Kampala International University
Uganda

BACHELOR OF COMMERCE

MODULE 10

BUSINESS STATISTICS

By

TABLE OF CONTENT

INTRODUCTION 6

Unit 1 7

Introduction to Statistical Methods 7

1.1 Introductions 7

1.2 Types of statistics 8

Inferential/Inductive statistics 8

1.3 Applications of Statistics 8

1.4 Roles of Statistic 10

1.5 Limitations of Statistics 11

1.6 Classification of Variables 12

1.7 Classification of Data 12

1.8 Sources of Data 12

1.9 Data Collection 13

1.9.1 Methods of Data Collection 13

1.10 Designing a Statistical Study 17

Review Questions 18

Unit 2 19

Populations and Samples 19

2.1 Introduction 19

A sample is the portion of the population that has been selected for analysis. 19

Statistic; is a quantitative measure that describes a characteristic of a sample. 19

A sampling frame is list of all the items in the population or some means of identifying a particular item in the population. The frame must be complete i.e. no item in the population should be left out. 19

The Census of population 19

The census of population as defined by the U.N is the total process of compiling and publishing demographic, economic and social data at a specified time relating to all the people in a country or an area. 19

Time of census 19

A population census is suitable when there is less population movement. It should be outside the normal holiday periods such as –mas, Easter and other festivals. 19

Basis of conducting census 19

There two broad basis 19

DEJURE; on this basis a person is counted in the place where she or he normally lives 19

DEFACTO; a person is counted in the place where she or he spends the census night 19

Usual errors in population census 19

Age; a great number of women state their age under 30 19

Education and occupation are normally overstated 19
Marital status; divorced women often call themselves widows 19
Fertility is usually understated. 19
Reasons for sampling 19
Cost; the expenses involved in data collection and analysis are smaller than for attempting a complete or nearly complete coverage 19
Time; less time is taken to collect and analyze the data and information may be required urgently 20
Scope; highly trained personnel and specialized equipments can be used thus providing a more detailed analysis 20
Accuracy; high quality personnel can be given intensive training and ore careful supervision of fieldwork and therefore produce more accurate results. 20
2.2 Sampling Methods 20
Review Questions 21
Unit 3 22
Data Presentation and Analysis 22
3.1 Introduction 22
3.2 The tables, bar chart, Line (Time series) graph, and Pie-chart 22
3.3 Frequency distribution 29
3.4 Graph Presentation of frequency (histogram, frequency polygon and Ogive) 29
Review Questions 31
Unit 4 32
Measures of Central Tendency 32
4.1 Introduction 32
4.2 Methods of Computing the Measures 32
Mean 32
Arithmetic Mean 32
Geometric Mean (G.M) 33
Harmonic mean (H.M) 34
4.3 Comparison of the mean, median and mode 34
Median 36
Mode 37
Review Questions 41
Unit 5 42
Measures of Variability 42
5.1 Range 42
5.2 Mean Deviation 43
5.3 Quartiles 44
The semi-interquartile range 48
5.4 Percentiles 49
5.5 The variance and standard deviation 50
5.6 Coefficient of Variation 54

Review Questions 57
Unit 6 58
Introduction to Probability Theory 58
6.1 Basic Set Concepts Sample Space, Sample Point, Event And Probability Of An Event 59
Approaches of Assigning Probabilities 62
Types of Probability 63
Methods to Use in Solving Probability Problems 64
6.2 Laws of Probability; Multiplication Law, Addition Law 64
Formulas 67
6.3 Types of Events 69
Complementary Events 72
6.4 Probability Tree Diagrams 77
Review Questions 79
7.1 Introduction 81
7.2 Construction of Probability Distributions 82
7.3 Binomial Distribution 90
7.4 Normal Distribution or Normal Curve 94
Review Questions 102
Unit 8 103
Inferential Statistics, Sampling and Sampling distribution 103
Properties of the Sampling Distribution of Means 103
Central Limit Theorem 103
Sampling From a Finite Population 106
8.2 Sampling Distribution of Sample Proportion Sample proportion is computed by dividing the number of items in a sample that possess the characteristic, X, by the number of items in the sample, n. 107
8.3 Hypothesis Test 108
Null Hypothesis 109
Alternative Hypothesis 110
Simple Hypothesis 111
Composite Hypothesis 111
8.4 Hypothesis Testing Errors 111
Type I Error 111
Type II Error 112
8.5 Test Statistic 113
Critical Value(s) 113
Critical Region 113
Significance Level 113
8.5 Review Questions 114
Review Questions 114
Unit 9 116
Sampling 116
9.1 Introduction 116

9.2 Probability and nonprobability sampling 118
9.3 Sampling methods 119
9.4 Line-intercept sampling 127
9.5 Sampling and data collection 129
9.6 Sampling Distribution Of The Mean 132
9.7 Central Limit Theorem 133
9.8 Confidence limits of a proportionate or percentage 134
9.9 Convergence to the limit 140
9.10 Using Sample Data For Estimation 149
9.11 Confidence Intervals For A Population Mean 154
9.12 Confidence Intervals for a Proportion 165
9.13 Sample Size Determination With A Given Margin Of Error Sample Size Needed For Specified Margin Of Error 166
9.14 Hypothesis Testing 168
9.15 Hypothesis Testing For A Population Mean Or Proportion 181
Review Questions 182
Unit 10 183
Measures of Relationship 183
10.1 Correlation Analysis 183
Correlation 183
10.2 Coefficient of Determination 183
Standard Error of Estimate (SEE) 183

INTRODUCTION

Unit 1

Introduction to Statistical Methods

1.1 Introductions

Formal definition of Statistics and Key statistical concepts

The word statistics have been derived from the Latin word status or the Italian word "statista". Both the word means a political state. The word statistic is also found used by Shakespeare and Milton in the same of states men i.e. a person well versed in the affairs of the state. Its originally meant information useful to the state.

Statistics is a mathematical science involving the collection, interpretation, analysis, and presentation of data. It is often used to make predictions based on data. It is widely applicable in various social and natural sciences as such as political science and medicine as well as in business such as the insurance industry.

Statistics are very important in various aspects of business; a terrific example is the insurance industry. It is the job of an actuarial scientist to determine how long people will live (statistically), how likely they are to have an accident, and how likely is it their home will burn down or be damaged in a hurricane? These risks are all rated based solely on statistical data and policies are priced accordingly. Anonymous

Statistics refers to the numerical information from which valid conclusions can be drawn on the basis of some analysis

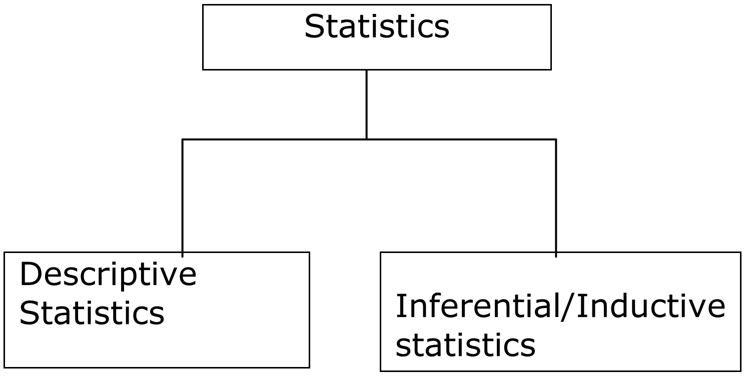
Statistics are quantifies computed or derived from a sample for example sample mean, standard deviation etc.

Statistics is the method of collecting, organizing, presenting, analyzing, interpreting and disseminating numerical information.

- **Collection of Data** is the process of obtaining measurements or counts or observations.
- **Organization of Data** is the task of presenting the collected data in a form suitable for deriving logical conclusions.
- **Analysis of Data** is the process of extracting from the given measurements, counts or observations relevant information from which a summarized and comprehensive numerical description can be formulated. The most important measures used for this purpose are the mean, median, range and standard deviation.
- **Interpretation of Data** is the task of drawing logical conclusions from the analysis of the data and usually involves the formulation of predictions concerning a large collection of objects from information available for a small collection of similar objects.

1.2 Types of statistics

Statistics can be broken down into two broad component parts



- **Descriptive statistics;** these are procedures used to organize and summarize masses of numerical information e.g. mean, mode, median, range etc.

- **Inferential/Inductive statistics;** these are methods used to draw conclusions about a population based on sample results. E.g. Comparisons and tests of hypotheses.

1.3 Applications of Statistics

The examples below illustrate some of the uses of statistics for business and economics.

Accounting

Public accounting firms use statistical sampling procedures when conducting audits for their clients. For instance, suppose an accounting firm wants to determine whether or not the amounts of accounts receivable shown on the client's balance sheet fairly represents the actual amounts of accounts receivable. Usually the number of individual accounts receivable is so large that reviewing and validating every account would be too time -consuming and expensive. In such situations, it is common practice for the audit staff to select a subset of the accounts called a sample. After reviewing the accuracy of the sample accounts, the auditors draw a conclusion as to whether or not the accounts receivable amount shown on the client's balance sheet is acceptable.

Finance

Financial advisors use a variety of statistical information to guide their investment recommendations. In the case of stocks, the advisers review a variety of financial data including price-earnings ratios and dividend yields. By comparing the information for an individual stock with information about the stock averages, a financial advisor can begin to draw a conclusion as to whether an individual stock is over- or undervalued.

Marketing

Electronic scanners at retail checkout counters are being used to date for a variety of marketing research applications. For example, data suppliers purchase point-of-sale scanner data from grocery stores, process the data, and sell the statistical summaries of the data to manufacturers. Manufacturers also purchase data and statistical summaries on promotional activities such as special pricing and the use of in-store displays. Product brand managers can review the scanner statistics and the promotional activity statistics to gain a better understanding of the relationship between promotional activities and sales. Such analyses are helpful in establishing future marketing strategies for the various products.

Production

With today's emphasis on quality, quality control is an important application of statistics in production. A variety of statistical quality control charts are used to monitor the output of a production process. In particular, an x-bar chart is used to monitor the average output. Suppose, for example, that a machine is being used to fill containers with 500 cm^3 of a well-known soft drink. Periodically, a sample of containers is selected and the average contents of the sample containers determined. This average or x-bar value, is plotted on the x-bar chart. A plotted value above the chart's upper control limit indicates overfilling and a plotted value below the chart's lower control limit, indicates under filling. Thus, the x-bar chart shows when adjustments are necessary to correct the production process. The process is termed " in control" and allowed to continue as long as the plotted x-bar values are between the chart's upper and lower control limits.

Economics

Economists are frequently asked to provide forecast about the future of the economy or some aspect of it. They use a variety of statistical information in making such forecasts. For example in forecasting inflation rates, economists use statistical information on such indicators as the Producer Price Index (PPI), Consumer Price Index (CPI), the unemployment rate, and the manufacturing capacity utilization. Often, statistical indicators are entered into computerized forecasting models that predict inflation rates.

1.4 Roles of Statistic

- Quality control, establishments normally set up acceptable quality limits. Management makes decisions on the quality of current production on the basis of these limits.

- Market research, the marketing department has the responsibility of making recommendations regarding the profitability of a new product or business location. The department therefore has to conduct consumer tests and make profit projections based on sample results.

- Planning; numerical information collected over a period of time normally shows some trend. On the basis of this trend a forecast or prediction can be made and this helps in planning future activities of the business e.g. sales during x-mas.

- Human activity, Statistics has come to play an important role in almost every field of life and human activity. There is hardly any field where statistical data or statistical methods are used for one purpose or the other our arrival in this world and departure from here are recorded as statistical data somewhere and in same form.

- Decision making, Statistics plays an important role in business, because it provides the quantitative basis for arriving at decisions in all matters. All types of banks make use of statistics for a number of purposes. Statistics has proved to be of immense use in physics and chemistry. It has given a new understanding to the essential qualities of the laws of nature.

- Statistics plays an important in psychology and education. In experimental psychology, whenever a problem has to be studied, it has to be based on a sample. Statistical methods are also used in analyzing the experimental data and drawing conclusions there from.

1.5 Limitations of Statistics

Statistics with all its wide application its limitations are as given below;

- Statistics is not suitable to the study of qualitative phenomenon: Since statistics is basically a science and deals with a set of numerical data, it is applicable to the study of only these subjects of enquiry, which can be expressed in terms of quantitative measurements. As a matter of fact, qualitative phenomenon like honesty, poverty, beauty, intelligence etc, cannot be expressed numerically and any statistical analysis cannot be directly applied on these qualitative phenomenons. Nevertheless, statistical techniques may be applied indirectly by first reducing the qualitative expressions to accurate quantitative terms. For example, the intelligence of a group of students can be studied on the basis of their marks in a particular examination.

- Statistics does not study individuals: Statistics does not give any specific importance to the individual items; in fact it deals with an aggregate of objects. Individual items, when they are taken individually do not constitute any statistical data and do not serve any purpose for any statistical enquiry.

- Statistical laws are not exact: It is well known that mathematical and physical sciences are exact. But statistical laws are not exact and statistical laws are only approximations. Statistical conclusions are not universally true. They are true only on an average.

- Statistics table may be misused: Statistics must be used only by experts; otherwise, statistical methods are the most dangerous tools on the hands of the inexpert. The use of statistical tools by the inexperienced and untraced persons might lead to wrong conclusions. Statistics can be easily misused by quoting wrong figures of data.

- Statistics is only, one of the methods of studying a problem: Statistical methods do not provide complete solution of the problems because problems are to be studied taking the background of the countries culture, philosophy or religion into consideration.

1.6 Classification of Variables

A variable is a characteristic that can assume different values and outcomes e.g. age, height, weight.

A characteristic that assumes the same valve under all circumstances is referred to as a constant e.g. pie, location, address, name etc.

Types of variables

Continuous variable, this variable assumes all numerical values of an interval or different intervals e.g. distance, time, height etc i.e. include decimal values e.g. 3.5 km

Discrete variable; this is limited to certain whole values e.g. number of people (20 or 21 but not 20.5)

1.7 Classification of Data

There are two classifications of data which include the following;

Quantitative data; a quantitative data is one which can be assigned a particular numerical value e.g. number of items, age, sale, height etc.

Qualitative Data, this is the data which can be identified or described but cannot be measured numerically e.g. colour, character, sex etc.

1.8 Sources of Data

Primary data; this is data which is collected and published by the same organization. It is original data that has been collected and published for the first time e.g. population census.

Secondary data; this is an extraction from an already existing source e.g. from books, magazines, newspapers etc.

Merits of primary data

- It is original and often includes definition of terms and units used
- Primary data is more detailed while secondary data omits part of the information
- Primary data includes the procedure of selecting the sample and collecting the data
- It is free from errors of transcription

Merits of secondary data

- It is easy to collect
- It is cheap in terms of money and manpower
- A lot of time is saved
- It convenient when reliable information is available.

1.9 Data Collection

Data are the facts and figures that are collected, analyzed, and summarized for presentation and interpretation. Together, the data collected in a particular study are referred to as a **data set** for that study.

1.9.1 Methods of Data Collection

1. A) Interview (personal)

Is an enumerator visits the respondent, asks the necessary questions and records the answers on a designed form

Advantages

- The right respondent is contacted

- Clarity can be made where necessary
- Non response is avoided

Disadvantages

- It is expensive in terms of staff salaries, transport, allowances and training cots
- Some respondents may not give accurate information in the presence of an interviewer especially on sensitive issues like educational standards, age, number of children etc
- Interviewer bias; the interviewer may influence the way the questions re answered by introducing his or her own ideas and ask questions that are not included in the schedule.

Qualities of interviewers

The interviewer needs long training since data collection involves probing hidden factors. The interviewer should possess the following desirable characteristics;

- Tact; she or he should be calm and avoid flattering and antagonizing respondents as this may lead to excitement or fear
- Accuracy; the interviewer should stick to the list of respondents; answers should be recorded accurately and any arithmetic calculations should also be done accurately and crosschecked.
- Amiability; the interviewer should be pleasant and sociable and should avoid investigating private lives and habits
- Neutrality; the interviewer should not segregate on matters of colour, sex, tribe, religion, politics etc as this will introduce bias.

b) Telephone Inquiry

In this method the interviews are conducted by telephone

Advantage

- It is a cheap method of collecting information

Disadvantage

- Not many people can be reached by telephone
- Information requiring documents cannot be obtained by telephone

2. Questionnaires

A questionnaire is a set of questions printed with blank/opinions space for answers and/ or pre –coded answers. In a mail questionnaire the question are printed and posted to the respondent to be answered at his/her own convenience.

Advantages

- The method is cheaper than interviews because only stamps are required.
- Large samples can be covered and the results are therefore more reliable

Disadvantages

- Wrong respondents may be contacted
- Explanations/clarifications cannot be made when required
- Some respondents take their own time to respond thus delaying the compilation and analysis of the information
- The method has the problem of non response unless there is incentive or legal obligation

Characteristic of a good questionnaire

- The questionnaire should be short

- The questions should be clear and simple
- The questions should not be ambiguous i.e. Every question should have a definite interpretation
- The questions should not require difficult calculations
- Unnecessary/irrelevant questions should be left out
- Instructions and definitions should be concise with terms and units clearly spelt out
- Leading questions should be avoided
- A question should naturally lead to the next, i.e. they should follow a logical order.

3. Direct Observation

This method involves examining, counting and measurements using physical means.

Advantages

- The method is more accurate compared to personal interviewer and mail questionnaire
- The respondent is not given the chance to give wrong information using pretence because the interviewer has to physically observe the events as they occur.

Disadvantages

- It is expensive in terms of money manpower and equipment
- It is time consuming
- Certain types of information cannot be (clearly) observed directly e.g. income, expenditure etc.

4. Registration

In this method the information is reported t the relevant authority when or shortly after the event has occurred. E.g. death, marriage, birth, accident, migration etc.

1.10 Designing a Statistical Study

Sometimes data are not readily available from existing sources. If the data are considered necessary a statistical study can be conducted to obtain them. Such statistical studies can be classified as **experimental or observational**.

- **In an experimental study**, the variables of interest are first identified. Then one or more factors in the study are controlled so that the data can be obtained about how the factors influence the variables.

For example, a pharmaceutical firm might be interested in conducting an experiment to learn about how a new drug affects blood pressure. The new drug is the factor that influences the blood pressure. To obtain data about the effect of the new drug, a sample of individuals will be selected. The dosage level of the new drug will be controlled with different groups of individuals being given different dosage levels. Data on blood pressure will be collected for each group. Statistical analysis of the experimental data will help determine how the new drug affects blood pressure.

- **In non-experimental, or observation statistical studies** no attempt is made to control or influence the variables of interest. A survey is the most common type of observational study.

  In a personal interview survey, research questions are first identified. The questionnaire is designed and administered to a sample of individuals. Data are obtained about the research variables, but no attempt is made to control the factors that influence the variables.

Summary

This unit has presented statistics as a scientific investigative and research technique. A number of applications of statistics have been presented in order to emphasis the importance of statistics. The unit has also looked at sources of data, methods of collecting data and the different sampling techniques. These methods will be looked at in more detail in our next units.

It is therefore important that researchers and decision makers know the different statistical techniques of collecting and analyzing data if they are to make logical inferences about the collected data.

Review Questions

1. Describe the uses of statistics and limitations of statistics.
2. Discuss the methods of data collection with 3 advantages and disadvantages of each
3. Explain the role of statistics in business
4. What factors are put into consideration when choosing the source of date to use?

Unit 2
Populations and Samples
2.1 Introduction

A population/ universe; is the totality of all the items or things under consideration. It is a set of all the elements on which information is desired.
Parameter; is a quantitative measure that describes a characteristic of a population
A **sample** is the portion of the population that has been selected for analysis.

Statistic; is a quantitative measure that describes a characteristic of a sample.

A **sampling frame** is list of all the items in the population or some means of identifying a particular item in the population. The frame must be complete i.e. no item in the population should be left out.

The Census of population

The census of population as defined by the U.N is the total process of compiling and publishing demographic, economic and social data at a specified time relating to all the people in a country or an area.

Time of census

A population census is suitable when there is less population movement. It should be outside the normal holiday periods such as –mas, Easter and other festivals.

Basis of conducting census

There two broad basis
DEJURE; on this basis a person is counted in the place where she or he normally lives
DEFACTO; a person is counted in the place where she or he spends the census night

Usual errors in population census

Age; a great number of women state their age under 30
Education and occupation are normally overstated
Marital status; divorced women often call themselves widows
Fertility is usually understated.

Reasons for sampling

Cost; the expenses involved in data collection and analysis are smaller than for attempting a complete or nearly complete coverage

Time; less time is taken to collect and analyze the data and information may be required urgently
Scope; highly trained personnel and specialized equipments can be used thus providing a more detailed analysis
Accuracy; high quality personnel can be given intensive training and ore careful supervision of fieldwork and therefore produce more accurate results.

2.2 Sampling Methods

There are many ways to collect a sample. The most commonly used methods are:

- **Simple Random Sampling**
  The sample is drawn unit by unit and each item in the population has an equal chance of being included in the sample e.g. gold fish bowl
- **Systematic Sampling**
  The first item is selected at random from the first k items and thereafter every k^x Item is included in the sample where k is the sampling frame given by

  $$K = \frac{N}{n}$$

  Advantages
  - Easy to draw
  - The sample is spread more evenly over the population

- **Cluster sampling**
  This is often referred to as area sampling because it is frequently used on a geographical basis. The area of interest is divided into smaller units e.g. city blocks. A simple random sample of the blocks is then selected and every item in the selected blocks is included in the sample.

- **Stratified sampling**
  The population is subdivided into non overlapping homogeneous subpopulations called strata. A simple random sample is then taken separately from each stratum. Stratification is applied when the population is heterogeneous and the heterogeneity has a bearing on the characteristic being studied e.g. fertility

- **Quota sampling**
  In this method the interviewer is given the names and addresses of the people (items) to be included or a fixed number form each category of items.

Advantages

- Speed
- Reduced cost

Disadvantages

- Risk of bias; the interviewer may discriminate against certain types of people.

- **Multistage sampling**
  This involves drawing a series of random samples at successive stages; it is commonly used when the population is widely scattered e.g. we can start by sampling the regions in Uganda (stage 1).

Review Questions

1. Give and discuss the main characteristics of the following types of sampling techniques:
   (i) Simple random sampling
   (ii) Systematic sampling
   (iii) Stratified sampling
   (iv) Cluster sampling

2. Describe the main uses of sampling in business

3. Explain what is meant by the following terms:
   (i) Population
   (ii) Sample
   (iii) Parameter
   (iv) Statistic

Unit 3
Data Presentation and Analysis

3.1 Introduction

Data organization, editing and presentation are essential tasks which must be carried out before the planning and decisions making processes.

Data organization helps to summaries a huge mass of data in a clear and orderly manner.

Data editing is necessary to identify outliers and possible errors in data collection and coding.

Data presentation helps to lay them out in an orderly manner in order to reveal the underlying patterns or salient features in the data set.

Data may be presented in form of tables, graphs/charts and pictures.

3.2 The tables, bar chart, Line (Time series) graph, and Pie-chart

A table is a systematic arrangement of data in a two dimensional layout in form of rows and columns. The columns are horizontal arrangements while the rows are vertical arrangements.

Tabulation is the process of condensing data in the form of a table.

Major reasons for tabulation

- The data can be more easily comprehended
- Comparisons can be made easily

Types of tables

Informative/classifying table

These are original tables with data systematically arranged for record purposes.

They are frequently referred to as schedules and present data relating to a given phenomenon. E.g. logarithm table, table of square root, area under the normal curve, sin, cos etc.

General/reference tables

These originate from the informative tables and contain highly summaries information. They are normally given as appendix. E.g. financial tables for 1%, 5%, 10% etc

Text/summary tables

Summary tables are derived tables which contain only the data required for analysis. These tables are interpretive and highlight significant observations relating to a phenomenon. They often include ratios, percentages, totals, averages and other derived measures and are found in the body of the text. e.g class composition by gender

Class	Male	female	Total
BBA I	24	12	36
BBA II	16	20	36
Total	40	32	72
%	55.6	44.4	100

Frequency distribution tables

Frequency distribution tables give the number of items of varying sizes or magnitudes. These tables categories data according to specific characteristics or magnitudes.

While qualitative frequency distribution tables group data on the basis of the nature or characteristic of the data, quantitative frequency distribution tables show classes of data categorized according to numerical size

Frequency distribution table

Distances (KM) recorded by 120 sales executives in a week

Distance	**frequency (number of executives)**
400-420	12
421-441	27
442-462	34
463-483	24
484-504	15
505-525	08

Time series tables

These tables show the magnitude of a variable over a specifc perid of time. E.g value of exports (mil shs)

Year	**Value**
2000	15
2001	17
2002	10

Classification of tables

Tabulation involves the use of rows and columns to illustrate relationships between data for comparison purposes. Tables can be broadly classified into 2 categories;

- **Simple table**

A simple table presents the characteristics of a single item and shows only one relationship.

Name	**Age**
Jane	20
John	22
James	19

- **Complex tables**

Complex tables present the characteristics of more than one group of items set up in additional columns and rows

	20-40		Age 41+	
	Married	Unmarried	Married	Unmarried
Male				
Smoker	31	28	41	20
Nonsmoker	30	42	27	18
Female				
Smoker	38	37	49	43
Nonsmoker	32	29	31	39

Major parts of a table

- Title; this is a brief statement, which appears at the top of the table and explains what type of information is contained in the table. A good title should be compact but complete.
- Head not; this is a statement below the title, which clarifies the content of the table or main parts of the table. E.g. units used such as figures in tones, million shs etc.

- Stub (rows) and captions (columns); the stub consists of the stub head and stub entries. The stub head describes the stub entries and each stub entry 1 tables the data found in this row of the table.
- The caption labels the data found in the columns of the table. It consists of one or more column heads.
- The body; the body contains the actual numerical information
- Footnote; this is a phrase or statement which clarifies some specific parts of the table. The asterisk symbol is normally used in the body of the table
- Source note; this states clearly where the data was obtained. This si necessary for the reader t cross check the figures and possibly gather additional information.

Format of a table

Title

Head note

Stub head	Caption	
	Column head	Column head
Stub Entries	Body	

Footnote

Source note

Guidelines for constructing tables

- The table should be as simple as possible
- The table should have a comprehensive explanatory title placed at the top and centered

- Abbreviations should be avoided especially in tittles and headings
- A dash, rather than a zero, should be used to indicate that information is not available
- If a figure is repeated it should be shown each time. Ditto marks (") must be avoided.
- Symbols, especially the asterisk (*), should be used for footnotes
- The source must be stated
- Totals should be shown where appropriate.

Graphs

A graph is a representation of data by a continuous curve. The vertical height of the curve measures one variable while the horizontal measures another so that the measurements are significant in both directions

Advantages of graphs compared to table s

- A graph can make a stronger visual impact
- It does not require and special training to assimilate information on graphs

Limitations of graphs

- A graph cannot show many sets of facts like complex table. They are limited to mainly two variables.
- Exact values are not easily shown on a graph
- The construction requires certain amount of time

Principles of graph construction

- A graph must have a clear and comprehensive title
- The scale should be chosen such that the presentation ensures that the correct impression is given
- The independent variable should always be placed on the horizontal axis

- The vertical scale should always start from zero. If this is not possible the scale may be such that the zero is shown at the bottom of the scale and definite break in the scale is shown.
- The horizontal scale need not start from zero
- The axes should be clearly labeled. This should include both the variable and the units used.
- Curves must be distinct. If two or more curves are drawn on the same graph different colours can be used to distinguish them to avoid curves being confused
- The graph must not be overcrowded with curves. This makes it difficult to see the pattern formed by any one curve and the major reason for graphical presentation is not achieved.
- The source of the data must always be given.

Types of graph

Arithmetic line graph

Both the vertical and the horizontal axes have arithmetic scales. The scales consist of consecutive units equally spaced.

Semi log graph

A semi log graph is a graph, which has one axis (the horizontal) arithmetically scaled and one axis (the vertical having logarithmic scales. The latter means that instead of the consecutive units the logarithms of the units are plotted. A semi log graph shows the rate of change of a variable. E.g percentage increase.

Features of a semi log graph

- The slope of the curve shows the rate at which the figures are changing. (Increasing or decreasing)
- If the curve is a straight line the rate of change is constant

Geometric forms

Bar chart

These are diagrams in which figures are represented by the length of rectangles called bars

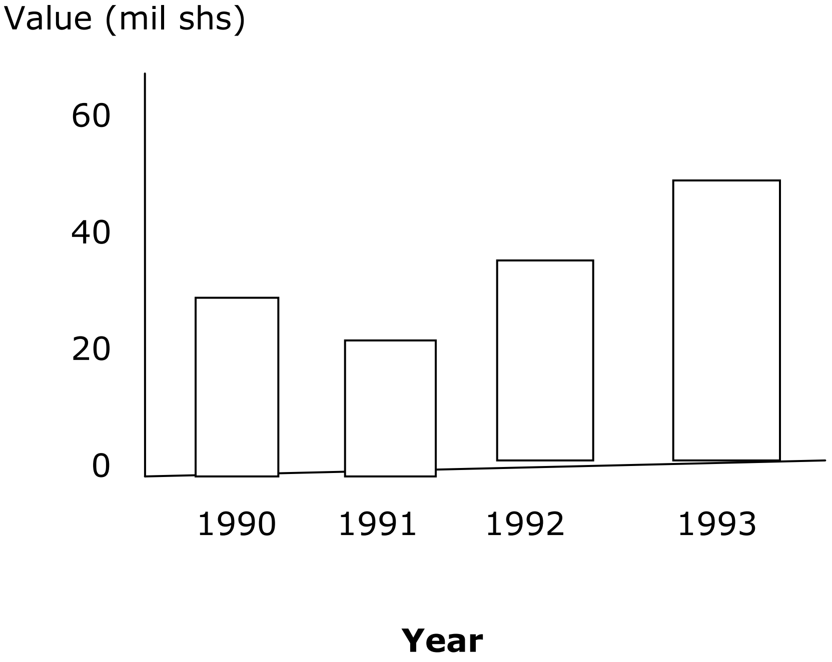
- **Simple bar chart**

  This is a graph that consists of a number of bars arranged vertically or horizontally whose heights or lengths vary with the magnitude of the figures represented but which are of equal width.

  The simple bar charts are used where change in totals only are required.

Year	**Products**
1990	28
1991	20
1992	40
1993	45

Simple bar chart

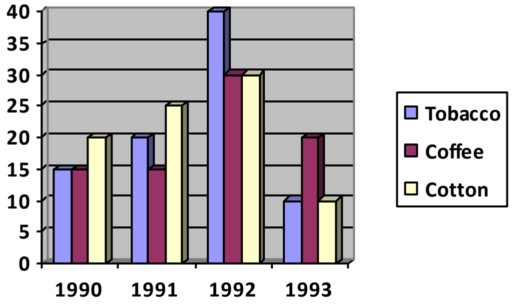


- **Component bar charts**

These are ordinary bar charts subdivided into component parts. The individual component lengths and the overall length f the bars however represent actual figures.

Component bar chars are used where changes in total and an indicating the size of each component figure are required.

- **Multiple bar charts**



3.3 Frequency distribution

3.4 Graph Presentation of frequency (histogram, frequency polygon and Ogive)

Determining the Mode from the Histogram

For grouped data, the mode can as well be obtained from the histogram using the tallest bar (rectangle).

After determining $L_{o,}$ c, d_1 and d_2, from the Histogram, the mode is computed using the following formula:

Mode(M_o)

$$M_o = L_o + c\left[\frac{d_1}{d_1 + d_2}\right]$$

where **L_o** = lower boundary of the modal class
c = class length/width of the modal class

Example 4.5

The following table shows the weights of 100 students measured to the nearest kg.

Weight(kg)	**Number of children**
10-14	5
15-19	9
20-24	12
25-29	18
30-34	25
35-39	15
40-44	10
45-49	6

Determining the Median From The O-Give

For grouped data, the median can as well be obtained from the plot of the greater-than and less-than O-give diagram. It occurs at the point of intersection.

Example 4.6

The table below gives the frequency of tail length of rats in a biology laboratory.

Length(cm)	**Frequency**
1-5	15
6-10	24
11-15	23
16-20	10
21-25	10
26-30	5
31-35	1
36-40	1

Draw a greater-than O-give and a less-than O-give and determine the median

Review Questions

1. The managing director (MD) of a bus company requires information about the distribution of the passengers for a managerial decision making regarding whether or not to purchase more bus. The data collected for the last 50 days of the business is shown in the table below:

Passengers	20-24	25-29	30-34	35-39	40-44	45-49	50-54
Days (f)	12	24	30	18	11	5	1

a) Draw a greater – than O-give and determine the first and third quartiles from it.

b) Determine the 25th, 50th and 75th percentile from the O-give

Unit 4
Measures of Central Tendency
4.1 Introduction

We are usually interested in the sample value around which the distribution of data is centred. Such a value is referred to as a measure of central tendency.

The common measures of location are; the arithmetic mean, the weighted mean, The Geometric and Harmonic mean, mode and median. A single value calculated is used to describe the distribution of the scores. They are called measures of central tendency because they estimate the tendency of the scores or the characteristics of people to focus on or cluster.

Measures of Location

Measures of location include the following

4.2 Methods of Computing the Measures

The arithmetic mean, the weighted mean, The Geometric and Harmonic mean

Mean

The mean is considered to be the most important measure of location. It is at times referred to as the average value of the variable. It is obtained by adding all the data values and dividing by the number of items.

If the data are from a sample, the mean is denoted by $\bar{x}$, and if the data is from a population the mean is denoted by μ

Let the number (n) of data items in a sample be denoted by : x_1 , x_2 , ..., x_n .
The mean is computed using the following formulae:

Arithmetic Mean

This is usually referred to as the mean and is the most widely unsed of all averages.

Ungrouped data

If observations are in a raw form, the mean is computed by summing all the observations and then dividing their sum by the total number of their observations

(a) Ungrouped data : $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$

This measure has one major disadvantage which is that it is affected by extreme values.

Example; consider the marks of 7 bstat students, 20, 38, 40, 60, 98, 95, 97

Solution;

$\overline{X}$ = means

$\overline{X} = \frac{20+38+40+60+98+95+97}{7}$

$\overline{X}$ = 64

Note: We shall suppress the index i on the summation symbol and on the variable x, frequency f etc.

