V = \frac{38 \, x \, 102}{155}$	38
2nd	9	$E_V = \frac{25 \, x \, 53}{155}$	16	$E_V = \frac{25 \, x \, 102}{155}$	25
3rd	15	$E_V = \frac{42 \, x \, 53}{155}$	27	$E_V = \frac{42 \, x \, 102}{155}$	42
4th	17	$E_V = \frac{50 \, x \, 53}{155}$	33	$E_V = \frac{50 \, x \, 102}{155}$	50
Column Total	53		102		155
Grand Total					155

Table 8.5.4 *Determined Expected Values for the Data used for the Test of Homogeneity.*

Births per animal per time	Male Observed values	Male Expected values	Female Observed Values	Female Expected values	Row Total
1st	12	12.99	26	25.01	38
2nd	9	8.54	16	16.45	25
3rd	15	14.36	27	27.64	42
4th	17	17.10	33	32.90	50
Column Totals	53		102		155
Grand Total					155

The chi-square value for the data can now be computed as follows:

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$x^2 = \sum_{i=1}^{n} \frac{(12 - 12.99)^2}{12.99} + \cdots + \frac{(17 - 17.10)^2}{17.10} + \frac{(26 - 25.01)^2}{25.01} + \cdots + \frac{(33 - 32.90)^2}{32.90}$$

$$x^2 = \sum_{i=1}^{n} \frac{(-0.994)^2}{12.99} + \cdots + \frac{(-0.097)^2}{17.10} + \frac{(0.994)^2}{25.01} + \cdots + \frac{(0.097)^2}{32.90}$$

$$x^2 = \sum_{i=1}^{n} \frac{0.987}{12.99} + \cdots + \frac{0.009}{17.10} + \frac{0.987}{25.01} + \cdots + \frac{0.009}{32.90}$$

$$x^2 = \sum_{i=1}^{n} 0.076 + \cdots + 0.001 + 0.040 + \cdots + 0.0003 = 0.20$$

$$df = n - 1 = 4 - 1 = 3$$

The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it from one, and then look it up (ie: 0.05 on the left is 0.95 on the right)

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838

Figure 8.5.2 *Chi-square Test Table.*

The read critical values from the table at both 1% and 5% levels of significance are:

$$x^2_{critical}(\,at\ df = 1, \alpha = 0.05\,) = 7.82$$

$$x^2_{critical}(\,at\ df = 1, \alpha = 0.01\,) = 11.35$$

Decision:

$$x^2_{calculated} = 0.20 < x^2_{critical}(\,at\ df = 1, \alpha = 0.05\,) = 7.82$$

$$x^2_{calculated} = 0.20 < x^2_{critical}(\,at\ df = 1, \alpha = 0.01\,) = 11.35$$

Conclusion:

It means the homogeneity of the data does not differ significantly for each trial. The data collected is similar for each trial.

Test for the independence of two groups attribute data

Considering another case where a researcher needs to establish whether a relationship exist between three different modes of admission of students into different programmes and the performance of the students. A data below can be subjected to the analysis:

Table 8.6.1 *Data used for Illustration of the Test for the Independence of Two Groups.*

Modes of students admission	Grade Points of students in each Programme		Row Total
	Mechanical Engineering	Electrical Engineering	
Regular	4.2	3.1	7.3
Access	3.2	3.4	6.6
Matured	3.6	2.1	5.7
Column Total	11	8.6	19.6

For the above data, the expected values for each observed value are computed as follows:

$$E_V = \frac{R_T \; x \; C_T}{G_T}$$

E_V=Expected value, R_T=Row Total, C_T=Column Total, G_T=Grand Total.

Table 8.6.2 *Determining the Expected Values for Data used for the Illustration of the Test for the Independence of Two Groups.*

Modes of students admission	**Grade Points of students in each Programme**				**Row Total**
	Mechanical Engineering		Electrical Engineering		
	Observed values	Expected values	Observed values	Expected Values	
Regular	4.2	$E_V = \frac{7.3 \; x \; 11}{19.6}$	3.1	$E_V = \frac{7.3 \; x \; 8.6}{19.6}$	7.3
Access	3.2	$E_V = \frac{6.6 \; x \; 11}{19.6}$	3.4	$E_V = \frac{6.6 \; x \; 8.6}{19.6}$	6.6
Matured	3.6	$E_V = \frac{5.7 \; x \; 11}{19.6}$	2.1	$E_V = \frac{5.7 \; x \; 8.6}{19.6}$	5.7
Column Total	11		8.6		19.6

Table 8.6.3 *Determined Expected Values for Data used for the Illustration of the Test for the Independence of Two Groups.*

Modes of students admission	**Grade Points of students in each Programme**				**Row Total**
	Mechanical Engineering		Electrical Engineering		
	Observed values	Expected values	Observed values	Expected Values	
Regular	4.2	4.1	3.1	3.2	7.3
Access	3.2	3.7	3.4	2.9	6.6
Matured	3.6	3.2	2.1	2.5	5.7
Column Total	11		8.6		19.6

The chi-square value for the data can now be computed as follows:

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$x^2 = \sum_{i=1}^{n} \frac{(4.2 - 4.1)^2}{12.99} + \cdots + \frac{(3.6 - 3.2)^2}{17.10} + \frac{(3.1 - 3.2)^2}{25.01} + \cdots + \frac{(2.1 - 2.5)^2}{32.90}$$

$$x^2 = \sum_{i=1}^{3} \frac{(0.1)^2}{12.99} + \cdots + \frac{(0.4)^2}{17.10} + \frac{(-0.1)^2}{25.01} + \cdots + \frac{(-0.4)^2}{32.90}$$

$$x^2 = \sum_{i=1}^{3} \frac{(0.1)^2}{12.99} + \cdots + \frac{(0.4)^2}{17.10} + \frac{(-0.1)^2}{25.01} + \cdots + \frac{(-0.4)^2}{32.90}$$

$$x^2 = \sum_{i=1}^{3} \frac{(0.1)^2}{12.99} + \cdots + \frac{(0.4)^2}{17.10} + \frac{(-0.1)^2}{25.01} + \cdots + \frac{(-0.4)^2}{32.90}$$

$$x^2 = \sum_{i=1}^{3} \frac{0.01}{12.99} + \cdots + \frac{0.16}{17.10} + \frac{0.01}{25.01} + \cdots + \frac{0.16}{32.90}$$

$$x^2 = \sum_{i=1}^{3} 0.002 + \cdots + 0.050 + 0.003 + \cdots + 0.020 = 0.23$$

$$x^2 = 0.23$$

$$df = (n_1 - 1)(n_2 - 1) = (3 - 1)(2 - 1) = 3$$

The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it from one, and then look it up (ie: 0.05 on the left is 0.95 on the right)

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838

Figure 8.6.1 *Chi-square Test Table.*

The read critical values from the table at both 1% and 5% levels of significance are:

$$x^2_{critical}(\,at\ df = 1, \alpha = 0.05\,) = 7.82$$

$$x^2_{critical}(\,at\ df = 1, \alpha = 0.01\,) = 11.35$$

It should be noted that with the testing of the independency of two variables or groups data attribute. These hypotheses are put forward:

$$H_0: the\ two\ groups\ are\ independent\ of\ eacher\ other$$

$$H_1: one\ variable\ or\ group\ depend\ on\ the\ other$$

Decision:

$$x^2_{calculated} = 0.23 < x^2_{critical}(\,at\ df = 1, \alpha = 0.05\,) = 7.82$$

$$x^2_{calculated} = 0.23 < x^2_{critical}(\ at\ df = 1, \alpha = 0.01\ ) = 11.35$$

Conclusion:

$$since\ x^2_{calculated}(0.23) < x^2_{critical}(7.82)\ at\ df = 1,$$
$$\alpha = 0.0\ and\ (11.35)\ at\ df = 3, \alpha = 0.01$$

we fail to reject the null hypothesis.

It means the data supports the null hypothesis which is the mode of admission is independent of the performance of the students. The data collected is similar for each trial.

Analysis of Variance (ANOVA)

Analysis of Variance has the acronym (ANOVA). It is used to test whether there is a difference between the means of two or more continuous dependent variables or to test the difference in the means of a single continuous dependent variable determined at two or more different times. Whereas the t-test employs the t-statistics, ANOVA employs the F-test or the F-statistics. When finding the difference between two groups, both the t-statistics and the F-statistics give the same results. However ANOVA can be used to test whether or not differences exist between two or more groups whereas t-test can be used to test whether there is a difference in the means of just two groups at a time.

Assuming a researcher collects data on the performances in terms of the final grade points obtained by students enrolled in mechanical, electrical and civil engineering at the end of their programmes and summarized it in a table 8.7.1.

Table 8.7.1 *Data for the Illustration of ANOVA.*

Grade Points of students in each Programme		
Mechanical Engineering (m)	Civil Engineering (c)	Electrical Engineering (e)
4.2	3.2	3.1
3.2	2.1	3.4
3.6	4.1	2.1
3.7	2.3	4.4
2.3	4.2	3.7

The researcher can test whether means of the performance of the students in the different programmes differ or are the same using ANOVA.

