

# PROBABILITY & STATISTICS

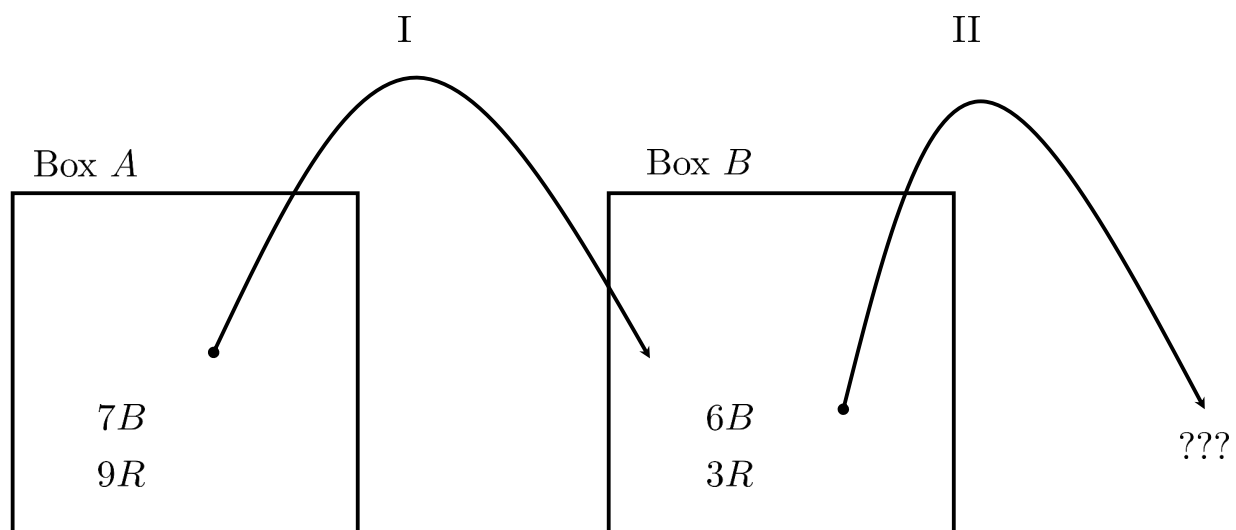
## MTH 1203

ADVENT SEMESTER

BSIT 1:2

BSCS 1:2

SEPTEMBER 2025



$$\begin{aligned}P(R_2) &= P(R_1 \cap R_2) + P(B_1 \cap R_2) \\&= P(R_2 | R_1) \cdot P(R_1) + P(R_2 | B_1) \cdot P(B_1) \\&= \frac{4}{10} \cdot \frac{9}{16} + \frac{3}{10} \cdot \frac{7}{16} \\&= \frac{57}{160}\end{aligned}$$

D.W. Ddumba, C. Muganga, P. Musisi & C. Namanya  
Department of Computing and Technology  
Uganda Christian University

# Contents

<b>1</b>	<b>Basics of Statistics</b>	<b>1</b>
1.1	Introduction . . . . .	1
1.1.1	The Role of Quantitative Methods in Business and Management . . . . .	1
1.1.2	Statistics . . . . .	2
1.1.3	Aim for the Chapter . . . . .	2
1.2	Descriptive and Inferential Statistics . . . . .	3
1.2.1	Descriptive Statistics . . . . .	4
1.2.2	Inferential Statistics . . . . .	5
1.3	Forms of Statistical Data: . . . . .	7
1.3.1	Primary Data . . . . .	7
1.3.2	Secondary Data . . . . .	7
1.4	Methods and Sources of Primary Data Collection . . . . .	8
1.4.1	Direct Observation . . . . .	8
1.4.2	Personal Interview . . . . .	8
1.4.3	Questionnaire Method . . . . .	9
1.5	Cross-Section and Time-Series Data . . . . .	9
1.6	Statistics in Research . . . . .	10
1.6.1	Experimental Method . . . . .	10
1.6.2	Quasi-Experimental Method . . . . .	12
1.6.3	Correlation Method . . . . .	12
1.7	Scales of Measurement . . . . .	13
1.7.1	Norminal Scales . . . . .	13
1.7.2	Ordinal Scales . . . . .	13
1.7.3	Interval Scales . . . . .	14
1.7.4	Ratio Scales . . . . .	15
1.8	Types of Statistical Data . . . . .	17
1.8.1	Continuous and Discrete Variables . . . . .	17
1.8.2	Quantitative and Qualitative Variable . . . . .	17
1.8.3	Recording and Sorting Data . . . . .	20
1.9	Methods of Data Presentation . . . . .	21
1.9.1	Tabular Representation . . . . .	21
1.9.2	Diagrammatic (Pictorial) Representation of Data . . . . .	22
1.9.3	Graphical Representation . . . . .	25
1.10	Frequency Distributions . . . . .	36
1.10.1	Graphical Presentation of Frequency Distribution . . . . .	40
1.11	Chapter Examples . . . . .	45