(b) Grouped data :

(I) $\overline{X} = \frac{\sum fx}{\sum f}$ where ∑f = n

(ii) $\overline{X} = \overline{x_A} + c\left(\frac{\sum fd}{\sum f}\right)$, where $\overline{X_A}$ = assumed mean

d = code (in coding method)
c = class length/width

Geometric Mean (G.M)

We define the geometric mean by

$$G.M = \sqrt[n]{x_1.x_2.......x_n}$$

Harmonic mean (H.M)

We define the harmonic mean(H.M) by:

$$H.M = \frac{n}{\sum \frac{1}{x}}$$

If we are working with the entire population with N data items, we substitute μ for $\overline{X}$ and N for n in the above expressions.

The Weighted Mean

$$\text{WM} = = \frac{\sum_{i}^{n} Xiwi}{wi}$$

Supposing the observations xi i=1,2,……n, have corresponding weights Wi, i==1,2,……n, the defined weighted mean is as below;

I	1	2	3	4	5	Total
X	10	15	20	32	12	
W	4	3	1	3	2	13
Xw	40	45	20	96	24	225

$$\text{Weighted mean} = \frac{225}{13}$$

Weighted mean = 17.3

4.3 Comparison of the mean, median and mode

Example 4.1

Given the following data find the mean and mean absolute deviation.

(a)11, 14, 17, 20, 16, 10

Solution

Mean $\overline{X}$ = $\frac{\sum_{i=1}^{n} x_i}{n}$ = $\frac{11+14+17+20+16+10}{6} = 14.67$ (corrected to 2dp))

Mean absolute deviation M.A.D = $\frac{\sum |x-\overline{x}|}{n}$ =
(3.67 + 0.67 + 2.33 + 5.33 + 1.33 + 4.67)/6 = 3.00 (corrected to 2 dp).

(b)

Goals(x)	0	1	2	3	4	5	6
Frequency(f)	1	2	5	8	6	3	0

Solution

The computations are summarized in the following table:

Goals X	frequency f	fx	(x - $\overline{x}$)	$\lvert x-\overline{x}\rvert$	f $\lvert x-\overline{x}\rvert$
0 1 2 3 4 5 6	1 2 5 8 6 3 0	0 2 10 24 24 15 0	-3 -2 -1 0 1 2 3	3 2 1 0 1 2 3	3 4 5 0 6 6 0
	$\sum f = 25$	$\sum fx = 75$			$\sum f\lvert x-\overline{x}\rvert = 24$

Hence mean $\overline{X}$ = $\frac{\sum fx}{\sum f}$ = 75/25 = 3

Mean absolute deviation M.A.D = $\frac{\sum f\lvert x-\overline{x}\rvert}{\sum f}$ = 24/25 = 0.96

(c)

Class	0.5-8.5	8.5-16.5	16.5-24.5	24.5-32.5	32.5-40.5
Frequency(f)	3	7	8	5	2

Solution

The computations are shown in the table below

Class	midpoint x	frequency f	fx	(x - $\bar{x}$)	$\lvert x-\bar{x}\rvert$	f$\lvert x-\bar{x}\rvert$
0.5-8.5 8.5-16.5 16.5-24.5 24.5-32.5 32.5-40.5	4.5 12.5 20.5 28.5 36.5	3 7 8 5 2	13.5 87.5 164.0 142.5 73.0	-14.7 -6.7 1.3 9.3 17.3	14.7 6.7 1.3 9.3 17.3	44.1 46.9 10.4 46.5 34.6
		$\sum f = 25$	$\sum fx = 480.5$			$\sum f\lvert x-\bar{x}\rvert =$ 182.5

Hence mean $\overline{X}$ = $\frac{\sum fx}{\sum f}$ = 480.5/25 = 19.22

Mean absolute deviation M.A.D = $\frac{\sum f\lvert x-\bar{x}\rvert}{\sum f}$ = 182.5/25 = 7.3

Median

Although, the mean is the most commonly used measure of central location, there are situations in which the median is preferred. In general, whenever there are extreme data values. The median is often the preferred measure of central location.

The median of a set of n numbers, x_1 , x_2 , ..., x_n , arranged in either ascending or descending order of magnitude, is computed using the following formula depending on whether n is even or odd.

Note: For ungrouped data,

(i) If there is an odd number of items, the median is the value of the middle item when all items are arranged in ascennding/decsending order.

(ii) if there is an even number of items, the median is the average value of the two middle items when all items are arranged in ascending/descending order.

Median(M_d)

(a)ungrouped data (In ascending/descending order)

$$M_d = \begin{cases} X_{\frac{n+1}{2}}, \text{when} \quad n \quad \text{is} \quad odd \\ \frac{1}{2}\left(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}\right), \text{when} \quad n \quad \text{is} \quad even \end{cases}$$

(b) grouped data

$$M_d = L_o + c\left[\frac{\sum\frac{f+1}{2} - C.F_b}{f_{m_d}}\right]$$

where L_o = lower class boundary of the median class

c = class length/width

$\sum f$ = total frequency = n

$C.F_b$ = cumulative frequency preceding (before) that of the median

Class

f_{md} =frequency of the median class

Note : $\frac{\sum f+1}{2}$ **= position of median**

Mode

The mode is the data value that occurs most often i.e. the data value with the highest frequency.

Note:

(i) Situations can arise for which the greatest frequency occurs at two or more different values. In these instances more than one mode exists.

(ii) If the data have exactly two modes, we say that the data are bimodal.

(iii) If the data have more than two modes, we say that the data are multimodal.

(Iv) In multimodal cases the mode is almost never reported, since listing three or more modes would not be helpful in describing a location for the data.

For grouped data the mode is computed using the following formula:

Mode(M_o)

$$M_o = L_o + c\left[\frac{d_1}{d_1 + d_2}\right]$$

where **L_o** = lower boundary of the modal class

c = class length/width of the modal class

d_1 = $f_{mo} - f_b$

$d_2 = f_{mo} - f_a$

f_{mo} = frequency of the modal class

f_b = frequency preceding(before) that of the modal class

f_a = frequency immediately after that of the modal class

Example 4.2

Compute the median and mode for the following data:

(a) 9, 2, 5, 10, 6

solution

In ascending order we have: 2, 5, 6, 9, 10

x_1, x_2, x_3, x_4, x_5

n = 5(odd), hence median $M_d = x_{\frac{n+1}{2}} = 6$

There is no mode

(b) 6, 7, 12, 9, 7, 12, 16, 20, 7, 4, 7, 12

Solution

In descending order we have:

20, 16, 12, 12, 12, 9, 7, 7, 7, 7, 6, 4

$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}$

n = 12 (even)

Hence median = $\frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)$ = (9 + 7)/2 = 8

Mode = 7

(c) The table below gives the frequency of tail length of rats in a biology laboratory.

Length(cm)	**Frequency**	
1-5	15	f_b

Modal class	6-10	24	f_{mo}
	11-15	23	f_a
	16-20	10	
	21-25	10	
	26-30	5	
	31-35	1	
	36-40	1	

Find the median and mode length of the tails.

(i)<u>Mode</u>

Step1
Determine the class(modal class) with the highest frequency

Highest frequency is 24. Hence modal class is 6 – 10

Step2
compute the mode

$M_o = L_o + c\left[\frac{d_1}{d_1 + d_2}\right]$

where $\mathbf{L_o}$ = lower boundary of the modal class = 5.5

c = class length/width of the modal class = 5

$\mathbf{d_1} = f_{mo} - f_b$

$d_2 = f_{mo} - f_a$

f_{mo} = frequency of the modal class = 24

f_b = frequency preceding(before) that of the modal class = 15

f_a = frequency immediately after that of the modal class = 23

Hence $\mathbf{d_1} = f_{mo} - f_b = 24 - 15 = 9$

$d_2 = f_{mo} - f_a = 24 - 23 = 1$

Hence $M_o = 5.5 + 5[9/(9 + 1)] = 10$

Note:

the mode can as well be obtained from the histogram using the tallest bar(rectangle)

(ii)<u>Median</u>

The computations are shown in the following table:

class/ length(cm)	class boundaries	midpoint x	frequency f	cumulative frequency(C.F)
1-5	0.5-5.5	3	15	15
6-10	5.5-10.5	8	24	39 $(C.F_b)$
11-15	10.5-15.5	13	23(f_m)	(position of M_d)
16-20	15.5-20.5	18	10	62
21-25	20.5-25.5	23	10	(higher)
26-30	25.5-30.5	28	5	72
31-35	30.5-36.6	33	1	82
36-40	35.5-40.5	38	1	87
				89
				90
			$\sum f = 90$	

Step 1:

Determine the position of the median : Position = $\frac{\sum f + 1}{2}$ **= (90 +1)/2** = 45.5

Step 2

Determine the two values between which the position in step1 lies in the column of cumulative frequency. Take the higher (bigger in magnitude)of the two values and use its row to determine the median class. Hence median class is 11-15.

Step 3

Compute the median

Now, $M_d = L_o + c \left[\frac{\sum \frac{f+1}{2} - C.F_b}{f_{m_d}} \right]$

where L_o = lower class boundary of the median class = 10.5

c = class length/width = 5

☐f = total frequency = n = 90

$C.F_b$ = cumulative frequency preceding (before) that of the median

class = 39

f_{md} =frequency of the median class = 23

$\frac{\sum f + 1}{2}$ **= position of median = 45.5**

Hence, M_d = 10.5 + 5(45.5 - 39)/23 = 11.91 (corrected to 2 dp)

Note:

The median can also be determined from the cumulative frequency curve (Ogive)

Review Questions

Unit 5
Measures of Variability

5.1 Range

Given the distance between the largest and the smallest observation in a given data act or ungrouped data, range is given as H-L for H is largest observation and L is the smallest observation.

For grouped data, it's given by

- **Use of class limits**; here the range is the difference between the upper class limit of the last class and the lower limit of the 1st class interval.
- **Use of class marks**; here the range is the difference between the class mark of the last class interval and the class mark of the 1st class interval.

Examples;

59-74

74.5-58.5

Range = 15

This is the simplest measure of dispersion.

Range = Largest value - smallest value

For example 4.3

(a) range = 2825 - 2210 = 615

(b) range = 49-14 =35(using upper class limits for largest and smallest class)

=45-10 =35(using lower class limits for largest and smallest class)

=49.5-14.5=35(using upper class boundaries for largest and smallest class)

=44.5-9.5=35(using lower class boundaries for largest and smallest class)

Interquartile range

A measure of dispersion that overcomes the dependence on extreme data values is the interquartile range.

This measure is simply the difference between the third quartile and the first quartile.

Interquartile range = $(Q_3 - Q_1)$

semi-interquartile range = $(Q_3 - Q_1)/2$

Example 4.4

Consider example 8. Find the the interquartile range and the semi-interquartile range.

Solution

(a) interquartile range = $Q_3 - Q_1$ = 2500 - 2365 = 135
semi-inetrquartile range = $(Q_3 - Q_1)/2$ = 135/2 = 67.5

(b) interquartile range =$(Q_3 - Q_1)$ = 36.75 - 24.19 = 12.56
semi-interquartile range =$(Q_3 - Q_1)/2$
= 12.56/2 = 6.28

Interpretation of the range

The wider the range, the greater the dispersion and vice versa provided the units are similar.

Characteristics of the range

- The range is simple easily understood and calculated. It is ideal for preliminary work or when a quick at measure of dispersion is required.
- It is sensitive to extreme or deviant scores, because of its sensitivity to the extreme scores, the range is often unrealizable (instable and inconsistent), and misleading. Calculation of the range does not put into consideration how the scores are distributed.
- Calculation of the range does not put into consideration how the scores are distributed.
- It is not used in any statistical test.

5.2 Mean Deviation

Mean deviation is the sum of the deviations of each score from the mean, without regard to the sign, divided by the number of scores. For a given data set xi, i= 1,2,3..n, we defined the mean deviation as the average of the absolute deviations of the observations from their mean.

Mean Absolute Deviation (M.A.D)

We define the mean absolute deviation by:

(a) Ungrouped data

$$\text{M.A.D} = \frac{\sum\left|x-\bar{x}\right|}{n}$$

Example;
Find the mean deviation for the follow data
2,3,6,8,11

Solution;

$$\text{M.D} = \frac{2+3+6+8+11}{5}$$

M.D = 6

(b) Grouped data

$$\text{M.A.D} = \frac{\sum f\left|x-\bar{x}\right|}{\sum f}$$

5.3 Quartiles

It is often desired to divide data into **four parts**, with each part containing approximately one - fourth. The division points are referred to as quartiles and are defined as:

Q_1 = first quartile, or 25th percentile

Q_2 = second quartile, or 50th percentile (also the median)

Q_3 = third quartile, or 75th percentile.

(a) Ungrouped data

The computational steps of the quartiles are the same as those of the percentiles with a few slight modifications as shown below:

For Q_1
position = (i/100)n = (25/100)n =(1/4)n

For Q_2
position = (i/100)n = (50/100)n = (1/2)n = position of median

For Q_3
position = (i/100)n = (75/100)n = (3/4)n

In general the position of the ith quartile, Q_i, = (i/4)n
i = 1, 2, 3

(b) Grouped data

The steps are the same as those for finding the median.

For grouped data the ith quartile is computed using the following formula.

$$\text{i-th quartile}(Q_i) = L_o + c\left[\frac{i(\sum \frac{f+1}{4}) - C.F_b}{f_{Q_i}}\right]$$

where L_o = lower class boundary of the quartile class
c = class length/width
$\sum$f = total frequency = n
$C.F_b$ = cumulative frequency preceding (before) that of the quartile class
f_{Q_i} =frequency of the percentile class

Note : $\frac{i(\sum f + 1)}{4}$ **= position of the ith quartile**

Example 4.3

(a) The table below shows the monthly starting salaries for a sample of 12 Business School Graduates

Graduate	**Monthly salary($)**	**Graduate**	**Monthly salary($)**
1	**2350**	**7**	**2390**
2	**2450**	**8**	**2630**
3	**2550**	**9**	**2440**
4	**2380**	**10**	**2825**
5	**2255**	**11**	**2420**
6	**2210**	**12**	**2380**

Find the (i) 50th percentile and the 85th percentile
(ii) first, second and third quartiles

(i)Solution

Step 1 : Arranging the data in ascending order we have :

2210 2255 2350 2380 2380 2390 2420 2440 2450 2550 2630 2825

Note:
n = 12 (even). Hence **median** = (2390 + 2420)/2 = 2405

Step 2:

Position of ith percentile = (i/100)n
Hence, postion of 50th percentile =(50/100)12 = 6.
position of 85th percentile = (85/100)12 = 10.2

Step 3

Since the position of the 50th percentile is an integer(6), then the 50th percentile is the average of the 6th and 7th data values.
i.e 50th percentile,P_{50} = (2390 + 2420)/2 = 2405 = median

For the 85th percentile, since the position (10.2) is not an integer, we round up. Hence the position of the 85th percentile is the next integer greater than 10.2, the 11th position.

Therefore, the 85th percentile corresponds to the 11th data value.
Thus, the 85th percentile, P_{85} = 2630.
(ii) Finding the quartiles

Solution

Step1: Arrange the data in ascending order:
2210 2255 2350 2380 2380 2390 2420 2440 2450 2550 2630 2825

Step2:
The position of the ith quartile(Q_i) is:
position = (25i/100)n = (i/4)n
Hence for Q_1, position = (1/4)12 = 3
For Q_2, position = (1/2)12 = 6
For Q_3, position = (3/4)12 = 9

Step3:
Note: All positions are integers.
Thus, Q_1 = (2350 + 2380)/2 = 2365
Q_2 = (2390 + 2420)/2 = 2405 = median.
Q_3 = (2450 + 2550) = 2500

(b) The following table shows the weights of 100 students measured to the nearest kg.

Weight(kg)	**Number of children**

10-14	5
15-19	9
20-24	12
25-29	18
30-34	25
35-39	15
40-44	10
45-49	6

Find:

(i) 65th percentile

(ii) first and third quartile

The computations are arranged in the following table

weight(kg)	Class boundaries	Midpoint x	frequency f	Cumulative frequency(C.F)
10-14	9.5-14.5	12	5	5
15-19	14.5-19.5	17	9	**14 C.F$_b$**
				25.25
20-24	19.5-24.5	22	12 f_{Q1}	26
25-29	24.5-29.5	27	18	**44 C.F$_b$**
				65.65
30-34	29.5-34.5	32	25 f_{Pi}	**69 C.F$_b$**
				75.75
35-39	34.5-39.5	37	15 fQ_3	84
40-44	39.5-44.5	42	10	94
45-49	44.5-49.5	47	6	100
			$\sum f = 100$	

(i) position of 65th percentile = $\frac{i(\sum f + 1)}{100}$ **= 65(100 +1)/100**

= 65.65

Hence 65th percentile class is 30-34

Now the ith percentile is given by:

$$\text{i-th percentile}(\ p_i) = L_o + c\left[\frac{i(\sum \frac{f+1}{100}) - C.F_b}{f_{p_i}}\right]$$

Where L_o = lower class boundary of the percentile class = 29.5

c = class length/width = 5
□f = total frequency = n = 100
C.F$_b$ = cumulative frequency preceding (before) that of the percentile
class = 44
f_{p_i} =frequency of the percentile class = 25

Hence P$_{65}$ = 29.5 + 5[65.65 - 44]/25 = 33.83

The semi-interquartile range

The semi inter-quartile range is given by: (Q$_3$ – Q$_1$)/2

Position of first quartile, Q$_1$, = $\frac{i(\sum f+1)}{4}$ **=(100 + 1)/4**
=25.25

Hence the first quartile class is: 20-24.

Now first quartile(Q$_1$) = L$_o$ + c $\left[\frac{(\sum \frac{f+1}{4}) - C.F_b}{f_{Q_1}}\right]$

where L$_o$ = lower class boundary of the quartile class = 19.5
c = class length/width = 5
$\sum$f = total frequency = n = 100
C.F$_b$ = cumulative frequency preceding (before) that of the quartile
class = 14
f_{Q_1} =frequency of the percentile class = 12

Hence Q$_1$ = 19.5 + 5[25.25 - 14]/12 = 24.19

Position of third quartile, Q$_3$, = $\frac{i(\sum f+1)}{4}$ **= 3(100 + 1)/4**
=75.75

Hence the third quartile class is: 35-39.

Now third quartile(Q$_3$) = L$_o$ + c $\left[\frac{3(\sum \frac{f+1}{4}) - C.F_b}{f_{Q_3}}\right]$

where L$_o$ = lower class boundary of the quartile class = 34.5
c = class length/width = 5

$\sum f$ = total frequency = n = 100

$C.F_b$ = cumulative frequency preceding (before) that of the quartile

class = 69

f_{Q_3} =frequency of the quartile class = 15

Hence Q_3 = 34.5 + 5[75.75 - 69]/15 = 36.75

Semi-interquartile range is: $(Q_3 – Q_1)/2$ = 36.75 – 24.19 = 12.56

5.4 Percentiles

A percentile is a measure that locates values in the data set that are not necessarily central locations. A percentile provides information about how the data items are spread over the interval from the smallest value to largest value.

(a) Ungrouped data

For data that do not have numerous repeated values, the ith percentile divides the data into two parts:

(i) Approximately i percent of the items have values less than the ith percentile

(ii) Approximately (100 - i)percent of the items have values greater than the ith percentile.

Thus, the ith percentile is a value that at least i percent of the items take this value or less and at least (100 - i) percent of the items take this value or more.

Step1

Arrange the data in ascending order(rank order from smallest value to largest value)

Step2

Compute the position of the percentile from

position = (i/100)n

where i is the percentile of interest and n is the number of items

Step3

(a) If the **position is not an integer**, round up. The next integer value greater than this position denotes the position of the ith percentile.

(b) If the **position is an integer**, the ith percentile is the average of the data values in **position and position + 1.**

(b) Grouped data

The steps are the same as those for finding the median.

For grouped data the ith percentile is computed from the following expression:

i-th percentile(P_i) = L_o + c $\left[\dfrac{i(\sum\frac{f+1}{100}) - C.F_b}{f_{p_i}}\right]$

where L_o = lower class boundary of the percentile class
c = class length/width
$\sum$f = total frequency = n
$C.F_b$ = cumulative frequency preceding (before) that of the percentile class
f_{p_i} =frequency of the percentile class

Note : $\dfrac{i(\sum f+1)}{100}$ = position of the ith percentile

5.5 The variance and standard deviation

Variance is defined as the mean of the squared deviations of individual observations from their arithmetic means denoted by δ^2 for population and S^2 for a sample.
The variance is the measure of dispersion that utilises all the data values.

The variance is based on the difference between each data value, x_i, and the mean ($\overline{X}$, for the sample and μ, for the population).

This difference is known as the **deviation about the mean(see mean absolute deviation).** The variance may be obtained using any of the following formulae:

(a) Ungrouped data

(i) s^2 = $\dfrac{\sum x^2}{n} - (\overline{x})^2 = \dfrac{\sum x^2}{n} - \left(\dfrac{\sum x}{n}\right)^2 = \dfrac{n\sum x^2 - (\sum x)^2}{n^2}$

(ii) s^2 = $\dfrac{\sum(x-\overline{x})^2}{n-1}$

(b) Grouped data

(i) s^2 = $\dfrac{\sum fx^2}{\sum f} - (\overline{x})^2 = \dfrac{\sum fx^2}{\sum f} - \left(\dfrac{\sum fx}{\sum f}\right)^2 = \dfrac{n\sum fx^2 - (\sum fx)^2}{n(n-1)}$

n = ∑f

(ii) $s^2 = \dfrac{\sum f\left(x-\bar{x}\right)^2}{n-1}$, where n = ∑f

(iii) $s^2 = c^2\left[\dfrac{\sum fd^2}{\sum f}-\left(\dfrac{\sum fd}{\sum f}\right)^2\right]$, using the assumed mean(coding method)

or $s^2 = c^2\left[\dfrac{n\sum fd^2-\left(\sum fd\right)^2}{n(n-1)}\right]$, where n = ∑f

Note:

(i) If we are working with the entire population with N data items, we substitute μ for $\overline{X}$ and N for n in the above expressions. However, for (a)(ii) and (b)(ii) we use N in the denominator instead of N-1!

(ii)Standard deviation is defined to be the positive square root of the variance.

Standard deviation = $\sqrt{\text{var}iance}$

For sample, standard deviation, s = $\sqrt{s^2}$

For population, standard deviation, σ = $\sqrt{s^2}$

Example

Find the variance and standard deviation for the following data.

(a)11, 14, 17, 20, 16, 10

Solution

Mean $\overline{X} = \dfrac{\sum_{i=1}^{n} x_i}{n} = \dfrac{11+14+17+20+16+10}{6} = 14.67$ (corrected to 2dp)

Variance, $s^2 = \dfrac{\sum\left(x-\bar{x}\right)^2}{n-1} =$
$[(-3.67)^2 + (-0.67)^2 + (2.33)^2 + (5.33)^2 + (1.33)^2 + (-4.67)^2]/5 = 11.89$ (corrected to 2 dp).

Standard deviation, s = $\sqrt{(11.89)} = 3.45$ (2dp)

(b)

Goals(x)	0	1	2	3	4	5	6
frequency(f)	1	2	5	8	6	3	0

The computations are summarized in the following table:

Goals x	frequency f	fx	(x - x)	$(x-x)^2$	f $(x-x)^2$
o	1	0	-3	9	9
1	2	2	-2	4	8
2	5	10	-1	1	5
3	8	24	0	0	0
4	6	24	1	1	6
5	3	15	2	4	12
6	0	0	3	9	0
	$\sum f = 25$	$\sum fx = 75$			$\sum f\left(x-\bar{x}\right)^2$ =4 0

Hence mean $\overline{X}$ = $\dfrac{\sum fx}{\sum f}$ = 75/25 = 3

Mean variance, s^2 = $\dfrac{\sum f\left(x-\bar{x}\right)^2}{n-1}$, where n = $\sum f = 25$

Hence, s^2 = 40/(25 - 1)= 1.67(2 dp)

Standard deviation, s = $\sqrt{(1.67)} = 1.29$ (2dp)

(c) The following table shows the weights of 100 students measured to the nearest kg.

Weight(kg)	**Number of children**
10-14	5
15-19	9
20-24	12
25-29	18
30-34	25
35-39	15
40-44	10
45-49	6

Solution

Weight	midpoint	frequency	code			
	x	f	d	d^2	fd	fd^2
10-14	12	5	-4	16	-20	80
15-19	17	9	-3	9	-27	81
20-24	22	12	-2	4	-24	48
25-29	27	18	-1	1	-18	18
30-34	32	25	0	0	0	0
35-39	37	15	1	1	15	15
40-44	42	10	2	2	20	40
45-49	47	6	3	9	18	54
		∑f=100			∑fd=-36	$\sum fd^2$= 336

Method of assigning codes (d)

Step 1

Choose the assumed mean, x_A = 32(use the column of mid-points).

Step 2

Assign code 0 to x_A

Step 3

Assign consecutive negative codes to values above x_A and consecutive positive codes to values below it.

The mean, $\overline{X} = \overline{x_A} + c\left(\frac{\sum fd}{\sum f}\right)$, where $\overline{X_A}$ = assumed mean = 32

d = code(in coding method)
c = class length/width = 5

Hence, $\overline{X}$ = 32 +5[-36/100] = 30.20

Variance, $s^2 = c^2 \left[\frac{n\sum fd^2 - \left(\sum fd\right)^2}{n(n-1)}\right]$, where n = ∑f = 100

Variance, $s^2 = 5^2 \left[\frac{100(336) - (-36)^2}{100(100-1)}\right]$ = 81.58 (2 dp)

Hence, standard deviation, s = $\sqrt{(81.58)} = 9.03$ (2 dp)

- Interpretation of the standard deviation

5.6 Coefficient of Variation

the **coefficient of variation (CV)** is a normalized measure of dispersion of a probability distribution. It is also known as **unitized risk** or the **variation coefficient.**

The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to the mean μ :

$$c_v = \frac{\sigma}{|\mu|}$$

which is the inverse of the signal-to-noise ratio. The CV is defined only for *non-zero* mean and the absolute value is taken for the mean to ensure it is always *positive*. It is sometimes expressed as a percent, in which case the CV is multiplied by 100%.[1]

The coefficient of variation should be computed only for data measured on a ratio scale. To demonstrate this using an example, if a group of temperatures are analyzed, the standard deviation does not depend on whether the Kelvin or Celsius scale is used since an object that changes its temperature by 1 K also changes its temperature by 1° C. However the mean temperature of the data set would differ in each scale by an amount of 273 and thus the coefficient of variation would differ. So the coefficient of variation may not have any meaning for data on an interval scale.[2]

[edit] Comparison to standard deviation

[edit] Advantages

The coefficient of variation is useful because the standard deviation of data must always be understood in the context of the mean of the data. The coefficient of variation is a dimensionless number. So for comparison between data sets with different units or widely different means, one should use the coefficient of variation instead of the standard deviation.

[edit] Disadvantages

- When the mean value is close to zero, the coefficient of variation will approach infinity and is hence sensitive to small changes in the mean.
- Unlike the standard deviation, it cannot be used to construct confidence intervals for the mean.

The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to the mean μ :

$$c_v = \frac{\sigma}{|\mu|}$$

which is the inverse of the signal-to-noise ratio. The CV is defined only for *non-zero* mean and the absolute value is taken for the mean to ensure it is always *positive*. It is sometimes expressed as a percent, in which case the CV is multiplied by 100%.[1]

The coefficient of variation should be computed only for data measured on a ratio scale. To demonstrate this using an example, if a group of temperatures are analyzed, the standard deviation does not depend on whether the Kelvin or Celsius scale is used since an object that changes its temperature by 1 K also changes its temperature by 1° C. However the mean temperature of the data set would differ in each scale by an amount of 273 and thus the coefficient of variation would differ. So the coefficient of variation may not have any meaning for data on an interval scale.[2]

[edit] Comparison to standard deviation

[edit] Advantages

The coefficient of variation is useful because the standard deviation of data must always be understood in the context of the mean of the data. The coefficient of variation is a dimensionless number. So for comparison between data sets with different units or widely different means, one should use the coefficient of variation instead of the standard deviation.

[edit] Disadvantages

- When the mean value is close to zero, the coefficient of variation will approach infinity and is hence sensitive to small changes in the mean.
- Unlike the standard deviation, it cannot be used to construct confidence intervals for the mean.

The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to the mean μ :

$$c_v = \frac{\sigma}{|\mu|}$$

which is the inverse of the signal-to-noise ratio. The CV is defined only for *non-zero* mean and the absolute value is taken for the mean to ensure it is always *positive*. It is sometimes expressed as a percent, in which case the CV is multiplied by 100%.[1]

The coefficient of variation should be computed only for data measured on a ratio scale. To demonstrate this using an example, if a group of temperatures are analyzed, the standard deviation does not depend on whether the Kelvin or Celsius scale is used since an object that changes its temperature by 1 K also changes its temperature by 1° C. However the mean

temperature of the data set would differ in each scale by an amount of 273 and thus the coefficient of variation would differ. So the coefficient of variation may not have any meaning for data on an interval scale.[2]

[edit] Comparison to standard deviation

[edit] Advantages

The coefficient of variation is useful because the standard deviation of data must always be understood in the context of the mean of the data. The coefficient of variation is a dimensionless number. So for comparison between data sets with different units or widely different means, one should use the coefficient of variation instead of the standard deviation.

[edit] Disadvantages

- When the mean value is close to zero, the coefficient of variation will approach infinity and is hence sensitive to small changes in the mean.
- Unlike the standard deviation, it cannot be used to construct confidence intervals for the mean.

Applications

The coefficient of variation is also common in applied probability fields such as renewal theory, queueing theory, and reliability theory. In these fields, the exponential distribution is often more important than the normal distribution. The standard deviation of an exponential distribution is equal to its mean, so its coefficient of variation is equal to 1. Distributions with CV < 1 (such as an Erlang distribution) are considered low-variance, while those with CV > 1 (such as a hyper-exponential distribution) are considered high-variance. Some formulas in these fields are expressed using the **squared coefficient of variation**, often abbreviated SCV. In modeling, a variation of the CV is the CV(RMSD). Essentially the CV(RMSD) replaces the standard deviation term with the Root Mean Square Deviation (RMSD).

[edit] Distribution

Provided that negative and small positive values of the sample mean occur with negligible frequency, the probability distribution of the coefficient of variation for a sample of size *n* has been shown by Hendricks and Robey [3] to be

$$dF_{c_v} = \frac{2}{\pi^{1/2}\Gamma\left(\frac{n-1}{2}\right)} e^{-\frac{n}{2\frac{\sigma}{\mu}^2}\frac{{c_v}^2}{1+{c_v}^2}} \frac{{c_v}^{n-2}}{(1+{c_v}^2)^{n/2}} \sum_{i=0}^{n-1}{}' \frac{(n-1)!\Gamma\left(\frac{n-i}{2}\right)}{(n-1-i)!i!} \frac{n^{i/2}}{2^{i/2}(\frac{\sigma}{\mu})^i} \frac{1}{(1+{c_v}^2}$$

where the symbol $\sum{}'$ indicates that the summation is over only even values of *n*-1-*i*.

This is useful, for instance, in the construction of hypothesis tests or confidence intervals.

Review Questions

Question 1

a) Briefly explain the following: -
   (i) Median
   (ii) Mode

b) Differentiate between the weighted mean and the arithmetic mean

c) The Table below shows the monthly allowances in Dollars of the employees in American Embassy in a certain Country.

Allowances ($)	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 - 69
No. of Employees	12	14	16	10	15	8

Calculate:
   (i) The mean
   (ii) The median
   (iii) The mode

Question 2

a) Distinguish between the following
   (i) Decile and Percentile
   (ii) Range and Inter-quartile range

b) Briefly explain Quartile?

c) The Table below shows the ages in years of the contestants for Miss World.

Age	20 - 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49	50 – 54
frequency	12	24	30	18	11	5	1

   (i) Draw the histogram and the frequency polygon of the above Data on the same axis.
   (ii) Calculate the Inter-quartile range.

Question 3

a) Briefly explain the following
   i. Arithmetic mean

ii. Weighted mean
iii. The Median
iv. Variance
v. Standard deviation

b) The Table below shows the age of participants in a certain workshop in Hotel Africana

Ages	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49	50 - 54
frequency	11	24	30	18	11	5	1

i. Calculate the inter-quartile Range
ii. Calculate the mean
iii. Calculate the median
iv. Calculate the mode
v. Calculate the standard deviation

Unit 6
Introduction to Probability Theory

This unit provides an introduction to probability theory. Experience resulting from repeated experiments or from recurrent observations, is frequently used to predict the outcome of future events. For example, from the past and present knowledge of weather conditions in a particular locality, we may say that it will probably be warm tomorrow or that there is a probability of rain. In this sense we are using the word probability to denote a belief founded on a certain amount of evidence and in many cases, no doubt, influenced by considerable wishful thinking.

- In mathematical usage the meaning of the word probability is established by definition and is not connected with beliefs or wishful thinking.
- Statistics and probability are so fundamentally interrelated that it is impossible to discuss statistics without an understanding of the meaning of probability.
- Knowledge of probability theory makes it possible to interpret statistical results, since many statistical procedures involve conclusions based on samples which are always affected by random variation, and it is by means of probability theory that we can express numerically the inevitable uncertainties.

6.1 Basic Set Concepts Sample Space, Sample Point, Event And Probability Of An Event

A **set** is a well defined collection of objects. Elements of a set are enclosed in the braces {} e.g. A = {Students of Kampala International University} is a set.

B = {…, -2, -1, 0, 1, 2, …} is a set of integers

If every element of a set A is also contained in a set B, then A is called a **subset of B** i.e. $A \subset B$.

A **universal set** is a set containing everything under the field of study. It is denoted by U or ε .

Note:

(i) All subsets belong to the universal set

(ii) If x is an element/member of the set A, we write: $x \in A$

An empty set, denoted by ϕ: **is a set with no elements**

The **intersection** of two sets A and B is denoted by : $A \cap B$, and is the set of all elements in **A and B**.

i.e. $\mathbf{A \cap B = \{ x \in A \text{ and } x \in B\}}$

An element qualifies for the **intersection** of A,B if it is in both A and B. For example, if A=(2, 8, 14, 18) and B=(4, 6, 8, 10, 12), then the intersection of (A,B)=8, i.e. **$A \cap B$ = { x $\in$A and x $\in$B}={8}**

The key word indicating the intersection of two or more events is and

Note:

If $A \cap B = \phi$, then the sets A and B are said to be disjoint i.e. A and B do not have any element(s) in common.