To use the data in the table above to test whether the mean performances of the students differ between the programmes, the researcher must do the calculations below.

After the data have been collected the researcher needs to calculate the following from the data:

- Sum of Scores of each group = $\sum x$.
- Mean of Scores of each group = $\bar{x}$.
- Sum of the Squares of Scores of each group = $\sum x^2$.
- Summation of the Sums of the Scores for all the groups = $\sum(\sum x)$.
- Summation of the Sums of squares of the Scores for all the groups = $\sum(\sum x^2)$.

Let us proceed to calculate these variables for the data:

First, we calculate the sums of the scores, the sums of squares and the means of each group

$$\sum x_m = m_1 + m_2 + m_3 \ldots\ldots + m_n$$

$$\sum x_m = 4.2 + 3.2 + 3.6 \ldots + 2.3 = 17$$

$$\bar{x}_m = \frac{\sum x_m}{n_m} = \frac{17}{5} = 3.40$$

$$\sum x_c = c_1 + c_2 + c_3 \ldots\ldots + c_n$$

$$\sum x_c = 3.2 + 2.1 + 4.1 \ldots + 4.2 = 15.9$$

$$\bar{x}_c = \frac{\sum x_c}{n_c} = \frac{15.9}{5} = 3.18$$

$$\sum x_e = e_1 + e_2 + e_1 \ldots\ldots + e_n$$

$$\sum x_e = 3.1 + 3.4 + 2.1 \ldots + 3.7 = 16.7$$

$$\bar{x}_e = \frac{\sum x_e}{n_e} = \frac{16.7}{5} = 3.34$$

$$\sum {x_m}^2 = {m_1}^2 + {m_2}^2 + {m_3}^2 \ldots + {m_n}^2$$

$$\sum {x_m}^2 = 4.2^2 + 3.2^2 + 3.6^2 \ldots + 2.3^2 = 59.82$$

$$\sum {x_c}^2 = {c_1}^2 + {c_2}^2 + {c_3}^2 \ldots + {c_n}^2$$

$$\sum {x_c}^2 = 3.2^2 + 2.1^2 + 4.1^2 \ldots + 4.2^2 = 54.39$$

$$\sum {x_e}^2 = {e_1}^2 + {e_2}^2 + {e_3}^2 \ldots + {e_n}^2$$

$$\sum {x_e}^2 = 4.2^2 + 3.2^2 + 3.6^2 \ldots + 2.3^2 = 58.63$$

The next step is to find the summation of the sums of scores of each group; and the summation of the sums of squares of scores for each group:

$$\sum(\sum x) = \sum x_m + \sum x_c + \sum x_e = 17 + 15.9 + 16.7 = 49.6$$

$$\sum(\sum x^2) = \sum {x_m}^2 + \sum {x_c}^2 + \sum {x_e}^2 = 59.82 + 54.39 + 58.63 = 172.84$$

Calculate for the Total Sum of Squares; Sum of Squares between groups; and the Sum of Squares within groups

$$SS_T = \sum\left(\sum x^2\right) - \frac{(\sum(\sum x)^2}{n_T}$$

$$SS_T = \sum\left(\sum x^2\right) - CF$$

$$CF = \frac{(\sum(\sum x)^2}{n_T}$$

$$SS_T = \sum 172.84 - \frac{(49.6)^2}{15}$$

$$SS_T = \sum 172.84 - \frac{2460.16}{15} = 172.84 - 164.01 = 8.83$$

CF=Correction factorm, n_T=n_m+n_c+n_e =Total number of scores, SS_T=Total Sum of Squares n=n_e=n_e=n_e=number scores for each group

$$SS_B = \sum\frac{(\sum x)^2}{n} - \frac{(\sum(\sum x)^2}{n_T}$$

$$SS_B = \sum(\frac{(x_m)^2}{n_m} + \frac{(x_c)^2}{n_c} + \frac{(x_e)^2}{n_e}) - \frac{(\sum(\sum x)^2}{n_T}$$

$$SS_B = \sum\left(\frac{(17)^2}{5} + \frac{(15.9)^2}{5} + \frac{(16.7)^2}{5}\right) - \frac{(49.6)^2}{15}$$

$$SS_B = \sum\left(\frac{289}{5} + \frac{252.81}{5} + \frac{278.89}{5}\right) - \frac{2460.16}{15}$$

$$SS_B = \sum(57.8 + 50.56 + 55.78) - 164.01 = 164.14 - 164.01 = 0.13$$

SS_B=Sum of Squares between the groups, n=n_m=n_c=n_e=number scores for each group

$$SS_T = SS_B + SS_W$$

Therefore

$$SS_W = SS_T - SS_B$$

$$SS_W = 8.83 - 0.13 = 8.7$$

Where SS_W=Sum of Squares within the groups

$$df_T = df_B + df_W$$

$$df_T = n - 1 = 15 - 1 = 14$$

$$df_B = n - 1 = 3 - 1 = 2 \;\; df_W = N_T - \; n = 15 - 3 = 12$$

The Means Squares of between and within groups are then computed:

$$MS_B = \frac{SS_B}{df_B}$$

$$MS_B = \frac{0.13}{2} = 0.07$$

$$MS_W = \frac{SS_W}{df_W}$$

$$MS_W = \frac{8.7}{12} = 0.73$$

where MS_B=Mean Square between groups; MS_W=Mean Squares within groups; df_B=between groups degree of freedom; df_W=within groups degree of freedom.

Finally compute the F-calculated value and read F-critical values from F-test table for comparison:

$$F = \frac{MS_B}{MS_W}$$

$$F_{calculated} = \frac{0.07}{0.73} = 0.10$$

Reading of F-critical values at 5% and 1% levels of significance shown below:

Critical values of F for the 0.05 significance level:

	1	2	3	4	5	6
1	161.45	199.50	215.71	224.58	230.16	233.99
2	18.51	19.00	19.16	19.25	19.30	19.33
3	10.13	9.55	9.28	9.12	9.01	8.94
4	7.71	6.94	6.59	6.39	6.26	6.16
5	6.61	5.79	5.41	5.19	5.05	4.95
6	5.99	5.14	4.76	4.53	4.39	4.28
7	5.59	4.74	4.35	4.12	3.97	3.87
8	5.32	4.46	4.07	3.84	3.69	3.58
9	5.12	4.26	3.86	3.63	3.48	3.37
10	4.97	4.10	3.71	3.48	3.33	3.22
11	4.84	3.98	3.59	3.36	3.20	3.10
12	4.75	3.89	3.49	3.26	3.11	3.00
13	4.67	3.81	3.41	3.18	3.03	2.92
14	4.60	3.74	3.34	3.11	2.96	2.85
15	4.54	3.68	3.29	3.06	2.90	2.79

Critical values of F for the 0.01 significance level:

	1	2	3	4	5	6
1	4052.19	4999.52	5403.34	5624.62	5763.65	5858.97
2	98.50	99.00	99.17	99.25	99.30	99.33
3	34.12	30.82	29.46	28.71	28.24	27.91
4	21.20	18.00	16.69	15.98	15.52	15.21
5	16.26	13.27	12.06	11.39	10.97	10.67
6	13.75	10.93	9.78	9.15	8.75	8.47
7	12.25	9.55	8.45	7.85	7.46	7.19
8	11.26	8.65	7.59	7.01	6.63	6.37
9	10.56	8.02	6.99	6.42	6.06	5.80
10	10.04	7.56	6.55	5.99	5.64	5.39
11	9.65	7.21	6.22	5.67	5.32	5.07
12	9.33	6.93	5.95	5.41	5.06	4.82
13	9.07	6.70	5.74	5.21	4.86	4.62
14	8.86	6.52	5.56	5.04	4.70	4.46
15	8.68	6.36	5.42	4.89	4.56	4.32

Figure 8.7.1 *F-critical Table.*

In order to read the F-critical values at the respective levels of significance, one must use the between groups degree of freedom along the row (horizontal) and then the within groups degree of freedom (df) along the column (vertical). The meeting points of these degrees of freedom on the table lies the F-critical values respectively for each chosen level of significance table.

Thus from the tables, the following are read as the F-critical values for the case being dealt with:

$$F_{critical}\,(2, 14\ df and\ 5\%) = 3.74;\ and\ F_{critical}\,(2, 14\ df and\ 1\%) = 6.52$$

Now we can complete the ANOVA table:

Table 8.7.2 *Completed ANOVA Table.*

Sources of Variations	df	SS	MS	F-calc.	Fcrit at 5%	Fcrit at 1%
Between groups	2	0.13	0.07	0.10	3.74	6.52
Within groups	14	8.7	0.73			
Total	14	8.83				

Now the researcher can compare the F-calculated values with the F-critical values at both levels of significance to take the decision and make conclusions on the case.

It should however be noted that whenever F-calculated is greater than F-critical at stated degrees of freedom and a given level of significance (α), the p-value (p-value < α) would also be less than the chosen level of significance. The contrary also holds i.e. if F-calculated is lesser than F-critical, p-value (p-value >α) would be less than the chosen level of significance.

Decision:

if $F_{(Calculated)} > F_{(Critical)}$ at (5% or 1%, df=2, 14); it means the performances of the students in the various programmes are significantly different.