<b>2</b>	<b>Descriptive Statistics</b>	<b>71</b>
2.1	Measures of Central Tendency . . . . .	71
2.1.1	The Mean . . . . .	71
2.1.2	The Median . . . . .	77
2.1.3	The Mode . . . . .	79
2.2	Measures of Position . . . . .	81
2.2.1	Quartiles . . . . .	81
2.3	Measures of Variation . . . . .	83
2.3.1	The Range . . . . .	83
2.3.2	The Variance, Standard Deviation and Mean Deviation of ungrouped data	84
2.3.3	Variance, Standard Deviation and Mean Deviation of Grouped Data . . .	86
2.4	Other Statistical Measures . . . . .	94
2.4.1	Quartile Deviation . . . . .	94
2.4.2	Percentiles . . . . .	97
2.4.3	Deciles . . . . .	100
2.5	Measures of Shape . . . . .	105
2.5.1	Skewness . . . . .	105
2.5.2	Kurtosis . . . . .	113
2.6	Moments . . . . .	119
2.6.1	Central Moments . . . . .	119
2.6.2	Raw Moments . . . . .	123
2.7	The Box And Whisker Plot . . . . .	125
2.7.1	Finding The Five-number Summary . . . . .	125
2.7.2	Interpreting Quartiles . . . . .	127
<b>3</b>	<b>Correlation &amp; Regression</b>	<b>131</b>
3.1	Introduction . . . . .	131
3.2	Scatter Diagrams . . . . .	131
3.2.1	Direct Correlation (Positive correlation) . . . . .	132
3.2.2	Inverse Correlation (Negative Correlation) . . . . .	133
3.2.3	Absence of Correlation . . . . .	134
3.3	The Line of Best Fit . . . . .	135
3.4	Regression Analysis . . . . .	141
3.4.1	Limitations of Scatter diagram . . . . .	141
3.4.2	Way Forward . . . . .	141
3.4.3	Regression of $y$ on $x$ . . . . .	142
3.4.4	The Regression of $x$ on $y$ . . . . .	145
3.5	Correlation Analysis . . . . .	148
3.6	Correlation Coefficients . . . . .	149
3.6.1	Pearson Product - Moment Correlation Coefficient . . . . .	149
3.7	Correlation by Ranks . . . . .	153
3.7.1	Spearman Rank Correlation Coefficient . . . . .	153
3.7.2	Kendall's Rank Correlation Coefficient . . . . .	157
3.8	Interpretation of the Coefficient . . . . .	162
3.8.1	Covariance . . . . .	162

<b>4</b>	<b>Probability Spaces</b>	<b>163</b>
4.1	Sample Space, Events and Experiment . . . . .	164
4.2	Operations with Events . . . . .	165
4.2.1	Intersection of Events . . . . .	165
4.2.2	Union of Events . . . . .	165
4.2.3	Complementary Events . . . . .	165
4.2.4	Null Event . . . . .	165
4.2.5	Disjoint Events . . . . .	166
4.2.6	Subsets . . . . .	166
4.2.7	Other Resulting Operations of Events . . . . .	166
4.3	Probabilities of Events . . . . .	168
4.4	Axioms of Probability . . . . .	169
4.5	Theorems of Probability . . . . .	170
4.5.1	Other Related Theorems of Probability . . . . .	170
4.6	Contingency Table . . . . .	171
4.7	De Morgan's Laws . . . . .	172
4.8	Special Events . . . . .	176
4.8.1	Mutually Exclusive Events . . . . .	176
4.8.2	Conditional Events . . . . .	178
4.8.3	Multiplicative Rule . . . . .	178
4.8.4	Independent events . . . . .	181
4.9	Total Probability Theorem . . . . .	188
4.10	Bayes' Theorem . . . . .	188
4.11	Chapter Examples . . . . .	195
<b>5</b>	<b>Random Variables</b>	<b>215</b>
5.1	Random Variables & Sample Spaces . . . . .	215
5.2	Discrete Random Variables . . . . .	219
5.2.1	Discrete Probability Distributions . . . . .	219
5.2.2	Cumulative Distribution Function (CDF) . . . . .	223
5.2.3	Expectation of a Discrete Random Variable X . . . . .	226
5.2.4	Variance of a Discrete Random Variable . . . . .	231
5.3	Discrete Random Variables Chapter Examples . . . . .	236
5.4	Continuous Random Variables . . . . .	242
5.4.1	Properties of Continuous Random Variables . . . . .	242
5.4.2	The Cumulative Distribution Function . . . . .	246
5.4.3	Expectation of a Continuous Random Variable . . . . .	250
5.4.4	Variance of a Continuous Random Variable . . . . .	250
5.4.5	Median of A continuous Random Variable . . . . .	250
5.5	Continuous Random Variables Chapter Examples . . . . .	253
5.6	Probability Distribution Functions . . . . .	255
<b>6</b>	<b>Common Discrete Distributions</b>	<b>256</b>
6.1	The Binomial Distribution . . . . .	256
6.1.1	Binomial Experiment . . . . .	256
6.1.2	Binomial Distribution . . . . .	257
6.1.3	Reading Binomial Tables . . . . .	259
6.1.4	The Mean (Expectation) of a Binomial Distribution . . . . .	265
6.1.5	The Variance of Binomial Distribution . . . . .	266
6.2	The Poisson Distribution . . . . .	273