The **union** of two sets A and B is denoted by : $A \cup B$ and is the set of all elements in **either A or B or both.**

i.e. $A \cup B$ = { **x $\in$A or x $\in$B}**

An element qualifies for the **union** of A, B if it is in either A or B or in both A and B. For example, if A=(2, 8, 14, 18) and B=(4, 6, 8, 10, 12), then the union of (A,B)=(2, 4, 6, 8, 10, 12, 14, 18), **i.e.** $A \cup B$ = { **x $\in$A or x $\in$B} = {2, 4, 6, 8, 10, 12, 14, 18}**

The key word indicating the union of two or more events is or.

If A is a set of the universal set U, then the **complement of the set A** with respect to U is denoted by: **A'** and is the set of all elements in U but not in A.

i.e. A' = { **x $\in$U and x $\notin$A}**

Note:

The above basic set concepts are a useful tool in understanding the probability concepts since there is a one-to-one correspondence between set theory and probability theory.

A **statistical experiment** is any process that generates raw data. Experiment is an activity that is either observed or measured

For example the following are statistical experiments:

- Tossing a die
- Flipping a coin
- Drawing a card

Any outcome of a statistical experiment is called an **event.** An event is a possible outcome of an experiment. For example, if the experiment is to sample six lamps coming off a production line, an event could be to get one defective and five good ones.

Elementary Events: Elementary events are those types of events that cannot be broken into other events. For example, suppose that the experiment is to roll a die. The elementary events for this experiment are to roll a 1 or a 2, and so on, i.e., there are six elementary events (1, 2, 3, 4, 5, 6). Note that rolling an even number is an event, but it is not an elementary event, because the even number can be broken down further into events 2, 4, and 6.

A set whose elements represent **all the possible outcomes** of a statistical experiment is called a **sample space,** denoted by S. A sample space is a complete set of all events of an experiment. The sample space for the roll of a single die is 1, 2, 3, 4, 5, and 6.

The sample space of the experiment of tossing a coin three times is:

First toss.........T	T	T	T	H	H	H	H
Second toss.....T	T	H	H	T	T	H	H
Third toss........T	H	T	H	T	H	T	H

Note:

Sample space can aid in finding probabilities. However, using the sample space to express probabilities is hard when the sample space is large. Hence, we usually use other approaches to determine probability.

Note:

An event is a subset of a sample space.
For example, tossing a die, S = {1, 2, 3, 4, 5, 6}
Flipping a coin, S = {h, t}, where h = head and t = tail
Any element of a sample space is called a **sample point.**

Note:

An event consists of one or more sample points.
The probability of an event A is the sum of weights assigned to all sample points in an event A. It is denoted by : P(A) (or Pr(A) or Prob(A))

Probability theory provides a way to find and express our uncertainty in making decisions about a population from sample information. Probability is a number between 0 and 1. The highest value of any probability is 1.

Probability reflects the long-run relative frequency of the outcome. A probability is expressed as a decimal, such as 0.7 or as a fraction, such as 7/10, or as percentage, such as 70%.

Note:

(i) The probability of an event A, P(A), lies between 0 and 1. That is, it can neither be negative nor greater than 1
i.e $0 \leq P(A) \leq 1$

(ii) $P(\phi) = 0$, and $P(S) = 1$. That is , the probability of an empty set is 0 and that of the sample space is 1.

(iii) If P(A) = 0, the event A cannot occur, and if P(A) = 1 , then the event will certainly occur.

(iv) Weights are assigned to sample points in the sample space, S.

(v) If A and B are two events, then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(vi) If A is an event and A' is its complement, then P(A) + P(A') = 1

Approaches of Assigning Probabilities

There are three approaches of assigning probabilities, as follows:

1. Classical Approach

Classical probability is predicated on the assumption that the outcomes of an experiment are equally likely to happen. The classical probability utilizes rules and laws. It involves an experiment. The following equation is used to assign classical probability:

P(A) = Number of favorable outcomes / Total number of possible outcomes

Note:

We can apply the classical probability when the events have the same chance of occurring (called equally likely events), and the set of events are mutually exclusive and collectively exhaustive.

2. Relative Frequency Approach

Relative probability is based on cumulated historical data. The following equation is used to assign this type of probability:

P(A) = Number of times an event occurred in the past/ Total number of opportunities for the event to occur.

Note :

The relative probability is not based on **rules** or **laws** but on what has happened in the past. For example, the company wants to decide on the probability that its inspectors are going to reject the next batch of raw materials from a supplier. Data collected from the company record books show that the supplier had sent your company 80 batches in the past, and inspectors had rejected 15 of them. By the method of relative probability, the probability of the inspectors rejecting the next batch is 15/80, or 0.19. If the next batch is rejected, the relative probability for the subsequent shipment would change to 16/81 = 0.20.

3. Subjective Approach

The subjective probability is based on personal judgment, accumulation of knowledge, and experience. For example, medical doctors sometimes assign subjective probabilities to the length of life expectancy for people having cancer. Weather forecasting is another example of subjective probability.

Types of Probability

Four types of probabilities are discussed in this unit:

1. Marginal Probability

A marginal probability is usually calculated by dividing some subtotal by the whole. For example, the probability of a person wearing glasses is calculated by dividing the number of people wearing glasses by the total number of people. Marginal probability is denoted P(A), where A is some event.

2. Union Probability

A union probability is denoted by P(A or B), where A and B are two events. P(A or B) is the probability that A will occur or that B will occur or that both A and B will occur. The probability of a person wearing glasses or having blond hair is an example of union probability. All people wearing glasses are included in the union, along with all blondes and all blond people who wear glasses.

3. Joint Probability

A joint probability is denoted by P(A and B). To become eligible for the joint probability, both events A and B must occur. The probability that a person is a blond head and wears glasses is an example of joint probability.

4. Conditional Probability

A conditional probability is denoted by P(A|B). This phrase is read: the probability that A will occur given that Y is known to have occurred. An example of conditional probability is the probability that a person wears glasses given that she is blond.

Methods to Use in Solving Probability Problems

There are indefinite numbers of ways which can be used in solving probability problems. These methods include the tree diagrams, laws of probability, sample space, insight, and contingency table. Because of the individuality and variety of probability problems, some approaches apply more readily in certain cases than in others.

There is no best method for solving all probability problems.

6.2 Laws of Probability; Multiplication Law, Addition Law

Three laws of probability are discussed in this unit:

- the additive law,
- the multiplication law, and
- The conditional law.

The Additive Law

A. General Rule of Addition

When two or more events will happen at the same time, and the events are not mutually exclusive, then:

P(A or B) = P(A) + P(B) - P(A and B)

For example, what is the probability that a card chosen at random from a deck of cards will either be a king or a heart?
P(King or Heart) = P(X or Y) = 4/52 + 13/52 - 1/52 = 30.77%

B. Special Rule of Addition:

When two or more events will happen at the same time, and the events are mutually exclusive, then:

P(X or Y) = P(X) + P(Y)

For example, suppose we have a machine that inserts a mixture of beans, peas, and other types of vegetables into a plastic bag. Most of the bags contain the correct weight, but because of slight variation in the size of the beans and other vegetables, a package might be slightly underweight or overweight. A check of many packages in the past indicate that:

Weight	**Event**	**No. of Packages**	**Probability**
Underweight	A	100	0.025
Correct weight	B	3600	0.9
Overweight	C	300	0.075
Total		**4000**	**1.00**

What is the probability of selecting a package at random and having the package be under weight or over weight? Since the events are mutually exclusive, a package cannot be underweight and overweight at the same time. The answer is:

P(A or C) = P(0.025 + 0.075) = 0.1

The Multiplication Law

A. General Rule of Multiplication:

When two or more events will happen at the same time, and the events are dependent, then the general rule of multiplication law is used to find the joint probability:

P(A and B) = P(A) . P(B|A)

For example, suppose there are 10 marbles in a bag, and 3 are defective. Two marbles are to be selected, one after the other without replacement. What is the probability of selecting a defective marble followed by another defective marble?

Probability that the first marble selected is defective:

P(A)=3/10

Probability that the second marble selected is defective:

P(B)=2/9

P(A and B) = (3/10) . (2/9) = 7%

This means that if this experiment were repeated 100 times, in the long run 7 experiments would result in defective marbles on both the first and second selections. Another example is selecting one card at random from a deck of cards and finding the probability that the card is an 8 and a diamond.

P(8 and diamond) = (4/52) . (1/4) = 1/52 which is = P(diamond and 8) = (13/52) . (1/13) = 1/52.

B. Special Rule of Multiplication:

when two or more events will happen at the same time, and the events are independent, then the special rule of multiplication law is used to find the joint probability:

P(A and B) = P(A) . P(B)

If two coins are tossed, what is the probability of getting a tail on the first coin and a tail on the second coin?
P(T and T) = (1/2) . (1/2) = 1/4 = 25%. This can be shown by listing all of the possible outcomes: T T, or T H, or H T, or H H. Games of chance in casinos, such as roulette and craps, consist of independent events. The next occurrence on the die or wheel should have nothing to do with what has already happened.

The Conditional Law

Conditional probabilities are based on knowledge of one of the variables. The conditional probability of an event, such as X, occurring given that another event, such as Y, has occurred is expressed as:

P(A|B) = P(A and B) / P(B) = {P(A) . P(B|A)} / P(B)

Note

When using the conditional law of probability, you always divide the joint probability by the probability of the event after the word given. Thus, to get P(A given B), you divide the joint probability of A and B by the unconditional probability of B. In other words, the above equation is used to find the conditional probability for any two **dependent events**. When two events, such as A and B, are **independent** their conditional probability is calculated as follows:

P(A|B) = P(A) and P(B|A) = P(B)

For example, if a single card is selected at random from a deck of cards, what is the probability that the card is a king given that it is a club?
P(king given club) = P (A|B) = {P(A) .P(B|A)} / P(B)
P(B) = 13/52, and P(king given club) = 1/52, thus
P(king given club) = P(A|B) = (1/52) / (13/52) = 1/13
Note that this example can be solved conceptually without the use of equations. Since it is given that the card is a club, there are only 13 clubs in the deck. Of the 13 clubs, only 1 is a king. Thus P(king given club) = 1/13.

Combination Rule:

The combination equation is used to find the number of possible arrangements when there is only **one group of objects and when the order of choosing is not important**. In other words, combinations are used to summarize all possible ways that outcomes can occur without listing the possibilities by hand. The combination equation is as follows:

$C = \frac{n!}{x!(n-x)!}$ **and 0<= x <="n"**

where: n = total number of objects, x= number of objects to be used at one time,
C = number of ways the object can be arranged, and ! stands for factorial.
Note: 0! = 1, and 3! means 3x2x1.

<u>Example 5.1</u>
For example, suppose that 4% of all TVs made by W&B Company in 1995 are defective. If eight of these TVs are randomly selected from across the country and tested, what is the probability that exactly three of them are defective? Assume that each TV is made independently of the others.

Using the combination equation to enumerate all possibilities yields:

C = 8!/ 3! (8-3)! = (8x7x6x5!)/ {(3x2x1)(5!) = 336/6 = 56

which means there are 56 different ways to get three defects from a total of eight TVs. Assuming D is a defective TV and G is a good TV, one way to get three defecs would be: P (D1 and D2 and D3 and G1 and G2 and G3 and G4 ang G5). Because the TVs are made independently, the probability of getting the first three defective and the last five good is:
(.04)(.04)(.04)(.96)(.96)(.96)(.96)(.96)=0.0000052 which is the probability of getting three defects in the above order. Now, multiplying the 56 ways by the probability of getting one of these ways gives:
(56)(0.0000052)=0.03%, which is the answer for drawing eight TVs and getting exactly three defectives (in above order).

Formulas

General Rule of Addition
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Special Law of Addition
$P(A \cup B) = P(A) + P(B)$

General Law of Multiplication

$$P(A \cap B) = P(A).P(B|A) = P(B).P(A|B)$$

Special Law of Multiplication

$$P(A \cap B) = P(A).P(B)$$

Law of Conditional Probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A).P(B \mid A)}{P(B)}$$

Combination Formula

$$^{n}C_{r} = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Example 5.2

A coin is tossed twice. What is the probability that at least one head will occur?

Note:

(i) Tossing a coin twice is the same as tossing two coins once.

(ii) If the coin is balanced, then all the sample points are equally likely, i.e. the sample points should be equally weighted.

Therefore, the sample space, S = { hh, ht, th, tt}

Let w be the weight of each sample point.

Hence w + w + w + w = P(S) = 1

4w = 1. Hence w = ¼

Hence the probability of each sample point is ¼

Let A be the event that **at least one head will occur.**

A = {hh, ht, th}
w, w, w

Hence, P(A) = w + w + w = ¼ + ¼ +1/4 = ¾

Alternatively, P(at least one head) + P(no head) = 1

i.e. P(A) + P(A') = 1

P(A) +P(tt) = 1. Hence P(A) = 1- P(A') = 1 - ¼ = ¾

Note:

If an experiment can result in any of the N different equally likely outcomes and if exactly n of these correspond to event A, then the probability of event A is P(A) = n/N

For the above example, the process of tossing a coin twice, results in four outcomes (N = 4). Exactly three (n = 3) of these correspond to event A. Thus P(A) = ¾

Example 5. 3

An even number is twice as likely to occur as an odd number when a die is tossed. What is the probability that a number less than 4 occurs?

Solution

The sample space, $S = \{1, 2, 3, 4, 5, 6\}$

Let w be the weight of an odd number. Then the weight of an even number is $2w$, since an even number is twice as likely to occur, as an odd number.

Hence, $S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$

$w,\ 2w,\ w,\ 2w,\ w,\ 2w$

Therefore, $w + 2w + w + 2w + w + 2w = 9w$

$w = 1/9$

Thus, $S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$

$1/9,\ 2/9,\ 1/9,\ 2/9,\ 1/9,\ 2/9$

Let A be the event that a number less than 4 occurs. Then, $A = \{1,\ 2,\ 3\}$

$1/9,\ 2/9,\ 1/9$

Hence, $P(A) = 1/9 + 2/9 + 1/9 = 4/9$.

Note:

The definition, $P(A) = n/N$, is not applicable here, since the outcomes are not "equally likely".

6.3 Types of Events

Mutually Exclusive Events

Those events that cannot happen together are called mutually exclusive events. For example, in the toss of a single coin, the events of heads and tails are mutually exclusive. The probability of two mutually exclusive events occurring at the same time is zero.

Two events A and B are said to be mutually exclusive events if

$$\mathbf{A \cap B = \phi, \quad therefore, \quad P(A \cap B) = 0.}$$

Thus, two events are mutually exclusive , if the occurrence of either event excludes the possibility of the occurrence of the other event i.e. either one or the other event, but not both, can occur.

Note:

If two events A and B are mutually exclusive, then, $P(A \cup B) = (A) + P(B)$.

Example 5.4

(a) If a die is tossed and A is the event of obtaining an even number
i.e. $A = \{2, 4, 6\}$
and B is the event of obtaining an odd number,
i.e. $B = \{1, 3, 5\}$
Then , $\mathbf{A \cap B = \phi}$, **therefore,** $\mathbf{P(A \cap B) = 0}$. **Therefore, A and B are mutually exclusive events.**

(b) The probability that a student passes Statistics is 2/3, and the probability that s(h)e passes Corporate Finance is 4/9. If the probability of passing at least one course is 4/5, what is the probability of passing both courses?

Solution

Let A = event of passing Statistics, $P(A) = 2/3$
Let B = event of passing Corporate Finance, $P(B) = 4/9$
Let $A \cup B$ = event of passing at least one of them, $P(A \cup B) = 4/5$
Let $A \cap B$ = event of passing both, $P(A \cap B) = ?$

Note: These are not mutually exclusive events: both A and B can occur.

Hence, using $\mathbf{P(A \cup B) = P (A) + P(B) - P(A \cap B)}$
We have: $\mathbf{4/5 = 2/3 + 4/9 - P(A \cap B)}$
$P(A \cap B) = 14/45$

(c) What is the probability of getting a total of 7 or 11 when a pair of dice is tossed?

Solution

The sample points are shown in the following diagram

		First die
	+	1 2 3 4 5 6

S	1	2	3	4	5	6
e	2	**7**				
c	3	3	4	5	6	**7**
die	4	8				
o	5	4	5	6	**7**	8
n	6	9				
d		5	6	**7**	8	9
		10				
		6	**7**	8	9	10
		11				
		7	8	9	10	**11**
		12				

Total number of sample points, in the sample space, S is 36.Thus, N(S) = 36

Let A = event of getting a total of 7, n(A) = 6

Hence, P(A) = n(A)/N(S) = 6/36

Let B = event of getting a total of 11, n(B) = 2

Hence, P(B) = n(B)/N(S) = 2/36

Note:

$A \cap B = \phi$, and thus $P(A \cap B) = 0$. Therefore, A and B are mutually exclusive.

Hence, **$P(A \cup B) = P(A) + P(B) = 6/36 + 8/36 = 2/9$**

Independent Events

Two or more events are called independent events when the occurrence or nonoccurrence of one of the events does not affect the occurrence or nonoccurrence of the others. Thus, when two events are independent, the probability of attaining the second event is the same regardless of the outcome of the first event. For example, the probability of tossing a head is always 0.5, regardless of what was tossed previously. Note that in these types of experiments, the events are independent if sampling is done with replacement.

Two events are said to be independent events if :

$P(A \cap B) = P(A) .P(B)$

Therefore, two events are said to be independent if the occurrence or non-occurrence of one event has no influence on the occurrence or non-occurrence of the other.

Conditional Events

Two events are said to be conditional events if :

$P(B/A) = P(B\cap A)/P(A)$, provided $P(A) \neq 0$.

That is, event B occurs given that event A has already occurred.

Collectively Exhaustive Events

A list of collectively exhaustive events contains all possible elementary events for an experiment. For example, for the die-tossing experiment, the set of events consists of 1, 2, 3, 4, 5, and 6. The set is collectively exhaustive because it includes all possible outcomes. Thus, all sample spaces are collectively exhaustive.

Complementary Events

The complement of an event such as A consists of all events not included in A. For example, if in rolling a die, event A is getting an odd number, the complement of A is getting an even number. Thus, the complement of event A contains whatever portion of the sample space that event A does not contain.

Contingency Table

This table is quite useful in obtaining probabilities especially if events are independent. It is illustrated by way of example below.

<u>Example 5.5</u>

(a) If a coin(with sides labelled: h and t) and a die (with sides numbered: 1 to 6) are
thrown, the way in which a die lands in no way does it affect the possible ways in which a coin lands and vice-versa.

Therefore, throwing a die with a six, for example, and throwing a coin with a head, are independent events. If both are fair/unbiased, then

P(head and a six) = P(six).P(head) =1/2 x 1/6 =1/12

(b) In an examination only two papers namely Microeconomics I and

Microeconomics II were done. The failure rates were 45% and 40% respectively.

The number of candidates who sat the examination was 2000. Find the probability that a candidate selected at random,

(i) Passed both Microeconomics I and Microeconomics II
(ii) failed both Microeconomics I and Microeconomics II
(iii) passed Microeconomics I and failed Microeconomics II

Solution

Let A = event of passing Microeconomics I . Then A' = event of failing.
Thus , $P(A') = 45/100 = 0.45$, $P(A) = 1 - P(A') = 0.55$
Let B = event of passing Microeconomics II. Then, B' = event of failing.
Thus, $P(B') = 40/100 = 0.40$, $P(B) = 1 - P(B') = 1 - 0.4 = 0.6$

Note:

Passing/failing Microeconomics I in no way affects passing/failing Microeconomics II and vice-versa.
Hence, the two events are independent.

(i) Thus, $P(\text{ A and B}) = P(A \cap B) = P(A).P(B) = 0.55 \times 0.6 = 0.33$

(ii) Note: It can be shown that if A and B are independent, so are A' and B' .
Hence, $P(A' \text{ and } B') = P(A' \square' B) = P(A') \times P(B') = 0.45 \times 0.4 = 0.18$

(iii) $P(A \text{ and } B') = P(A \cap B') =?$

This may be obtained from the following contingency table:

	A	A'	
B	$P(A \cap B) = 0.33$	$P(A' \cap B) = 0.27$	$P(B) = 0.6$
B'	$P(A \cap B') =0.22$	$P(A' \cap B') = 0.18$	$P(B') = 0.4$
	$P(A) = 0.55$	$P(A') = 0.45$	1

Note: The table above is filled by using the following equations:

(a) Adding row-wise we have:
(i) $P(A \cap B) + P(A' \cap B) = P(B)$
(ii) $P(A \cap B') + P(A' \cap B') = P(B')$
(iii) $P(A) + P(A') = 1$

(b) Adding column -wise we have:
(i) $P(A \cap B) + P(A \cap B') = P(A)$
(ii) $P(A' \cap B) + P(A' \cap B') = P(A')$
(iii) $P(B) + P(B') = 1$

Therefore, $P(A \cap B') = 0.22$.

Conditional Probability

The **conditional probability of B given A, denoted by: P(B/A),** is the probability that B occurs given that A has already occurred, and it is defined by:

$P(B/A) = P(B\cap A)/P(A)$, provided $P(A) \neq 0$.

Note:

(i) $P(B\cap A) = P(A \cap B)$, using commutative laws.

(ii) Similarly, $P(A/B) = P(A\cap B)/P(B)$, provided $P(B) \neq 0$.

Example 5.6

A loaded (biased) die with an even number twice as likely as an odd number is tossed

(i) What is the probability of getting a four?

(ii) What is the probability of getting a four given that a number greater than three occurred?

Solution

Using the results of example 2, we have:

S = {1, 2, 3, 4, 5, 6}
1/9, 2/9, 1/9, 2/9, 1/9, 2/9

Let A = event of obtaining a four

(i) A = {4} , hence P(A) = 2/9

(ii) Let B = event of obtaining a number greater than 3

B = {4, 5, 6} , hence, P(B) = 2/9 + 1/9 + 2/9 = 5/9

Note:

$A\cap B =\{4\}$, hence, $P(A\cap B) = 2/9$

Therefore, $P(A/B) = P(A\cap B)/P(B) = (2/9)/(5/9) = 2/9$

Baye's Theorem

This is a useful Theorem used to simplify and obtain conditional probabilities. It is illustrated by way of example below:

Example 5.7

The table below shows the number of males and females who are employed and those who are not employed.

	Employed	Unemployed
Males	140	260
Females	460	40

(i) Determine the probability that a female is selected given that she is unemployed

(ii) If the following additional information is available:

36 employed and 12 unemployed belong to a club, determine the probability that the selected person belongs to a club.

Solution:

Let F = event a female is selected i.e. a man is not selected = M'

Then, F' = event female is not selected i.e. a man is selected = M

Let E = event a person is employed.

Then, E' = event person is unemployed

(i) Required to find P(F/E')

	Employed	Unemployed	**Total**
Males	140	260	**400**
Females	460	40	**500**
Total	**600**	**300**	**900**

Let S = sample space , N(S) = 900, also n(E) = 600, n(E') = 300, n(M) = 400

n(F) = 500

Now, P(F/E') = P(F $\cap$ E')/P(E'), provided P(E') is not zero

But n(F $\cap$ E ') = 300, thus, P(F $\cap$ E ') = n(F$\cap$ E ')/ N(S)

= 40/900 = 4/90

also P(E') = n(E')/N(S) = 300/900 = 1/3

Hence, P(F/E') = (4/90)/(1/3) = 12/90

(ii) Required to find the probability that the selected person belongs to a club. This may be rephrased as: " the selected person is employed given that s(h)e belongs to a club".

Let C = event that the selected person belongs to a club.

Thus, required to find: P(E/C).

Now, P(E/C) = P(E $\cap$ C) /P(C), and P (C) is the probability that : " a person belongs to a club and is employed" or " a person belongs to a club and s(h)e is unemployed"

i.e. P (C) =P [(E and C) or (E' and C)] = P [(E$\cap$ C) $\cup$ (E'$\cap$ C)]

Note:

(i) The events (E and C) and (E' and C) are mutually exclusive.

Hence, P [(E $\cap$ C) $\cup$ (E'$\cap$ C)] = P (E $\cap$ C) + P(E' $\cap$ C)

Therefore,

P (C) = P (C$\cap$ E) + P(C $\cap$ E')

Hence, P(E/C) = P(E $\cap$C) /P(C)

$= P(E \cap C) / [P (E\cap C) + P(E'\cap C)]$ -----------------------------
--(1)

(ii) The result in equation (1) is known as : Baye's Theorem", a useful tool in evaluating conditional probabilities.

(iii) The following multiplicative law is useful in simplifying Baye's theorem:

$$P(E \cap C) = P(E). P(C/E)$$

Similarly, $P(E' \cap C) = P(E'). P(C/E')$

Hence, equation (1) may as well be written as:

$P(E/C) = P(E \cap C) / [P (E\cap C) + P(E'\cap C)]$

$= P(E). P(C/E) /[P(E). P(C/E) + P(E '). P(C/E')]$ ……………………..(2)

Now, $P (E\cap C) = n(E\cap C) /N(S) = 36/900$,

and $P(E' \cap C) = n(E' \cap C)/N(S) =12/900$

Hence , using equation (1) we have: $P(E/C) = (36/900)/[36/900 + 12/900] =3/4$

<u>Example 5.8</u>

Tom is to travel from Kampala to Lira for an interview. The probabilities that he will be on time for the interview given that he travels by bus and taxi are respectively 0.1 and 0.2. The probability that he will travel by bus and taxi are respectively 0.6 and 0.4.

(i) Find the probability that he will be on time

(ii) Find the probability that he travelled by a taxi given that he is on time

(iii) Find the probability that he travelled by bus given that he is not on time.

Solution

Let A = event that he travels by bus, $P(A) = 0.6$

B = event that he travels by taxi. $P(B) = 0.4$

Let T = event that he is on time =?

Given: $P(T/A) = 0.1$ and $P(T/B) = 0.2$

Required to find P(T).

Now T = event that: " he is on time **and** he has travelled by bus" **or** "he is on time **and** he has travelled by taxi"

i.e. T = (T and A) or (T and B)

$= (T \cap A) \cup (T \cap B)$

Hence, $P(T) = P[(T \cap A) \cup (T \cap B)]$

But the events $(T \cap A)$ and $(T \cap B)$ are mutually exclusive events.

Therefore, $P(T) = P(T \cap A) +P (T \cap B)$

$= P(A).P(T/A) + P(B).P(T/B)$

$= (0.1 \times 0.6) + (0.2 \times 0.4) = 0.14$

(ii) Required to find the probability that he travelled by taxi given that he is on time

i.e. required to find P(B/T)

Solution

$P(B/T) = P(B \cap T) / P(T)$

$= P(B \cap T) / [P(B \cap T) + P(A \cap T)]$

$= P(B).P(T/B) / [P(B).P(T/B) + P(A).P(T/A)]$

$= (0.4 \times 0.2)/[(0.4 \times 0.2) + (0.6 \times 0.1)] = 8/14$

(iii) Required to find the probability that he travelled by bus given that he is not on time i.e. required to find P(B/T')

Solution

Now $P(A/T') = P(A \cap T') / P(T')$

$= [P(A).P(T'/A)] / P(T')$

But $P(T) + P(T') = 1$. Hence, $P(T') = 1 - P(T)$

$= 1 - 0.14 = 0.86$

and $P(T/A) + P(T'/A) = 1$. Hence, $P(T'/A) = 1 - P(T/A)$

$= 1 - 0.1 = 0.9$

Therefore, $P(A/T) = (0.6 \times 0.9) / 0.86 = 54/86$

6.4 Probability Tree Diagrams

A probability tree is a diagram used to obtain probabilities of events, especially if there is picking with/without replacement, and if the events are conditional such that Baye's Theorem may be used. The method is illustrated in the example below:

Example

A bag contains 2 white and 4 red balls. Another bag contains 2 white balls and 1 red ball. A ball is chosen at random from the first bag and put in the second bag, and a ball is randomly selected from the second bag.

(i) Find the probability that the selected ball from the second bag is white

(ii) Find the probability that a white ball was transferred from the first bag given that a white ball was selected from the second bag.

Solution

Let W_1 = white ball in bag1 and R_1 = red ball in bag1

Let W_2 = white ball in bag2 and R_2 = red ball in bag2

$2W_1$
$4R_1$

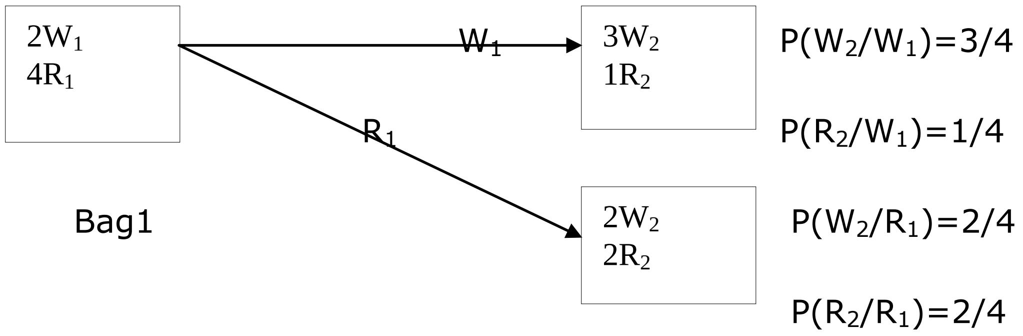
$2W_2$
$1R_2$

Bag 2

The following is what happens after a ball is transferred from bag1

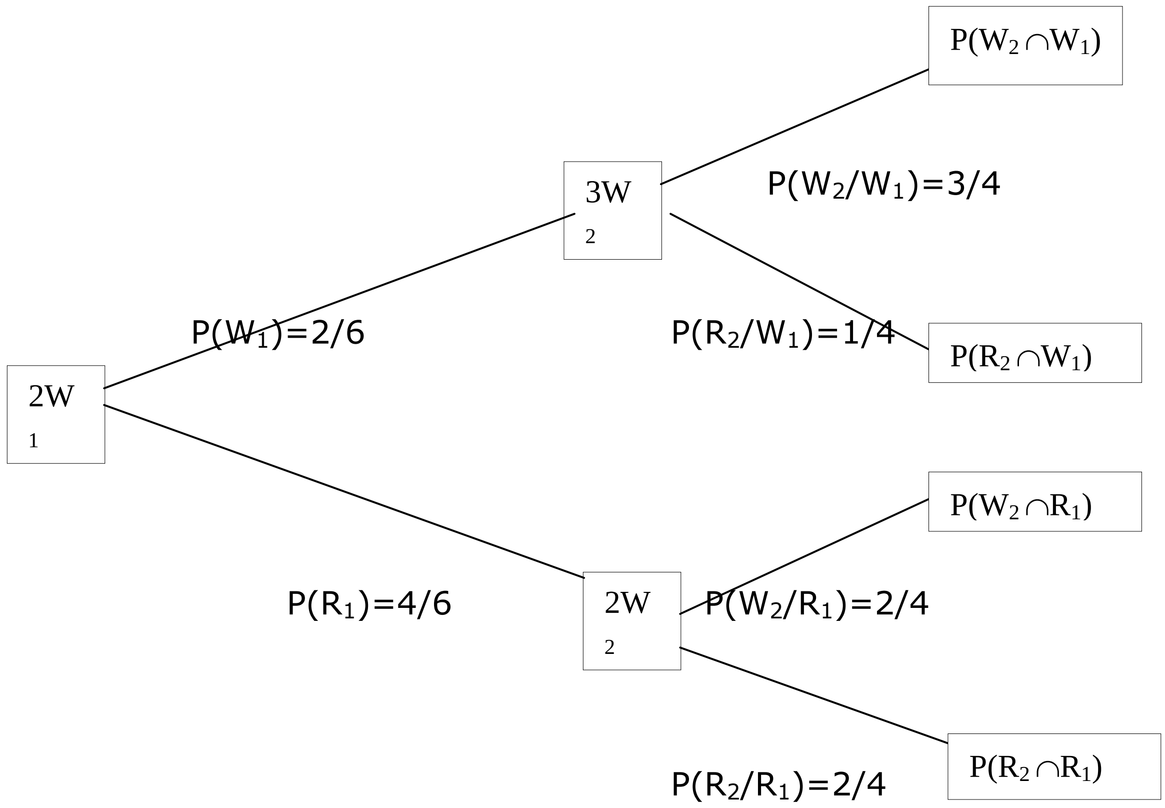
$P(W_1)=2/6$

$P(R_1)=4/6$



Bag2

Transferring a ball from bag1 Picking a ball from bag2



(i) Required to find $P(W_2)$

$P(W_2) = P(W_2 \cap W_1) + P(W_2 \cap R_1)$

$= P(W_1).P(W_2/W_1) + P(R_1).P(W_2/R_1) = (2/6 \times ¾) + (4/6 \times 2/4)$

$=7/12$

(ii) Required to find $P(W_1/W_2)$

By definition $P(W_1/W_2) = P(W_1 \cap W_2) /P(W_2)$

$= P(W_1).P(W_2/W_1) /[P(W_1).P(W_2/W_1) + P(R_1).P(W_2/R_1)]$

$=(6/24)/(7/12) = 3/7$

Review Questions

Question 1

a) (i) What are the properties of probability?

(ii) Find the probability that either a Club or a Spade is drawn in a single draw from a deck of cards

b) Prove that for any event A $P(A^c) = 1 - P(A)$

c) Given that Events A, B and C are mutually exclusive with the probabilities

(A) = 0.2, P(B) = 0.3 and P(C) = 0.2

Find:

i) $P (A^{|})$

ii) P(BUC)

iii) $P(AUC^{|})$

iv) P(AUBUC)

Question 2

a) Write brief notes on the following: -

(i) Mutually exclusive events

(ii) Independent events

(iii) Complementary events

b) Three Coins are tossed simultaneously. Determine the probability of getting more than one Head.

c) Jack and Jill sell Insurance in a family business. Jack sells 80% of the policies and Jill sells 20%. Of the 80% policies sold by Jack, 10% of the policyholders file a claim, and of the 20% policies sold by Jill 25% of the policyholders file a claim in a year. A client announces his intention to file a claim. What is the probability that Jack sold him the policy [p(Jack/Claim)]

Question 3

a) List three (3) properties of probability

b) Distinguish between Mutual events and Independent events

c) Two Workers A and B assemble parts of a production plant. The probability that Worker A makes a mistake is 0.02 and the probability that Worker B makes a mistake is 0.03. However, Worker A assembles 55% of the parts and Worker B assembles

45% of the parts. An assembled part is selected randomly and is found to be defective. What is the probability that Worker B assembles it?

Question 4

a) Let A and B be events with P (A) = 3/8, P (B) = ½, and P (A $\cap$ B) = ¼

Find:

i. P(A$\cup$ B)

ii. P(A′) and P(B′)

iii. P(A′$\cap$ B′)

iv. P(A′$\cup$ B′)

v. P(A $\cup$ B′)

vi. P(B $\cap$ A′)

b) Find the number of permutations of letters in the word STATISTICS.