However, if $F_{(Calculated)} < F_{(Critical)}$ (at 5% or 1%, df=2, 14); it suggests that the performances of the students are not significantly different.

For the case understudy:

$$F_{Calculated}\,(0.10) < F_{Critical}\,(at\ 5\%, df = 2, 14) = 3.74$$

$$F_{Calculated}\,(0.10) < F_{Critical}\,(at\ 1\%, df = 2, 14) = 6.52$$

Conclusion:

It can therefore be concluded that there is enough evidence presented by the data collected from the study that the performances of the students in the respective programmes are not significantly different or in other words are the same at both levels of significance.

Table I: Chi-Square Probabilities

The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it from one, and then look it up (ie: 0.05 on the left is 0.95 on the right)

Df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Table II: Standard Normal (Z)

Area between 0 and z

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Table III: Student's T Table

T Table With Right Tail Probabilities

df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
inf	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

Table IV: Chi-Square Table

df\ area	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05
1	0.00004	0.00016	0.00098	0.00393	0.01579	0.10153	0.45494	1.3233	2.70554	3.84146
2	0.01003	0.0201	0.05064	0.10259	0.21072	0.57536	1.38629	2.77259	4.60517	5.99146
3	0.07172	0.11483	0.2158	0.35185	0.58437	1.21253	2.36597	4.10834	6.25139	7.81473
4	0.20699	0.29711	0.48442	0.71072	1.06362	1.92256	3.35669	5.38527	7.77944	9.48773
5	0.41174	0.5543	0.83121	1.14548	1.61031	2.6746	4.35146	6.62568	9.23636	11.0705
6	0.67573	0.87209	1.23734	1.63538	2.20413	3.4546	5.34812	7.8408	10.64464	12.59159
7	0.98926	1.23904	1.68987	2.16735	2.83311	4.25485	6.34581	9.03715	12.01704	14.06714
8	1.34441	1.6465	2.17973	2.73264	3.48954	5.07064	7.34412	10.21885	13.36157	15.50731
9	1.73493	2.0879	2.70039	3.32511	4.16816	5.89883	8.34283	11.38875	14.68366	16.91898
10	2.15586	2.55821	3.24697	3.9403	4.86518	6.7372	9.34182	12.54886	15.98718	18.30704
11	2.60322	3.05348	3.81575	4.57481	5.57778	7.58414	10.341	13.70069	17.27501	19.67514
12	3.07382	3.57057	4.40379	5.22603	6.3038	8.43842	11.34032	14.8454	18.54935	21.02607
13	3.56503	4.10692	5.00875	5.89186	7.0415	9.29907	12.33976	15.98391	19.81193	22.36203
14	4.07467	4.66043	5.62873	6.57063	7.78953	10.16531	13.33927	17.11693	21.06414	23.68479
15	4.60092	5.22935	6.26214	7.26094	8.54676	11.03654	14.33886	18.24509	22.30713	24.99579
16	5.14221	5.81221	6.90766	7.96165	9.31224	11.91222	15.3385	19.36886	23.54183	26.29623
17	5.69722	6.40776	7.56419	8.67176	10.08519	12.79193	16.33818	20.48868	24.76904	27.58711
18	6.2648	7.01491	8.23075	9.39046	10.86494	13.67529	17.3379	21.60489	25.98942	28.8693
19	6.84397	7.63273	8.90652	10.11701	11.65091	14.562	18.33765	22.71781	27.20357	30.14353
20	7.43384	8.2604	9.59078	10.85081	12.44261	15.45177	19.33743	23.82769	28.41198	31.41043
21	8.03365	8.8972	10.2829	11.59131	13.2396	16.34438	20.33723	24.93478	29.61509	32.67057
22	8.64272	9.54249	10.98232	12.33801	14.04149	17.23962	21.33704	26.03927	30.81328	33.92444
23	9.26042	10.19572	11.68855	13.09051	14.84796	18.1373	22.33688	27.14134	32.0069	35.17246
24	9.88623	10.85636	12.40115	13.84843	15.65868	19.03725	23.33673	28.24115	33.19624	36.41503
25	10.51965	11.52398	13.11972	14.61141	16.47341	19.93934	24.33659	29.33885	34.38159	37.65248
26	11.16024	12.19815	13.8439	15.37916	17.29188	20.84343	25.33646	30.43457	35.56317	38.88514
27	11.80759	12.8785	14.57338	16.1514	18.1139	21.7494	26.33634	31.52841	36.74122	40.11327
28	12.46134	13.56471	15.30786	16.92788	18.93924	22.65716	27.33623	32.62049	37.91592	41.33714
29	13.12115	14.25645	16.04707	17.70837	19.76774	23.56659	28.33613	33.71091	39.08747	42.55697
30	13.78672	14.95346	16.79077	18.49266	20.59923	24.47761	29.33603	34.79974	40.25602	43.77297

Table V: F Distribution Tables

F - Table for alpha = 0.10

df2/df1	1	2	3	4	5	6	7	8	9	10	12	15	20
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61

F -Table for alpha = 0.05

df2/df1	1	2	3	4	5	6	7	8	9	10	12	15
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75

F -Table for alpha = 0.025

df2/df1	1	2	3	4	5	6	7	8	9	10	12	15
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	976.71	984.87
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.95
Inf	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83

F -Table for alpha = 0.01

df2/df1	1	2	3	4	5	6	7	8	9	10	12
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6106.32
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89
6	13.75	10.93	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96
14	8.86	6.52	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.90	3.81	3.67
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46
18	8.29	6.01	5.09	4.58	4.25	4.02	3.84	3.71	3.60	3.51	3.37
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12
23	7.88	5.66	4.77	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96
27	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.15	3.06	2.93
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90
29	7.60	5.42	4.54	4.05	3.73	3.50	3.33	3.20	3.09	3.01	2.87
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.67
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34
inf	6.64	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.19

Table VI: Standard Normal (Z) Table

Values in the table represent areas under the curve to the left of Z quantiles along the margins.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Bibliography

[1] Anders, M. E. (1975). Statistical Information as a Basis for Cooperative Planning. *Library Trends*, 24: 229-244.

[2] Berdie, D. R., Anderson, J. F. (1974) *Questionnaires: Design and Use.* Metuchen, New Jersey, Scarecrow.

[3] Bruning, J. L., Kintz, B. L. (1977). *Computational Handbook of Statistics.* 2nd ed. Glenview, Illinois, Scott, Foresman.

[4] Childers, T. A. (1975). Statistics That Describe Libraries and Library Service. *In:* Voigt, M. J., ed.: *Advances in Librarianship.* New York, Academic Press.

[5] McHugh, M. L. (2003). Descriptive statistics, part I: level of measurement. *JSPN*, 8: 35-37.

[6] Norton, H. W. (1976). Review of Basic Statistics for Librarians, by I. S. Simpson. *Newsletter on Library Research*, 18: 8-10.

[7] Overholser, B. R., Sowinski, K. M. (2007). Biostatistics primer: part I. *Nutr Clin Pract.* 22: 629-635.

[8] Simpson, I. S. (1975). *Basic Statistics* for *Librarians.* London, Clive Bingley.

[9] Wallis, W. A., Roberts, H. V. (1956). *Statistics: A New Approach.* Glencoe, Illinois, Free Press.

Chapter 9

Transformation of Data

9 Transformation of Data

Felix Kutsanedzie[1]; Sylvester Achio[1]; Edmund Ameko[1]

[1]Accra Polytechnic, Accra, Ghana

Abstract

Data is collected for every research conducted. The collected date needs to be analysed using the appropriate statistical test. Every data collected can be either normally distributed or not. Those that are normally distributed are subjected to parametric tests in their analyses while those that are not are intend subjected to non parametric test. Parametric test are said to be more reliable than the parametric test because less assumptions are made on then compared with non parametric. However whenever data is collected from an experiment or a study and it is not normally distributed but a parametric test is subjected to it, the results become misleading and therefore unreliable. In order to use a parametric test on data that is not normally distributed, it is converted or transformed into normally distributed data. This chapter explains and demonstrate how to transform data that is not normally distributed to normally distributed data in order to apply parametric tests to analyse them. The SPSS software has been used to demonstrate how to transform data in that regard.

Keywords

Normal Distribution, Parametric, Non Parametric, Statistical Tests, Data

9.1 Introduction

Data collection and analysis are an indispensable aspect of research which great skills and expertise is needed to handle. In order to make meaningful and reliable findings in a research study, the accuracy of the data and the appropriateness of the tools used in analyzing the data must be employed. Statistical tests are grouped broadly into parametric and non parametric tests. Each of these tests has their strengths and weaknesses. For benefits to be derived from the use of these tests for data analysis data, the research must adopt the use of the right tests i.e. either parametric or non parametric for the analysis.

Parametric test are said to be stronger compared to non parametric tests because the latter are based on less assumptions and are less quantitative in nature. The main assumption that parametric tests is based on is that they must be applied on normally distributed data. This chapter would not focus on the assumptions underpinning both tests but on when to use both tests.