6.2.1	The Poisson Experiment and Distribution . . . . .	273
6.2.2	Poisson Probability Distribution . . . . .	273
6.2.3	Examples of Poisson Distribution . . . . .	274
6.2.4	Expectation of Poisson Distribution . . . . .	278
6.2.5	Variance of Poisson Distribution . . . . .	279
<b>7</b>	<b>Probability Tables</b>	<b>281</b>
7.1	Cumulative Binomial Probabilities $P(X \leq c)$ Table . . . . .	282
7.2	Poisson Distribution Table . . . . .	289
7.3	Normal Distribution Table . . . . .	291
7.3.1	Negative $Z$ -values Table . . . . .	291
7.3.2	Positive $Z$ -values Table . . . . .	292
7.4	Student's $t$ Distribution Table . . . . .	293
<b>8</b>	<b>Probability and Statistics By Python</b>	<b>294</b>

# Chapter 1

## Basics of Statistics

### 1.1 Introduction

#### 1.1.1 The Role of Quantitative Methods in Business and Management

Quantitative methods play an important role both in business research and in the practical solution of business problems. Managers have to take decisions on a wide range of issues, such as:

- 1.) how much to produce
- 2.) what prices to charge
- 3.) how many staff to employ
- 4.) whether to invest in new capital equipment
- 5.) whether to fund a new marketing initiative
- 6.) whether to introduce a new range of products
- 7.) whether to employ an innovative method of production.

In all of these cases, it is clearly highly desirable to be able to compute the likely effects of the decisions on the company's costs, revenues and, most importantly, profits.

Similarly, it is important in business research to be able to use data from samples to estimate parameters relating to the population as a whole (for example, to predict the effect of introducing a new product on sales throughout the country from a survey conducted in a few selected regions).

These sorts of business problems require the application of statistical methods such as:

- 1.) time-series analysis and forecasting
- 2.) correlation and regression analysis
- 3.) estimation and significance testing
- 4.) decision-making under conditions of risk and uncertainty

5.) break-even analysis.

These methods in turn require an understanding of a range of summary statistics and concepts of probability. These topics therefore form the backbone of this course.

### 1.1.2 Statistics

Most of the quantitative methods mentioned above come under the general heading of statistics. The term "statistics" of course is often used to refer simply to a set of data – so, for example, we can refer to a country's unemployment statistics (which might be presented in a table or chart showing the country's unemployment rates each year for the last few years, and might be broken down by gender, age, region and/or industrial sector, etc.).

However, we can also use the term "Statistics" (preferably with a capital letter) to refer to the academic discipline concerned with the collection, description, analysis and interpretation of numerical data. As such, the subject of Statistics may be divided into two main categories:

- 1.) Descriptive Statistics
- 2.) Statistical Inference

### 1.1.3 Aim for the Chapter

The aim of this chapter is essentially

- 1.) Distinguish between descriptive and inferential statistics.
- 2.) Explain how samples and populations, as well as a sample statistic and population parameter, differ.
- 3.) Describe three research methods commonly used in behavioral science.
- 4.) State the four scales of measurement and provide an example for each.
- 5.) Distinguish between qualitative and quantitative data.
- 6.) Determine whether a value is discrete or continuous.
- 7.) Enter data into SPSS, excel or any other statistical package (coding is required).

## 1.2 Descriptive and Inferential Statistics

**Definition 1.2.1** Statistics is a body of concepts and methods which deal with collection, organization, presentation, analysis and interpretation of data using different phenomena, to draw valid conclusions and making reasonable decisions on the basis of research analysis.

Or simply statistics refers to the data itself and all numbers (statistics obtained from it for example the means that is Arithmetic mean, Harmonic mean and the geometric mean, mode, median etc. Or

**Definition 1.2.2** Statistics is a branch of mathematics used to summarize, analyze, and interpret what we observe-to make sense or meaning of our observations.

**Example 1.2.1** A family counselor may use statistics to describe patient behavior and the effectiveness of a treatment program.

**Example 1.2.2** A social psychologist may use statistics to summarize peer pressure among teenagers and interpret the causes.

**Example 1.2.3** A college professor may give students a survey to summarize and interpret how much they like (or dislike) the course.

In each case, the counselor, psychologist, and professor make use of statistics to do their job. Statistics are part of your everyday life (Statistics are all around you), and they are subject to interpretation. The interpreter, of course, is YOU.

**Remark 1.2.1** Statistics, while originally developed for the physical sciences, now pervade the social sciences, politics, sports, and business. Most articles in the main section of the newspaper (as well as the sports section) contain statistics (especially now that polling data fills the news).

**Note 1.2.1** But statistics can be misleading in a couple of ways, and one of the main goals of this course is for you to be equipped to judge the mathematical significance of statistics you encounter (statistics can be deceiving-and so can interpreting them).

There are two general types of statistics:

- 1.) Descriptive (quantitative) statistics: statistics that summarize observations.
- 2.) Inferential (qualitative) statistics: statistics used to interpret the meaning of descriptive statistics.



### 1.2.1 Descriptive Statistics

Researchers can measure many behavioral variables, such as love, anxiety, memory, and thought. Often, hundreds or thousands of measurements are made, and procedures were developed to organize, summarize, and make sense of these measures. These procedures, referred to as **descriptive statistics**, are specifically used to describe or summarize numeric observations, referred to as **data**.

Descriptive (quantitative) statistics comprises of the methods concerned with collecting and describing the set of data so as to yield meaningful information without inferring (concluding) anything beyond the given set of data. These methods include,

- 1.) Summarizing and describing the over all pattern of the data by presentation of tables and graphs and by examining the over all shape of the graph /pictorial for important features such as symmetry and departure from symmetry (skewness).
- 2.) Computation of numerical data measures such as measures of central tendency and measures of spread (dispersion)

**Definition 1.2.3** Descriptive statistics are procedures used to summarize, organize, and make sense of a set of scores or observations or data.