Unit 7

Probability Distribution and Random variables

7.1 Introduction

Probability distributions are related to frequency distributions. Probability distributions may be thought of as theoretical frequency distributions

A probability distribution is similar to the frequency distribution of a quantitative population because both provide a long-run frequency for outcomes. In other words, a probability distribution is listing of all the possible values that a random variable can take along with their probabilities.

Example 6.1:

Suppose we want to find out the probability distribution for the number of heads on three tosses of a coin:

First toss	T	T	T	T	H	H	H	H
Second toss	T	T	H	H	T	T	H	H
Third toss	T	H	T	H	T	H	T	H

the probability distribution of the above experiment is as follows (columns 1, and 2 in the following table)

(Column1)	(Column2)	(Column3)
Number of heads	Probability	(1)(2)
X	P(X)	(X)P(X)
0	1/8	0.0
1	3/8	0.375
2	3/8	0.75
3	1/8	0.375
Total		1.5 = E(X)

Note:

(i) In probability distribution we describe how outcomes in an experiment are expected to vary using probabilities instead of frequencies

(ii) Because these distributions deal with expectations, they are useful models in making inferences and decisions under conditions of uncertainty.

(iii) The main difference between frequency distributions and probability distributions is that a frequency distribution is a listing of observed frequencies of all outcomes of an experiment that actually occurred when the experiment was done, whereas a probability distribution is a listing of all the probabilities of all possible outcomes that could result when the experiment was done.

A variable is said to be random if it takes on different values as the result of the outcomes of a random experiment.

In a **discrete probability distribution** the random variable is allowed to take on countable values.

In a **continuous probability distribution,** the random variable is allowed to take on values within an interval (the values are infinite).

7.2 Construction of Probability Distributions

Example 6.2

An urn contains 4 red balls and 3 black balls. 2 balls are drawn in succession without replacement. Construct the probability distribution for the number of red balls in the sample of the 2 balls drawn.

Solution

Let R = red ball, B = black ball

4 R
3 B

Hence, P(R) = 4/7, P(B) = 3/7

The following are the possible number of red balls in the sample of the 2 balls drawn: **RR, RB, BR, BB**

Hence the sample points in the sample space S are :RR, RB, BR, BB

i.e. S = {RR, RB, BR, BB}

Let X be the random variable of the number of red balls in the sample of the 2 balls drawn.

Then X takes on values: x = 2, 1, 1, 0

i.e. for S = {RR, RB, BR, BB}
x = 2, 1, 1, 0

Possible outcomes from the experiment

1st ball **2nd ball**	**Number of red balls in Sample of 2 balls**	**Probability of the four possible outcomes**
R R R B B R B B	2 1 1 0	4/7 x 4/7 = 16/49 4/7 x 3/7 = 12/49 3/7 x 4/7 = 12/49 3/7 x 3/7 = 9/49
		Sum: 1.00

Hence probability distribution is :

X	0 1 2
P(X = x)	9/49 (12/49 + 12/49) = 24/49 16/49

Mathematical Expectations

Let X be a discrete random variable with the following probability distribution

X	x_1	x_2	x_3	---	x_n
f(X=x)	$f(x_1)$	$f(x_2)$	$f(x_3)$	---	$f(x_n)$

The **expected value of a random variable X or the mathematical expectation of X** is denoted by E(X) and is defined by:

$$E(X) = \sum_{i=1}^{n} x_i f(x_i) \quad or \quad E(X) = \sum_{i=1}^{n} x_i P(PX = x_i)$$

The **variance of the random variable X is denoted by Var.(X)** and is defined by

$$\text{Var.}(X) = \sum_{i=1}^{n} x^2{}_i \; f(x_i) - [E(X)]^2$$

The **standard deviation of the random variable X is denoted by s = √ Var.(X)** and is defined by s = $\sqrt{\sum_{i=1}^{n} x^{2}{}_{i}\ f(x_i) - [E(X)]^{2}}$

<u>Example</u> For example 6.1, the mathematical expectation is computed as shown below:

(Column 1)	(Column 2)	(Column 3)	(Column 4)	(Column 5)
Number of heads	Probability	(1)(2)	$(1)^2$	(2)(4)
X	P(X)	(X)P(X)	X^2	$X^2P(X)$
0	1/8	0.0	0	0.0
1	3/8	0.375	1	0.375
2	3/8	0.75	4	1..5
3	1/8	0.375	9	1.125
Total		1.5 = E(X)		3.000 = Var(X)

i.e., Expected value, E(X) = $\sum_{i=1}^{n} x_i f(x_i)\quad or\quad E(X) = \sum_{i=1}^{n} x_i P(PX = x_i)$ = 1.5

Variance, Var.(X) = $\sum_{i=1}^{n} x^{2}{}_{i}\ f(x_i) - [E(X)]^{2}$ = 3.0 – $(1.5)^2$ = 3.0 – 2.25 = 0.75

Standard deviation = √(0.75)

<u>Example</u>

Suppose a charity organization is mailing printed return-address stickers to over one million homes. Each recipient is asked to donate either \$1, \$2, \$5, \$10, \$15, or \$20. Based on past experience, the amount a person donates is believed to follow the following probability distribution:

X:	\$1	\$2	\$5	\$10	\$15	\$20
P(X)	0.1	0.2	0.3	0.2	0.15	0.05

The question is, what is expected that an average donor to contribute, and what is the standard devation.

Solution
The solution is as follows.

(1)	(2)	(3)	(4)	(5)	(6)
X	P(X)	X.P(X)	$(X - \mu)$	$[(X - \mu)]^2$	(5)x(2)
1	0.1	0.1	-6.25	39.06	3.906
2	0.2	0.4	-5.25	27.56	5.512
5	0.3	1.5	-2.25	5.06	1.518
10	0.2	2.0	2.75	7.56	1.512
15	0.15	2.25	7.75	60.06	9.009
20	0.05	1.0	12.75	162.56	8.125
Total		**7.25 = E(X)**			**29.585**

Thus, the expected value is $7.25, and standard deviation is the square root of $29.585, which is equal to $5.55. In other words, an average donor is expected to donate $7.25 with a standard deviation of $5.55.

Example 6.5:
A committee of 3 LC officials is to be formed from 4 men and 3 women of Rubaga Division to represent that division on Kampala City Council. If X is the random variable of the number of men on the committee, find the expected number of men on the committee, and the standard deviation of this number.(Correct all the answers to one whole number).

Solution
The probability distribution of X is given by:

$P(X = x) = {}^4C_x \cdot {}^3C_{3-x} / {}^7C_3$, x = 0, 1, 2, 3.(it can be proved)

The student (need to know about permutations/combinations and factorials)

Hence,

X	0	1	2	3
P(X=x)	$\frac{1}{35}$	$\frac{12}{35}$	$\frac{18}{35}$	$\frac{4}{35}$

Expectation, E(X) = $\sum_{i=1}^{4} XP(X = X_i)$

$$= 0\left(\frac{1}{35}\right) + 1\left(\frac{12}{35}\right) + 2\left(\frac{18}{35}\right) + 3\left(\frac{4}{35}\right) = 1.71$$

= 2 (to the nearest whole number)

Variance, Var.(X) = $\sum_{i=1}^{n} X_{i^2} P(X = X_i) - [E(X)]^2$

$$= 0\left(\frac{1}{35}\right) + 1^2\left(\frac{12}{35}\right) + 2^2\left(\frac{18}{35}\right) + 3^2\left(\frac{4}{35}\right) - (1.71)^2 = 0.51$$

= 1(to the nearest whole number)

standard deviation = $\sqrt{\text{var}(x)}$ = 0.71 = 1 (to the nearest whole number)

Example

3 coins are tossed. A man gets 5 shillings when all heads or all tails appear and he pays 3 shillings if either 1 or 2 head(s) show. What is his average gain or loss?

Sample space for a single coin, S = {h,t}. The sample space for the 2 coins is computed as shown in the table below:

Solution

	1st coins	
	h	**t**
h	hh	ht
2nd coin t	th	tt

Sample space for 2 coins, S = {hh, ht, th, tt}

The sample space for the 3 coins is computed as shown in the table below:

		First two coins			
		hh	**ht**	**th**	**tt**
3rdcoin	**h**	hhh	hht	hth	htt
	t	thh	tht	tth	ttt

Sample space for the 3 coins is

S = $\left\{\begin{array}{cccccccc} hhh, & hht, & hth, & htt, & thh, & tht, & tth, & ttt \\ w, & w, & w, & w, & w, & w, & w, & w \end{array}\right\}$, where w is the weight.

Hence, 8w = 1 , w = $\frac{1}{8}$

$$S = \left\{\begin{array}{cccccccc} hhh, & hht, & hth, & htt, & thh, & tht, & tth, & ttt \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{array}\right.$$

Let A = event : all heads appear, A = $\left\{\begin{array}{c} hhh \\ \frac{1}{8} \end{array}\right\}$, P(A) = $\frac{1}{8}$

Let B = event: all tails appear, B = $\left\{\begin{array}{c} ttt \\ \frac{1}{8} \end{array}\right\}$, P(B) = $\frac{1}{8}$.

Let C = event: one head appears,

$$C = \begin{cases} hht, tht, tth, \\ \frac{1}{8} \quad \frac{1}{8} \quad \frac{1}{8} \end{cases}$$

$$P(C) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

Let D = event: two heads appear,
D =
$hht, tht, tth,$

$\frac{1}{8} \quad \frac{1}{8} \quad \frac{1}{8}$

P(D) = $\frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$
Hence, P(either all heads or all tails appear) = P(A or B) = P(A ∪B)
Since A and B are mutually exclusive events, P(A or B) = P(A U B) = P(A) +P(B)
$= \frac{1}{8} + \frac{1}{8} = \frac{2}{8} = \frac{1}{4}$
P(either 1 or 2 heads show) = P(C or D) = P(C □ D) = P(C) + P(D), since C and D are mutually exclusive events.

P(CorD = P(C ∪ D) = P(C) + P(D)
$= \frac{3}{8} + \frac{3}{8} = \frac{6}{8} = \frac{3}{4}$

Let X be the random variable of the amount he gains

Hence, X takes on values x = 5 or x = -3

Hence the probability distribution of X is:
P(X = 5) = $\frac{1}{4}$, P(X=-3) = $\frac{3}{4}$

E(X) = $\sum XP(X = x)$
$= -3\left(\frac{3}{4}\right) + 5\left(\frac{1}{4}\right) = -1$

Hence, on the average his total loss is +1

Example

A box contains 10 coloured discs of which two are red. A man pays 10 shillings to play a game in which discs are drawn one at a time without replacement.. He will receive 25 shillings if the first disc drawn is red, 20 shillings if the second disk drawn is red, and 5 shillings if the third disc drawn is red but nothing otherwise.
Find the man's expected profit or loss in any game.

Solution:

If the random variable X is the amount in shillings received in any game, then X takes on values 25,20,5,and 0.

P(X = 25) = Prob(1st disc is red) = $\frac{2}{10} = \frac{1}{5}$

P(X = 20) = Prob(2nd disc is red)= $\frac{8}{10} x \frac{2}{9} = \frac{8}{45}$

P(X = 5) = Prob(3rd disc is red) = $\frac{8}{10} x \frac{7}{9} x \frac{2}{8} = \frac{7}{45}$

P(X = 0) = Prob (1st 3 discs are not red) = 1-$\left(\frac{1}{5} + \frac{8}{45} + \frac{7}{45}\right)$= 1- $\left(\frac{24}{25}\right)$ = $\frac{21}{25}$

Hence probability distribution of X is:

X	0	5	20	25
P(X = x)	$\frac{21}{45}$	$\frac{7}{45}$	$\frac{8}{45}$	$\frac{1}{5}$

E(X) = $\sum XP(X = x)$

= 0.P(X = 0) + 5.P(X = 5) + 20.P(X= 20) + 25.P(X = 25)

$$= 0 + 5\left(\frac{7}{45}\right) + 20\left(\frac{8}{45}\right) + \left(\frac{1}{5}\right) \qquad = \frac{39}{9} + \frac{160}{45} + \frac{25}{5}$$

$$= \frac{39}{9} + \frac{25}{5} = \frac{39}{9} + 5 = \frac{84}{9} = 9\frac{3}{9} = 9\frac{1}{3}$$

Since the man pays 10 shillings for the game, his expected loss is = 10 - $9\frac{1}{3}$

= 10 - $\frac{28}{3} = \frac{30-28}{3} = \frac{2}{3}$ shillings

7.3 Binomial Distribution

One of the most widely known of all discrete probability distributions is the binomial distribution. Several characteristics underlie the use of the binomial distribution.

Characteristics of the Binomial Distribution:

- The experiment consists of n identical trials.
- Each trial has only one of the two possible mutually exclusive outcomes, success or a failure.
- The probability of each outcome does not change from trial to trial, and
- The trials are independent, thus we must sample with replacement.

Note

If the sample size, n, is less than 5% of the population, the independence assumption is not of great concern. Therefore the acceptable sample size for using the binomial distribution with samples taken without replacement is [n<5% N] where n is equal to the sample size, and N stands for the size of the population. The birth of children (male or female), true-false or multiple-choice questions (correct or incorrect answers) are some examples of the binomial distribution.

BinomialEquation

When using the binomial formula to solve problems, all that is necessary is that we be able to identify three things:

- the number of trials (n)
- the probability of a success on any one trial (p), and
- The number of successes desired (x).

The formulas used to compute the probability, the mean, and the standard deviation of a binomial distribution are as follows.

If X is binomially distributed with n trial and r successes with probability of succeeding p then:

$$\mathbf{P(X = x) =} \frac{n!}{(n-x)!x!}p^{x}q^{n-x}\mathbf{, \ for\ x = 0, 1, 2, \ldots, n}$$

where: n = the sample size or the number of trials, x = the number of successes desired,
p = probability of getting a success in one trial, and q = (1 - p) = the probability of getting a failure in one trial.

In general n! = n(n - 1)(n - 2)(n - 3)…3x2x1

Note:
E(X)= μ = np, and Var(X) = σ^2 = npq are the expected value(mean) and variance respectively of X.

Example

Let's go back to Unit 5 and solve the probability problem of defective TVs by applying the binomial equation once again. We said, suppose that 4% of all TVs made by W&B Company in 1995 are defective. If eight of these TVs are randomly selected from across the country and tested, what is the probability that exactly three of them are defective? Assume that each TV is made independently of the others.

In this problem, n=8, x=3, p=0.04, and q=(1-p)=0.96. Plugging these numbers into the binomial formula (see the above equation) we get:

$$\mathbf{P(X = 3) =} \frac{8!}{(8-3)!3!}(0.04)^{3}(0.96)^{5}\mathbf{, \ for\ x = 0, 1, 2, \ldots, 8}$$

P(X=3) = P(3) = 0.0003 or 0.03% , which is the same answer as in Unit 5.

The **mean** is equal to (n) x (p) = (8)(0.04)=0.32, the **variance** is equal to np (1 - p) = (0.32)(0.96) = 0.31, and the **standard deviation** is the square root of 0.31, which is equal to0.6.

The Binomial Table

Mathematicians constructed a set of binomial tables containing presolved probabilities. Binomial distributions are a family of distributions. In other words, every different value of n and/or every different value of p gives a different binomial distribution. Tables are available for different combinations of n and p values. For the tables, refer to the text. Each table is headed by a value of n, and values of p are presented in the top row of each table of size n. In the column below each value of p is the binomial distribution for that value of n and p. The binomial tables are easy to use. Simply look up n and p, then find X (located in the first column of each table), and read the corresponding probability. The following table is the binomial probabilities for n = 6. Note that the probabilities in each column of the binomial table must add up to 1.0.

Binomial Probability Distribution Table
(n = 6)

```
-----------------------------------------------------------------------------------
Probability
X.....0.1........0.2.....0.3.....0.4.....0.5.....0.6.....0.7.....0.8.....0.9
-----------------------------------------------------------------------------------
0.....0.531............0.118...................................................0.000
1.....0.354............0.303...................................................0.000
2.....0.098............0.324...................................................0.001
3.....0.015............0.185...................................................0.015
4.....0.001............0.060...................................................0.098
5.....0.000............0.010...................................................0.354
6.....0.000............0.001...................................................0.531
-----------------------------------------------------------------------------------
```

<u>Example</u>

Suppose that an examination consists of six true and false questions, and assume that a student has no knowledge of the subject matter. The probability that the student will guess the correct answer to the first question is 30%. Likewise, the probability of guessing each of the remaining questions correctly is also 30%. What is the probability of getting more than three correct answers?

Solution

For the above problem, n = 6, p = 0.30, and X >3. In the above table, search along the row of p values for 0.30. The problem is to locate the P(X > 3). Thus, the answer involves summing the probabilities for X = 4, 5, and 6. These values appear in the X column at the intersection of each X value and p = 0.30, as follows:
P (X > 3) = Summation of {P (X=4) + P(X=5) +P(X=6)} = (0.060)+(0.010)+(0.001) = 0.071 or 7.1%

Thus, we may conclude that if 30% of the exam questions are answered by guessing, the probability is 0.071 (or 7.1%) that more than four of the questions are answered correctly by the student.

Graphing the Binomial Distribution

The graph of a binomial distribution can be constructed by using all the possible X values of a distribution and their associated probabilities. The X values are graphed along the X axis, and the probabilities are graphed along the Y axis. Note that the graph of the binomial distribution has three shapes: If p<0.5, **the graph is positively skewed**, if p>0.5, **the graph is negatively skewed**, **and if p=0.5, the graph is symmetrical**. The skewness is eliminated as n gets large. In other words, if n remains constant but p becomes larger and larger up to 0.50, the shape of the binomial probability distribution becomes more symmetrical. If p remains the same but n becomes larger and larger, the shape of the binomial probability distribution becomes more symmetrical.

The Poisson Distribution

The poisson distribution is another discrete probability distribution. It is named after Simeon-Denis Poisson (1781-1840), a French mathematician. The poisson distribution depends only on the average number of occurrences per unit time of space. There is no n, and no p. The poisson probability distribution provides a close approximation to the binomial probability distribution when n is large and p is quite small or quite large. In other words, if n>20 and np<=5 [or n(1-p)<="5]," then we may use poisson distribution as an approximation to binomial distribution. for detail discussion of the poisson probability distribution, refer to any statistics textbook.

The Hyper-geometric Distribution

Another discrete probability distribution is the hypergeometric distribution. The binomial probability distribution assumes that the population from which the sample is selected is very large. For this reason, the probability of success does not change with each trial. The hypergeometric distribution is used to determine the probability of a specified number of successes and/or failures when:

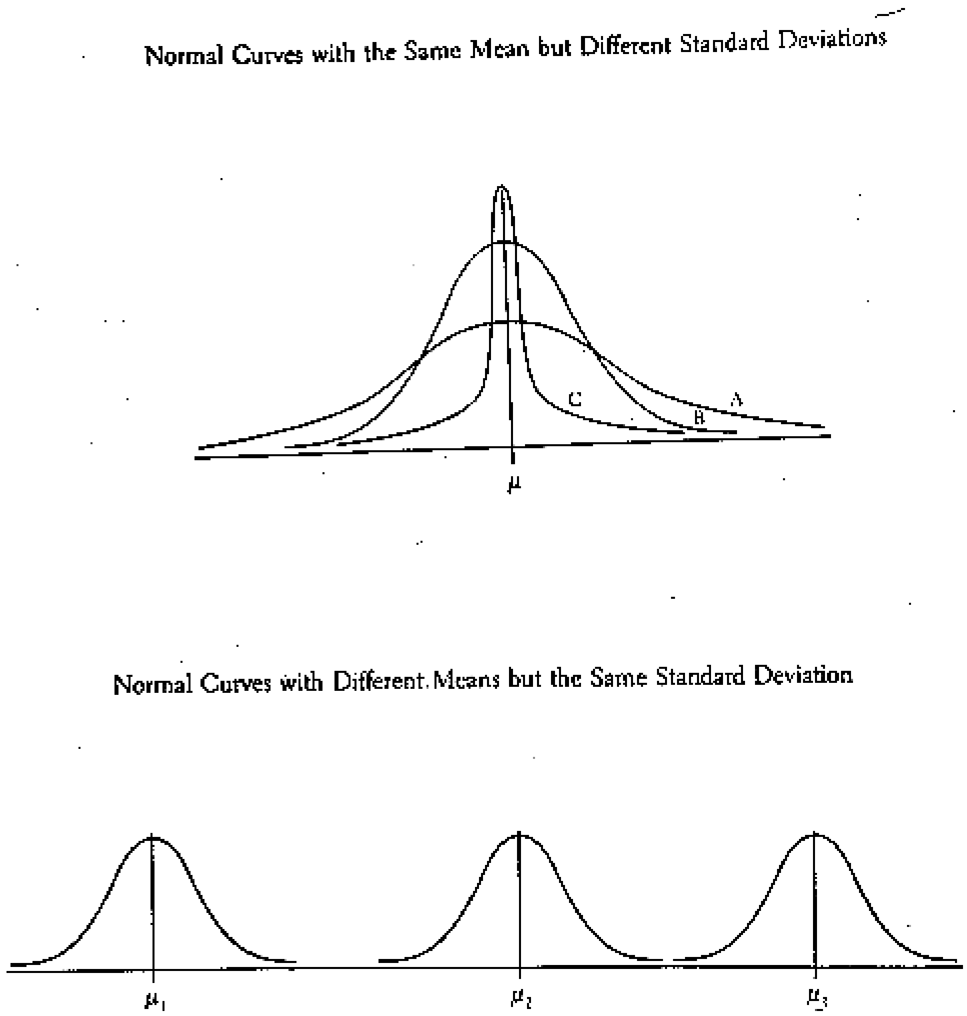
(1) a sample is selected from a finite population without replacement and/or
(2) when the sample size, n, is greater than or equal to 5% of the population size,
N, i.e., [n>=5% N].

7.4 Normal Distribution or Normal Curve

Normal distribution is probably one of the most important and widely used continuous distributions. It is known as a normal random variable, and its probability distribution is called a normal distribution. The following are the characteristics of the normal distribution:

Characteristics of the Normal Distribution:

1. It is bell shaped and is symmetrical about its mean.
2. It is asymptotic to the axis, i.e., it extends indefinitely in either direction from the mean.
3. It is a continuous distribution.
4. It is a family of curves, i.e., every unique pair of mean and standard deviation defines a different normal distribution. Thus, the normal distribution is completely described by two parameters: mean and standard deviation. See the following figure.
5. Total area under the curve sums to 1, i.e., the area of the distribution on each side of the mean is 0.5.
6. It is unimodal, i.e., values mound up only in the center of the curve.
7. The probability that a random variable will have a value between any two points is equal to the area under the curve between those


Normal Curves with the Same Mean but Different Standard Deviations

Normal Curves with Different Means but the Same Standard Deviation

points.

Figure 6.3

Note that the **integral calculus** is used to find the area under the normal distribution curve. However, this can be avoided by transforming all normal distribution to fit the standard normal distribution. This conversion is done by **rescaling** the normal distribution axis from its true units (time, weight, dollars, and...) to a standard measure **called Z score or Z value.** A Z score is the number of standard deviations that a value, X, is away from the mean. If the value of X is greater than the mean, the Z score is positive; if the value of X is less than the mean, the Z score is negative. The Z score or equation is as follows:

Z = (X - Mean) /Standard deviation

That is, $Z = \frac{X - \mu}{\sigma}$

A standard Z table can be used to find probabilities for any normal curve problem that has been converted to Z scores. For the table, refer to any statistics textbook.

The Z distribution is a normal distribution with a mean of 0 and a standard deviation of 1.

The following steps are helpful when working with the normal curve problems:

1. Graph the normal distribution, and shade the area related to the probability you want to find.
2. Convert the boundaries of the shaded area from X values to the standard normal random variable Z values using the Z formula above.
3. Use the standard Z table to find the probabilities or the areas related to the Z values in step 2.

Example 6.1:

Graduate Management Aptitude Test (GMAT) scores are widely used by graduate schools of business as an entrance requirement. Suppose that in one particular year, the mean score for the GMAT was 476, with a standard deviation of 107. Assuming that the GMAT scores are normally distributed, answer the following questions:

a) What is the probability that a randomly selected score from this GMAT falls between 476 and 650? <= x <="650)"
b) What is the probability of receiving a score greater than 750 on a GMAT test that has a mean of 476 and a standard deviation of 107? i.e., P(X >= 750) =?
c) What is the probability of receiving a score of 540 or less on a GMAT test that has a mean of 476 and a standard deviation of 107? i.e., P(X <= 540)="?."
d) What is the probability of receiving a score between 440 and 330 on a GMAT test that has a mean of 476 and a standard deviation of 107?

Solutions

a). What is the probability that a randomly selected score from this GMAT falls between 476 and 650? <= x <="650)"

The following figure shows a graphic representation of this problem.

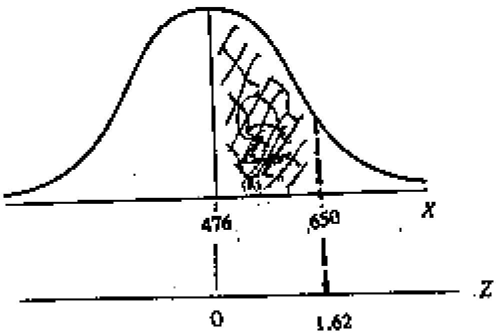


Figure 6.4

Applying the Z equation, we get:
Z = (650 - 476)/107 = 1.62.

The **Z value of 1.62** indicates that the GMAT score of 650 is 1.62 standard deviation above the mean. The standard normal table gives the probability of value falling between 650 and the mean.

Note:

> The whole number and tenths place portion of the Z score appear in the first column of the table. Across the top of the table are the values of the hundredths place portion of the Z score. Thus the answer is that 0.4474 or 44.74% of the scores on the GMAT fall between a score of 650 and 476.

b) What is the probability of receiving a score greater than 750 on a GMAT test that has a mean of 476 and a standard deviation of 107? i.e., P(X >= 750) = ?.

Solution
This problem is asking for determining the area of the upper tail of the distribution.

The Z score is:
Z = (750 - 476)/107 = 2.56.

From the table, the probability for this Z score is 0.4948. This is the probability of a GMAT with a score between 476 and 750.

Note:

> The rule is that when we want to find the probability in either tail, we must substract the table value from 0.50. Thus, the answer to this problem is: 0.5 - 0.4948 = 0.0052 or 0.52%. Note that P(X >= 750) is

the same as P(X >750), because, in continuous distribution, the area under an exact number such as X=750 is zero.

The following figure shows a graphic representation of this problem.

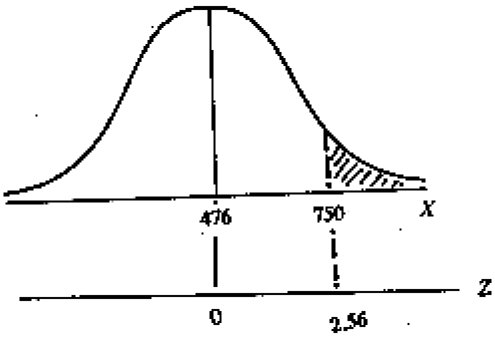


Figure 6.5

c) What is the probability of receiving a score of 540 or less on a GMAT test that has a mean of 476 and a standard deviation of 107? i.e., P(X <= 540)="?."

Solution:

We are asked to determine the area under the curve for all values less than or equal to 540. the z score is:

z="(540" 476)/107="0.6."

From the table, the probability for this z score is 0.2257 which is the probability of getting a score between the mean (476) and 540.

Note:

The rule is that when we want to find the probability between two values of x on either side of the mean, we just add the two areas together. Thus, the answer to this problem is: 0.5 + 0.2257 = 0.73 or 73%.

The following figure shows a graphic representation of this problem.

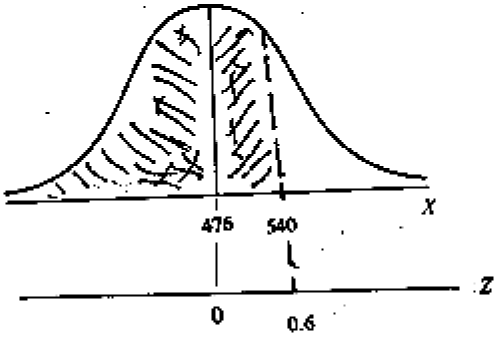


Figure 6.6

d) What is the probability of receiving a score between 440 and 330 on a GMAT test that has a mean of 476 and a standard deviation of 107? i.e., P(330 < X <440)

<u>**Solution:**</u>

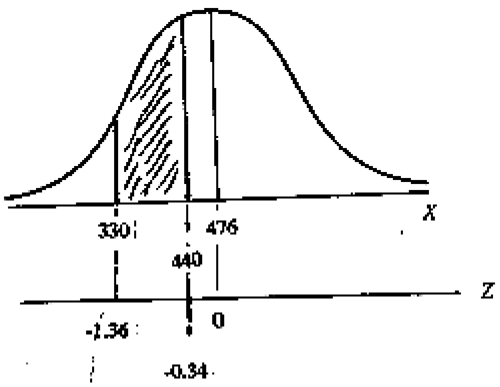


Figure 6.7

In this problem, the two values fall on the same side of the mean. The Z scores are:

Z1 = (330 - 476)/107 = -1.36,

and

Z2 = (440 - 476)/107 = -0.34.

The probability associated with Z = -1.36 is 0.4131, and the probability associated with
Z = -0.34 is 0.1331.

Note:

The rule is that when we want to find the probability between two values of X on one side of the mean, we just subtract the smaller area from the larger area to get the probability between the two values. Thus, the answer to this problem is:
0.4131 - 0.1331 = 0.28 or 28%.

<u>**Example 6.12:**</u>

Suppose that a tire factory wants to set a mileage guarantee on its new model called LA 50 tire. Life tests indicated that the mean mileage is 47,900, and standard deviation of the normally distributed distribution of mileage is 2,050 miles. The factory wants to set the guaranteed mileage so that no more than 5% of the tires will have to be replaced.

What guaranteed mileage should the factory announce? i.e., P(X <= ?)="5%.

Solution

In this problem, the mean and standard deviation are given, but X and Z are unknown. The problem is to solve for an X value that has 5% or 0.05 of the X values less than that value. If 0.05 of the values are less than X, then 0.45 lie between X and the mean (0.5 - 0.05), see the following graph.

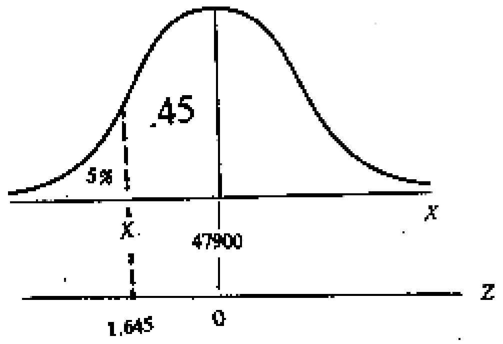


Figure 6.8

Note:
Refer to the standard normal distribution table and search the body of the table for 0.45. Since the exact number is not found in the table, search for the closest number to 0.45. There are two values equidistant from 0.45-- 0.4505 and 0.4495. Move to the left from these values, and read the Z scores in the margin, which are: 1.65 and 1.64. Take the average of these two Z scores, i.e., (1.65 + 1.64)/2 = 1.645. Plug this number and the values of the mean and the standard deviation into the Z equation, you get:

> **Z =(X - mean)/standard deviation or -1.645 =(X - 47,900)/2,050 = 44,528 miles.**
>
> Thus, the factory should set the guaranteed mileage at 44,528 miles if the objective is not to replace more than 5% of the tires.

The Normal Approximation to the Binomial Distribution
Earlier, we talked about the binomial probability distribution, which is a discrete distribution. You remember that we said as sample sizes get larger, binomial distribution approach the normal distribution in shape regardless of the value of p (probability of success). For large sample values, the binomial distribution is cumbersome to analyze without a computer. Fortunately, the normal distribution is a good approximation for binomial distribution problems for large values of n. The commonly accepted guidelines for using the normal approximation to the binomial probability distribution is when (n x p) and [n(1 - p)] are both greater than 5.

Example 6.13:

Suppose that the management of a restaurant claimed that 70% of their customers returned for another meal. In a week in which 80 new (first-time) customers dined at the restaurant, what is the probability that 60 or more of the customers will return for another meal?, ie., P(X >= 60) =?.

Solution

The solution to this problem can be illustrated as follows:

- First, the two guidelines that (n x p) and [n(1 - p)] should be greater than 5 are satisfied: (n x p) = (80 x 0.70) = 56 > 5, and [n(1 - p)] = 80(1 - 0.70) = 24 > 5.
- Second, we need to find the mean and the standard deviation of the binomial distribution. The mean is equal to (n x p) = (80 x 0.70) = 56 and standard deviation is square root of [(n x p)(1 - p)], i.e., square root of 16.8, which is equal to 4.0988.

Using the Z equation we get,

Z = (X - mean)/standard deviation = (59.5 - 56)/4.0988 = 0.85.

From the table, the probability for this Z score is 0.3023 which is the probability between the mean (56) and 60. We must substract this table value 0.3023 from 0.5 in order to get the answer, i.e., P(X >= 60) = 0.5 - 0.3023 = 0.1977. Therefore, the probability is 19.77% that 60 or more of the 80 first-time customers will return to the restaurant for another meal. See the following graph.

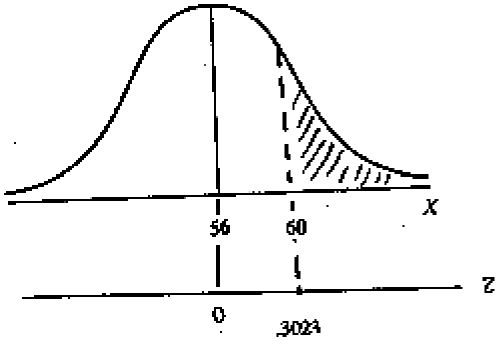


Figure 6.9

Correction Factor

The value 0.5 is added or subtracted, depending on the problem, to the value of X when a binomial probability distribution is being approximated by a normal distribution. This correction ensures that most of the binomial problem's information is correctly transferred to the normal curve analysis. This correction is called the correction for continuity. The decision as to how to correct for continuity depends on the equality sign and the direction of the desired outcomes of the binomial distribution. The following table shows some rules of thumb that can help in the application of the correction for continuity, see the above example.