Whenever data is collected from a research study and inferential statistics is expected to be used in the analysis of the data, such data must be tested to ascertain whether they are normally distributed or not before an appropriate parametric or non parametric test is used for their analysis. When the data is proven to be normally distributed, then the appropriate test for the analysis would be a parametric tests if not then a non parametric test should be selected for the analysis.

For data confirmed to be normally distributed, it must satisfy some requirements or conditions such as: it must have the mean, mode and median to be equal; kurtosis and skewness must range between -1.96 to 1.96; a histogram of such data must follow the shape of the normal curve; data must follow a normal Q-Q plot trend; and a box-plot of the data must be symmetrical. Any

data that does not follow or approximately satisfy these conditions cannot be said to be normally distributed.

However, when the normality of data is tested and it proves not to be normally distributed, various methods can be used to transform the data to that which is anormally distributed in order for a parametric test to be used for its analysis. It is only when the transformed data still does not conform to a normally distributed one that the researcher can go ahead to analyse it with an appropriate non parametric test. The process of converting data that is not normally distributed to that which is normally distributed is referred to as *transformation.*

This chapter uses SPSS software to demonstrate how to check whether data collected from a researcher study is normally distributed considering the conditions and requirements for normality as well as show how to transform data that is not normally distributed.

9.2 Testing the Data Normality

It should be noted that it is the data collected on the dependent variables that are tested to be either normally distributed or not. The acronym SPSS stands for Statistical Package for Social Scientist and it is an application used for data summarizing and analyzing. SPSS has two windows (views) for handling of data – Data and Variable views where the collected data can be labeled and entered. The data view is the window where data is entered whiles the variable view is where the levels of measurement, label and codes of variables whose data had been entered are specified.

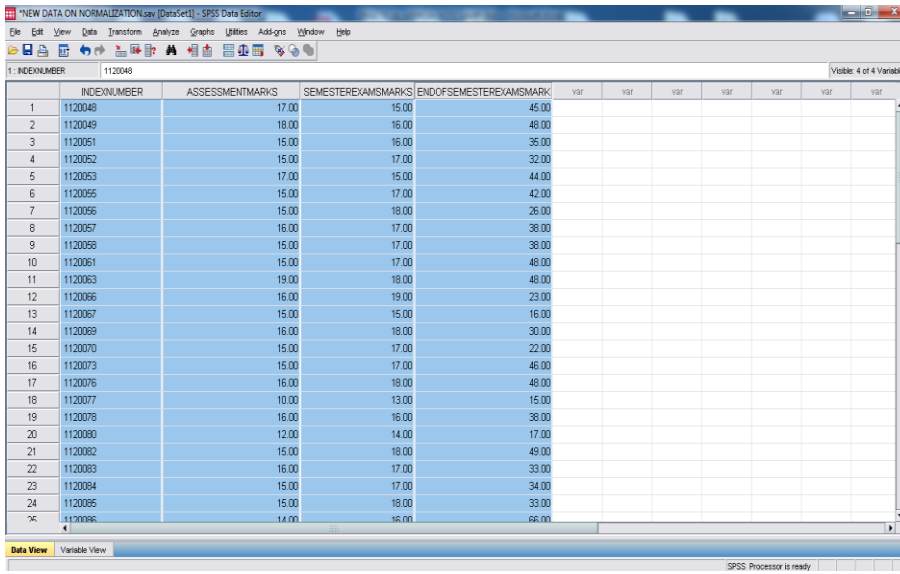


Figure 9.1 *Data view page.*

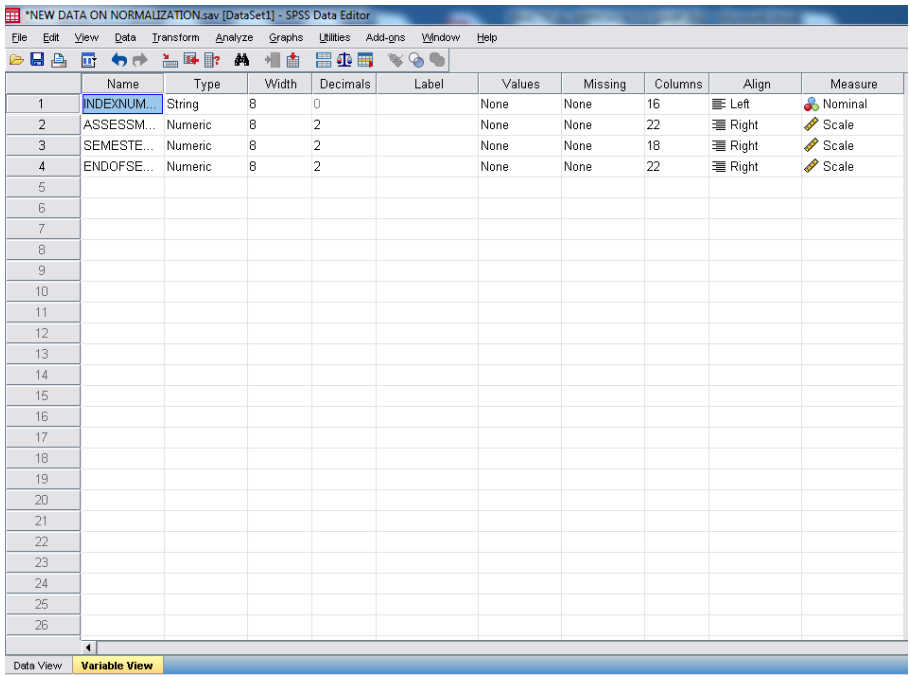


Figure 9.2 *Variable view page.*

The researcher can copy or import the data collected from Excel application into the Data view page, or better still if the data is not already typed, it can be typed directly into the cells in the Data view window as seen in Fig. 9.1. Once the data has been entered, the researcher must click on the variable view button of the page to open its window and then label appropriately the variables with their specifications – level of measurement, type of variable i.e either numeric or string etc. as shown in Figure 9.2.

Let us proceed to use the data below to demonstrate how to test for the normality of a collected data from a research study:

Table 9.1 *Data used for Illustration on the Test for Normality.*

Index No	Assessment Marks	Semester Exam Marks	End of Semester Exam Marks
1120048	17	15	45
1120049	18	16	48
1120051	15	16	35
1120052	15	17	32
1120053	17	15	44
1120055	15	17	42
1120056	15	18	26
1120057	16	17	38
1120058	15	17	38
1120061	15	17	48
1120063	19	18	48
1120066	16	19	23
1120067	15	15	16
1120069	16	18	30
1120070	15	17	22
1120073	15	17	46
1120076	16	18	48
1120077	10	13	15
1120078	16	16	38

Index No	Assessment Marks	Semester Exam Marks	End of Semester Exam Marks
1120080	12	14	17
1120082	15	18	49
1120083	16	17	33
1120084	15	17	34
1120085	15	18	33
1120086	14	16	66
1120087	19	17	28
1120090	15	18	34
1120091	16	16	44
1120094	15	18	46
1120097	17	18	53
1120098	15	17	46
1120099	15	18	26
1120103	15	18	35

After the data has been entered and label, the *analyze command button* at the top part is pressed, and cursor moved to descriptive statistics and then to *'explore'* as indicated in Figure 9.3.

The '*explore'* button is then clicked to open another window shown below where the cursor is placed on dependent variable whose normality is to be tested and then moved by clicking on the '*arrow'* to send into the dialogue box provided for the dependent list as shown in Figure 9.4. One can highlight on all the dependent variables in the box if he or she needs to test their normality and then move them all into the dialogue box provided by click on the '*arrow'*. But for this illustration let us use only the dependent variable "End of Semester Exams Mark".

Once the dependent variable whose normality is to be tested is moved into box of the dependent list, click on '*plots'* and then use the cursor to check the

boxes for ‘histogram’, ‘dependents together’ and ‘normality plots with tests’ after which one must click on the continue button as shown in Figure 9.5.