**Definition 1.2.4** Descriptive statistics are typically presented graphically, in tabular form (in tables), or as summary statistics (single values).

**Definition 1.2.5** Data (plural) are measurements or observations that are typically numeric.

**Definition 1.2.6** A datum (singular) is a single measurement or observation, usually referred to as a score or raw score.

Data are generally presented in summary. Typically, this means that data are presented *graphically*, in *tabular form* (in tables), or as *summary statistics* (e.g., an average). For example, the number of times each individual fidgeted is not all that meaningful, whereas the average (mean), middle (median), or most common (mode) number of times among all individuals is more meaningful. Tables and graphs serve a similar purpose to summarize large and small sets of data.

Most often, researchers collect data from a portion of individuals in a group of interest. For example, the 50, 100, or 1,000 students in the anxiety example would not constitute all students in college. Hence, these researchers collected anxiety data from some students, not all. So researchers require statistical procedures that allow them to infer what the effects of anxiety are among all students of interest using only the portion of data they measured.

### 1.2.2 Inferential Statistics

The problem described in the last paragraph is that most scientists have limited access to the phenomena they study, especially behavioral phenomena. As a result, researchers use procedures that allow them to interpret or infer the meaning of data. These procedures are called inferential statistics.

**Definition 1.2.7** Qualitative (Inferential ) statistics comprises methods of analysis that yield conclusions and predictions about the entire set of data and even beyond the data ( forecasting).

**Definition 1.2.8** Inferential statistics are procedures used that allow researchers to infer or generalize observations made with samples to the larger population from which they were selected.

**Example 1.2.4** To illustrate, let's continue with the college student anxiety example. All students enrolled in college would constitute the **population**. This is the group that researchers want to learn more about. Specifically, they want to learn more about characteristics in this population, called **population parameters**. The characteristics of interest are typically some descriptive statistic. In the anxiety example, the characteristic of interest is anxiety, specifically measured as the number of times students fidget during a class presentation.

Unfortunately, in behavioral research, scientists rarely know what these population parameters are since they rarely have access to an entire population. They simply do not have the time, money, or other resources to even consider studying all students enrolled in college.

The alternative is to select a portion or **sample** of individuals in the population. Selecting a sample is more practical, and most scientific research you read comes from samples and not populations. Going back to our example, this means that selecting a portion of students from the larger population of all students enrolled in college would constitute a sample. A characteristic that describes a sample is called a **sample statistic**-this is similar to a parameter, except it describes characteristics in a sample and not a population.

Inferential statistics use the characteristics in a sample to infer what the unknown parameters are in a given population. In this way, a sample is selected from a population to learn more about the characteristics in the population of interest.

**Definition 1.2.9** A **population** is defined as the set of all individuals, items, or data of interest. This is the group about which scientists will generalize.

If a set of data consists of all possible observations of a given phenomenon under study then we refer to such a set as a population.

**Example 1.2.5** A set of all students of a Ugandan university constitute a population.

**Definition 1.2.10** A characteristic (usually numeric) that describes a population is referred to as a **population parameter**.

Its any numerical value that describes the characteristics of the population. Such parameters include population mean, population variance particularly the mean age of all students in the University. Characteristic or measure obtained from a population.

**Definition 1.2.11** A **sample** is defined as a set of selected individuals, items, or data taken from a population of interest.

It's a set of observations selected from a population that is a sample in a subset or a representative of the population, and important conclusions about the population can often be inferred from the analysis of the sample.

**Example 1.2.6** An example of a sample all students in a faculty or college at a university.

**Definition 1.2.12** A characteristic (usually numeric) that describes a sample is referred to as a **sample statistic**. It's any numerical value that describes the characteristics of the sample.

**Example 1.2.7** As an example, the sample mean is a statistic. Characteristic or measure obtained from a sample.

**Example 1.2.8** On the basis of the following example, we will identify the population, sample, population parameter, and sample statistic: Suppose you read an article in the local college newspaper citing that the average college student plays 2 hours of video games per week. To test whether this is true for your school, you randomly approach 20 fellow students and ask them how long (in hours) they play video games per week. You find that the average student, among those you asked, plays video games for 1 hour per week. Distinguish the population from the sample.

**Solution :** *In this example, all college students at your school constitute the population of interest, and the 20 students you approached is the sample that was selected from this population of interest. Since it is purported that the average college student plays 2 hours of video games per week, this is the population parameter (2 hours). The average number of hours playing video games in the sample is the sample statistic (1 hour).* ■

### Exercise 1.1

- 1.) ..... are techniques used to summarize or describe numeric data.
- 2.) ..... describe(s) how a population is characterized, whereas ..... describe(s) the characteristics of samples.
  - A. Statistics; parameters
  - B. Parameters; statistics
  - C. Descriptive; inferential
  - D. Inferential; descriptive
- 3.) A psychologist wants to study a small population of 40 students in a local private school. If the researcher was interested in selecting the entire population of students for this study, then how many students must the psychologist include?
  - A. None, since it is not possible to study an entire population in this case.
  - B. At least half, since this would constitute the majority of the population.
  - C. All 40 students, since all students constitute the population.
- 4.) *True or false:* Inferential statistics are used to help the researcher infer whether observations made with samples are reflective of the population.