Value Being Determined..............................Correction
X >...+0.50
X > =...-0.50
X <...-0.50
X <=...+0.50
<= X <="..................................-0.50" & +0.50

<.......................................+0.50>X =...-
0.50 & +0.50

Review Questions

Unit 8
Inferential Statistics, Sampling and Sampling distribution

Definition

The distribution of all possible sample means and their related probability is called the **sampling distribution of the means.**

Properties of the Sampling Distribution of Means

If a population is normally distributed, then:

1. The mean of the sampling distribution of means equals the population mean.
2. The standard deviation of the sampling distribution of means (or standard error of the mean) is smaller than the population standard deviation, see the following equations

$$\mu_{\bar{X}} = \mu$$

…………………………………………………….. (1)

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

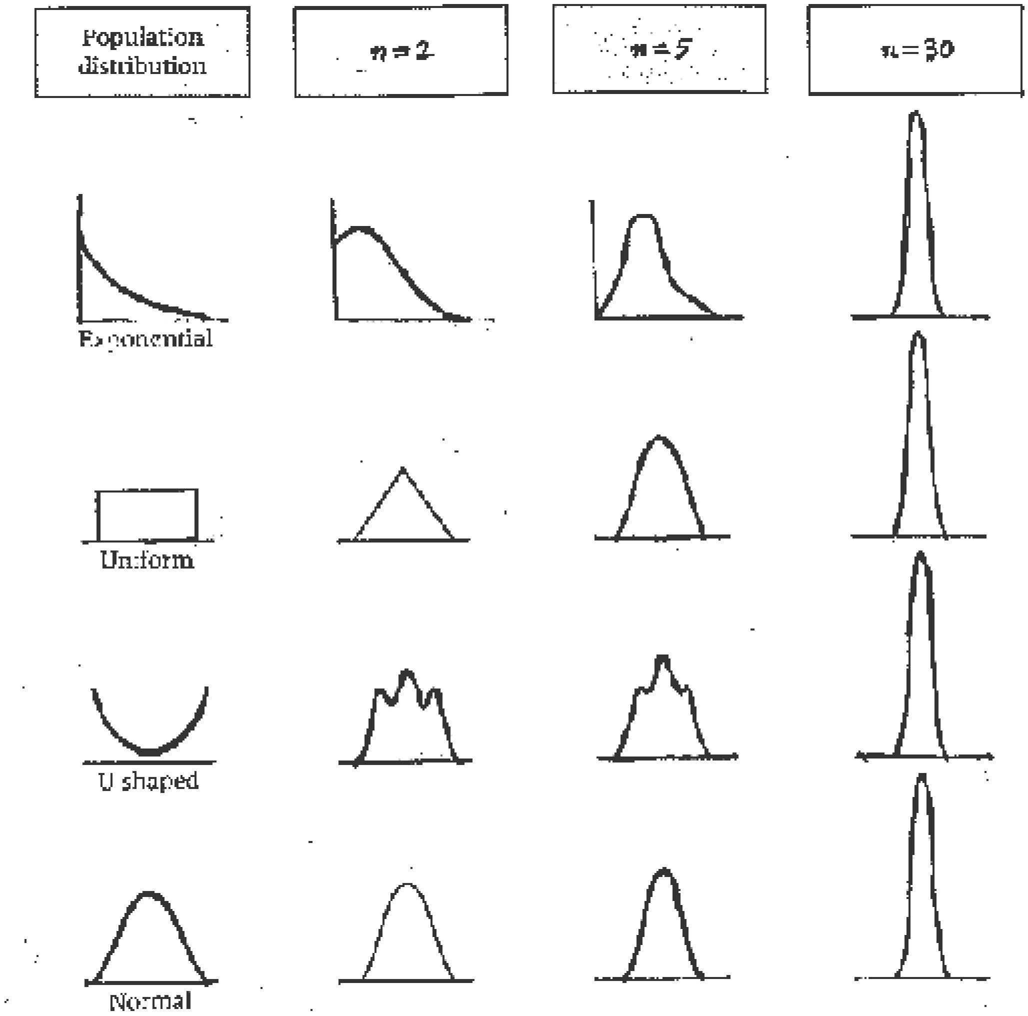
For example, from the above table, the mean of the means is equal to 8% which is same as the population mean, and standard error of the mean is equal to 3.26% which is less than the population standard deviation of 8.15%.

Central Limit Theorem

If a random sample of n observation is selected from **any population,** then, when the sample size is sufficiently large **(n>=30)** the sampling distribution of the mean tends to approximate the normal distribution. The larger the sample size, n, the better will be the normal approximation to the sampling distribution of the mean. Then, again in this case it can be shown that the mean of the sample means is same as population mean, and the standard error of the mean is smaller than the population standard deviation, see the equation above.

The real advantage of the central limit theorem is that sample data drawn from populations not normally distributed or from populations of unknown shape also can be analysised by using the normal distribution, because the sample means are normally distributed for sample sizes of n>=30.

Column 1 of the following figure shows four different population distributions. Each ensuing column displays the shape of the distribution of the sample means for a particular sample size. Note that the distribution of the sample means begins to approximate the normal curve as the sample size, n, gets larger.



Figure

Since the central limit theorem states that sample means are normally distributed regardless of the shape of the population for large samples and for any sample size with normally distributed population, thus sample means can be analysised by using Z scores. Recall:

$$Z = \frac{X - \mu}{\sigma} \quad \text{..(2)}$$

If sample means are normally distributed, the Z score equation applied to sample means would be:

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}}$$

OR ....(3)

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Example 6.15:

You are the director of transporation safety.. You are concerned because the average highway speed of all trucks may exceed the 60 mph speed limit. A random sample of 120 trucks show a mean speed of 62 mph. Assuming that the population mean is 60 mph and population standard deviation is 12.5 mph, find the probability of the average of the
speed greater than or equal to 62 mph.

In this problem, n= 120, the mean of the means = population mean = 60 mph, and standard error of the mean = population standard deviation /square root of sample size = 12.5/10.95 = 1.14.

Plugging these numbers into the Z score equation (equation 3) we get,

Z = (62 - 60)/1.14 = 1.75.

From the standard normal distribution table, this Z value yields a probability of 0.4599. This is the probability of getting a mean between 62 mph and the population mean 60 mph. Therefore, the probability of getting a sample average speed grater than 62 mph is (0.5 - 0.4599) = 0.04. That is, 4% of the time, a random sample of 120 trucks from the population will yield a mean speed of 62 mph or more. The following figure shows the problem.

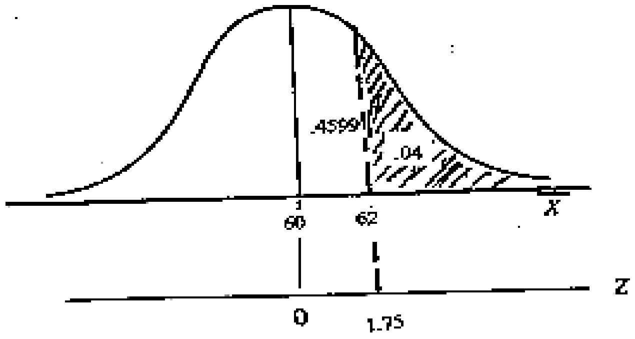


Figure 6.11

Sampling From a Finite Population

A **finite population** is a population which has a fixed upper bound. For example, there are 5,124 students enrolled in MBA. In cases of a finite population, an adjustment is made to the Z equation for sample means (equation 3 above). The adjustment is called **correction factor, or finite population multiplier.**

$$\sqrt{\frac{N-n}{N-1}}$$

Correction Factor

Note:

A rule of thumb is that if sampling is done without replacement from a finite population and the sample size n is greater than 5% of the population size N, i.e., n/N>0.05, then the correction factor should be used to adjust the standard deviation (or standard error) of the mean. Thus, the following Z equation is used when samples are drawn from finite population.

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \qquad \text{...(4)}$$

Example 6.16:

A production company's 250 hourly employees average 39.5 years of age, with a standard deviation of 9.3 years. If a random sample of 35 hourly employees is taken, what is the probability that the sample will have an average age less than 43 years?

In this problem, the population mean is 39.5, with a population standard deviation of 9.3. The sample size is 35 which is drawn from a finite population of 250. The sample mean is 43. The following graph shows the problem on a normal curve.

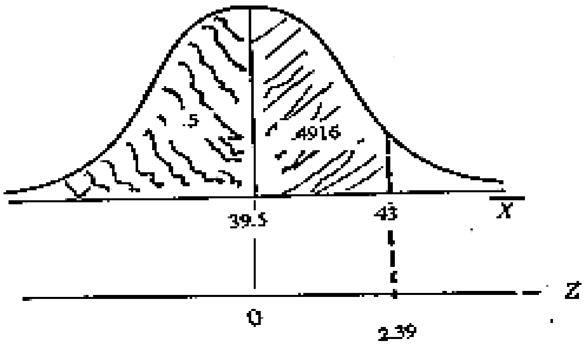


Figure 6.11

Using the Z equation with the correction factor (equation 4) gives a Z score of 2.39. From the standard normal distribution table, this Z value yields a probability of 0.4916. Therefore, the probability of getting a sample average age less than 43 years is (0.5 + 0.4916) = 0.9916 or 99.16%. Had the correction factor not been used, the Z value would have been 2.23, and the probability of getting a sample average age less than 43 years would have been 98.71%.

8.2 Sampling Distribution of Sample Proportion

Sample proportion is computed by dividing the number of items in a sample that possess the characteristic, X, by the number of items in the sample, n.

$$\hat{p} = \frac{X}{n} \qquad \text{...(5)}$$

The central limit theorem also applies to sample proportions in that the normal distribution approximates the shape of the distribution of sample proportion if (n x p) > 5 and [n (1 - p)] > 5, where p is the population proportion.

The mean of sample proportion for all samples of size n randomly drawn from a population is p (the population proportion) and the standard deviation of the sampling distribution of sample proportions (or the standard error of the proportion) is the square root of (p . q)/n, where q = 1 - p. The Z equation for the sample proportion is as follows:

$$Z = \frac{\hat{p} - P}{\sqrt{\frac{P \cdot Q}{n}}} \quad \text{..(6)}$$

Note that equation 6 is used when we are counting discrete items, such people or defectives, and we are interested in percentages or proportions.

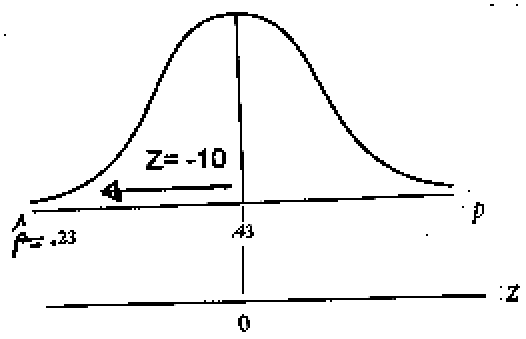
Example 6.17:

Suppose that fourty-three percent of all the country households had a telephone-answering machine in 1994. Marie believes that this proportion may not be true for her state. If she takes a random sample of 600 households and finds that only 135 have an answering machine, what is the probability of getting a sample proportion this small or smaller if the population proportion really is 0.43?

For this problem, p = 0.43, n = 600, X = 135, and sample proportion = X/n = 135/600 = 0.23. Using equation 6, and solving for Z gives

Z **= (0.23 - 0.43)/square root of [(0.43) . (0.57)]/600 = - 10**

Almost all the area under the curve lies to the right of this Z value. The probability of getting this sample proportion or a smaller one is virtually zero. That is, the results obtained from this sample are almost too different from the 43% proportion for Marie to accept the national figure for her state. The following graph shows this problem.



8.3 Hypothesis Test

Setting up and testing hypotheses is an essential part of statistical inference. In order to formulate such a test, usually some theory has been put forward, either because it is **believed to be true** or because it is to be used as a **basis for argument,** but has not been proved, for example, claiming that a new drug is better than the current drug for treatment of the same symptoms.

In each problem considered, the question of interest is simplified into two competing claims / hypotheses between which we have a choice:

the **null hypothesis,** denoted **H0,** against the **alternative hypothesis,** denoted **H1.**

These two competing claims / hypotheses are not, however, treated on an equal basis, special consideration is given to the null hypothesis. We have two common situations:

1) The experiment has been carried out in an attempt to **disprove or reject** a particular hypothesis, the null hypothesis, thus we give that one priority so it cannot be rejected unless the evidence against it is **sufficiently strong**.

For example**, H0**: there is no difference in taste between coke and diet coke against **H1**: there is a difference.

2) If one of the two hypotheses is **'simpler'** we give it priority so that a more **'complicated'** theory is not adopted unless there is sufficient evidence against the simpler one.

For example, it is 'simpler' to claim that there is no difference in flavour between coke and diet coke than it is to say that there is a difference.

The hypotheses are often statements about **population parameters** like expected **value and variance**, for example H0 might be that the expected value of the height of ten year old boys in the Ugandan population is not different from that of ten year old girls? A hypothesis might also be a statement about the distributional form of a characteristic of interest, for example that the height of ten year old boys is normally distributed within the Ugandan population?

The outcome of a hypothesis test is **'reject H0' or 'do not reject H0'.**

Null Hypothesis

The null hypothesis, H0 represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved.

For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug.
We would write:

H0: there is no difference between the two drugs on average.

Note:

(i) We give special consideration to the null hypothesis. This is due to the fact that the null hypothesis relates to the statement being tested, whereas the alternative hypothesis relates to the statement to be accepted if / when the null is rejected.

(ii) The final conclusion once the test has been carried out is always given in terms of the null hypothesis. We either 'reject H0 in favour of H1' or 'do not reject H0'**; we never conclude 'reject H1', or even 'accept H1'.**

(iii) If we conclude 'do not reject H0', this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against H0 in favour of H1; **rejecting the null hypothesis then, suggests that the alternative hypothesis may be true.**

Alternative Hypothesis

The alternative hypothesis, H1, is a statement of what a statistical hypothesis test is set up to establish.

For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect, on average, compared to that of the current drug.

We would write:

H1: the two drugs have different effects, on average.

The alternative hypothesis might also be that the new drug is better, on average, than the current drug.

In this case we would write:

H1: the new drug is better than the current drug, on average.

Note:

(i) The final conclusion once the test has been carried out is always given in terms of the null hypothesis. We either 'reject H0 in favour of H1' or 'do not reject H0'; we never conclude 'reject H1', or even 'accept H1'.

(ii) If we conclude 'do not reject H0', this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against H0 in favour of H1; rejecting the null hypothesis then, suggests that the alternative hypothesis may be true.

Simple Hypothesis

A simple hypothesis is a hypothesis which specifies the population distribution completely.

Examples:

1. H0: X~Bi(100,1/2) i.e. p is specified
2. H0: X~N(5,20) i.e. μ and σ are specified

Composite Hypothesis

A composite hypothesis is a hypothesis which does not specify the population distribution completely.

Examples

1. X~Bi(100,p) H1: p > 0.5
2. X~N(0, σ^2) H1: σ unspecified

8.4 Hypothesis Testing Errors

Type I Error

In a hypothesis test, a **type I error** occurs when **the null hypothesis is rejected when it is in fact true**; that is, H0 is wrongly rejected.

For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is

H0: there is no difference between the two drugs on average.

A type I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them.

The following table gives a summary of possible results of any hypothesis test:

		Decision	
		Reject H0	Don't Reject H0
Truth	H0	Type I Error	Right Decision
	H1	Right Decision	Type II Error

Note:

a) A type I error is often considered to be more serious, and therefore more important to avoid, than a type II error. The hypothesis test procedure is therefore adjusted so that there is a **guaranteed 'low' probability of rejecting the null hypothesis wrongly; this probability is never 0.** This probability of a type I error can be precisely computed as:

i. P(type I error) = significance level = α

b) The exact probability of a type II error is generally unknown.

(i) If we do not reject the null hypothesis, it may still be false (a type II error) as the sample may not be big enough to identify the falseness of the null hypothesis (especially if the truth is very close to hypothesis).
(ii) For any given set of data, type I and type II errors are inversely related; the smaller the risk of one, the higher the risk of the other.
(iii) A type I error can also be referred to as an error of the first kind.

Type II Error

In a hypothesis test, a type II error occurs when **the null hypothesis H0, is not rejected when it is in fact false**.

For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is

H0: there is no difference between the two drugs on average.

A type II error would occur if it was concluded that the two drugs produced the same effect, that is, there is no difference between the two drugs on average, when in fact they produced different ones.

Note:

(i) A type II error is frequently due to sample sizes being too small.
(ii) The probability of a type II error is symbolised by β and written:

P(type II error) = β (but is generally unknown).

(iii) A type II error can also be referred to as an error of the second kind.

8.5 Test Statistic

A **test statistic** is a quantity calculated from our sample of data. Its value is used to decide whether or not the null hypothesis should be rejected in the hypothesis test.

The choice of a test statistic will depend on the assumed **probability model** and the **hypotheses under question**.

Critical Value(s)

The **critical value(s)** for a hypothesis test is a threshold to which the value of the test statistic in a sample is compared to determine whether or not the null hypothesis is rejected.

The critical value for any hypothesis test depends on the significance level at which the test is carried out, and whether the test is one-sided or two-sided.

Critical Region

The **critical region CR**, or **rejection region RR**, is a set of values of the test statistic for which the null hypothesis is rejected in a hypothesis test; that is, the sample space for the test statistic is partitioned into two regions; one region (the critical region) will lead us to reject the null hypothesis H0', the other not. So, if the observed value of the test statistic is a member of the critical region, we conclude 'reject H0'; if it is not a member of the critical region then we conclude 'do not reject H0.

Significance Level

The **significance level** of a statistical hypothesis test is a fixed probability of **wrongly rejecting the null hypothesis H0, if it is in fact true.**
It is the probability of a **type I Error** and is set by the investigator in relation to the consequences of such an error. That is, we want to make the significance level as small as possible in order to protect the null hypothesis and to prevent, as far as possible, the investigator from inadvertently making false claims.

The significance level is usually denoted by α

Significance Level = P(type I error) = α

Note:

Usually, the significance level is chosen to be = 0.05 = 5%.

8.5 Review Questions

Question 1

Distinguish between the following terms

(i) Null Hypothesis and Alternative Hypothesis
(ii) Type I error and Type II error
(iii) One-Tailed test and Two-Tailed test

b) Explain the following terms

(i) Upper tailed test
(ii) Lower tailed test

c) The Headmistress of a certain school claims that the mean height of her candidates is 2 meters. To test this claim, a random sample of 10 candidates was selected and the following summary statistics is obtained: $\Sigma x = 26$, $\Sigma(x - x)^2 = 210.6$; Test the Headmistress' claim at 2% level of significance.

Review Questions

Question 2

a) Briefly explain the following

i. Type I error
ii. Type II error

b) Distinguish between One-tailed test and Two-tailed test.

c) The manager of a certain Bar in Kabalagala claims that 75% of his female customers take V & A. To test this claim, a sample of 90 female customers showed that 65 female customers take V & A. Test this claim at 5% level of significance.

Question 3

(a) Distinguish between the following terms

(i) Null Hypothesis and Alternative Hypothesis
(ii) Type I error and Type II error
(iii) One-Tailed test and Two-Tailed test

(b) The manger of a payless supermarket in Bugolobi claims that 85% of his customers take Sugar. To test the manager's claim a random sample of 120 customers showed that 95 customers take sugar. Test this claim at 5% level of significance.

Question 4

(a) Distinguish between

(I) a null hypothesis and an alternative hypothesis

(ii) Type - I error and Type - II error

(iii)a one - tailed test and a two - tailed test

(b) A company claims that the average life of a certain type of batteries is μ = 21.5 hours. To test this claim, a laboratory tests 6 batteries manufactured by this company and obtains the following : sample mean $\overline{X}$ =20, sample variance s^2 = 10. Using a 5% level of ignificance determine whether or not the results indicate that the batteries of this type have a shorter life than claimed by the company.

Question 5

(a) (i) Distinguish between a One - tailed test and a Two - tailed test

(ii) Define the following terms :

- . Null hypothesis
- . Alternative hypothesis
- . Test Statistic

(b) Just before a referendum , the movement supporters claim that 60% of the electorate support a movement type of governance whereas the multi-partysts claim that 80% of the electorate support multiparty democracy. In an opinion poll 281 out of a random sample of 500 voters state that they will vote for the movement type of governance. Decide whether there is sufficient evidence at the 5 % level to suggest that less than 60% of the electorate will vote the movement type of governance.

Unit 9
Sampling
9.1 Introduction

Sampling is that part of statistical practice concerned with the selection of a subset of individuals from within a population to yield some knowledge about the whole population, especially for the purposes of making predictions based on statistical inference.

Researchers rarely survey the entire population for two reasons (Adèr, Mellenbergh, & Hand, 2008): the cost is too high, and the population is dynamic in that the individuals making up the population may change over time. The three main advantages of sampling are that the cost is lower, data collection is faster, and since the data set is smaller it is possible to ensure homogeneity and to improve the accuracy and quality of the data.

Each observation measures one or more properties (such as weight, location, color) of observable bodies distinguished as independent objects or individuals. In survey sampling, survey weights can be applied to the data to adjust for the sample design. Results from probability theory and statistical theory are employed to guide practice. In business and medical research, sampling is widely used for gathering information about a population

The sampling process comprises several stages:

- Defining the population of concern
- Specifying a sampling frame, a set of items or events possible to measure
- Specifying a sampling method for selecting items or events from the frame
- Determining the sample size
- Implementing the sampling plan
- Sampling and data collecting

Population definition

Successful statistical practice is based on focused problem definition. In sampling, this includes defining the population from which our sample is drawn. A population can be defined as including all people or items with the characteristic one wishes to understand. Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population.

Sometimes that which defines a population is obvious. For example, a manufacturer needs to decide whether a batch of material from production is of high enough quality to be released to the customer, or should be

sentenced for scrap or rework due to poor quality. In this case, the batch is the population.
Although the population of interest often consists of physical objects, sometimes we need to sample over time, space, or some combination of these dimensions. For instance, an investigation of supermarket staffing could examine checkout line length at various times, or a study on endangered penguins might aim to understand their usage of various hunting grounds over time. For the time dimension, the focus may be on periods or discrete occasions.

In other cases, our 'population' may be even less tangible. For example, Joseph Jagger studied the behaviour of roulette wheels at a casino in Monte Carlo, and used this to identify a biased wheel. In this case, the 'population' Jagger wanted to investigate was the overall behaviour of the wheel (i.e. the probability distribution of its results over infinitely many trials), while his 'sample' was formed from observed results from that wheel. Similar considerations arise when taking repeated measurements of some physical characteristic such as the electrical conductivity of copper.

This situation often arises when we seek knowledge about the cause system of which the *observed* population is an outcome. In such cases, sampling theory may treat the observed population as a sample from a larger 'superpopulation'. For example, a researcher might study the success rate of a new 'quit smoking' program on a test group of 100 patients, in order to predict the effects of the program if it were made available nationwide. Here the superpopulation is "everybody in the country, given access to this treatment" - a group which does not yet exist, since the program isn't yet available to all.

Note also that the population from which the sample is drawn may not be the same as the population about which we actually want information. Often there is large but not complete overlap between these two groups due to frame issues etc. (see below). Sometimes they may be entirely separate - for instance, we might study rats in order to get a better understanding of human health, or we might study records from people born in 2008 in order to make predictions about people born in 2009.

Time spent in making the sampled population and population of concern precise is often well spent, because it raises many issues, ambiguities and questions that would otherwise have been overlooked at this stage.

Sampling frame

In the most straightforward case, such as the sentencing of a batch of material from production (acceptance sampling by lots), it is possible to identify and measure every single item in the population and to include any one of them in our sample. However, in the more general case this is not possible. There is no way to identify all rats in the set of all rats. Where voting is not compulsory, there is no way to identify which people will actually vote at a forthcoming election (in advance of the election). These imprecise populations are not amenable to sampling in any of the ways below and to which we could apply statistical theory.

As a remedy, we seek a sampling frame which has the property that we can identify every single element and include any in our sample.[1] The most straightforward type of frame is a list of elements of the population (preferably the entire population) with appropriate contact information. For example, in an opinion poll, possible sampling frames include an electoral register and a telephone directory.

9.2 Probability and nonprobability sampling

A **probability sampling** scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

Example: We want to estimate the total income of adults living in a given street. We visit each household in that street, identify all adults living there, and randomly select one adult from each household. (For example, we can allocate each person a random number, generated from a uniform distribution between 0 and 1, and select the person with the highest number in each household). We then interview the selected person and find their income. People living on their own are certain to be selected, so we simply add their income to our estimate of the total. But a person living in a household of two adults has only a one-in-two chance of selection. To reflect this, when we come to such a household, we would count the selected person's income twice towards the total. (In effect, the person who is *selected from that household is taken as representing the person who* isn't *selected.)*

In the above example, not everybody has the same probability of selection; what makes it a probability sample is the fact that each person's probability is known. When every element in the population *does* have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

Probability sampling includes: Simple Random Sampling, Systematic Sampling, Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling. These various ways of probability sampling have two things in common:

1. Every element has a known nonzero probability of being sampled and
2. involves random selection at some point.

Nonprobability sampling is any sampling method where some elements of the population have *no* chance of selection (these are sometimes referred to as 'out of coverage'/'undercovered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, nonprobability sampling does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population.

Example: We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant, this is a nonprobability sample, because some people are more likely to answer the door (e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities.

Nonprobability Sampling includes: Accidental Sampling, Quota Sampling and Purposive Sampling. In addition, nonresponse effects may turn *any* probability design into a nonprobability design if the characteristics of nonresponse are not well understood, since nonresponse effectively modifies each element's probability of being sampled.

9.3 Sampling methods

Within any of the types of frame identified above, a variety of sampling methods can be employed, individually or in combination. Factors commonly influencing the choice between these designs include:

- Nature and quality of the frame
- Availability of auxiliary information about units on the frame
- Accuracy requirements, and the need to measure accuracy
- Whether detailed analysis of the sample is expected
- Cost/operational concerns

Simple random sampling

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

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

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

Simple random sampling is always an EPS design, but not all EPS designs are simple random sampling.

Systematic sampling

Systematic sampling relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every *k*th element from then onwards. In this case, *k*=(population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the *k*th element in the list. A simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10').

As long as the starting point is randomized, systematic sampling is a type of probability sampling. It is easy to implement and the stratification induced can make it efficient, *if* the variable by which the list is ordered is correlated with the variable of interest. 'Every 10th' sampling is especially useful for efficient sampling from databases.

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

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

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

Another drawback of systematic sampling is that even in scenarios where it is more accurate than SRS, its theoretical properties make it difficult to *quantify* that accuracy. (In the two examples of systematic sampling that are given above, much of the potential sampling error is due to variation between neighbouring houses - but because this method never selects two neighbouring houses, the sample will not give us any information on that variation.)

As described above, systematic sampling is an EPS method, because all elements have the same probability of selection (in the example given, one in ten). It is *not* 'simple random sampling' because different subsets of the same size have different selection probabilities - e.g. the set {4,14,24,...,994} has a one-in-ten probability of selection, but the set {4,13,24,34,...} has zero probability of selection.

Systematic sampling can also be adapted to a non-EPS approach; for an example, see discussion of PPS samples below.

Stratified sampling

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

First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample.

Second, utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). Even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less efficiency than would simple random sampling, provided that each stratum is proportional to the group's size in the population.

Third, it is sometimes the case that data are more readily available for individual, pre-existing strata within a population than for the overall population; in such cases, using a stratified sampling approach may be more convenient than aggregating data across groups (though this may potentially be at odds with the previously noted importance of utilizing criterion-relevant strata).

Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.

There are, however, some potential drawbacks to using stratified sampling. First, identifying strata and implementing such an approach can increase the cost and complexity of sample selection, as well as leading to increased complexity of population estimates. Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata. Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods (although in most cases, the required sample size would be no larger than would be required for simple random sampling.

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

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

Advantages over other sampling methods

1. Focuses on important subpopulations and ignores irrelevant ones.
2. Allows use of different sampling techniques for different subpopulations.
3. Improves the accuracy/efficiency of estimation.
4. Permits greater balancing of statistical power of tests of differences between strata by sampling equal numbers from strata varying widely in size.

Disadvantages

1. Requires selection of relevant stratification variables which can be difficult.
2. Is not useful when there are no homogeneous subgroups.
3. Can be expensive to implement.

Post stratification

Stratification is sometimes introduced after the sampling phase in a process called "post stratification". This approach is typically implemented due to a lack of prior knowledge of an appropriate stratifying variable or when the experimenter lacks the necessary information to create a stratifying variable during the sampling phase. Although the method is susceptible to the pitfalls of post hoc approaches, it can provide several benefits in the right situation. Implementation usually follows a simple random sample. In addition to allowing for stratification on an ancillary variable, post-stratification can be used to implement weighting, which can improve the precision of a sample's estimates.

Oversampling

Choice-based sampling is one of the stratified sampling strategies. In choice-based sampling, the data are stratified on the target and a sample is taken from each stratum so that the rare target class will be more represented in the sample. The model is then built on this biased sample. The effects of the input variables on the target are often estimated with more precision with the choice-based sample even when a smaller overall sample size is taken, compared to a random sample. The results usually must be adjusted to correct for the oversampling.

Probability proportional to size sampling

In some cases the sample designer has access to an "auxiliary variable" or "size measure", believed to be correlated to the variable of interest, for each element in the population. These data can be used to improve accuracy in sample design. One option is to use the auxiliary variable as a basis for stratification, as discussed above.

Another option is probability-proportional-to-size ('PPS') sampling, in which the selection probability for each element is set to be proportional to its size measure, up to a maximum of 1. In a simple PPS design, these selection probabilities can then be used as the basis for Poisson sampling. However, this has the drawback of variable sample size, and different portions of the population may still be over- or under-represented due to chance variation in selections. To address this problem, PPS may be combined with a systematic approach.

Example: Suppose we have six schools with populations of 150, 180, 200, 220, 260, and 490 students respectively (total 1500 students), and we want to use student population as the basis for a PPS sample of size three. To do this, we could allocate the first school numbers 1 to 150, the second school 151 to 330 (= 150 + 180), the third school 331 to 530, and so on to the last school (1011 to 1500). We then generate a random start between 1 and 500 (equal to 1500/3) and count through the school populations by multiples of 500. If our random start was 137, we would select the schools which have been allocated numbers 137, 637, and 1137, i.e. the first, fourth, and sixth schools.

The PPS approach can improve accuracy for a given sample size by concentrating sample on large elements that have the greatest impact on population estimates. PPS sampling is commonly used for surveys of businesses, where element size varies greatly and auxiliary information is often available - for instance, a survey attempting to measure the number of guest-nights spent in hotels might use each hotel's number of rooms as an auxiliary variable. In some cases, an older measurement of the variable of interest can be used as an auxiliary variable when attempting to produce more current estimates.

Cluster sampling

Sometimes it is cheaper to 'cluster' the sample in some way e.g. by selecting respondents from certain areas only, or certain time-periods only. (Nearly all samples are in some sense 'clustered' in time - although this is rarely taken into account in the analysis.)

Cluster sampling is an example of 'two-stage sampling' or 'multistage sampling': in the first stage a sample of areas is chosen; in the second stage a sample of respondents *within* those areas is selected.
This can reduce travel and other administrative costs. It also means that one does not need a sampling frame listing all elements in the target population. Instead, clusters can be chosen from a cluster-level frame, with an element-level frame created only for the selected clusters. Cluster sampling generally increases the variability of sample estimates above that of simple random sampling, depending on how the clusters differ between themselves, as compared with the within-cluster variation.

Nevertheless, some of the disadvantages of cluster sampling are the reliance of sample estimate precision on the actual clusters chosen. If clusters chosen are biased in a certain way, inferences drawn about population parameters from these sample estimates will be far off from being accurate.

Multistage sampling Multistage sampling is a complex form of cluster sampling in which two or more levels of units are embedded one in the other. The first stage consists of constructing the clusters that will be used to sample from. In the second stage, a sample of primary units is randomly selected from each cluster (rather than using all units contained in all selected clusters). In following stages, in each of those selected clusters, additional samples of units are selected, and so on. All ultimate units (individuals, for instance) selected at the last step of this procedure are then surveyed.

This technique, thus, is essentially the process of taking random samples of preceding random samples. It is not as effective as true random sampling, but it probably solves more of the problems inherent to random sampling. Moreover, It is an effective strategy because it banks on multiple randomizations. As such, it is extremely useful.

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

Matched random sampling
A method of assigning participants to groups in which pairs of participants are first matched on some characteristic and then individually assigned randomly to groups.

The procedure for matched random sampling can be briefed with the following contexts,

1. Two samples in which the members are clearly paired, or are matched explicitly by the researcher. For example, IQ measurements or pairs of identical twins.

2. Those samples in which the same attribute, or variable, is measured twice on each subject, under different circumstances. Commonly called repeated measures. Examples include the times of a group of athletes for 1500m before and after a week of special training; the milk yields of cows before and after being fed a particular diet.

Quota sampling

In **quota sampling**, the population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion. For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.

It is this second step which makes the technique one of non-probability sampling. In quota sampling the selection of the sample is non-random. For example interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance of selection. This random element is its greatest weakness and quota versus probability has been a matter of controversy for many years.

Convenience sampling or Accidental Sampling

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

Several important considerations for researchers using convenience samples include:

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

In social science research, snowball sampling is a similar technique, where existing study subjects are used to recruit more subjects into the sample.

9.4 Line-intercept sampling

Line-intercept sampling is a method of sampling elements in a region whereby an element is sampled if a chosen line segment, called a "transect", intersects the element.

Panel sampling

Panel sampling is the method of first selecting a group of participants through a random sampling method and then asking that group for the same information again several times over a period of time. Therefore, each participant is given the same survey or interview at two or more time points; each period of data collection is called a "wave". This sampling methodology is often chosen for large scale or nation-wide studies in order to gauge changes in the population with regard to any number of variables from chronic illness to job stress to weekly food expenditures. Panel sampling can also be used to inform researchers about within-person health changes due to age or help explain changes in continuous dependent variables such as spousal interaction. There have been several proposed methods of analyzing panel sample data, including MANOVA, growth curves, and structural equation modeling with lagged effects. For a more thorough look at analytical techniques for panel data, see Johnson (1995).

Event sampling methodology

Event sampling methodology (**ESM**) is a new form of sampling method that allows researchers to study ongoing experiences and events that vary across and within days in its naturally-occurring environment. Because of the frequent sampling of events inherent in ESM, it enables researchers to measure the typology of activity and detect the temporal and dynamic fluctuations of work experiences. Popularity of ESM as a new form of research design increased over the recent years because it addresses the shortcomings of cross-sectional research, where once unable to, researchers can now detect intra-individual variances across time. In ESM, participants are asked to record their experiences and perceptions in a paper or electronic diary.