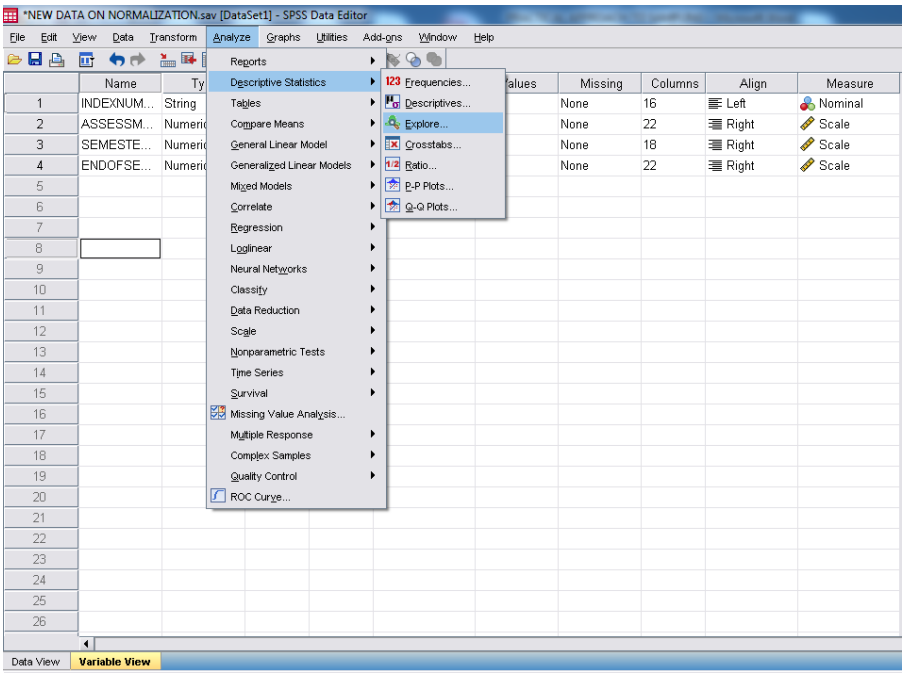
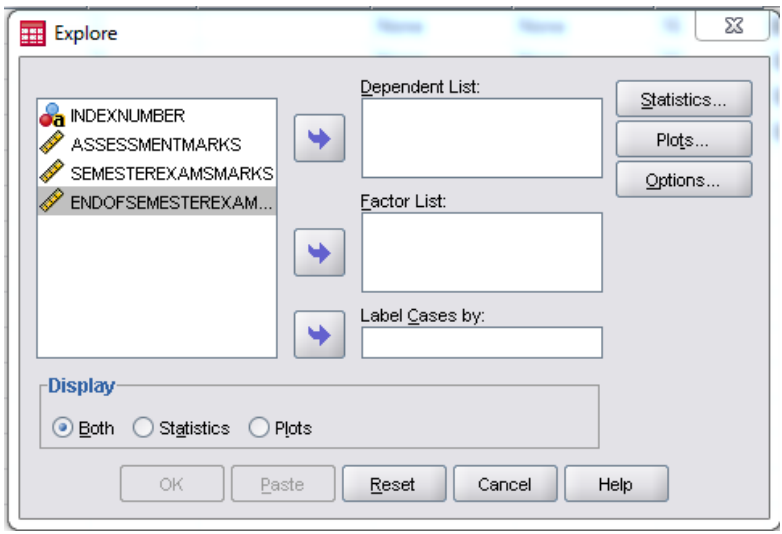


Figure 9.3 *Outcome of Clicking the Analyze Button.*



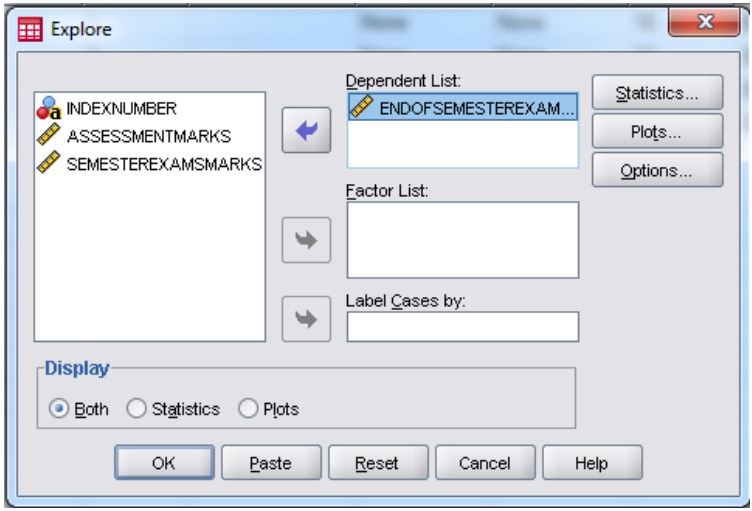


Figure 9.4 *Outcome of Clicking the Explore button.*

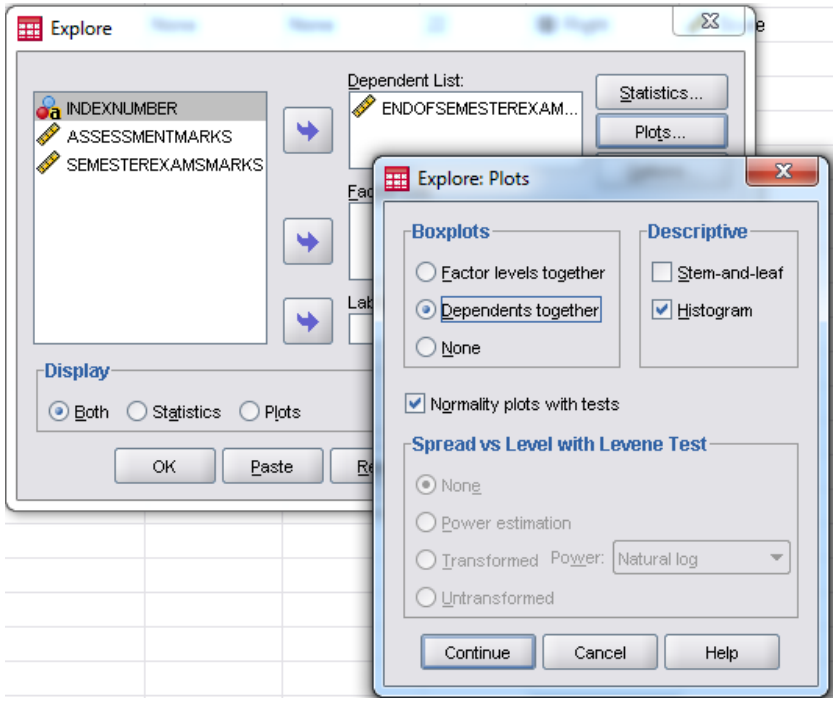


Figure 9.5 *Outcome of Clicking on the 'Plots' button.*

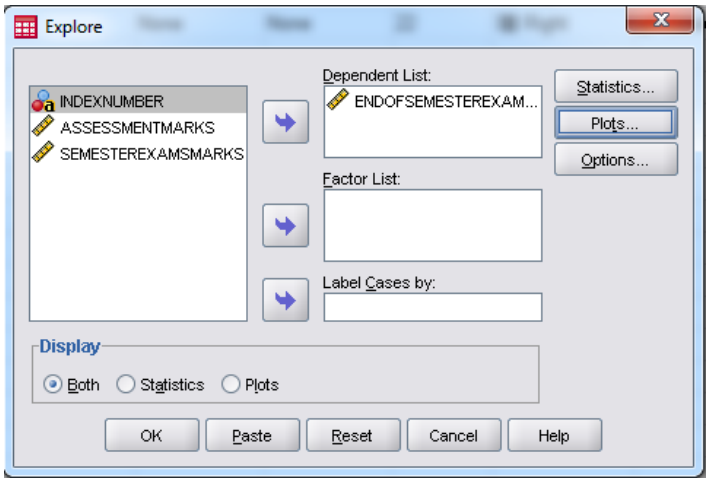


Figure 9.6 *Explore window with 'both' checked in the display box.*

By clicking on the *'continue'* button in the opened window sends one back to the explore window. The researcher can then click check *'both'* in the display box within the explore window (shown in Figure 9.5) and then proceed to click on *'Ok'* to obtain the output in an output window.

Now proceed to analyse the outcomes of the output box one-by-one juxtaposing them with the requirements for a data to be considered as normally distributed as follows:

Check the values for the mean, median and mode of the Data

Table 9.2.1 *Statistics.*

Mean		**37.1515**
Median		38.0000
Mode		48.00

From the table above obtained from the output of the SPSS application on the data being tested for, the mean and the median are not equal though close to each other. Since the output generated from the normality shown in the table

above reveals that the mean (37.15), median (38.0) and mode (48.00) of the data are not equal. Therefore the data is likely not to be normally distributed. It should be noted that the modal value of the data was obtained by clicking on the analyze – frequency – statistics -'check mode box' buttons in that logical sequence.

Compare the value of Skewness and Kurtosis for the Data

Table 9.2.2 *Descriptives.*

			Statistic	Std. Error
Endofsemesterexamsmark	Mean		37.1515	2.03153
	95% Confidence Interval for Mean	Lower Bound	33.0134	
		Upper Bound	41.2896	
	5% Trimmed Mean		37.0421	
	Median		38.0000	
	Variance		136.195	
	Std. Deviation		1.16703E1	
	Minimum		15.00	
	Maximum		66.00	
	Range		51.00	
	Interquartile Range		17.00	
	Skewness		-.021	.409
	Kurtosis		-.054	.798

For data to be considered as normally distributed, the z-scores for skewness and kurtosis must range between -1.96 to 1.96. In order to convert the skewness and kurtosis values to z-scores, the researcher must divide the kurtosis and skewness values by their respective standard error besides them as shown in the table above.

$$Z - scores\ for\ Skewness = \frac{-0.021}{0.409} = -0.05$$

$$Z - scores\ for\ Kurtosis = \frac{-0.054}{0.798} = -0.07$$

The values fall within the z-scores of -1.96 to 1.96 but though the data meets this requirement, this alone cannot indicate that the data is normally distributed or not.

The Normality test values

Table 9.2.3 *Tests of Normality.*

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
Endofsemesterexamsmark	.115	33	.200*	.968	33	.435
Lilliefors Significance Correction						
*. This is a lower bound of the true significance.						