1.) Descriptive    2.) B    3.) C    4.) True

Statistics

## 1.3 Forms of Statistical Data:

### 1.3.1 Primary Data

If data is collected for a specific purpose then it is known as *primary data*.

For example, the information collected direct from householders' television sets through a microcomputer link-up to a mainframe computer owned by a television company is used to decide the most popular television programmes and is thus primary data. The Census of Population, which is taken every ten years, is another good example of primary data because it is collected specifically to calculate facts and figures in relation to the people living in the country.

This refers to data collected directly from the field by carrying out the survey on any particular topic, here the enumerators are sent to the respondents or the field with questionnaires and records of data are taken.

### 1.3.2 Secondary Data

Secondary data is data which has been collected for some purpose other than that for which it is being used.

For example, if a company has to keep records of when employees are sick and you use this information to tabulate the number of days employees had flu in a given month, then this information would be classified as *secondary data*.

Most of the data used in compiling business statistics is secondary data because the source is the accounting, costing, sales and other records compiled by companies for administration purposes.

Secondary data must be used with great care; as the data was collected for another purpose, and you must make sure that it provides the information that you require. To do this you must look at the sources of the information, find out how it was collected and the exact definition and method of compilation of any tables produced.

Secondary data refer to data collected from published or unpublished compilations such as news papers, Text books and other already existing data sources.

Each of the above sources is important in statistics but the primary source is preferred to secondary source due to the following reasons,

- 1.) Secondary sources may contain mistakes due to errors encountered during the copying from the primary source.
- 2.) Primary source includes a schedule and description of the procedure used in collecting the sample data primary which enables the user to ascertain how much confidence they may attach to the findings of the study.
- 3.) A primary source usually shows data in detail which can be put to a greater number of uses.
- 4.) Primary sources of data often contain the definition of terms and units used ,this makes the data more user friendly compared to secondary source.

## 1.4 Methods and Sources of Primary Data Collection

The most common methods of data collection include direct observations, personal interview and questionnaire method.

### 1.4.1 Direct Observation

This involves enumerators taking observations directly from the sampling units of interest.

Some of the advantages of direct observation are:

- 1.) It's free from errors due to little memory lapse between the time of recording and hearing from the sampling unit as the enumerators record everything as they happen.
- 2.) Non response errors are not encountered as compared to questionnaire method.

Some of the disadvantages of direct observation are:

- 1.) Its expensive and time consuming as it involves moving from one place to another.
- 2.) Transport and communication is a problem especially in places where means of transport and communication network is very poor.
- 3.) Its not easy to get people's attitudes towards anything by mere observation.
- 4.) The technique is always not feasible especially when observing human behavior, this is because people have a tendency of changing behavior during the process of observation.

### 1.4.2 Personal Interview

This method involves the data collector (enumerator) being brought into contact with the respondents and asks him /her questions about the subject under study.

Advantages:

- 1.) The interaction creates an opportunity for an on spot clarification of concepts on which one is collecting information about.
- 2.) It is very useful where the respondents are not very sure of the kind of response to give.

Disadvantages:

- 1.) It involves high expenses on transport and other field related exercises.
- 2.) It's prone to interviewers bias, often at times interviewers may ask leading or misleading questions or suggestive questions and such questions may bias the respondent's answer.
- 3.) There is likely to be a problem of language barrier between the interviewer and the interviewed.

### 1.4.3 Questionnaire Method

The questionnaire method can be split up into self administered and mail questionnaire. Where mail questionnaire involves writing a set of questions on a paper, list of instructions and a letter explaining objectives or purposes of the study.

#### Advantages

- 1.) Correct and accurate information can be got since consultations about the topic under study can be made.
- 2.) It reduces errors due to the interviewers bias
- 3.) It's a very effective method where sample units are scattered
- 4.) It is Speed and cost effective as it does not involve movement of people.
- 5.) Its possible to get correct information about sensitive issues since people fill the questionnaires privately and send them to the researcher.

#### Disadvantages:

- 1.) It assumes high level of literacy among the respondents which is usually not the case in most African countries.
- 2.) It assumes the existence of a good postal system which is efficient and can reach every person in the areas of research.
- 3.) Response is usually slow because people fill in the questionnaires at their own pace.
- 4.) There is a high rate of non-response since follow ups are very difficult and expensive to conduct.
- 5.) If the questionnaires reach to a wrong respondent then the data required will be biased eg female based questionnaires filled by a male respondent may bias the information to be given.

## 1.5 Cross-Section and Time-Series Data

Data collected from a sample of units (e.g. individuals, firms or government departments) for a single time period is called cross-section data. For example, the test scores obtained by 20 management trainees in a company in 2007 would represent a sample of cross-section data.

On the other hand, data collected for a single unit (e.g. a single individual, firm or government department) at multiple time periods are called time-series data. For example, annual data on the UG inflation rate from 1985–2007 would represent a sample of time-series data. Sometimes it is possible to collect cross section over two or more time periods – the resulting data set is called a panel data or longitudinal data set.