There are three types of ESM:

1. Signal contingent – random beeping notifies participants to record data. The advantage of this type of ESM is minimization of recall bias.
2. Event contingent – records data when certain events occur
3. Interval contingent – records data according to the passing of a certain period of time

ESM has several disadvantages. One of the disadvantages of ESM is it can sometimes be perceived as invasive and intrusive by participants. ESM also leads to possible self-selection bias. It may be that only certain types of individuals are willing to participate in this type of study creating a non-random sample. Another concern is related to participant cooperation. Participants may not be actually fill out their diaries at the specified times. Furthermore, ESM may substantively change the phenomenon being studied. Reactivity or priming effects may occur, such that repeated measurement may cause changes in the participants' experiences. This method of sampling data is also highly vulnerable to common method variance.

Further, it is important to think about whether or not an appropriate dependent variable is being used in an ESM design. For example, it might be logical to use ESM in order to answer research questions which involve dependent variables with a great deal of variation throughout the day. Thus, variables such as change in mood, change in stress level, or the immediate impact of particular events may be best studied using ESM methodology. However, it is not likely that utilizing ESM will yield meaningful predictions when measuring someone performing a repetitive task throughout the day or when dependent variables are long-term in nature (coronary heart problems).

Replacement of selected units

Sampling schemes may be *without replacement* ('WOR' - no element can be selected more than once in the same sample) or *with replacement* ('WR' - an element may appear multiple times in the one sample). For example, if we catch fish, measure them, and immediately return them to the water before continuing with the sample, this is a WR design, because we might end up catching and measuring the same fish more than once. However, if we do not return the fish to the water (e.g. if we eat the fish), this becomes a WOR design.

Sample size

Formulas, tables, and power function charts are well known approaches to determine sample size.

Formulas

Where the frame and population are identical, statistical theory yields exact recommendations on sample size. However, where it is not straightforward to define a frame representative of the population, it is more important to understand the cause system of which the population are outcomes and to ensure that all sources of variation are embraced in the frame. Large number of observations are of no value if major sources of variation are neglected in the study. In other words, it is taking a sample group that matches the survey category and is easy to survey.

Bartlett, Kotrlik, and Higgins (2001) published a paper titled *Organizational Research: Determining Appropriate Sample Size* in Survey Research Information Technology, Learning, and Performance Journal that provides an explanation of Cochran's (1977) formulas. A discussion and illustration of sample size formulas, including the formula for adjusting the sample size for smaller populations, is included. A table is provided that can be used to select the sample size for a research problem based on three alpha levels and a set error rate.

Steps for using sample size tables

1. Postulate the effect size of interest, α, and β.
2. Check sample size table
    1. Select the table corresponding to the selected α
    2. Locate the row corresponding to the desired power
    3. Locate the column corresponding to the estimated effect size.
    4. The intersection of the column and row is the minimum sample size required.

9.5 Sampling and data collection

Good data collection involves:

a) Following the defined sampling process
b) Keeping the data in time order
c) Noting comments and other contextual events
d) Recording non-responses

Most sampling books and papers written by non-statisticians focus only in the data collection aspect, which is just a small though important part of the sampling process.

Errors in sample surveys

Survey results are typically subject to some error. Total errors can be classified into sampling errors and non-sampling errors. The term "error" here includes systematic biases as well as random errors.

Sampling errors and biases

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

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

Non-sampling error

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

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

After sampling, a review should be held of the exact process followed in sampling, rather than that intended, in order to study any effects that any divergences might have on subsequent analysis. A particular problem is that of *non-response*.

Two major types of nonresponse exist: unit nonresponse (referring to lack of completion of any part of the survey) and item nonresponse (submission or participation in survey but failing to complete one or more components/questions of the survey). In survey sampling, many of the individuals identified as part of the sample may be unwilling to participate, not have the time to participate (opportunity cost), or survey administrators may not have been able to contact them. In this case, there is a risk of differences, between respondents and nonrespondents, leading to biased estimates of population parameters. This is often addressed by improving survey design, offering incentives, and conducting follow-up studies which make a repeated attempt to contact the unresponsive and to characterize their similarities and differences with the rest of the frame. The effects can also be mitigated by weighting the data when population benchmarks are available or by imputing data based on answers to other questions.

Nonresponse is particularly a problem in internet sampling. Reasons for this problem include improperly designed surveys, over-surveying (or survey fatigue), and the fact that potential participants hold multiple e-mail addresses, which they don't use anymore or don't check regularly. Web-based surveys also tend to demonstrate nonresponse bias; for example, studies have shown that females and those from a white/Caucasian background are more likely to respond than their counterparts.

Survey weights

In many situations the sample fraction may be varied by stratum and data will have to be weighted to correctly represent the population. Thus for example, a simple random sample of individuals in the United Kingdom might include some in remote Scottish islands who would be inordinately expensive to sample. A cheaper method would be to use a stratified sample with urban and rural strata. The rural sample could be under-represented in the sample, but weighted up appropriately in the analysis to compensate.

More generally, data should usually be weighted if the sample design does not give each individual an equal chance of being selected. For instance, when households have equal selection probabilities but one person is interviewed from within each household, this gives people from large households a smaller chance of being interviewed. This can be accounted for using survey weights. Similarly, households with more than one telephone line have a greater chance of being selected in a random digit dialing sample, and weights can adjust for this.

Weights can also serve other purposes, such as helping to correct for non-response.

Random sampling by using lots is an old idea, mentioned several times in the Bible. In 1786 Pierre Simon Laplace estimated the population of France by using a sample, along with ratio estimator. He also computed probabilistic estimates of the error. These were not expressed as modern confidence intervals but as the sample size that would be needed to achieve a particular upper bound on the sampling error with probability 1000/1001. His estimates used Bayes' theorem with a uniform prior probability and it assumed his sample was random. The theory of small-sample statistics developed by William Sealy Gossett put the subject on a more rigorous basis in the 20th century.

However, the importance of random sampling was not universally appreciated and in the USA the 1936 *Literary Digest* prediction of a Republican win in the presidential election went badly awry, due to severe bias. More than two million people responded to the study with their names obtained through magazine subscription lists and telephone directories. It was not appreciated that these lists were heavily biased towards Republicans and the resulting sample, though very large, was deeply flawed.

9.6 Sampling Distribution Of The Mean

The sampling distribution of the mean is a very important distribution. In later chapters you will see that it is used to construct confidence intervals for the mean and for significance testing.

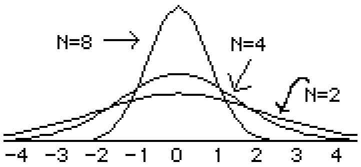
The **sampling distribution** of a statistic is distribution of that statistic, considered as a random variable, when derived from a random sample of size *n*. It may be considered as the distribution of the statistic for *all possible samples from the same population* of a given size. The sampling distribution depends on the underlying distribution of the population, the statistic being considered, the sampling procedure employed and the sample size used. There is often considerable interest in whether the sampling distribution can be approximated by an asymptotic distribution, which corresponds to the limiting case as *n* → ∞.

For example, consider a normal population with mean μ and variance σ^2. Assume we repeatedly take samples of a given size from this population and calculate the arithmetic mean $\bar{x}$for each sample — this statistic is called the sample mean. Each sample has its own average value, and the distribution of these averages is called the "sampling distribution of the sample mean". This distribution is normal $\mathcal{N}(\mu, \sigma^2/n)$since the underlying population is normal, although sampling distributions will also often be close to normal when the population distribution is not (see central limit theorem). An alternative to the sample mean is the sample median. When calculated from the same population, it has a different sampling distribution to that of the mean and is generally not normal (but it may be close for large sample sizes).

The mean of a sample from a population having a normal distribution is an example of a simple statistic taken from one of the simplest statistical populations. For other statistics and other populations the formulas are more complicated, and often they don't exist in closed-form. In such cases the sampling distributions may be approximated through Monte-Carlo simulations, bootstrap methods, or asymptotic distribution theory.

Given a population with a mean of μ and a standard deviation of σ, the sampling distribution of the mean has a mean of μ and a standard deviation of

$\sigma_M = \frac{\sigma}{\sqrt{n}}$, where n is the sample size. The standard deviation of the sampling distribution of the mean is called the standard error of the mean. It is designated by the symbol: σ_M . Note that the spread of the sampling distribution of the mean decreases as the sample size increases.



An example of the effect of sample size is shown above. Notice that the mean of the distribution is not affected by sample size. Click here for an interactive demonstration of sampling distributions.

9.7 Central Limit Theorem

The **central limit theorem** (**CLT**) states conditions under which the mean of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed. The central limit theorem (in its common form) requires the random variables to be identically distributed. Since real-world quantities are often the balanced sum of many unobserved random events, this theorem provides a partial explanation for the prevalence of the normal probability distribution. The CLT also justifies the approximation of large-sample statistics to the normal distribution in controlled experiments.

A simple example of the central limit theorem is given by the problem of rolling a large number of dice, each of which is weighted unfairly in some unknown way. The distribution of the sum (or average) of the rolled numbers will be well approximated by a normal distribution, the parameters of which can be determined empirically.

In more general probability theory, a **central limit theorem** is any of a set of weak-convergence theories. They all express the fact that a sum of many independent random variables will tend to be distributed according to one of a small set of "attractor" (i.e. stable) distributions. When the variance of the variables is finite, the "attractor" distribution is the normal distribution. Specifically, the sum of a number of random variables with power law tail distributions decreasing as $1/|x|^{\alpha + 1}$ where $0 < \alpha < 2$ (and therefore having infinite variance) will tend to a stable distribution with stability parameter (or index of stability) of α as the number of variables grows. This article is concerned only with the classical (i.e. finite variance) central limit theorem

Central limit theorem

The means of a large number of samples taken randomly from the same population will themselves be normally distributed, and the axis of symmetry will be the population means.

Calculating the standard error of mean

$$S.E. = \frac{\text{sample standard deviation}}{\sqrt{\text{Number of sampling units}}} = \frac{s}{\sqrt{n}}$$

Using the wing length measurements of 100 Robins (we determined in section 7.5 that the sample mean was 74.00mm and the standard deviation 2.34mm)

$$S.E. = \frac{s}{\sqrt{N}} = \frac{2.34}{\sqrt{100}} = \frac{2.34}{10} = 0.234.$$

9.8 Confidence limits of a proportionate or percentage

$$S.E. = \sqrt{\frac{P\ (1-\ p)}{N\ -\ 1}}$$

Where **p** = sample proportion;
n = Number of sampling units.

Thus, in this example**,**

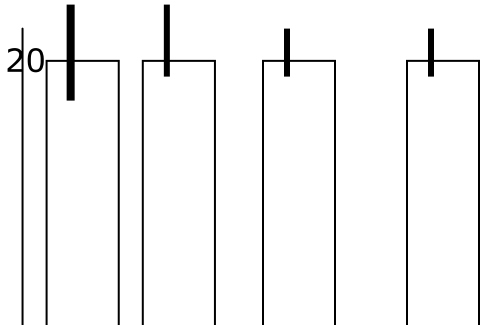
$$S.E. = \sqrt{\frac{0.75\ (1-\ 0.75)}{80 - 1}} = 0.049$$

The 95% confidence limits are therefore 0.75 ± (1.96 x 0.049)
= 0.75 ± 0.09

The graphical display of variation

Graphs or histograms depicting mean values of sets of counts or measurements often show “error” or deviation” bars. These bars may indicate any of the following

(a) range
(b) standard deviation
(c) standard error of the mean
(d) 95% confidence limits.



Mean weighted 10

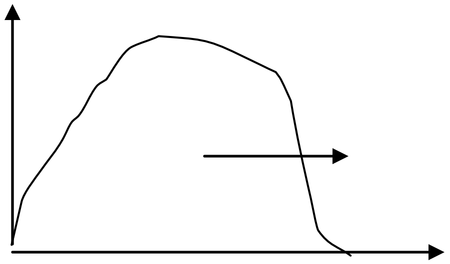
0 ± range ± s ± S.E

95% CL

Clumped Distribution

A clump distribution (also called "aggregated" over dispersed by a frequency distribution diagram of the data which has a strong positive skew (that is, the tail extends to the right) statistically, a sample of clumped data has a variance which is considerably greater than the mean. An example will illustrate this.

Positive skewed variance (s^2) >mean (x)



- S2 > x
- Not normally distributed
- They skewed
- Log transformation is done to normalize the data.

Log transformation

Is the most suitable transformation for clumped counts, but is not appropriate if there are any zero counts: Log (O) = infinity. "Tail" or skew has been squashed up by the transformation and the curve does, indeed, appear to be symmetrical.

The transformed counts the variance is now less than the mean and a satisfactory transformation to Normal maybe assumed.

When the analyses are complete, the arithmetic means of the transformed counts has to be transformed back to the original scale and thus becomes a derived mean. For a log (x + 1) transformation, the antilog of the mean transformed count x must be obtained and 1 subtracted.

Random distribution – the square root transformation.

A population which furnishes samples whose variances are about equal to the means is said to exhibit a random, or Poisson, diction and, because a random distribution, some transformation is necessary in circumstances when normality is required or assumed.

The variance is very similar to the men and so a random distribution may be assumed. In this case transformations of x by √x will covert the data to an appropriately Normal distribution.

The variance of the transformed counts is now well below the confidence limits or the use of parametric tests which assumes a normal or t distribution. At the end of the procedure, the transformed statistics are converted back to the original scale by squaring.

Is data transformation really necessary?

Data transformation should not be regarded as somehow "cheating", It is a way of making sure that the statistical methods can be validly applied. However, there is an alternative to data transformation

Central limit theorems for independent sequences

Classical CLT

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample of size n — that is, a sequence of independent and identically distributed random variables with expected values μ and variances σ^2. Suppose we are interested in the behavior of the sample average of these random variables: $S_n = 1\boxed{n}(X_1 + \ldots + X_n)$. Then the central limit theorem asserts that for large n's, the distribution of S_n is approximately normal with mean μ and variance $1\boxed{n}\sigma^2$. The true strength of the theorem is that S_n approaches normality regardless of the shapes of the distributions of individual X_i's. Formally, the theorem can be stated as follows:

> **Lindeberg–Lévy CLT:** suppose $\{X_i\}$ is a sequence of iid random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$. Then as n approaches infinity, the random variable $\sqrt{\boxed{n}}(S_n - \mu)$ converges in distribution to a normal $N(0, \sigma^2)$:
>
> $$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} \mathcal{N}(0,\,\sigma^2).$$

Convergence in distribution means that the cumulative distribution function of $\sqrt{\boxed{n}}(S_n - \mu)$ converges pointwise to the cdf of the $N(0, \sigma^2)$ distribution: for any real number z,

$$\lim_{n\to\infty} \Pr[\sqrt{n}(S_n - \mu) \le z] = \Phi(z/\sigma),$$

where $\Phi(x)$ is the standard normal cdf.

Lyapunov CLT

The theorem is named after a Russian mathematician Aleksandr Lyapunov. In this variant of the central limit theorem the random variables X_i have to be independent, but not necessarily identically distributed. The theorem also requires that random variables $|X_i|$ have moments of some order $(2 + \delta)$, and that the rate of growth of these moments is limited by the Lyapunov condition given below.

> **Lyapunov CLT:** let $\{X_i\}$ be a sequence of independent random variables, each having a finite expected value μ_i and variance σ 2
> *i* . Define *s* 2
> *n* = ∑*n*
> *i* = 1 σ 2
> *i* . If for some $\delta > 0$, the *Lyapunov's condition*
>
> $$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathrm{E}\big[\,|X_i - \mu_i|^{2+\delta}\,\big] = 0$$
>
> is satisfied, then a sum of $(X_i - \mu_i)/s_n$ converges in distribution to a standard normal random variable, as n goes to infinity:
>
> $$\frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0,\ 1).$$

In practice it is usually easiest to check the Lyapunov's condition for $\delta = 1$. If a sequence of random variables satisfies Lyapunov's condition, then it also satisfies Lindeberg's condition. The converse implication, however, does not hold.

Lindeberg CLT

In the same setting and with the same notation as above, we can replace the Lyapunov condition with the following weaker one (from Lindeberg in 1920). For every $\varepsilon > 0$

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{n} \mathrm{E}\big[(X_i - \mu_i)^2 \cdot \mathbf{1}_{\{|X_i-\mu_i| > \varepsilon s_n\}}\big] = 0$$

where $\mathbf{1}_{\{\ldots\}}$ is the indicator function. Then the distribution of the standardized sum Z_n converges towards the standard normal distribution N(0,1).

Multidimensional CLT

We can easily extend proofs using characteristic functions for cases where each individual $X_1, X_2, X_3, \ldots, X_n$ is an independent and identically distributed random vector in $\mathbb{R}^k$, with mean vector $\mu = E\left(X_i\right)$and covariance matrix **Σ** (amongst the individual components of the vector). Now, if we take the summations of these vectors as being done componentwise, then the Multidimensional central limit theorem states that when scaled, these converge to a multivariate normal distribution[5].

Let

$$\mathbf{X_i} = \begin{bmatrix} X_{i(1)} \\ \vdots \\ X_{i(k)} \end{bmatrix}$$

be the i-vector. The bold in $\mathbf{X_i}$ means that it is a random vector, not a random (univariate) variable.

Then the sum of the random vectors will be

$$\begin{bmatrix} X_{1(1)} \\ \vdots \\ X_1(k) \end{bmatrix} + \begin{bmatrix} X_{2(1)} \\ \vdots \\ X_{2(k)} \end{bmatrix} + \ldots + \begin{bmatrix} X_{n(1)} \\ \vdots \\ X_{n(k)} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \left[X_{i(1)}\right] \\ \vdots \\ \sum_{i=1}^{n} \left[X_{i(k)}\right] \end{bmatrix} = \sum_{i=1}^{n} \left[\mathbf{X_i}\right]$$

and the average will be

$$\left(\frac{1}{n}\right) \sum_{i=1}^{n} \left[\mathbf{X_i}\right] = \frac{1}{n} \begin{bmatrix} \sum_{i=1}^{n} \left[X_{i(1)}\right] \\ \vdots \\ \sum_{i=1}^{n} \left[X_{i(k)}\right] \end{bmatrix} = \begin{bmatrix} \bar{X}_{i(1)} \\ \vdots \\ \bar{X}_{i(k)} \end{bmatrix} = \mathbf{\bar{X}_n}$$

and therefore

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathbf{X_i} - E\left(X_i\right)\right] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathbf{X_i} - \mu\right] = \sqrt{n}\left(\overline{\mathbf{X}}_n - \mu\right).$$

The multivariate central limit theorem states that

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \mu\right) \xrightarrow{D} \mathcal{N}_k(0, \Sigma)$$

where the covariance matrix Σ is equal to

$$\Sigma = \begin{bmatrix} Cov\left(X_1, X_1\right) & Cov\left(X_1, X_2\right) & Cov\left(X_1, X_3\right) & \cdots & Cov\left(X_1, X_n\right) \\ Cov\left(X_2, X_1\right) & Cov\left(X_2, X_2\right) & Cov\left(X_2, X_3\right) & \cdots & Cov\left(X_2, X_n\right) \\ Cov\left(X_3, X_1\right) & Cov\left(X_3, X_2\right) & Cov\left(X_3, X_3\right) & \cdots & Cov\left(X_3, X_n\right) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov\left(X_n, X_1\right) & Cov\left(X_n, X_2\right) & Cov\left(X_n, X_3\right) & \cdots & Cov\left(X_n, X_n\right) \end{bmatrix} =$$

$$= \begin{bmatrix} Var\left(X_1\right) & Cov\left(X_1, X_2\right) & Cov\left(X_1, X_3\right) & \cdots & Cov\left(X_1, X_n\right) \\ Cov\left(X_2, X_1\right) & Var\left(X_2\right) & Cov\left(X_2, X_3\right) & \cdots & Cov\left(X_2, X_n\right) \\ Cov\left(X_3, X_1\right) & Cov\left(X_3, X_2\right) & Var\left(X_3\right) & \cdots & Cov\left(X_3, X_n\right) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov\left(X_n, X_1\right) & Cov\left(X_n, X_2\right) & Cov\left(X_n, X_3\right) & \cdots & Var\left(X_n\right) \end{bmatrix}$$

Central limit theorems for dependent processes

CLT under weak dependence

A useful generalization of a sequence of independent, identically distributed random variables is a mixing random process in discrete time; "mixing" means, roughly, that random variables temporally far apart from one another are nearly independent. Several kinds of mixing are used in ergodic theory and probability theory. See especially strong mixing (also called α-mixing) defined by $\alpha(n) \to 0$ where $\alpha(n)$ is so-called strong mixing coefficient.

A simplified formulation of the central limit theorem under strong mixing is given in (Billingsley 1995, Theorem 27.4):

Theorem. Suppose that X_1, X_2, … is stationary and α-mixing with $\alpha_n = O(n^{-5})$ and that $E(X_n) = 0$ and $E(X_n^{12}) < \infty$. Denote $S_n = X_1 + \ldots + X_n$, then the limit $\sigma^2 = \lim_n n^{-1}E(S_n^2)$ exists, and if $\sigma \neq 0$ then $S_n/(\sigma\sqrt{n})$converges in distribution to N(0, 1).

In fact, $\sigma^2 = E(X_1^2) + 2\sum_{k=1}^{\infty}E(X_1X_{1+k})$, where the series converges absolutely.

The assumption $\sigma \neq 0$ cannot be omitted, since the asymptotic normality fails for $X_n = Y_n - Y_{n-1}$ where Y_n are another stationary sequence.

For the theorem in full strength see (Durrett 1996, Sect. 7.7(c), Theorem (7.8)); the assumption $E(X_n^{12}) < \infty$ is replaced with $E(|X_n|^{2+\delta}) < \infty$, and the assumption $\alpha_n = O(n^{-5})$ is replaced with $\sum_n \alpha_n^{\frac{\delta}{2(2+\delta)}} < \infty.$ Existence of such $\delta > 0$ ensures the conclusion. For encyclopedic treatment of limit theorems under mixing conditions see (Bradley 2005).

Martingale difference CLT

Theorem. Let a martingale M_n satisfy

- $\frac{1}{n}\sum_{k=1}^{n} E((M_k - M_{k-1})^2 | M_1, \ldots, M_{k-1}) \to 1$ in probability as n tends to infinity,
- for every $\varepsilon > 0$, $\frac{1}{n}\sum_{k=1}^{n} E\Big((M_k - M_{k-1})^2; |M_k - M_{k-1}| > \varepsilon\sqrt{n}\Big) \to 0$ as n tends to infinity,

then $M_n/\sqrt{n}$converges in distribution to N(0,1) as n tends to infinity.

See (Durrett 1996, Sect. 7.7, Theorem (7.4)) or (Billingsley 1995, Theorem 35.12).

Caution: The restricted expectation E(X; A) should not be confused with the conditional expectation E($X|A$) = E(X; A)/**P**(A).

Remarks

Proof of classical CLT

For a theorem of such fundamental importance to statistics and applied probability, the central limit theorem has a remarkably simple proof using characteristic functions. It is similar to the proof of a (weak) law of large numbers. For any random variable, Y, with zero mean and a unit variance (var(Y) = 1), the characteristic function of Y is, by Taylor's theorem,

$$\varphi_Y(t) = 1 - \frac{t^2}{2} + o(t^2), \quad t \to 0$$

where o (t^2) is "little o notation" for some function of t that goes to zero more rapidly than t^2. Letting Y_i be (X_i − μ)/σ, the standardized value of X_i, it is easy to see that the standardized mean of the observations X_1, X_2, ..., X_n is

$$Z_n = \frac{n\overline{X}_n - n\mu}{\sigma\sqrt{n}} = \sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}}.$$

By simple properties of characteristic functions, the characteristic function of Z_n is

$$\left[\varphi_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n \to e^{-t^2/2}, \quad n \to \infty.$$

But this limit is just the characteristic function of a standard normal distribution N(0, 1), and the central limit theorem follows from the Lévy continuity theorem, which confirms that the convergence of characteristic functions implies convergence in distribution.

9.9 Convergence to the limit

The central limit theorem gives only an asymptotic distribution. As an approximation for a finite number of observations, it provides a reasonable approximation only when close to the peak of the normal distribution; it requires a very large number of observations to stretch into the tails.

If the third central moment E((X_1 − μ)^3) exists and is finite, then the above convergence is uniform and the speed of convergence is at least on the order of 1/$n^{1/2}$.

The convergence to the normal distribution is monotonic, in the sense that the entropy of Z_n increases monotonically to that of the normal distribution, as proven in Artstein, Ball, Barthe and Naor (2004).

The central limit theorem applies in particular to sums of independent and identically distributed discrete random variables. A sum of discrete random variables is still a discrete random variable, so that we are confronted with a sequence of discrete random variables whose cumulative probability distribution function converges towards a cumulative probability distribution function corresponding to a continuous variable (namely that of the normal distribution). This means that if we build a histogram of the realisations of the sum of *n* independent identical discrete variables, the curve that joins the centers of the upper faces of the rectangles forming the histogram converges toward a Gaussian curve as *n* approaches infinity, this relation is known as de Moivre–Laplace theorem. The binomial distribution article details such an application of the central limit theorem in the simple case of a discrete variable taking only two possible values.

Relation to the law of large numbers

The law of large numbers as well as the central limit theorem are partial solutions to a general problem: "What is the limiting behavior of S_n as n approaches infinity?" In mathematical analysis, asymptotic series are one of the most popular tools employed to approach such questions.

Suppose we have an asymptotic expansion of $f(n)$:

$$f(n) = a_1\varphi_1(n) + a_2\varphi_2(n) + O(\varphi_3(n)) \qquad (n \to \infty).$$

Dividing both parts by $\varphi_1(n)$ and taking the limit will produce a_1, the coefficient of the highest-order term in the expansion, which represents the rate at which $f(n)$ changes in its leading term.

$$\lim_{n\to\infty} \frac{f(n)}{\varphi_1(n)} = a_1.$$

Informally, one can say: "$f(n)$ grows approximately as $a_1\ \varphi(n)$". Taking the difference between $f(n)$ and its approximation and then dividing by the next term in the expansion, we arrive at a more refined statement about $f(n)$:

$$\lim_{n\to\infty} \frac{f(n) - a_1\varphi_1(n)}{\varphi_2(n)} = a_2$$

Here one can say that the difference between the function and its approximation grows approximately as $a_2\ \varphi_2(n)$. The idea is that dividing the function by appropriate normalizing functions, and looking at the limiting behavior of the result, can tell us much about the limiting behavior of the original function itself.

Informally, something along these lines is happening when the sum, S_n, of independent identically distributed random variables, $X_1, ..., X_n$, is studied in classical probability theory. If each X_i has finite mean μ, then by the law of large numbers, $S_n/n \to \mu$. If in addition each X_i has finite variance σ^2, then by the central limit theorem,

$$\frac{S_n - n\mu}{\sqrt{n}} \to \xi$$

where ξ is distributed as N(0, σ^2). This provides values of the first two constants in the informal expansion

$$S_n \approx \mu n + \xi\sqrt{n}.$$

In the case where the X_i's do not have finite mean or variance, convergence of the shifted and rescaled sum can also occur with different centering and scaling factors:

$$\frac{S_n - a_n}{b_n} \rightarrow \Xi,$$

or informally

$$S_n \approx a_n + \Xi b_n.$$

Distributions Ξ which can arise in this way are called *stable*. Clearly, the normal distribution is stable, but there are also other stable distributions, such as the Cauchy distribution, for which the mean or variance are not defined. The scaling factor b_n may be proportional to n^c, for any $c \geq 1/2$; it may also be multiplied by a slowly varying function of n.

The law of the iterated logarithm tells us what is happening "in between" the law of large numbers and the central limit theorem. Specifically it says that the normalizing function $\sqrt{n \log \log n}$intermediate in size between n of the law of large numbers and √n of the central limit theorem provides a non-trivial limiting behavior.

Illustration

Main article: Illustration of the central limit theorem

Given its importance to statistics, a number of papers and computer packages are available that demonstrate the convergence involved in the central limit theorem.

Alternative statements of the theorem

Density functions

The density of the sum of two or more independent variables is the convolution of their densities (if these densities exist). Thus the central limit theorem can be interpreted as a statement about the properties of density functions under convolution: the convolution of a number of density functions tends to the normal density as the number of density functions increases without bound, under the conditions stated above.

Characteristic functions

Since the characteristic function of a convolution is the product of the characteristic functions of the densities involved, the central limit theorem has yet another restatement: the product of the characteristic functions of a number of density functions becomes close to the characteristic function of the normal density as the number of density functions increases without bound, under the conditions stated above. However, to state this more precisely, an appropriate scaling factor needs to be applied to the argument of the characteristic function.

An equivalent statement can be made about Fourier transforms, since the characteristic function is essentially a Fourier transform.

Extensions to the theorem

Products of positive random variables

The logarithm of a product is simply the sum of the logarithms of the factors. Therefore when the logarithm of a product of random variables that take only positive values approaches a normal distribution, the product itself approaches a log-normal distribution. Many physical quantities (especially mass or length, which are a matter of scale and cannot be negative) are the products of different random factors, so they follow a log-normal distribution.

Whereas the central limit theorem for sums of random variables requires the condition of finite variance, the corresponding theorem for products requires the corresponding condition that the density function be square-integrable (see Rempala 2002).

Multivariate central limit theorem

If the i.i.d. random variate is an *m*-dimensional vector, represented as X_j where j=0,1,...,m and the mean vector is:

$$\mu_j = E(X_j)$$

and the covariance matrix for the *m* components is

$$\Sigma_{jk} = E(X_j X_k) - E(X_j)E(X_k)$$

and if X_{ij} is the j-th component in the i-th sample, then the sample mean for *N* trials is

$$\overline{X_j} = \frac{1}{N}\sum_{i=1}^{N} X_{ij}$$

and the central limit theorem for the distribution of $\overline{X_j}$ will be:

$$\overline{X_j} - \mu_j \xrightarrow{d} \mathcal{N}_N(0, \Sigma_{jk}/N) = \frac{\exp\left[-\frac{N}{2}\Sigma_{jk}^{-1}(\overline{X_j} - \mu_j)(\overline{X_k} - \mu_k)\right]}{\sqrt{|2\pi\Sigma_{jk}/N|}}$$

where $\mathcal{N}_N()$is the multivariate normal distribution for *N* trials, "| |" specifies the determinant, and summation is assumed over products with repeated indices. For *N=1*, $X_j \to X,\ \mu_j \to \mu, \Sigma_{jk} \to \sigma^2$and the univariate central limit theorem is recovered.

Beyond the classical framework

Asymptotic normality, that is, convergence to the normal distribution after appropriate shift and rescaling, is a phenomenon much more general than the classical framework treated above, namely, sums of independent random variables (or vectors). New frameworks are revealed from time to time; no single unifying framework is available for now.

Convex body

Theorem (Klartag 2007, Theorem 1.2). There exists a sequence $\varepsilon_n \downarrow 0$ for which the following holds. Let $n \geq 1$, and let random variables $X_1, \ldots, X_n$ have a log-concave joint density f such that $f(x_1, \ldots, x_n) = f(|x_1|, \ldots, |x_n|)$ for all $x_1, \ldots, x_n$, and $E(X_k^2) = 1$ for all $k = 1, \ldots, n$. Then the distribution of $(X_1 + \ldots + X_n)/\sqrt{n}$is ε_n-close to N(0, 1) in the total variation distance.

These two ε_n-close distributions have densities (in fact, log-concave densities), thus, the total variance distance between them is the integral of the absolute value of the difference between the densities. Convergence in total variation is stronger than weak convergence.

An important example of a log-concave density is a function constant inside a given convex body and vanishing outside; it corresponds to the uniform distribution on the convex body, which explains the term "central limit theorem for convex bodies".

Another example: $f(x_1, \ldots, x_n) = \text{const} \cdot \exp(-(|x_1|^\alpha + \ldots + |x_n|^\alpha)^\beta)$ where $\alpha > 1$ and $\alpha\beta > 1$. If $\beta = 1$ then $f(x_1, \ldots, x_n)$ factorizes into $\text{const} \cdot \exp(-|x_1|^\alpha)\ldots\exp(-|x_n|^\alpha)$, which means independence of $X_1, \ldots, X_n$. In general, however, they are dependent.

The condition $f(x_1, \ldots, x_n) = f(|x_1|, \ldots, |x_n|)$ ensures that $X_1, \ldots, X_n$ are of zero mean and uncorrelated; still, they need not be independent, nor even pairwise independent. By the way, pairwise independence cannot replace independence in the classical central limit theorem (Durrett 1996, Section 2.4, Example 4.5).

Here is a Berry-Esseen type result.

Theorem (Klartag 2008, Theorem 1). Let $X_1, \ldots, X_n$ satisfy the assumptions of the previous theorem, then

$$\left|\mathbb{P}\Big(a \leq \frac{X_1 + \cdots + X_n}{\sqrt{n}} \leq b\Big) - \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-t^2/2}\,\mathrm{d}t\right| \leq \frac{C}{n}$$

for all $a < b$; here C is a universal (absolute) constant. Moreover, for every $c_1, \ldots, c_n \in \mathbf{R}$ such that $c_1^2 + \ldots + c_n^2 = 1$,

$$\left|\mathbb{P}(a \le c_1X_1+\cdots+c_nX_n \le b) - \frac{1}{\sqrt{2\pi}}\int_a^b \mathrm{e}^{-t^2/2}\,\mathrm{d}t\right| \le C(c_1^4+\cdots+c_n^4).$$

A more general case is treated in (Klartag 2007, Theorem 1.1). The condition $f(x_1, \ldots, x_n) = f(|x_1|, \ldots, |x_n|)$ is replaced with much weaker conditions: $E(X_k) = 0$, $E(X_k^2) = 1$, $E(X_kX_\ell) = 0$ for $1 \le k < \ell \le n$. The distribution of $(X_1+\ldots+X_n)/\sqrt{n}$need not be approximately normal (in fact, it can be uniform). However, the distribution of $c_1X_1 + \ldots + c_nX_n$ is close to N(0,1) (in the total variation distance) for most of vectors $(c_1, \ldots, c_n)$ according to the uniform distribution on the sphere $c_1^2 + \ldots + c_n^2 = 1$.