For data to pass the Kolmogorov-Smirnov test and be considered as normally distributed, the p-value (indicated in the table above as sigma) must be less than the level of significance (α) which in this case is 5% (0.05). Since the p-value or sigma is 0.20 and greater than an (α) = 5% (0.05), then the data is not normally distributed. For Sharpiro-Wilk test of normality, a data can be considered as normally distributed only when the p-value (sigma) is greater than the level of significance (α) which in this case is 5% (0.05). For this data though the p-value (sigma) according to Shapiro-Wilk test (0.435) is greater than the level of significance (α) which in this case is 5% (0.05), the data is not normally distributed. It thus should be noted that the Shapiro-Wilk test is not always right and therefore cannot be used as the only conclusive test to determine the normality of data.

Examine the trend of Normal Q-Q Plot for the Data

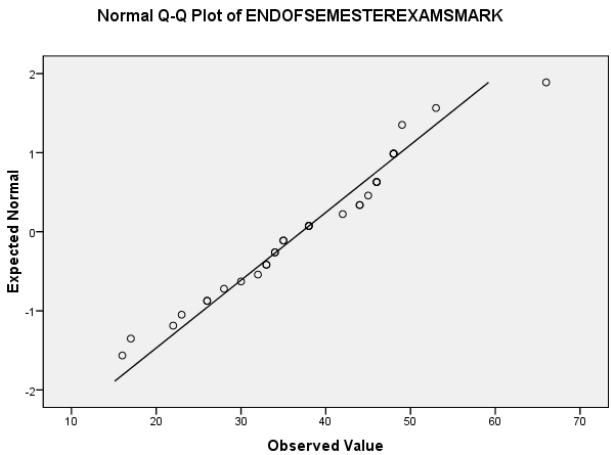


Figure 9.7.1 *Normal Q-Q Plot for the Data.*

For a normally distributed data, all or most of data points must lie exactly on or almost close to the trend line shown in a Normal Q-Q Plot. From the Normal Q-Q Plot for the data tested as shown above, most of the points do not fall on or close to the trend line and hence the data cannot be said to be normally distributed.

Examine the shape of the Histogram plot for the Data

When data is normally distributed, a drawn histogram for such data must look and have a shape of a normal curve with the data histogram formed being bilaterally symmetrical. However, the histogram for the data being tested as shown above does not show such characteristics and therefore cannot be said to be normally distributed.

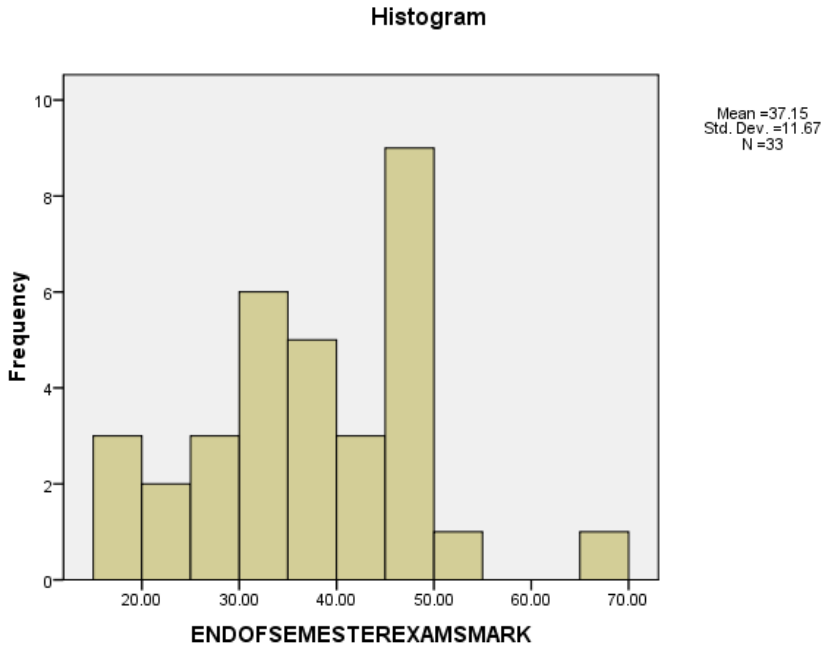


Figure 9.7.2 *Histogram for the Data.*

Examine the Box –Plot for the Data

The Box-Plot just like the histogram must show that the data have two bilateral symmetrical halves for it to be considered as normally distributed. The upper end to the middle portion or the mid section is more stretched as compared with the lower end to the mid section of the plot. This shows that the data is not normally distributed.

Thus since majority of the tests do not meet the conditions or the requirements for the data to be passed as normally distributed, the researcher can proceed to try and see whether the data can be transformed to become normally distributed before a parametric test can be used for its analysis. If all transformations applied do not succeed in converting the data to a normally

distributed one, then the researcher can select an appropriate non parametric test for its analysis.

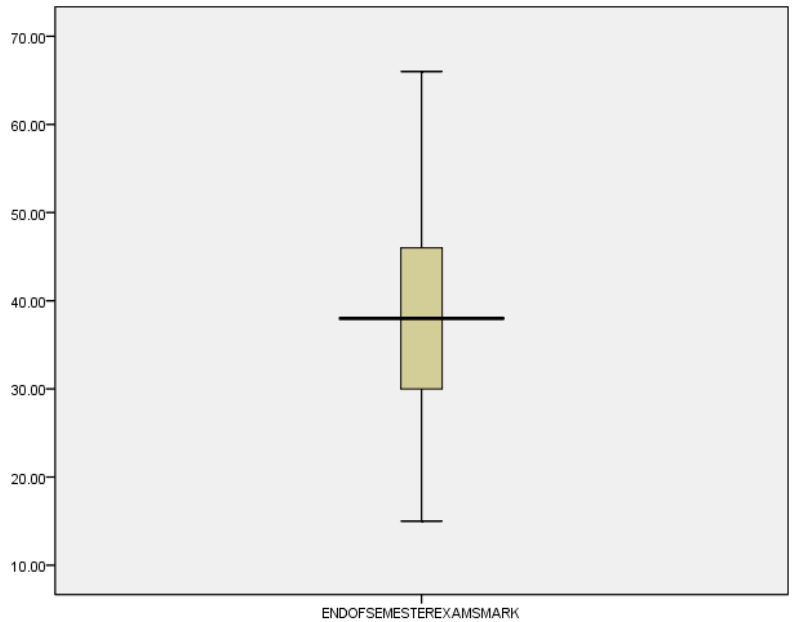


Figure 9.7.3 *Box-Plot for the Data.*

9.3 Transformation of Data

Since all the requirements for the normality test suggest that the data being considered is not normally distributed, the researcher must now try and transform the data. For one to transform the data using the SPSS application, the following described procedures must be used:

Description of Data Transformation Procedure

To transform the existing data, click on the *'transform'* button and then move to the *'compute'* variable button and click it to open a new window – compute variable window as shown in the Figure 9.8.1.

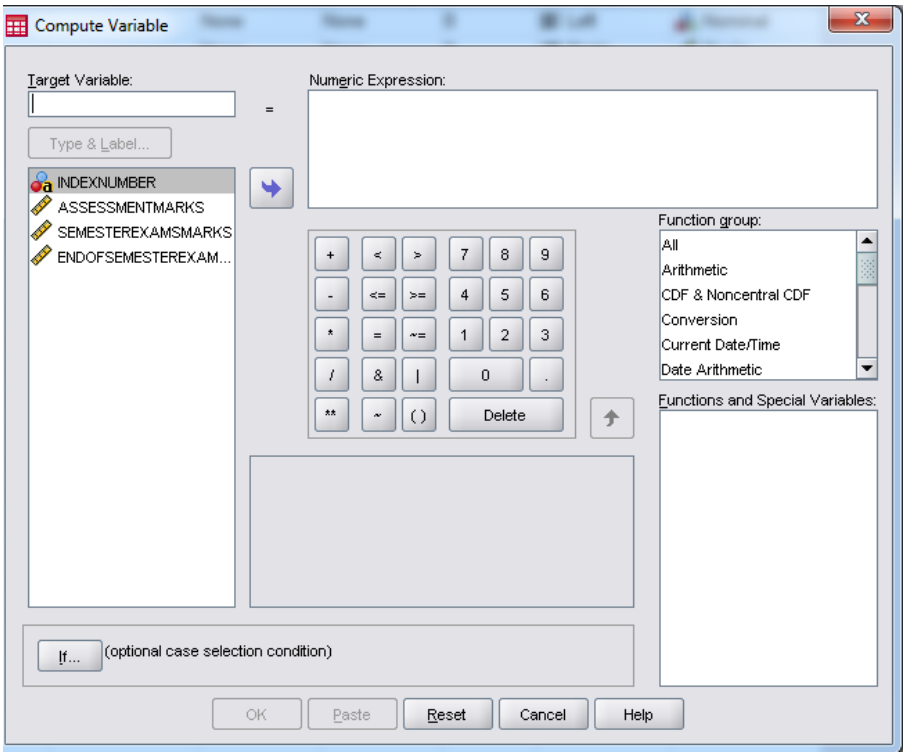


Figure 9.8.1 *Outcome of clicking on the 'compute variable' button.*

Type the variable to be transformed in the 'Target Variable' bar and then click on the 'type and label dialogue box' that appears to give the new label for the data to be transformed. One must then move the cursor to function group dialogue box and click on arithmetic, which will then enable the various 'functions and special variables box' from which one can select the function to use to perform the transformation. For this example, let us use the function 'log_{10}' as shown in Figure 9.8.2.