## 1.6 Statistics in Research

This course will describe many ways of *measuring and interpreting data*. Yet, simply collecting data does not make you a scientist. To engage in science, you must follow specific procedures for collecting data. Think of this as playing a game. Without the rules and procedures for playing, the game itself would be lost. The same is true in science; without the rules and procedures for collecting data, the ability to draw scientific conclusions would be lost. Ultimately, statistics are used in the context of **science**, and so it is necessary to introduce you to the basic procedures of scientific inquiry.

**Definition 1.6.1** Science is the study of phenomena, such as behavior, through strict observation, evaluation, interpretation, and theoretical explanation.

### 1.6.1 Experimental Method

Any study that demonstrates cause is called an **experiment**. To demonstrate cause, though, an experiment must follow strict procedures to ensure that the possibility of all other possible causes have been minimized or eliminated. So researchers must control the conditions under which observations are made to isolate cause-and-effect relationships between variables.

Three requirements must be satisfied for a study to be regarded as an experiment:

- 1.) Randomization (of assigning participants to conditions)
- 2.) Manipulation (of variables that operate in an experiment)
- 3.) Comparison (or a control group)

**Example 1.6.1** An experiment to determine the effect of distraction on student test scores. A sample of students was selected from a population of all undergraduates. In one group, the professor sat quietly while students took the exam (low-distraction group); in the other, the professor rattled papers, tapped her foot, and made other sounds during the exam (high-distraction group). Exam scores in both groups were measured and compared.

For this distraction example, the independent variable was distraction. The researchers first manipulated the levels of this variable (low, high), meaning that they created the conditions. They then assigned each student at random to experience one of the levels of distraction.

Notice also that there are two groups in the experiment. So tests scores for students experiencing high levels of distraction were compared to those experiencing low levels of distraction. By comparing test scores between groups, we can determine whether high levels of distraction caused lower scores (compared to scores in the low-distraction group). This satisfies the requirement of comparison (Requirement 3), which requires that at least two groups be observed in an experiment. This allows scores in one group to be compared to those in at least one other group.

In this example, test scores were measured in each group. The measured variable in an experiment is referred to as the **dependent variable** (dependent variable). Dependent variables can often be measured in many ways, and therefore require an **operational definition**. This is where a dependent variable is defined in terms of how it will be measured. For example, here we operationally defined exam performance as a score between 0 and 100 on a test. So

to summarize the experiment, levels of distraction (independent variable) were presumed to cause an effect or difference in exam grades (dependent variable) between groups. This is an experiment since the researchers satisfied the requirements of randomization, manipulation, and comparison, thereby allowing them to draw cause-and-effect conclusions.

The dependent variable is the variable that is believed to change in the presence of the independent variable. It is the “presumed effect.” An operational definition is a description of some observable event in terms of the specific process or manner by which it was observed or measured.

**Definition 1.6.2 Random assignment** is a random procedure used to ensure that participants in a study have an equal chance of being assigned to a particular group or condition.

**Definition 1.6.3 An independent variable** is the variable that is manipulated in an experiment. This variable remains unchanged (or “independent”) between conditions being observed in an experiment. It is the “presumed cause.”

**Definition 1.6.4** The specific conditions of an independent variable are referred to as the **levels of the independent variable**.

**Definition 1.6.5** The **dependent variable** is the variable that is believed to change in the presence of the independent variable. It is the “presumed effect.”

**Definition 1.6.6** An **operational definition** is a description of some observable event in terms of the specific process or manner by which it was observed or measured.

**Example 1.6.2** A researcher conducts the following study: Participants are presented with a list of words written on a white background on a PowerPoint slide. In one group, the words are written in red (Group Color); in a second group, the words are written in black (Group Black). Participants are allowed to study the words for 1 minute. After that time, the slide is removed and participants are allowed 1 minute to write down as many words as they can recall. The number of words correctly recalled will be recorded for each group. Explain how this study can be an experiment.

**Solution :** *To create an experiment, we must satisfy the three requirements for demonstrating cause and effect: randomization, manipulation, and comparison. To satisfy each requirement, the researcher can*

- 1.) Randomly assign participants to experience one of the conditions. This ensures that some participants read colored words and others read black words entirely by chance.*
- 2.) Create the two conditions. The researcher could write 20 words on a PowerPoint slide. On one slide, the words are written in red; on the second slide, the same words are written in black.*
- 3.) Include a comparison group. In this case, the number of colored words correctly recalled will be compared to the number of black words correctly recalled, so this study has a comparison group.*

*Remember that each requirement is necessary to demonstrate that the levels of an independent variable are causing changes in the value of a dependent variable. If any one of these requirements is not satisfied, then the study is not an experiment.*





### 1.6.2 Quasi-Experimental Method

A study that lacks randomization, manipulation, or comparison is called a quasi-experiment. This most often occurs in one of two ways:

- 1.) The study includes a quasi-independent variable.
- 2.) The study lacks a comparison group.

In a typical quasi-experiment, the variables being studied can't be manipulated, which makes random assignment impossible. This occurs when variables are preexisting or inherent to the participants themselves. These types of variables are called quasi-independent variables.

**Example 1.6.3** Since the levels of gender (male, female) can't be randomly assigned (it is a **quasi-independent variable**), this study is regarded as a quasi-experiment.

**Definition 1.6.7** A quasi-independent variable is a variable whose levels are not randomly assigned to participants (nonrandom). This variable differentiates the groups or conditions being compared in a quasi-experiment.