Lacunary trigonometric series

Theorem (Salem - Zygmund). Let U be a random variable distributed uniformly on $(0, 2\pi)$, and $X_k = r_k \cos(n_kU + a_k)$, where

- n_k satisfy the lacunarity condition: there exists $q > 1$ such that $n_{k+1} \ge qn_k$ for all k,
- r_k are such that
$$r_1^2+r_2^2+\cdots=\infty \text{ and } \frac{r_k^2}{r_1^2+\cdots+r_k^2}\to 0,$$
- $0 \le a_k < 2\pi$.

Then

$$\frac{X_1+\cdots+X_k}{\sqrt{r_1^2+\cdots+r_k^2}}$$

converges in distribution to $N(0, 1/2)$.

See (Zygmund 1959, Sect. XVI.5, Theorem (5-5)) or (Gaposhkin 1966, Theorem 2.1.13).

Gaussian polytopes

Theorem (Barany & Vu 2007, Theorem 1.1). Let $A_1, \ldots, A_n$ be independent random points on the plane $\mathbf{R}^2$ each having the two-dimensional standard normal distribution. Let K_n be the convex hull of these points, and X_n the area of K_n Then

$$\frac{X_n-\mathrm{E}X_n}{\sqrt{\mathrm{Var}X_n}}$$

converges in distribution to $N(0,1)$ as n tends to infinity.

The same holds in all dimensions (2, 3, ...).

The polytope K_n is called Gaussian random polytope.

A similar result holds for the number of vertices (of the Gaussian polytope), the number of edges, and in fact, faces of all dimensions (Barany & Vu 2007, Theorem 1.2).

Linear functions of orthogonal matrices

A linear function of a matrix M is a linear combination of its elements (with given coefficients), $M \mapsto \mathrm{tr}(AM)$ where A is the matrix of the coefficients; see Trace_(linear_algebra)#Inner product.

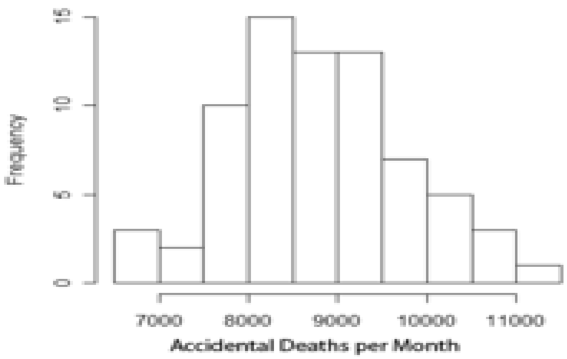
A random orthogonal matrix is said to be distributed uniformly, if its distribution is the normalized Haar measure on the orthogonal group $O(n,\mathbf{R})$; see Rotation matrix#Uniform random rotation matrices.

Theorem (Meckes 2008). Let M be a random orthogonal $n \times n$ matrix distributed uniformly, and A a fixed $n \times n$ matrix such that $\mathrm{tr}(AA^*) = n$, and let $X = \mathrm{tr}(AM)$. Then the distribution of X is close to $N(0,1)$ in the total variation metric up to $2\sqrt{3}/(n-1)$.

Subsequences

Theorem (Gaposhkin 1966, Sect. 1.5). Let random variables $X_1, X_2, \ldots \in L_2(\Omega)$ be such that $X_n \to 0$ weakly in $L_2(\Omega)$ and $X_n^2 \to 1$ weakly in $L_1(\Omega)$. Then there exist integers $n_1 < n_2 < \ldots$ such that $(X_{n_1} + \cdots + X_{n_k})/\sqrt{k}$ converges in distribution to $N(0, 1)$ as k tends to infinity.

Applications and examples



A histogram plot of monthly accidental deaths in the US, between 1973 and 1978 exhibits normality, due to the central limit theorem

There are a number of useful and interesting examples and applications arising from the central limit theorem (Dinov, Christou & Sanchez 2008).

The probability distribution for total distance covered in a random walk (biased or unbiased) will tend toward a normal distribution.

- Flipping a large number of coins will result in a normal distribution for the total number of heads (or equivalently total number of tails).

From another viewpoint, the central limit theorem explains the common appearance of the "Bell Curve" in density estimates applied to real world data. In cases like electronic noise, examination grades, and so on, we can often regard a single measured value as the weighted average of a large number of small effects. Using generalisations of the central limit theorem, we can then see that this would often (though not always) produce a final distribution that is approximately normal.

In general, the more a measurement is like the sum of independent variables with equal influence on the result, the more normality it exhibits. This justifies the common use of this distribution to stand in for the effects of unobserved variables in models like the linear model.

Signal processing

Signals can be smoothed by applying a Gaussian filter, which is just the convolution of a signal with an appropriately scaled Gaussian function. Due to the central limit theorem this smoothing can be approximated by several filter steps that can be computed much faster, like the simple moving average.

The central limit theorem implies that to achieve a Gaussian of variance σ^2 n filters with windows of variances σ_1^2, …, σ_n^2 with $\sigma^2 = \sigma_1^2 + \cdots + \sigma_n^2$ must be applied.

History

Tijms (2004, p. 169) writes:

The central limit theorem has an interesting history. The first version of this theorem was postulated by the French-born mathematician Abraham de Moivre who, in a remarkable article published in 1733, used the normal distribution to approximate the distribution of the number of heads resulting from many tosses of a fair coin. This finding was far ahead of its time, and was nearly forgotten until the famous French mathematician Pierre-Simon Laplace rescued it from obscurity in his monumental work *Théorie Analytique des Probabilités*, which was published in 1812.

Laplace expanded De Moivre's finding by approximating the binomial distribution with the normal distribution. But as with De Moivre, Laplace's finding received little attention in his own time. It was not until the nineteenth century was at an end that the importance of the central limit theorem was discerned, when, in 1901, Russian mathematician Aleksandr Lyapunov defined it in general terms and proved precisely how it worked mathematically. Nowadays, the central limit theorem is considered to be the unofficial sovereign of probability theory

Sir Francis Galton (*Natural Inheritance*, 1889) described the Central Limit Theorem as:
I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error". The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along

The actual term "central limit theorem" (in German: "zentraler Grenzwertsatz") was first used by George Pólya in 1920 in the title of a paper.[Pólya referred to the theorem as "central" due to its importance in probability theory. According to Le Cam, the French school of probability interprets the word *central* in the sense that "it describes the behaviour of the centre of the distribution as opposed to its tails". The abstract of the paper *On the central limit theorem of calculus of probability and the problem of moments* by Pólya in 1920 translates as follows.

The occurrence of the Gaussian probability density e^{-x2} in repeated experiments, in errors of measurements, which result in the combination of very many and very small elementary errors, in diffusion processes etc., can be explained, as is well-known, by the very same limit theorem, which plays a central role in the calculus of probability. The actual discoverer of this limit theorem is to be named Laplace; it is likely that its rigorous proof was first given by Tschebyscheff and its sharpest formulation can be found, as far as I am aware of, in an article by Liapounoff.

A thorough account of the theorem's history, detailing Laplace's foundational work, as well as Cauchy's, Bessel's and Poisson's contributions, is provided by Hald. Two historical accounts, one covering the development from Laplace to Cauchy, the second the contributions by von Mises, Pólya, Lindeberg, Lévy, and Cramér during the 1920s, are given by Hans Fischer. Le Cam describes a period around 1935.

A curious footnote to the history of the Central Limit Theorem is that a proof of a result similar to the 1922 Lindeberg CLT was the subject of Alan Turing's 1934 Fellowship Dissertation for King's College at the University of Cambridge. Only after submitting the work did Turing learn it had already been proved. Consequently, Turing's dissertation was never published.

9.10 Using Sample Data For Estimation

Sample size determination is the act of choosing the number of observations to include in a statistical sample. The sample size is an important feature of any empirical study in which the goal is to make inferences about a population from a sample. In practice, the sample size used in a study is determined based on the expense of data collection, and the need to have sufficient statistical power. In complicated studies there may be several different sample sizes involved in the study: for example, in as survey sampling involving stratified sampling there would be different sample sizes for each population. In a census, data are collected on the entire population, hence the sample size is equal to the population size. In experimental design, where a study may be divided into different treatment groups, there may be different sample sizes for each group.

Sample sizes may chosen in several different ways:

- expedience: for example, include those items it is convenient to collect within a given time period
- using a target variance for an estimate to be derived from the sample eventually obtained
- using a target for the power of a statistical test to be applied once the sample is collected.

How samples are collected is discussed in sampling (statistics) and survey data collection
Larger sample sizes generally lead to increased precision when estimating unknown parameters. For example, if we wish to know the proportion of a certain species of fish that is infected with a pathogen, we would generally have a more accurate estimate of this proportion if we sampled and examined 200, rather than 100 fish. Several fundamental facts of mathematical statistics describe this phenomenon, including the law of large numbers and the central limit theorem.

In some situations, the increase in accuracy for larger sample sizes is minimal, or even non-existent. This can result from the presence of systematic errors or strong dependence in the data, or if the data follow a heavy-tailed distribution.

Sample sizes are judged based on the quality of the resulting estimates. For example, if a proportion is being estimated, one may wish to have the 95% confidence interval be less than 0.06 units wide. Alternatively, sample size may be assessed based on the power of a hypothesis test. For example, if we are comparing the support for a certain political candidate among women with the support for that candidate among men, we may wish to have 80% power to detect a difference in the support levels of 0.04 units.

Estimating proportions and means

A relatively simple situation is estimation of a proportion. For example, we may wish to estimate the proportion of residents in a community who are at least 65 years old.

The estimator of a proportion is $\hat{p} = X/n$, where *X* is the number of 'positive' observations (e.g. the number of people out of the *n* sampled people who are at least 65 years old). When the observations are independent, this estimator has a (scaled) binomial distribution (and is also the sample mean of data from a Bernoulli distribution). The maximum variance of this distribution is 0.25/*n*, which occurs when the true parameter is *p* = 0.5. In practice, since *p* is unknown, the maximum variance is often used to for sample size assessments.

For sufficiently large *n*, the distribution of $\hat{p}$will be closely approximated by a normal distribution with the same mean and variance.Using this approximation, it can be shown that around 95% of this distribution's probability lies within 2 standard deviations of the mean. Because of this, an interval of the form

$$(\hat{p} - 2\sqrt{0.25/n}, \hat{p} + 2\sqrt{0.25/n})$$

will form a 95% confidence interval for the true proportion. If this interval needs to be no more than *W* units wide, the equation

$$4\sqrt{0.25/n} = W$$

can be solved for *n*, yielding $n = 4/W^2 = 1/B^2$ where *B* is the error bound on the estimate, i.e., the estimate is usually given as *within* ± *B*. So, for *B* = 10% one requires *n* = 100, for *B* = 5% one needs *n* = 400, for *B* = 3% the requirement approximates to *n* = 1000, while for *B* = 1% a sample size of *n* = 10000 is required. These numbers are quoted often in news reports of opinion polls and other sample surveys.

Estimation of means

A proportion is a special case of a mean. When estimating the population mean using an independent and identically distributed (iid) sample of size *n*, where each data value has variance σ^2, the standard error of the sample mean is:

$$\sigma/\sqrt{n}.$$

This expression describes quantitatively how the estimate becomes more precise as the sample size increases. Using the central limit theorem to justify approximating the sample mean with a normal distribution yields an approximate 95% confidence interval of the form

$$(\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n}).$$

If we wish to have a confidence interval that is *W* units in width, we would solve

$$4\sigma/\sqrt{n} = W$$

for *n*, yielding the sample size $n = 16\sigma^2/W^2$.

For example, if we are interested in estimating the amount by which a drug lowers a subject's blood pressure with a confidence interval that is six units wide, and we know that the standard deviation of blood pressure in the population is 15, then the required sample size is 100.

Required sample sizes for hypothesis tests

A common problem facing statisticians is calculating the sample size required to yield a certain power for a test, given a predetermined Type I error rate α. As follows, this can be estimated by pre-determined tables for certain values, by Mead's resource equation, or, more generalized, but the cumulative distribution function:

By tables

The table shown at right can be used to in a two-sample t-test to estimate the sample sizes of an experimental group and a control group that are of equal size, that is, the total number of individuals in the trial is twice that of the number given, and the desired significance level is 0.05. The parameters used are:

- The desired statistical power of the trial, shown in column to the left.
- Cohen's d, which is the expected difference between the means of the target values between the experimental group and the control group, divided by the expected standard deviation.

[4]	**Cohen's d**		
Power	**0.2**	**0.5**	**0.8**
0.25	84	14	6
0.50	193	32	13
0.60	246	40	16
0.70	310	50	20
0.80	393	64	26
0.90	526	85	34
0.95	651	105	42
0.99	920	148	58

Mead's resource equation

Mead's resource equation is often used for estimating sample sizes of laboratory animals, as well as in many other laboratory experiments. It may not be as accurate as using other methods in estimating sample size, but gives a hint of what is the appropriate sample size where parameters such as expected standard deviations or expected differences in values between groups are unknown or very hard to estimate.

All the parameters in the equation are in fact the degrees of freedom of the number of their concepts, and hence, their numbers are subtracted by 1 before insertion into the equation.

The equation is:

$E = N - B - T$,

where:

- N is the total number of individuals or units in the study (minus 1)
- B is the *blocking component*, representing environmental effects allowed for in the design (minus 1)
- T is the *treatment component*, corresponding to the number of treatment groups (including control group) being used, or the number of questions being asked (minus 1)
- E is the degrees of freedom of the *error component*, and should be somewhere between 10 and 20.

For example, if a study using laboratory animals is planned with four treatment groups (T=3), with eight animals per group, making 32 animals total (N=31), without any further stratification (B=0), then E would equal 28, which is above the cutoff of 20, indicating that sample size may be a bit too large, and six animals per group might be more appropriate.

By cumulative distribution function

Let X_i, i = 1, 2, ..., n be independent observations taken from a normal distribution with unknown mean μ and known variance σ^2. Let us consider two hypotheses, a null hypothesis:

H_0:μ = 0

and an alternative hypothesis:

H_a:μ = μ^*

for some 'smallest significant difference' μ^* >0. This is the smallest value for which we care about observing a difference. Now, if we wish to (1) reject H_0 with a probability of at least 1-β when H_a is true (i.e. a power of 1-β), and (2) reject H_0 with probability α when H_0 is true, then we need the following:

If z_α is the upper α percentage point of the standard normal distribution, then

$$\Pr(\bar{x} > z_\alpha\sigma/\sqrt{n}|H_0 \text{ true}) = \alpha$$

and so

'Reject H_0 if our sample average ($\bar{x}$) is more than $z_\alpha\sigma/\sqrt{n}$ '

is a decision rule which satisfies (2). (Note, this is a 1-tailed test)

Now we wish for this to happen with a probability at least 1-β when H_a is true. In this case, our sample average will come from a Normal distribution with mean μ^*. Therefore we require

$$\Pr(\bar{x} > z_\alpha\sigma/\sqrt{n}|H_a \text{ true}) \geq 1 - \beta$$

Through careful manipulation, this can be shown to happen when

$$n \geq \left(\frac{\Phi^{-1}(1-\beta) + z_\alpha}{\mu^*/\sigma}\right)^2$$

where Φ is the normal cumulative distribution function.

Stratified sample size

With more complicated sampling techniques, such as stratified sampling, the sample can often be split up into sub-samples. Typically, if there are *k* such sub-samples (from *k* different strata) then each of them will have a sample size n_i, $i = 1, 2, ..., k$. These n_i must conform to the rule that $n_1 + n_2 + ... + n_k = n$ (i.e. that the total sample size is given by the sum of the sub-sample sizes). Selecting these n_i optimally can be done in various ways, using (for example) Neyman's optimal allocation.

There are many reasons to use stratified sampling: to decrease variances of sample estimates, to use partly non-random methods, or to study strata individually. A useful, partly non-random method would be to sample individuals where easily accessible, but, where not, sample clusters to save travel costs.

In general, for *H* strata, a weighted sample mean is

$$\bar{x}_w = \sum_{h=1}^{H} W_h \bar{x}_h,$$

with

$$\mathrm{Var}(\bar{x}_w) = \sum_{h=1}^{H} W_h^2 \,\mathrm{Var}(\bar{x}_h).$$

The weights, *W*(*h*), frequently, but not always, represent the proportions of the population elements in the strata, and *W*(*h*)=*N*(*h*)/*N*. For a fixed sample size, that is n=Sum{n(h)},

$$\mathrm{Var}(\bar{x}_w) = \sum_{h=1}^{H} W_h^2 \,\mathrm{Var}(h) \left(\frac{1}{n_h} - \frac{1}{N_h}\right),$$

which can be made a minimum if the sampling rate within each stratum is made proportional to the standard deviation within each stratum: $n_h / N_h = kS_h$.

An "optimum allocation" is reached when the sampling rates within the strata are made directly proportional to the standard deviations within the strata and inversely proportional to the square roots of the costs per element within the strata:

$$\frac{n(h)}{N(h)} = \frac{KS(h)}{\sqrt{C(h)}},$$

or, more generally, when

$$n(h) = \frac{K'W(h)S(h)}{\sqrt{C(h)}}.$$

9.11 Confidence Intervals For A Population Mean

A **confidence interval** (**CI**) is a particular kind of interval estimate of a population parameter and is used to indicate the reliability of an estimate. It is an observed interval (i.e it is calculated from the observations), in principle different from sample to sample, that frequently includes the parameter of interest, if the experiment is repeated. How frequently the observed interval contains the parameter is determined by the **confidence level** or **confidence coefficient**.

A confidence interval with a particular confidence level is intended to give the assurance that, if the statistical model is correct, then taken over all the data that *might* have been obtained, the procedure for constructing the interval would deliver a confidence interval that included the true value of the parameter the proportion of the time set by the confidence level. More specifically, the meaning of the term "confidence level" is that, if confidence intervals are constructed across many separate data analyses of repeated (and possibly different) experiments, the proportion of such intervals that contain the true value of the parameter will approximately match the confidence level; this is guaranteed by the reasoning underlying the construction of confidence intervals.

A confidence interval does *not* predict that the true value of the parameter has a particular probability of being in the confidence interval given the data actually obtained. (An interval intended to have such a property, called a credible interval, can be estimated using Bayesian methods; but such methods bring with them their own distinct strengths and weaknesses).

Interval estimates can be contrasted with point estimates. A point estimate is a single value given as the estimate of a population parameter that is of interest, for example the mean of some quantity. An interval estimate specifies instead a range within which the parameter is estimated to lie. Confidence intervals are commonly reported in tables or graphs along with point estimates of the same parameters, to show the reliability of the estimates.

For example, a confidence interval can be used to describe how reliable survey results are. In a poll of election voting-intentions, the result might be that 40% of respondents intend to vote for a certain party. A 90%

confidence interval for the proportion in the whole population having the same intention on the survey date might be 38% to 42%. From the same data one may calculate a 95% confidence interval, which might in this case be 36% to 44%. A major factor determining the length of a confidence interval is the size of the sample used in the estimation procedure, for example the number of people taking part in a survey.

Let X be a random sample from a probability distribution with parameters θ, which is a quantity to be estimated, and φ, representing quantities not of immediate interest. A *confidence interval* for the parameter θ, with confidence level or confidence coefficient γ, is an interval with random endpoints $(u(X),v(X))$, determined by the pair of statistics (i.e., observable random variables) $u(X)$ and $v(X)$, with the property:

$$\gamma = \Pr_{\theta,\phi}(u(X) < \theta < v(X)).$$

The quantities φ in which there is no immediate interest are called nuisance parameters, as statistical theory still needs to find some way to deal with them. The number γ, with typical values close to but not greater than 1, is sometimes given in the form $1 - \alpha$ (or as a percentage $100\%\cdot(1 - \alpha)$), where α is a small nonnegative number, close to 0.

Here $\Pr_{\theta,\varphi}$ is used to indicate the probability when the random variable X has the distribution characterised by (θ, φ). An important part of this specification is that the random interval (U, V) covers the unknown value θ with a high probability no matter what the true value of θ actually is.
Note that here $\Pr_{\theta,\varphi}$ need not refer to an explicitly given parameterised family of distributions, although it often does. Just as the random variable X notionally corresponds to other possible realizations of x from the same population or from the same version of reality, the parameters (θ, φ) indicate that we need to consider other versions of reality in which the distribution of X might have different characteristics.

In a specific situation, when x is the outcome of the sample X, the interval $(u(x),v(x))$ is also referred to as a confidence interval for θ. Note that it is no longer possible to say that the (observed) interval $(u(x),v(x))$ has probability γ to contain the parameter θ. This observed interval is just one realization of all possible intervals for which the probability statement holds.

Intervals for random outcomes

Confidence intervals can be defined for random quantities as well as for fixed quantities as in the above. See prediction interval. For this, consider an additional single-valued random variable Y which may or may not be statistically dependent on X. Then the rule for constructing the interval $(u(x), v(x))$ provides a confidence interval for the as-yet-to-be observed value y of Y if

$$\Pr_{\theta,\phi}(u(X) < Y < v(X)) = 1 - \alpha \text{ for all } (\theta, \phi).$$

Here $\Pr_{\theta,\varphi}$ is used to indicate the probability over the joint distribution of the random variables (*X*, *Y*) when this is characterised by parameters (*θ*, *φ*).

Approximate confidence intervals

For non-standard applications it is sometimes not possible to find rules for constructing confidence intervals that have exactly the required properties. But practically useful intervals can still be found. The coverage probability *c*(*θ*, *φ*) for a random interval is defined by

$$\Pr_{\theta,\phi}(u(X) < \theta < v(X)) = c(\theta, \phi)$$

and the rule for constructing the interval may be accepted as providing a confidence interval if

$$c(\theta, \phi) \approxeq 1 - \alpha \text{ for all } (\theta, \phi)$$

to an acceptable level of approximation.

Comparison to Bayesian interval estimates

A Bayesian interval estimate is called a credible interval. Using much of the same notation as above, the definition of a credible interval for the unknown true value of *θ* is, for a given *α*,

$$\Pr(u(x) < \Theta < v(x) | X = x) = 1 - \alpha.$$

Here Θ is used to emphasize that the unknown value of *θ* is being treated as a random variable. The definitions of the two types of intervals may be compared as follows.

- The definition of a confidence interval involves probabilities calculated from the distribution of *X* for given (*θ*, *φ*) (or conditional on these values) and the condition needs to hold for all values of (*θ*, *φ*).

- The definition of a credible interval involves probabilities calculated from the distribution of Θ conditional on the observed values of *X* = *x* and marginalised (or averaged) over the values of Φ, where this last quantity is the random variable corresponding to the uncertainty about the nuisance parameters in *φ*.

Note that the treatment of the nuisance parameters above is often omitted from discussions comparing confidence and credible intervals but it is markedly different between the two cases.

In some simple standard cases, the intervals produced as confidence and credible intervals from the same data set can be identical. They are very different if informative prior information is included in the Bayesian analysis; and may be very different for some parts of the space of possible data even if the Bayesian prior is relatively uninformative.

Desirable properties

When applying standard statistical procedures, there will often be standard ways of constructing confidence intervals. These will have been devised so as to meet certain desirable properties, which will hold given that the assumptions on which the procedure rely are true. These desirable properties may be described as: validity, optimality and invariance. Of these "validity" is most important, followed closely by "optimality". "Invariance" may be considered as a property of the method of derivation of a confidence interval rather than of the rule for constructing the interval. In non-standard applications, the same desirable properties would be sought.

- *Validity.* This means that the nominal coverage probability (confidence level) of the confidence interval should hold, either exactly or to a good approximation.

- *Optimality.* This means that the rule for constructing the confidence interval should make as much use of the information in the data-set as possible. Recall that one could throw away half of a dataset and still be able to derive a valid confidence interval. One way of assessing optimality is by the length of the interval, so that a rule for constructing a confidence interval is judged better than another if it leads to intervals whose lengths are typically shorter.

- *Invariance.* In many applications the quantity being estimated might not be tightly defined as such. For example, a survey might result in an estimate of the median income in a population, but it might equally be considered as providing an estimate of the logarithm of the median income, given that this is a common scale for presenting graphical results. It would be desirable that the method used for constructing a confidence interval for the median income would give equivalent results when applied to constructing a confidence interval for the logarithm of the median income: specifically the values at the ends of the latter interval would be the logarithms of the values at the ends of former interval.

Methods of derivation

For non-standard applications, there are several routes that might be taken to derive a rule for the construction of confidence intervals. Established rules for standard procedures might be justified or explained via several of these routes. Typically a rule for constructing confidence intervals is closely tied to a particular way of finding a point estimate of the quantity being considered.

Statistics

This is closely related to the method of moments for estimation. A simple example arises where the quantity to be estimated is the mean, in which case a natural estimate is the sample mean. The usual arguments indicate that the sample variance can be used to estimate the variance of the sample mean. A naive confidence interval for the true mean can be constructed centered on the sample mean with a width which is a multiple of the square root of the sample variance.

Likelihood theory

Where estimates are constructed using the maximum likelihood principle, the theory for this provides two ways of constructing confidence intervals or confidence regions for the estimates.

Estimating equations

The estimation approach here can be considered as both a generalization of the method of moments and a generalization of the maximum likelihood approach. There are corresponding generalizations of the results of maximum likelihood theory that allow confidence intervals to be constructed based on estimates derived from estimating equations.[*citation needed*]

Via significance testing

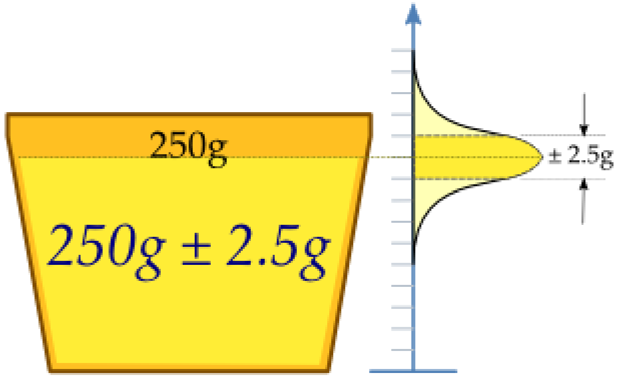
If significance tests are available for general values of a parameter, then confidence intervals/regions can be constructed by including in the 100p% confidence region all those points for which the significance test of the null hypothesis that the true value is the given value is not rejected at a significance level of (1-p).

Bootstrapping

In situations where the distributional assumptions for that above methods are uncertain or violated, resampling methods allow construction of confidence intervals or prediction intervals. The observed data distribution and the internal correlations are used as the surrogate for the correlations in the wider population.

Examples

Practical example



A machine fills cups with margarine, and is supposed to be adjusted so that the content of the cups is 250 g of margarine. As the machine cannot fill every cup with exactly 250 g, the content added to individual cups shows some variation, and is considered a random variable X. This variation is assumed to be normally distributed around the desired average of 250 g, with a standard deviation of 2.5 g. To determine if the machine is adequately calibrated, a sample of *n* = 25 cups of margarine is chosen at random and the cups are weighed. The resulting measured masses of margarine are $X_1, \ldots, X_{25}$, a random sample from *X*.

To get an impression of the expectation *μ*, it is sufficient to give an estimate. The appropriate estimator is the sample mean:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The sample shows actual weights $x_1, \ldots, x_{25}$, with mean:

$$\bar{x} = \frac{1}{25}\sum_{i=1}^{25} x_i = 250.2\,\text{grams}.$$

If we take another sample of 25 cups, we could easily expect to find mass values like 250.4 or 251.1 grams. A sample mean value of 280 grams however would be extremely rare if the mean content of the cups is in fact close to 250 grams. There is a whole interval around the observed value 250.2 grams of the sample mean within which, if the whole population mean actually takes a value in this range, the observed data would not be considered particularly unusual. Such an interval is called a confidence interval for the parameter *μ*. How do we calculate such an interval? The endpoints of the interval have to be calculated from the sample, so they are statistics, functions of the sample $X_1, \ldots, X_{25}$ and hence random variables themselves.

In our case we may determine the endpoints by considering that the sample mean *X* from a normally distributed sample is also normally distributed, with the same expectation *μ*, but with a standard error of:

$$\frac{\sigma}{\sqrt{n}} = 0.5 \text{ grams}$$

By standardizing, we get a random variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{0.5}$$

dependent on the parameter *μ* to be estimated, but with a standard normal distribution independent of the parameter *μ*. Hence it is possible to find numbers −*z* and *z*, independent of *μ*, between which *Z* lies with probability 1 − ɑ, a measure of how confident we want to be. We take 1 − ɑ = 0.95. So we have:

$$P(-z \le Z \le z) = 1 - \alpha = 0.95.$$

The number *z* follows from the cumulative distribution function, in this case the cumulative normal distribution function:

$$\Phi(z) = P(Z \le z) = 1 - \tfrac{\alpha}{2} = 0.975,$$

$$z = \Phi^{-1}(\Phi(z)) = \Phi^{-1}(0.975) = 1.96,$$

and we get:

$$0.95 = 1 - \alpha = P(-z \le Z \le z) = P\left(-1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96\right)$$

$$= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(\bar{X} - 1.96 \times 0.5 \le \mu \le \bar{X} + 1.96 \times 0.5\right)$$

$$= P\left(\bar{X} - 0.98 \le \mu \le \bar{X} + 0.98\right).$$

This might be interpreted as: with probability 0.95 we will find a confidence interval in which we will meet the parameter *μ* between the stochastic endpoints

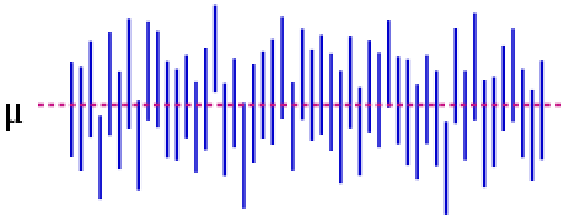
$$\bar{X} - 0.98$$

and

$$\bar{X} + 0.98.$$

This does not mean that there is 0.95 probability of meeting the parameter *μ* in the calculated interval. Every time the measurements are repeated, there will be another value for the mean *X* of the sample. In 95% of the cases *μ* will be between the endpoints calculated from this mean, but in 5% of the cases it will not be. The actual confidence interval is calculated by entering the measured masses in the formula. Our 0.95 confidence interval becomes:

$$(\bar{x} - 0.98; \bar{x} + 0.98) = (250.2 - 0.98; 250.2 + 0.98) = (249.22; 251.18).$$



The vertical line segments represent 50 realizations of a confidence interval for *μ*.

As the desired value 250 of *μ* is within the resulted confidence interval, there is no reason to believe the machine is wrongly calibrated.

The calculated interval has fixed endpoints, where μ might be in between (or not). Thus this event has probability either 0 or 1. One **cannot** say: "with probability (1 − α) the parameter *μ* lies in the confidence interval." One only knows that by repetition in 100(1 − α) % of the cases, *μ* will be in the calculated interval. In 100α % of the cases however it does not. And unfortunately one does not know in which of the cases this happens. That is why one can say: "with **confidence level** 100(1 − α) %, *μ* lies in the confidence interval."

The figure on the right shows 50 realizations of a confidence interval for a given population mean *μ*. If we randomly choose one realization, the probability is 95% we end up having chosen an interval that contains the parameter; however we may be unlucky and have picked the wrong one. We will never know; we are stuck with our interval.

Theoretical example

Suppose X_1, ..., X_n are an independent sample from a normally distributed population with (parameters) mean *μ* and variance σ^2. Let

$$\overline{X} = (X_1 + \cdots + X_n)/n,$$
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

be the well known statistics, sample mean and sample variance. Then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a Student's t-distribution with *n* − 1 degrees of freedom. Note that the distribution of *T* does not depend on the values of the unobservable parameters *μ* and σ^2; i.e., it is a pivotal quantity. Suppose we wanted to calculate a 90% confidence interval for μ. Then, denoting c as the 95th percentile of this distribution,

$$\Pr\left(-c < T < c\right) = 0.9.$$

(Note: "95th" and "0.9" are correct in the preceding expressions. There is a 5% chance that T will be less than $-c$ and a 5% chance that it will be larger than $+c$. Thus, the probability that T will be between $-c$ and $+c$ is 90%.)
Consequently

$$\Pr\left(\overline{X} - \frac{cS}{\sqrt{n}} < \mu < \overline{X} + \frac{cS}{\sqrt{n}}\right) = 0.9$$

and we have a theoretical (stochastic) 90% confidence interval for μ.
After observing the sample we find values x for X and s for S, from which we compute the confidence interval

$$\left[\overline{x} - \frac{cs}{\sqrt{n}}, \overline{x} + \frac{cs}{\sqrt{n}}\right],$$

an interval with fixed numbers as endpoints, of which we can no more say there is a certain probability it contains the parameter μ. Either μ is in this interval or isn't.

Relation to hypothesis testing

While the formulations of the notions of confidence intervals and of statistical hypothesis testing are distinct they are in some senses related and to some extent complementary. While not all confidence intervals are constructed in this way, one general purpose approach to constructing confidence intervals is to define a 100(1 − α)% confidence interval to consist of all those values θ_0 for which a test of the hypothesis $\theta = \theta_0$ is not rejected at a significance level of 100α%. Such an approach may not always be available since it presupposes the practical availability of an appropriate significance test. Naturally, any assumptions required for the significance test would carry over to the confidence intervals.

It may be convenient to make the general correspondence that parameter values within a confidence interval are equivalent to those values that would not be rejected by an hypothesis test, but this would be dangerous. In many instances the confidence intervals that are quoted are only approximately valid, perhaps derived from "plus or minus twice the standard error", and the implications of this for the supposedly corresponding hypothesis tests are usually unknown.

Meaning and interpretation

For users of frequentist methods, various interpretations of a confidence interval can be given.

- The confidence interval can be expressed in terms of samples (or repeated samples): "*Were this procedure to be repeated on multiple* *samples, the calculated confidence interval (which would differ for each sample) would encompass the true population parameter 90% of the time."* Note that this need not be repeated sampling from the same population, just repeated sampling.

- The explanation of a confidence interval can amount to something like: "*The confidence interval represents values for the population parameter for which the difference between the parameter and the observed estimate is not statistically significant at the 10% level*". In fact, this relates to one particular way in which a confidence interval may be constructed.

- The probability associated with a confidence interval may also be considered from a pre-experiment point of view, in the same context in which arguments for the random allocation of treatments to study items are made. Here the experimenter sets out the way in which they intend to calculate a confidence interval and know, before they do the actual experiment, that the interval they will end up calculating has a certain chance of covering the true but unknown value. This is very similar to the "repeated sample" interpretation above, except that it avoids relying on considering hypothetical repeats of a sampling procedure that may not be repeatable in any meaningful sense. See Neyman construction.