Click on the function 'log_{10}' and then move the cursor to the 'type and label' dialogue box and click on the target variable and then use the enabled 'arrow' to

move the target variable into the numeric expression dialogue box by clicking on it as shown in Figure 9.8.3.

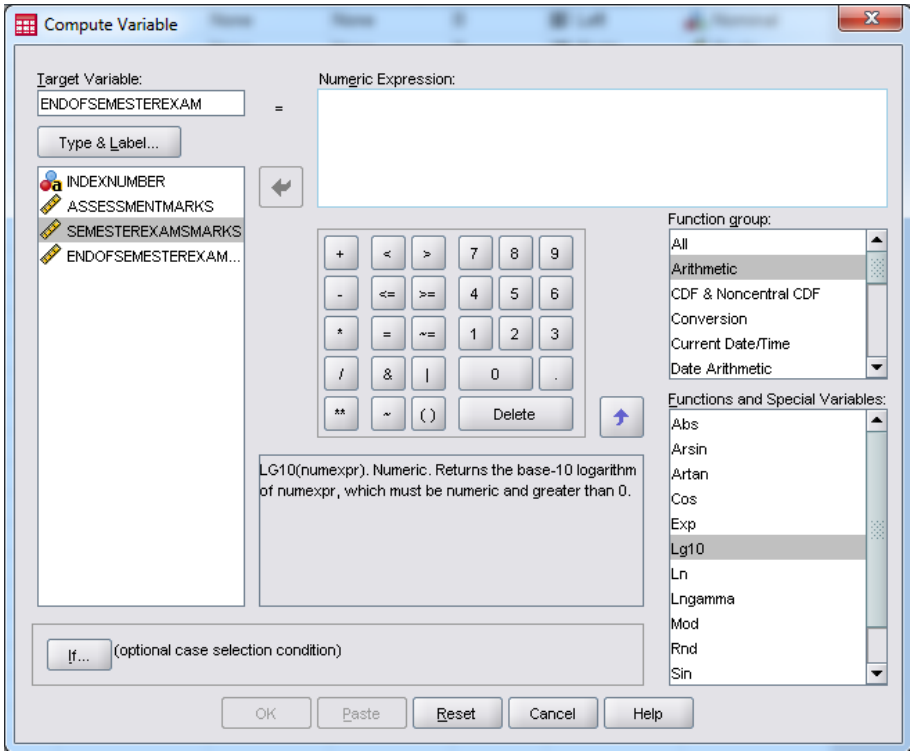


Figure 9.8.2 *Outcome of clicking the Arithmetic under the 'function group dialogue box'.*

Once this has been done the researcher can then move the cursor and click on the 'Ok' button to transform the target variable data to the indicated new label that have been assigned to it. The transformed data with the new label can be seen to be added to the other variables already in the data view window as shown in Figure 9.8.4.

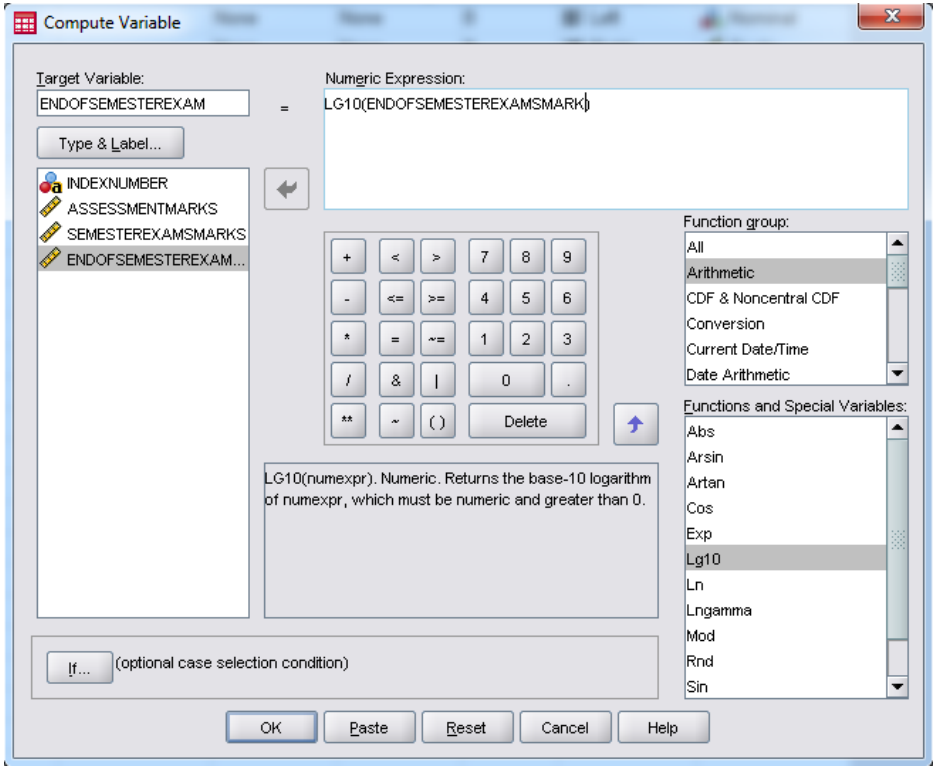


Figure 9.8.3 *Outcome of clicking on the function '$\log_{10}$' and moving the target variable into the numeric expression dialogue box.*

Now that the data on one dependent variable has been transformed, there is a need to test if the transformed variable would pass the variables normality tests and requirement as done previously. Thus needs to go through the process of testing the transformed data.

	INDEXNUMBER	ASSESSMENTMARKS	SEMESTEREXAMSMARKS	ENDOFSEMESTEREXAMSMARK	ENDOFSEMESTEREXAM	var
1	1120048	17.00	15.00	45.00	1.65	
2	1120049	18.00	16.00	48.00	1.68	
3	1120051	15.00	16.00	35.00	1.54	
4	1120052	15.00	17.00	32.00	1.51	
5	1120053	17.00	15.00	44.00	1.64	
6	1120055	15.00	17.00	42.00	1.62	
7	1120056	15.00	18.00	26.00	1.41	
8	1120057	16.00	17.00	38.00	1.58	
9	1120058	15.00	17.00	38.00	1.58	
10	1120061	15.00	17.00	48.00	1.68	
11	1120063	19.00	18.00	48.00	1.68	
12	1120066	16.00	19.00	23.00	1.36	
13	1120067	15.00	15.00	16.00	1.20	
14	1120069	16.00	18.00	30.00	1.48	
15	1120070	15.00	17.00	22.00	1.34	
16	1120073	15.00	17.00	46.00	1.66	
17	1120076	16.00	18.00	48.00	1.68	
18	1120077	10.00	13.00	15.00	1.18	
19	1120078	16.00	16.00	38.00	1.58	
20	1120080	12.00	14.00	17.00	1.23	
21	1120082	15.00	18.00	49.00	1.69	
22	1120083	16.00	17.00	33.00	1.52	
23	1120084	15.00	17.00	34.00	1.53	

Data View | Variable View

Figure 9.8.4 *Transformed data shown in the Data view window.*

Testing to check if the transformed data passed the normality tests and requirements:

Click on the analyse button, move the cursor to descriptive and then to explore. Click on the explore button to open the explore window. Place the cursor on the label of the transformed data and then use the enabled arrow to move it into the dependent list dialogue box. Continue by clicking on the ‘plots’ button in the window and then check the histogram and the normality tests boxes. Proceed by clicking the ‘Continue’ button to return to the explore window. In the explore window, check ‘both’ in the display dialogue box and

then click on ‘Ok’ for the results or the outputs of the normality tests. Now let us compare the results and check whether the transformation has succeeded in converting the data of the transformed variable to become normally distributed.

Check the values for the mean, median and standard deviation of the Data

Table 9.3.1 *Statistics.*

Mean		**1.5458**
Median		1.5798
Mode		1.68

Since the output generated from the normality shown in the table above reveals that the mean (1.55≈2), median (1.58≈2) and mode (1.68≈2) for the data are not equal. This shows that is not normally distributed.

Compare the value of Skewness and Kurtosis for the Data

Table 9.3.2 *Descriptives.*

			Statistic	**Std. Error**
ENDOFSEMESTEREXAMS2	Mean		1.5458	.02687
	95% Confidence Interval for Mean	Lower Bound	1.4911	
		Upper Bound	1.6006	
	5% Trimmed Mean		1.5526	
	Median		1.5798	
	Variance		.024	
	Std. Deviation		.15436	
	Minimum		1.18	
	Maximum		1.82	
	Range		.64	
	Interquartile Range		.20	
	Skewness		-.857	.409
	Kurtosis		.356	.798

For data to be considered as normally distributed, the z-scores for skewness and kurtosis must range between -1.96 to 1.96.

$$Z - scores\ for\ Skewness = \frac{-0.857}{0.409} = -2.09$$

$$Z - scores\ for\ Kurtosis = \frac{0.356}{0.798} = 0.45$$

The skewness falls outside the range whiles the kurtosis Z-score value values fall within the range. Since skewness fell within the range and the kurtosis outside it, it means the data is not normally distributed. They should all fall within the range before they can be said to be normally distributed.