A study is also regarded as a quasi-experiment when only one group is observed. Since only one group is observed, there is no comparison group. So differences between two levels of an independent variable can't be compared. In this way, failing to satisfy any of the three requirements for an experiment (randomization, manipulation, or comparison) makes the study a quasi-experiment.

### 1.6.3 Correlation Method

Another method for examining the relationship between variables is to measure pairs of scores for each individual. This method can determine whether a relationship exists between variables, but it lacks the appropriate control needed to demonstrate cause and effect. To illustrate, suppose you test for a relationship between time spent using a computer and exercising per week. Using the correlational method, we can examine the extent to which two variables change in a related fashion. For example we might show, as computer use increases, time spent exercising decreases. This pattern suggests that computer use and time spent exercising are related.

#### Exercise 1.2

- 1.) ..... is the study of phenomena through strict observation, evaluation, interpretation, and theoretical explanation. *Science*
- 2.) State whether each of the following describes an experiment, quasi-experiment, or correlational method.
  - A. A researcher tests whether dosage level of some drug (low, high) causes significant differences in health. *Experiment*
  - B. A researcher tests whether political affiliation (Republican, Democrat) is associated with different attitudes toward morality. *Quasi-Experiment*
  - C. A researcher measures the relationship between income and life satisfaction. *Correlation*
- 3.) True or false: An experiment is the only method that can demonstrate cause-and-effect relationships between variables. *True*

## 1.7 Scales of Measurement

We do require that variables in a study be measured on a certain scale of measurement.

In the early 1940s, Harvard psychologist S. S. Stevens coined the terms *nominal*, *ordinal*, *interval*, and *ratio* to classify the scales of measurement.

Scales of measurement are rules that describe the properties of numbers. These rules imply that a number is not just a number in science. Instead, the extent to which a number is informative depends on how it was used or measured.

In all, scales of measurement are characterized by three properties: order, differences, and ratios. Each property can be described by answering the following questions:

- 1.) Order: Does a larger number indicate a greater value than a smaller number?
- 2.) Differences: Does subtracting two numbers represent some meaningful value?
- 3.) Ratio: Does dividing (or taking the ratio of) two numbers represent some meaningful value?

**Definition 1.7.1** Scales of measurement refer to how the properties of numbers can change with different uses.

### 1.7.1 Norminal Scales

Numbers on a **nominal scale** identify something or someone; they provide no additional information. Common examples of nominal numbers include ZIP codes, license plate numbers, credit card numbers, country codes, telephone numbers, and Social Security numbers. These numbers simply identify locations, vehicles, or individuals and nothing more. One credit card number, for example, is not greater than another; it is simply different.

**Definition 1.7.2** Nominal scales are measurements where a number is assigned to represent something or someone.

In science, nominal variables are typically categorical variables that have been coded—converted to numeric values. Examples of nominal variables include a person's race, gender, nationality, sexual orientation, hair and eye color, season of birth, marital status, or other demographic or personal information. A researcher may code men as 1 and women as 2. They may code the seasons of birth as 1, 2, 3, and 4 for spring, summer, fall, and winter, respectively. These numbers are used to identify gender or the seasons and nothing more. We often code words with numeric values when entering them into statistical programs such as SPSS.

### 1.7.2 Ordinal Scales

An **ordinal scale** of measurement is one that conveys order alone. This scale indicates that some value is greater or less than another value. Examples of ordinal scales include finishing order in a competition, education level, and rankings. These scales only indicate that one value is greater or less than another, so differences between ranks do not have meaning. Consider, for example, ranking or class end of term position. Based on ranks alone, can we say that the difference between the psychology graduate programs ranked 1 and 11 is the same as the difference between those ranked 13 and 23? In both cases, 10 ranks separate the students. Yet,

if you look at the actual scores for determining rank, you find that the difference between ranks 1 and 11 is different from that of ranks 13 and 23. So the difference in points is not the same. Ranks alone don't convey this difference. They simply indicate that one rank is greater or less than another rank.

**Definition 1.7.3** **Ordinal scales** are measurements where values convey order or rank alone.

### 1.7.3 Interval Scales

An interval scale measurement, on the other hand, can be understood readily by two defining principles: equidistant scales and no true zero. A common example for this in behavioral science is the rating scale. Rating scales are taught here as an interval scale since most researchers report these as interval data in published research. This type of scale is a numeric response scale used to indicate a participant's level of agreement or opinion with some statement.

**Example 1.7.1** A table showing satisfaction rating. An example of a 7-point rating scale for satisfaction used for scientific investigation.

1	2	3	4	5	6	7
Completely Unsatisfied						Completely Satisfied

**Definition 1.7.4** **Interval scales** are measurements where the values have no true zero and the distance between each value is equidistant.

An **equidistant scale** is a scale distributed in units that are equidistant from one another. Many behavioral scientists assume that scores on a rating scale are distributed in equal intervals. For example, if you are asked to rate your satisfaction with a spouse or job on a 7-point scale from 1 (completely unsatisfied) to 7 (completely satisfied), like in the scale shown in Table above, then you are using an interval scale. Since the distance between each point (1 to 7) is assumed to be the same or equal, it is appropriate to compute differences between scores on this scale. So a statement such as, "The difference in job satisfaction among men and women was 2 points," is appropriate with interval scale measurements.