In each of the above, the following applies: If the true value of the parameter lies outside the 90% confidence interval once it has been calculated, then an event has occurred which had a probability of 10% (or less) of happening by chance.

Meaning of the term "confidence"

There is a difference in meaning between the common usage of the word "confidence" and its statistical usage, which is often confusing to the layman, and this is one of the critiques of confidence intervals, namely that in application by non-statisticians, the term "confidence" is misleading.

In common usage, a claim to 95% confidence in something is normally taken as indicating virtual certainty. In statistics, a claim to 95% confidence simply means that the researcher has seen something occur that happens only one time in 20 or less. If one were to roll two dice and get double six (which happens 1/36th of the time, or about 3%), few would claim this as proof that the dice were fixed, although statistically speaking one could have 97% confidence that they were. Similarly, the finding of a statistical link at

95% confidence is not proof, nor even very good evidence, that there is any real connection between the things linked.

When a study involves multiple statistical tests, people tend to assume that the confidence associated with individual tests is the confidence one should have in the results of the study itself. In fact, the results of all the statistical tests conducted during a study must be judged as a whole in determining what confidence one may place in the positive links it produces. For example, say a study is conducted which involves 40 statistical tests at 95% confidence, and which produces 3 positive results. Each test has a 5% chance of producing a false positive, so such a study will produce 3 false positives about two times in three. Thus the confidence one can have that any of the study's positive conclusions are correct is only about 32%, well below the 95% the researchers have set as their standard of acceptance.

Alternatives and critiques

Main articles: Interval estimation, Credible interval, and Prediction interval
Confidence intervals are one method of interval estimation, and the most widely used in frequentist statistics. An analogous concept in Bayesian statistics is credible intervals, while an alternative frequentist method is that of prediction intervals which, rather than estimating *parameters,* estimate the outcome of *future* samples. For other approaches to expressing uncertainty using intervals, see interval estimation.

There is disagreement about which of these methods produces the most useful results: the mathematics of the computations are rarely in question – confidence intervals being based on sampling distributions, credible intervals being based on Bayes' theorem – but the application of these methods, the utility and interpretation of the produced statistics, is debated.

Users of Bayesian methods, if they produced an interval estimate, would in contrast to confidence intervals, want to say "*My degree of* belief *that the parameter is in fact in this interval is 90%,*" while users of prediction intervals would instead say "I *predict* that the *next sample* will fall in this interval 90% of the time."

Confidence intervals are an expression of probability and are subject to the normal laws of probability. If several statistics are presented with confidence intervals, each calculated separately on the assumption of independence, that assumption must be honoured or the calculations will be rendered invalid. For example, if a researcher generates a set of statistics with intervals and selects some of them as significant, the act of selecting invalidates the calculations used to generate the intervals.

9.12 Confidence Intervals for a Proportion

An approximate confidence interval for a population mean can be constructed for random variables that are not normally distributed in the population, relying on the central limit theorem, if the sample sizes and counts are big enough. The formulae are identical to the case above (where the sample mean is actually normally distributed about the population mean).

The approximation will be quite good with only a few dozen observations in the sample if the probability distribution of the random variable is not too different from the normal distribution (e.g. its cumulative distribution function does not have any discontinuities and its skewness is moderate).

One type of sample mean is the mean of an indicator variable, which takes on the value 1 for true and the value 0 for false. The mean of such a variable is equal to the proportion that have the variable equal to one (both in the population and in any sample). This is a useful property of indicator variables, especially for hypothesis testing. To apply the central limit theorem, one must use a large enough sample. A rough rule of thumb is that one should see at least 5 cases in which the indicator is 1 and at least 5 in which it is 0. Confidence intervals constructed using the above formulae may include negative numbers or numbers greater than 1, but proportions obviously cannot be negative or exceed 1. Additionally, sample proportions can only take on a finite number of values, so the central limit theorem and the normal distribution are not the best tools for building a confidence interval. See "Binomial proportion confidence interval" for better methods which are specific to this case.

A researcher wants to estimate the proportion of people who report the side effect of nausea when taking a drug to reduce anxiety. Of 25 people who take the drug, 8 report nausea. In this sample, therefore, 0.32 of the patients reported nausea. Most likely the researcher would construct a confidence interval on the population proportion. The procedure for constructing a confidence interval assumes that the sampling distribution of p is normal. Since the sample proportion, p, can be thought of as the mean of N scores, each score being either zero or one, the central limit theorem is applicable. This theorem states that as N increases, the sampling distribution of the mean (p in this case) approaches a normal distribution. But how large an N is big enough? The population proportion, Pi, is another factor that affects the shape of the distribution. The closer Pi is to 0.5, the more normal the sampling distribution.

This applet allows you to explore the validity of confidence intervals on a proportion with various values of N and Pi. After you specify N, Pi, the level of confidence, and the number of simulations you wish to perform, the applet samples data according to your specification and computes a confidence interval for each simulation. The proportion of simulations for which the confidence interval contains Pi is recorded. If the method for constructing confidence intervals is valid, then about 95% of the 95% confidence intervals should contain Pi.

9.13 Sample Size Determination With A Given Margin Of Error Sample Size Needed For Specified Margin Of Error

The general formula for a **confidence interval** is
estimate **plus/minus** (*critical value*)(*standard deviation of the estimate*)
A **95% confidence interval** for proportions has the form p(hat) **plus/minus** 1.96 ÷[**(**(p(hat)**)(**1-p(hat)**)**/N] where N is the sample size and p(hat) is the sample proportion.
Since 1.96 is approximately 2, we will use 2 in what follows to simply computations.
If the population proportion parameter is p, the margin of error, *m*, for a **95% confidence interval** can be calculated using the formula *m* = 2 ÷[p(1-p)/N]

When sampling, p is replaced by p(hat), the sample proportion, to compute *m.*
We now ask the question:

What sample size is needed if one wants a specific margin of error?
Solving the above equation for N yields $m^2/4$ = p(1-p)/N ==> N = 4p(1-p)/m^2.
YIKE! We face a "Catch 22" situation. We want N, and we know *m*, but we don't know a value for p, and we can't get such a value until we actually take a sample.

We get around this dilemma by finding the value of p that will maximize N. Since 4 and m^2 are known constants, we need only maximize y = p(1-p) = p - p^2. This is simply a parabola that opens downward. We need only find the vertex. We can take a derivative and note that dy/dp = 1 - 2p which has value of 0 when p = 1/2. In other words, looking at the equation
N = 4p(1-p)/m^2
we will get the largest possible value of N when we substitute p = 1/2. Note that is the substitution is made, we get N = 4(1/2)(1/2)/m^2 = 1/m^2, a very

simple formula. In other words, if we want a 95% confidence interval and know *m*, margin of error, we can determine the sample size needed for the specified *m*. For instance, if we want a margin of error = 2%, then the sample size required is $1/(.02)^2$ = 2,500.

What is shown in the box below is a published survey related to the Persian Gulf War some years ago.

Would you support or oppose U.S. forces resuming action to force Saddam from power?

54% Support
37% Oppose

For this Newsweek Poll, the Gallop Organization interviewed a national sample of 751 adults by telephone April 4-5. The margin of error is plus or minus 4 percentage points. Some "Don't Know" and other responses not shown

Let's do some computations:

If we were to compute the margin of error using 54%, we would get 2 ÷[(.54)(.46)/751] = 0.0363736. Rounding "out" to the nearest integer percent, we would get the 4% stated in the survey results. If one calculates the margin of error using 37%, one obtains 2 ÷[(.37)(.63)/751] = 0.0352356. Again, if we round "out," we get 4%.

If we wanted a margin of error = 4%, the sample size needed would be $1/(.04)^2$ = 625. A margin of error of 3% would require a sample size = $1/(.03)^2$ = 1,111. What is reported in the survey "jives" with these calculations.

While published surveys such as the one above do not generally talk about a 95% confidence interval, the reported margin of error does relate to such an interval, as has been demonstrated. Using the information provided in the survey above, the 95% confidence interval for those **support** using action to remove Saddam from power is [50%, 58%]. The corresponding 95% confidence interval for those who **oppose** is [33%,41%].

Calculating the Sample Size

The sample size, in this case, refers to the number of children to be included in the survey.

Step 1: Base Sample-size Calculation

The appropriate sample size for a population-based survey is determined largely by three factors:

(i) the estimated prevalence of the variable of interest – chronic malnutrition in this instance,

(ii) the desired level of confidence and
(iii) the acceptable margin of error.

For a survey design based on a simple random sample, the sample size required can be calculated according to the following formula.

Formula:

$$\mathbf{n} = \frac{\mathbf{t^2} \times \mathbf{p(1-p)}}{\mathbf{m^2}}$$

Description:

n = required sample size
t = confidence level at 95% (standard value of 1.96)
p = estimated prevalence of malnutrition in the project area
m = margin of error at 5% (standard value of 0.05)

Example

In the Al Haouz project in Morocco, it has been estimated that roughly 30% (0.3) of the children in the project area suffer from chronic malnutrition. This figure has been taken from national statistics on malnutrition in rural areas. Use of the standard values listed above provides the following calculation.

Calculation:

9.14 Hypothesis Testing

A **statistical hypothesis test** is a method of making decisions using data, whether from a controlled experiment or an observational study (not controlled). In statistics, a result is called statistically significant if it is unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the significance level. The phrase "*test of significance*" was coined by Ronald Fisher: "Critical tests of this kind may be called tests of significance, and when such tests are available we may discover whether a second sample is or is not significantly different from the first."

Hypothesis testing is sometimes called **confirmatory data analysis**, in contrast to exploratory data analysis. In frequency probability, these decisions are almost always made using null-hypothesis tests (i.e., tests that answer the question *Assuming that the null hypothesis is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed?*) One use of hypothesis testing is deciding whether experimental results contain enough information to cast doubt on conventional wisdom.

A result that was found to be statistically significant is also called a **positive result**; conversely, a result that is not unlikely under the null hypothesis is called a **negative result** or a **null result**.

Statistical hypothesis testing is a key technique of frequentist statistical inference. The Bayesian approach to hypothesis testing is to base rejection of the hypothesis on the posterior probability. Other approaches to reaching a decision based on data are available via decision theory and optimal decisions.

The following examples should solidify these ideas.

Example 1 – Courtroom trial

A statistical test procedure is comparable to a criminal trial; a defendant is considered not guilty as long as his guilt is not proven. The prosecutor tries to prove the guilt of the defendant. Only when there is enough charging evidence the defendant is convicted.

In the start of the procedure, there are two hypotheses H_0: "the defendant is not guilty", and H_1: "the defendant is guilty". The first one is called *null hypothesis*, and is for the time being accepted. The second one is called *alternative (hypothesis)*. It is the hypothesis one tries to prove.

The hypothesis of innocence is only rejected when an error is very unlikely, because one doesn't want to convict an innocent defendant. Such an error is called *error of the first kind* (i.e. the conviction of an innocent person), and the occurrence of this error is controlled to be rare. As a consequence of this asymmetric behaviour, the *error of the second kind* (acquitting a person who committed the crime), is often rather large.

	Null Hypothesis (H_0) is true **He truly is not guilty**	**Alternative Hypothesis (H_1) is true** **He truly is guilty**
Accept Null Hypothesis Acquittal	Right decision	Wrong decision Type II Error
Reject Null Hypothesis Conviction	Wrong decision Type I Error	Right decision

A criminal trial can be regarded as either or both of two decision processes: guilty vs not guilty or evidence vs a threshold ("beyond a reasonable doubt"). In one view, the defendant is judged; in the other view the performance of the prosecution (which bears the burden of proof) is judged. A hypothesis test can be regarded as either a judgment of a hypothesis or as a judgment of evidence.

Example 2 – Clairvoyant card game
A person (the subject) is tested for clairvoyance. He is shown the reverse of a randomly chosen playing card 25 times and asked which of the four suits it belongs to. The number of hits, or correct answers, is called *X*.

As we try to find evidence of his clairvoyance, for the time being the null hypothesis is that the person is not clairvoyant. The alternative is, of course: the person is (more or less) clairvoyant.

If the null hypothesis is valid, the only thing the test person can do is guess. For every card, the probability (relative frequency) of any single suit appearing is 1/4. If the alternative is valid, the test subject will predict the suit correctly with probability greater than 1/4. We will call the probability of guessing correctly *p*. The hypotheses, then, are:

- null hypothesis : $H_0 : p = \frac{1}{4}$ (just guessing) and
- alternative hypothesis :$H_1 : p > \frac{1}{4}$ (true clairvoyant).

When the test subject correctly predicts all 25 cards, we will consider him clairvoyant, and reject the null hypothesis. Thus also with 24 or 23 hits. With only 5 or 6 hits, on the other hand, there is no cause to consider him so. But what about 12 hits, or 17 hits? What is the critical number, *c*, of hits, at which point we consider the subject to be clairvoyant? How do we determine the critical value *c*? It is obvious that with the choice *c*=25 (i.e. we only accept clairvoyance when all cards are predicted correctly) we're more critical than with *c*=10. In the first case almost no test subjects will be recognized to be clairvoyant, in the second case, certain number will pass the test. In practice, one decides how critical one will be. That is, one decides how often one accepts an error of the first kind – a false positive, or Type I error. With *c* = 25 the probability of such an error is:

$$P(\text{reject } H_0|H_0 \text{ is valid}) = P(X \geq 25|p = \tfrac{1}{4}) = \left(\tfrac{1}{4}\right)^{25} \approx 10^{-15},$$

and hence, very small. The probability of a false positive is the probability of randomly guessing correctly all 25 times.

Being less critical, with *c*=10, gives:

$$P(\text{reject } H_0|H_0 \text{ is valid}) = P(X \geq 10|p = \tfrac{1}{4}) \approx 0.07.$$

Thus, *c* = 10 yields a much greater probability of false positive.
Before the test is actually performed, the desired probability of a Type I error is determined. Typically, values in the range of 1% to 5% are selected.

Depending on this desired Type 1 error rate, the critical value *c* is calculated. For example, if we select an error rate of 1%, *c* is calculated thus:

$$P(\text{reject } H_0 | H_0 \text{ is valid}) = P(X \geq c | p = \tfrac{1}{4}) \leq 0.01.$$

From all the numbers c, with this property, we choose the smallest, in order to minimize the probability of a Type II error, a false negative. For the above example, we select: *c* = 12.

But what if the subject did not guess any cards at all? Having zero correct answers is clearly an oddity too. The probability of guessing incorrectly once is equal to *p*' = (1 − *p*) = 3/4. Using the same approach we can calculate that probability of randomly calling all 25 cards wrong is:

$$P(\text{reject } H_0 | H_0 \text{ is valid}) = P(X \geq 25 | p' = \tfrac{3}{4}) = \left(\tfrac{3}{4}\right)^{25} \approx 0.00075.$$

This is highly unlikely (less than 1 in a 1000 chance). While the subject can't guess the cards correctly, dismissing H_0 in favour of H_1 would be an error. In fact, the result would suggest a trait on the subject's part of avoiding calling the correct card. A test of this could be formulated: for a selected 1% error rate the subject would have to answer correctly at least twice, for us to believe that card calling is based purely on guessing.

Example 3 – Radioactive suitcase

As an example, consider determining whether a suitcase contains some radioactive material. Placed under a Geiger counter, it produces 10 counts per minute. The null hypothesis is that no radioactive material is in the suitcase and that all measured counts are due to ambient radioactivity typical of the surrounding air and harmless objects. We can then calculate how likely it is that we would observe 10 counts per minute if the null hypothesis were true. If the null hypothesis predicts (say) on average 9 counts per minute and a standard deviation of 1 count per minute, then we say that the suitcase is compatible with the null hypothesis (this does not guarantee that there is no radioactive material, just that we don't have enough evidence to suggest there is). On the other hand, if the null hypothesis predicts 3 counts per minute and a standard deviation of 1 count per minute, then the suitcase is not compatible with the null hypothesis, and there are likely other factors responsible to produce the measurements.

The test described here is more fully the null-hypothesis statistical significance test. The null hypothesis represents what we would believe by default, before seeing any evidence. Statistical significance is a possible finding of the test, declared when the observed sample is unlikely to have occurred by chance if the null hypothesis were true. The name of the test describes its formulation and its possible outcome. One characteristic of the test is its crisp decision: to reject or not reject the null hypothesis. A calculated value is compared to a threshold, which is determined from the tolerable risk of error.

Example 4 – Lady tasting tea

The following example is summarized from Fisher, and is known as the *Lady tasting tea* example. Fisher thoroughly explained his method in a proposed experiment to test a Lady's claimed ability to determine the means of tea preparation by taste. The article is less than 10 pages in length and is notable for its simplicity and completeness regarding terminology, calculations and design of the experiment.

The example is loosely based on an event in Fisher's life. The Lady proved him wrong.

- The experiment provided the Lady with 8 cups of tea at one time, 4 prepared with each method, presented in random order. She was to select the 4 cups prepared by one method.
    - This offered the Lady the advantage of judging cups by comparison.
    - The Lady was fully informed of the experimental method.
- The null hypothesis was that the Lady had no such ability.
- The test statistic was a simple count of the number of successes in selecting the 4 cups.
- The null hypothesis distribution was computed by the number of permutations. The number of selected permutations equaled the number of unselected permutations.

Tea-Tasting Distribution		
Success count	**Permutations of selection**	**Number of permutations**
0	oooo	1 × 1 = 1
1	ooox, ooxo, oxoo, xooo	4 × 4 = 16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = 36
3	oxxx, xoxx, xxox, xxxo	4 × 4 = 16
4	xxxx	1 × 1 = 1
	Total	70

- The critical region was the single case of 4 successes of 4 possible based on a conventional probability criterion (< 5%; 1 of 70 ≈ 1.4%).
- Fisher asserted that no alternative hypothesis was (ever) required.

If and only if the Lady properly categorized all 8 cups was Fisher willing to reject the null hypothesis – effectively acknowledging the Lady's ability with > 98% confidence (but without quantifying her ability). Fisher later discussed the benefits of more trials and repeated tests.

The testing process

In the statistical literature, statistical hypothesis testing plays a fundamental role.[6] The usual line of reasoning is as follows:

1. We start with a research hypothesis of which the truth is unknown.

- The first step is to state the relevant **null and alternative hypotheses**. This is important as mis-stating the hypotheses will muddy the rest of the process. Specifically, the null hypothesis allows to attach an attribute: it should be chosen in such a way that it allows us to conclude whether the alternative hypothesis can either be accepted or stays undecided as it was before the test.

- The second step is to consider the statistical assumptions being made about the sample in doing the test; for example, assumptions about the statistical independence or about the form of the distributions of the observations. This is equally important as invalid assumptions will mean that the results of the test are invalid.

- Decide which test is appropriate, and stating the relevant **test statistic** *T*.

- Derive the distribution of the test statistic under the null hypothesis from the assumptions. In standard cases this will be a well-known result. For example the test statistics may follow a Student's t distribution or a normal distribution.

- The distribution of the test statistic partitions the possible values of *T* into those for which the null-hypothesis is rejected, the so called critical region, and those for which it is not.

- Compute from the observations the observed value t_{obs} of the test statistic *T*.

- Decide to either **fail to reject** the null hypothesis or **reject** it in favor of the alternative. The decision rule is to reject the null hypothesis H_0 if the observed value t_{obs} is in the critical region, and to accept or "fail to reject" the hypothesis otherwise.

It is important to note the philosophical difference between accepting the null hypothesis and simply failing to reject it. The "fail to reject" terminology highlights the fact that the null hypothesis is assumed to be true from the start of the test; if there is a lack of evidence against it, it simply continues to be assumed true. The phrase "accept the null hypothesis" may suggest it has been proved simply because it has not been disproved, a logical fallacy known as the argument from ignorance. Unless a test with particularly high power is used, the idea of "accepting" the null hypothesis may be dangerous. Nonetheless the terminology is prevalent throughout statistics, where its meaning is well understood.

Alternatively, if the testing procedure forces us to reject the null hypothesis (H-null), we can accept the alternative hypothesis (H-alt) and we conclude that the research hypothesis is supported by the data. This fact expresses that our procedure is based on probabilistic considerations in the sense we accept that using another set could lead us to a different conclusion.

Definition of terms

The following definitions are mainly based on the exposition in the book by Lehmann and Romano:

Statistical hypothesis

A statement about the parameters describing a population (not a sample).

Statistic

A value calculated from a sample, often to summarize the sample for comparison purposes.

Simple hypothesis

Any hypothesis which specifies the population distribution completely.

Composite hypothesis

Any hypothesis which does *not* specify the population distribution completely.

Null hypothesis

A simple hypothesis associated with a contradiction to a theory one would like to prove.

Alternate hypothesis

A hypothesis (often composite) associated with a theory one would like to prove.

Statistical test

A decision function that takes its values in the set of hypotheses.

Region of acceptance

The set of values for which we fail to reject the null hypothesis.

Region of rejection / Critical region

The set of values of the test statistic for which the null hypothesis is rejected.

Power of a test ($1 - \beta$)

The test's probability of correctly rejecting the null hypothesis. The complement of the false negative rate, β.

Size / Significance level of a test (α)

For simple hypotheses, this is the test's probability of *incorrectly* rejecting the null hypothesis. The false positive rate. For composite hypotheses this is the upper bound of the probability of rejecting the null hypothesis over all cases covered by the null hypothesis.

p-value

The probability, assuming the null hypothesis is true, of observing a result at least as extreme as the test statistic.

Statistical significance test

A predecessor to the statistical hypothesis test. An experimental result was said to be statistically significant if a sample was sufficiently inconsistent with the (null) hypothesis. This was variously considered common sense, a pragmatic heuristic for identifying meaningful experimental results, a convention establishing a threshold of statistical evidence or a method for drawing conclusions from data. The statistical hypothesis test added mathematical rigor and philosophical consistency to the concept by making the alternative hypothesis explicit. The term is loosely used to describe the modern version which is now part of statistical hypothesis testing.

A statistical hypothesis test compares a test statistic (z or t for examples) to a threshold. The test statistic (the formula found in the table below) is based on optimality. For a fixed level of Type I error rate, use of these statistics minimizes Type II error rates (equivalent to maximizing power). The following terms describe tests in terms of such optimality:

Most powerful test

For a given *size* or *significance level*, the test with the greatest power.

Uniformly most powerful test (UMP)

A test with the greatest *power* for all values of the parameter being tested.

Consistent test

When considering the properties of a test as the sample size grows, a test is said to be consistent if, for a fixed size of test, the power against any fixed alternative approaches 1 in the limit.

Unbiased test

For a specific alternative hypothesis, a test is said to be **unbiased** when the probability of rejecting the null hypothesis is not less than the significance level when the alternative is true *and* is less than or equal to the significance level when the null hypothesis is true.

Conservative test

A test is conservative if, when constructed for a given nominal significance level, the true probability of *incorrectly* rejecting the null hypothesis is never greater than the nominal level.

Uniformly most powerful unbiased (UMPU)

A test which is UMP in the set of all unbiased tests.

Interpretation

The direct interpretation is that if the p-value is less than the required significance level, then we say the null hypothesis is rejected at the given level of significance. Criticism on this interpretation can be found in the corresponding section.

Common test statistics

In the table below, the symbols used are defined at the bottom of the table. Many other tests can be found in other articles.

Name	**Formula**	**Assumptions or notes**
One-sample z-test	$z=\frac{\overline{x}-\mu_0}{(\ s/\sqrt{n})}$	(Normal population **or** *n* > 30) **and** σ known. (*z* is the distance from the mean in relation to the standard deviation of the mean). For non-normal distributions it is possible to calculate a minimum proportion of a population that

		falls within *k* standard deviations for any *k* :
Two-sample z-test	$z = \dfrac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}}}$	Normal population **and** independent observations **and** σ_1 and σ_2 are known
Two-sample pooled t-test, equal variances*	$t = \dfrac{(\overline{x}_1 - \overline{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$ $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$ $df = n_1 + n_2 - 2$	(Normal populations **or** $n_1 + n_2 > 40$) **and** independent observations **and** $\sigma_1 = \sigma_2$ **and** σ_1 and σ_2 unknown
Two-sample unpooled t-test, unequal variances*	$t = \dfrac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n1} + \frac{s_2^2}{n2}}},$ $df = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$	(Normal populations **or** $n_1 + n_2 > 40$) **and** independent observations **and** $\sigma_1 \neq \sigma_2$ **and** σ_1 and σ_2 unknown
One-proportion z-test	$z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$n \cdot p_0 > 10$ **and** $n\ (1 - p_0) > 10$ **and** it is a SRS (Simple Random Sample), see notes.
Two-proportion z-test, pooled for $d_0 = 0$	$z = \dfrac{(\hat{p}_1 - \hat{p}_2) - d_0}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$ $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$	$n_1\ p_1 > 5$ **and** $n_1(1 - p_1) > 5$ **and** $n_2\ p_2 > 5$ **and** $n_2(1 - p_2) > 5$ **and** independent observations, see notes.
Two-proportion z-test, unpooled for $\lvert\ d_0\ \rvert > 0$	$z = \dfrac{(\hat{p}_1 - \hat{p}_2) - d_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n2}}}$	$n_1\ p_1 > 5$ **and** $n_1(1 - p_1) > 5$ **and** $n_2\ p_2 > 5$ **and** $n_2(1 - p_2) > 5$ **and** independent observations, see notes.
One-sample chi-square test	$\chi^2 = \dfrac{(n - 1)s^2}{\sigma_0^2}$	One of the following • All expected counts are at least 5 • All expected counts are > 1 and no more than 20% of expected counts are less than 5

*Two-sample F test for equality of variances	$F = \frac{s_1^2}{s_2^2}$	Arrange so s_1^2> $s_2^2$and reject H_0 for $F > F(\alpha / 2, n_1 - 1, n_2 - 1)$

In general, the subscript 0 indicates a value taken from the null hypothesis, H_0, which should be used as much as possible in constructing its test statistic. *... Definitions of other symbols:*

- α, the probability of Type I error (rejecting a null hypothesis when it is in fact true)
- n = sample size
- n_1 = sample 1 size
- n_2 = sample 2 size
- $\overline{x}$= sample mean
- μ_0 = hypothesized population mean
- μ_1 = population 1 mean
- μ_2 = population 2 mean
- σ = population standard deviation
- σ^2 = population variance

- s = sample standard deviation
- s^2 = sample variance
- s_1 = sample 1 standard deviation
- s_2 = sample 2 standard deviation
- t = t statistic
- df = degrees of freedom
- $\overline{d}$= sample mean of differences
- d_0 = hypothesized population mean difference
- s_d = standard deviation of differences

- $\hat{p}$= x/n = sample proportion, unless specified otherwise
- p_0 = hypothesized population proportion
- p_1 = proportion 1
- p_2 = proportion 2
- d_p = hypothesized difference in proportion
- $\min\{n_1, n_2\}$ = minimum of n_1 and n_2
- $x_1 = n_1 p_1$
- $x_2 = n_2 p_2$
- χ^2 = Chi-squared statistic
- F = F statistic

Origins

Hypothesis testing is largely the product of Ronald Fisher, Jerzy Neyman, Karl Pearson and (son) Egon Pearson. Fisher was an agricultural statistician who emphasized rigorous experimental design and methods to extract a result from few samples assuming Gaussian distributions. Neyman (who teamed with the younger Pearson) emphasized mathematical rigor and methods to obtain more results from many samples and a wider range of distributions. Modern hypothesis testing is an (extended) hybrid of the

Fisher vs Neyman/Pearson formulation, methods and terminology developed in the early 20th century.

Importance

Statistical hypothesis testing plays an important role in the whole of statistics and in statistical inference. For example, Lehmann (1992) in a review of the fundamental paper by Neyman and Pearson (1933) says: "Nevertheless, despite their shortcomings, the new paradigm formulated in the 1933 paper, and the many developments carried out within its framework continue to play a central role in both the theory and practice of statistics and can be expected to do so in the foreseeable future".

Significance testing has been the favored statistical tool in some experimental social sciences (over 90% of articles in the Journal of Applied Psychology during the early 1990s). Other fields have favored the estimation of parameters. Editors often consider significance as a criterion for the publication of scientific conclusions based on experiments with statistical results.

Controversy

Since significance tests were first popularized many objections have been voiced by prominent and respected statisticians. The volume of criticism and rebuttal has filled books with language seldom used in the scholarly debate of a dry subject. Much of the criticism was published more than 40 years ago. The fires of controversy have burned hottest in the field of experimental psychology. Nickerson surveyed the issues in the year 2000. He included 300 references and reported 20 criticisms and almost as many recommendations, alternatives and supplements. The following section greatly condenses Nickerson's discussion, omitting many issues.

Selected criticisms

- There are numerous persistent misconceptions regarding the test and its results.
- The test is a flawed application of probability theory.
- While the data can be unlikely given the null hypothesis, the alternative hypothesis can be even more unlikely. (Nobody can be that lucky. vs. Clairvoyance is impossible.)
- The test result is a function of sample size.
- The test result is uninformative.
- Statistical significance does not imply practical significance.
- Using statistical significance as a criterion for publication results in problems collectively known as publication bias.
- Published Type I errors are difficult to correct.
- Published effect sizes are biased upward.

- Meta-studies are biased by the invisibility of tests which failed to reach significance.
- Type II errors (false negatives) are common.

Each criticism has merit, but is subject to discussion.

Misuses and abuses

The characteristics of significance tests can be abused. When the test statistic is close to the chosen significance level, the temptation to carefully treat outliers, to adjust the chosen significance level, to pick a better statistic or to replace a two-tailed test with a one-tailed test can be powerful. If the goal is to produce a significant experimental result:

- Conduct a few tests with a large sample size.
- Rigorously control the experimental design.
- Publish the successful tests; Hide the unsuccessful tests.
- Emphasize the statistical significance of the results if the practical significance is doubtful.

If the goal is to fail to produce a significant effect:

- Conduct a large number of tests with inadequate sample size.
- Minimize experimental design constraints.
- Publish the number of tests conducted that show "no significant result".

Results of the controversy

The controversy has produced several results. The American Psychological Association has strengthened its statistical reporting requirements after review, medical journal publishers have recognized the obligation to publish some results that are not statistically significant to combat publication bias and a journal has been created to publish such results exclusively. Textbooks have added some cautions and increased coverage of the tools necessary to estimate the size of the sample required to produce significant results. Major organizations have not abandoned use of significance tests although they have discussed doing so.

Alternatives to significance testing

The numerous criticisms of significance testing do not lead to a single alternative or even to a unified set of alternatives. A unifying position of critics is that statistics should not lead to a conclusion or a decision but to a probability or to an estimated value with confidence bounds. The Bayesian statistical philosophy is therefore congenial to critics who believe that an experiment should simply alter probabilities and that conclusions should only be reached on the basis of numerous experiments.

One strong critic of significance testing suggested a list of reporting alternatives: effect sizes for importance, prediction intervals for confidence, replications and extensions for replicability, meta-analyses for generality. None of these suggested alternatives produces a conclusion/decision. Lehmann said that hypothesis testing theory can be presented in terms of conclusions/decisions, probabilities, or confidence intervals. "The distinction between the ... approaches is largely one of reporting and interpretation."

On one "alternative" there is no disagreement: Fisher himself said, "In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result." Cohen, an influential critic of significance testing, concurred, "...don't look for a magic alternative to NHST... It doesn't exist." "...given the problems of statistical induction, we must finally rely, as have the older sciences, on replication." The "alternative" to significance testing is repeated testing. The easiest way to decrease statistical uncertainty is by more data, whether by increased sample size or by repeated tests. Nickerson claimed to have never seen the publication of a literally replicated experiment in psychology.

While Bayesian inference is a possible alternative to significance testing, it requires information that is seldom available in the cases where significance testing is most heavily used.

Future of the controversy

It is unlikely that this controversy will be resolved in the near future. The flaws and unpopularity of significance testing do not eliminate the need for an objective and transparent means of reaching conclusions regarding experiments that produce statistical results. Critics have not unified around an alternative. Other forms of reporting confidence or uncertainty will probably grow in popularity.

Improvements

Jones and Tukey suggested a modest improvement in the original null-hypothesis formulation to formalize handling of one-tail tests. They conclude that, in the "Lady Tasting Tea" example, Fisher ignored the 8-failure case (equally improbable as the 8-success case) in the example test involving tea, which altered the claimed significance by a factor of 2

9.15 Hypothesis Testing For A Population Mean Or Proportion

Step 1: Set up the null Hypothesis and alternative hypothesis based on the context of the problem
Step 2: Set up the rejection region based on

1. the alternative hypothesis
2. given level
3. sample size *n*.

Review Questions

Unit 10
Measures of Relationship
10.1 Correlation Analysis

Correlation

When the value of one variable is related to the value of another, they are said to be correlated. There are three types of correlation:

(i) perfectly correlated
(ii) (ii) partially correlated
(iii) (iii) uncorrelated

Coefficient of Correlation (r) measures such a relationship, and it is given by

$$r = \sqrt{\frac{\sum\left(\hat{Y}-\overline{Y}\right)^2}{\sum\left(Y-\overline{Y}\right)^2}} = \frac{n\sum XY - \sum X\sum Y}{\sqrt{n\sum X^2-\left(\sum X\right)^2}\times\sqrt{n\sum Y^2-\left(\sum Y\right)^2}}$$

Note:

- The value of r ranges from -1 (perfectly correlated in the negative direction) to +1 (perfectly correlated in the positive direction)
- When r = 0, the two variables are not correlated

10.2 Coefficient of Determination

This calculates the proportion of the variation in the actual values which can be predicted by changes in the values of the independent variable. It is denoted by r^2 and the square of the coefficient of correlation is given by:

$$r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\sum\left(\hat{Y}-\overline{Y}\right)^2}{\sum\left(Y-\overline{Y}\right)^2}$$

Note:

- r^2 ranges from 0 to 1 (r ranges from -1 to +1)
- expressed as a percentage, it represents the proportion that can be predicted by the regression line
- the value 1 - r^2 is therefore the proportion contributed by other factors

Standard Error of Estimate (SEE)

It is a measure of the **variability** of the regression line, i.e. the **dispersion** around the regression line. It tells how much variation there is in the dependent variable between the raw value and the expected value in the regression :

$$s_{yx} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

The SEE allows us to generate the confidence interval on the regression line.

- Expressing correlation
- Calculation of Spearman correlation
- Calculation of Pearson correlation
- Interpretation of correlation coefficients
- The coefficient of Determination (R^2)

Regression and Prediction

- Importance of linear Regression, scatter diagram
- Simple Linear regression using the method of Ordinary Least Squares
- Interpretation of results

Unit 10: Examples

Unit 11: Examination sample questions