The Normality test values

Table 9.3.3 *Tests of Normality.*

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ENDOFSEMESTEREXAMS2	.130	33	.167	.925	33	.026
a. Lilliefors Significance Correction						

For data to pass the Kolmogorov-Smirnov test, since the p-value or sigma is 0.167 and greater than $\alpha = 5\%$ (0.05), then the data is not normally distributed. For Sharpiro-Wilk test of normality, since the p-value (sigma) = 0.026 is less than the level of significance $(\alpha) = 0.05$, the data is not normally distributed.

Examine the trend of Normal Q-Q Plot for the Data

For a normally distributed data, all or most of data points must lie exactly on or almost close to the trend line shown in a Normal Q-Q Plot. From the Normal Q-Q Plot for the data tested as shown above, most of the points do not fall on or close to the trend line and hence the data cannot be said to be normally distributed.

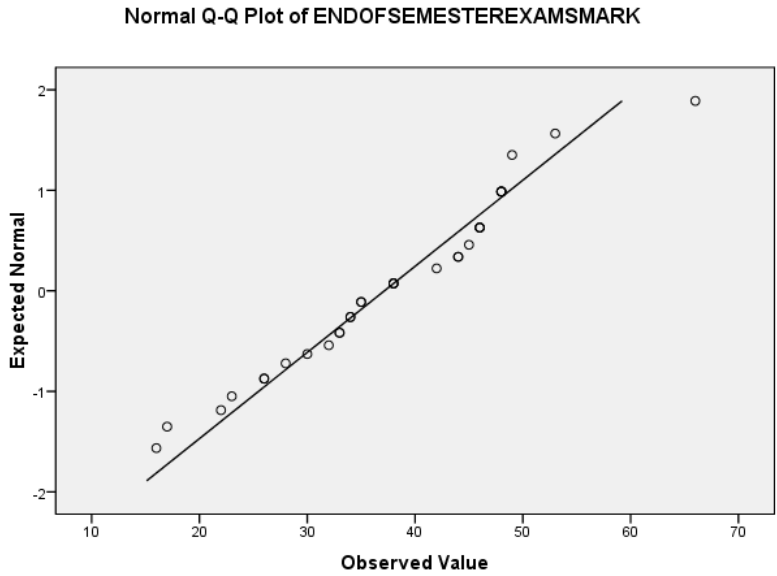


Figure 9.8.5 *Normal Q-Q Plot for the Data.*

Examine the shape of the Histogram plot for the Data

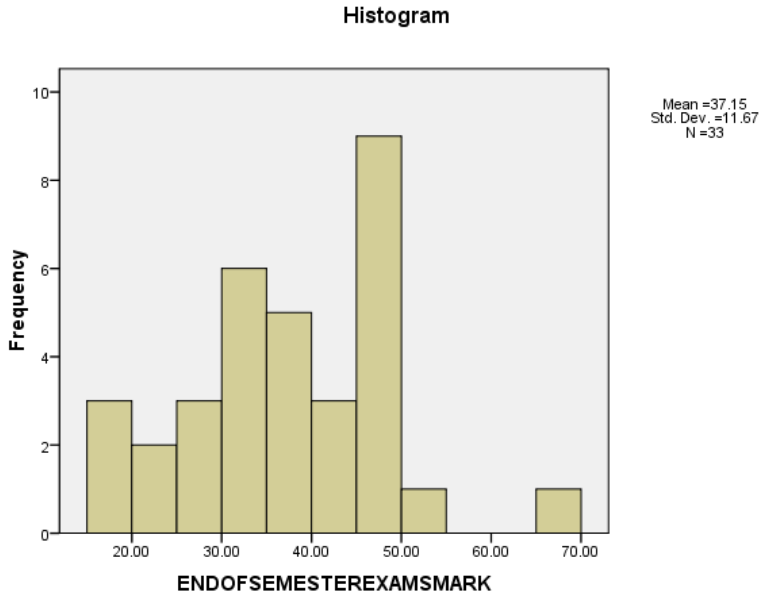


Figure 9.8.6 *Histogram for the Data.*

When a data is normally distributed, a drawn histogram for such data must look and have a shape of a normal curve with the data histogram formed being bilaterally symmetrical. However, the histogram for the data being tested as shown above does not show such characteristics and therefore cannot be said to be normally distributed.

Examine the Box –Plot for the Data

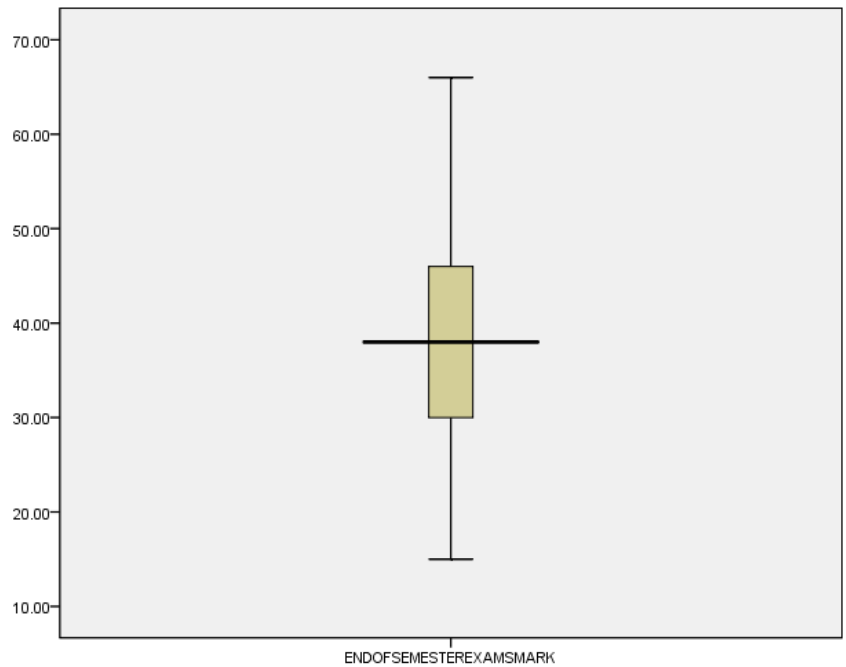


Figure 9.8.7 *Box-Plot for the Data.*

The Box-Plot just like the histogram must have two symmetrical halves for a data to be considered normally distributed. Since the histogram for the data does not show this characteristic, the data is not normally distributed.

If all transformations applied do not succeed, then the researcher can proceed to select an appropriate non parametric test to analyse the data.

Bibliography

[1] Box, G. E. P., Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society B* 26:211-252.

[2] Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68*, 155-165.

[3] Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

[4] Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson Wadsworth.

[5] John, J. A., Draper, N. R. (1980). An alternative family transformations. *Applied Statistics* 29:190-197.

[6] Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

[7] Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

[8] Tukey, J. W. (1957). On the comparative anatomy of transformations. *Annals of Mathematical Statistics* 28:602-632.

[9] Swinscow, T. D., Campbell, M. J. (2003). Statistics at square one. 10th ed. New Delhi: Viva Books Private limited.

[10] Whittaker, J., Whitehead, J., Somers, M. (2005). The neglogtransformation and quantile regression for the analysis of a large credit scoring database. *Applied Statistics* 54:863-878.

[11] Yeo, I., Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* 87:954-959.

Mr. Felix Kutsanedzie

Mr. Felix Kutsanedzie, the lead author of this book, is a Senior Research Fellow / lecturer and currently the Head of the Accra Polytechnic Research and Innovations Centre. He holds a MSc. degree in Bio-Engineering; a BSc. (Agric-Mech); and an Adv. Dip. (Project Management). He is an Associate Editor of Directory Open Access Journals (DOAJ) and a prolific writer with several peer-reviewed publications to his credit.

Professor Sylvester Achio

Professor Sylvester Achio, is currently the Rector of Accra Polytechnic and a professor at the Department of Science Laboratory Technology of Accra Polytechnic, who holds MSc. Hons. (Research) Agric (Agronomy), PGC Ed. (Russian Language), PhD. Bio. Sci. (Microbiology). He has authored several technical books and also has presented papers at many national and international conferences, workshops, seminars as well as several peer-reviewed journal publications to his credit.

Mr. Edmund Ameko

Mr. Edmund Ameko is a Senior Lecturer at the Department of Science Laboratory Technology of Accra Polytechnic. He holds a BSc. in Biochemistry and a MSc. degree in Food Science and Technology, and currently the Dean of School of Applied Sciences at Accra Polytechnic. He is a prolific writer with several peer-reviewed publications.

To order additional copies of this book, please contact:
Science Publishing Group
book@sciencepublishinggroup.com
www.sciencepublishinggroup.com

ISBN 978-1-940366-58-6

Price: US $125