**Definition 1.7.5** Equidistant scales are those values whose intervals are distributed in equal units.

However, an interval scale does not have a **true zero**. A common example of a scale without a true zero is temperature. A temperature equal to zero for most measures of temperature does not mean that there is no temperature; it is just an arbitrary zero point. Values on a rating scale also have no true zero. In the example in Table above, a 1 was used to indicate no satisfaction, not 0. Each value (including 0) is arbitrary. That is, we could use any number to represent none of something. Measurements of latitude and longitude also fit this criterion. The implication is that without a true zero, there is no value to indicate the absence of the phenomenon you are observing (so a zero proportion is not meaningful). For this reason, stating a ratio such as, "Satisfaction ratings were three times greater among men compared to women," is not appropriate with interval scale measurements.

**Definition 1.7.6** A **true zero** describes values where the value 0 truly indicates nothing. Values on an interval scale do not have a true zero.

### 1.7.4 Ratio Scales

**Ratio scales** are similar to interval scales in that scores are distributed in equal units. Yet, unlike interval scales, a distribution of scores on a ratio scale has a true zero. This is an ideal scale in behavioral research because any mathematical operation can be performed on the values that are measured. Common examples of ratio scales include counts and measures of length, height, weight, and time. For scores on a ratio scale, order is informative. For example, a person who is 30 years old is older than another who is 20. Differences are also informative. For example, the difference between 70 and 60 seconds is the same as the difference between 30 and 20 seconds (the difference is 10 seconds). Ratios are also informative on this scale because a true zero is defined—it truly means nothing. Hence, it is meaningful to state that 60 pounds is twice as heavy as 30 pounds.

**Definition 1.7.7** **Ratio scales** are measurements where a set of values has a true zero and are equidistant.

In science, researchers often go out of their way to measure variables on a ratio scale. For example, if they want to measure eating, they may choose to measure the amount of time between meals or the amount of food consumed (in ounces). If they measure memory, they may choose to measure the amount of time it takes to memorize some list or the number of errors made. If they measure depression, they may choose to measure the dosage (in milligrams) that produces the most beneficial treatment or the number of symptoms reported. In each case, the behaviors were measured using values on a ratio scale, thereby allowing researchers to draw conclusions in terms of order, differences, and ratios—there are no restrictions with ratio scale variables.

**Example 1.7.2** The table below summarizes the answers to the questions for each scale of measurement. You can think of each scale as a gradient of the informativeness of data. In this section, we begin with the least informative scale (nominal) and finish with the most informative scale (ratio).

		Scale of Measurement			
		Nominal	Ordinal	Interval	Ratio
Property	Order	No	Yes	Yes	Yes
	Difference	No	No	Yes	Yes
	Ratio	No	No	No	Yes

#### Exercise 1.3

- 1.) The ..... refer to how the properties of numbers can change with different uses.
- 2.) In 2010, Fortune 500 magazine ranked Apple as the most admired company in the world. This ranking is on a(n) ..... scale of measurement.

- 3.) What are two characteristics of rating scales that allow some researchers to use these values on an interval scale of measurement?
- A. Values on an interval scale have order and differences.
  - B. Values on an interval scale have differences and a true zero.
  - C. Values on an interval scale are equidistant and have a true zero.
  - D. Values on an interval scale are equidistant and do not have a true zero.
- 4.) Which of the following is not an example of a ratio scale variable?
- A. Age (in days)
  - B. Speed (in seconds)
  - C. Height (in inches)
  - D. Movie ratings (1 to 4 stars)

1.) Scale of measurement  
2.) Ordinal  
3.)  
4.)

## 1.8 Types of Statistical Data

The scales of measurement reflect the informativeness of data. With nominal scales, researchers can conclude little; with ratio scales, researchers can conclude just about anything in terms of order, differences, and ratios. Researchers also distinguish between the types of data they measure. The types of data researchers measure fall into two categories:

- 1.) Continuous or Discrete
- 2.) Quantitative or Qualitative

### 1.8.1 Continuous and Discrete Variables

Variables can be categorized as continuous or discrete. A **continuous variable** is measured along a continuum. So continuous variables are measured at any place beyond the decimal point. Consider, for example, that Olympic sprinters are timed to the nearest hundredths place (in seconds), but if the Olympic judges wanted to clock them to the nearest millionths place, they could.

**Definition 1.8.1** A **continuous variable** is measured along a continuum at any place beyond the decimal point. Continuous variables can be measured in whole units or fractional units.

A **discrete variable**, on the other hand, is measured in whole units or categories. So discrete variables are not measured along a continuum. For example, the number of brothers and sisters you have and your family's socioeconomic class (working class, middle class, upper class) are examples of discrete variables. Refer to Table 1.3 for more examples of continuous and discrete variables.

**Definition 1.8.2** A **discrete variable** is measured in whole units or categories that are not distributed along a continuum.

**Note 1.8.1** Continuous variables are measured along a continuum, whereas discrete variables are measured in whole units or categories.

### 1.8.2 Quantitative and Qualitative Variable

Variables can be categorized as quantitative or qualitative. A **quantitative variable** varies by amount. The variables are measured in numeric units, and so both continuous and discrete variables can be quantitative. For example, we can measure food intake in calories (a continuous variable) or we can count the number of pieces of food consumed (a discrete variable). In both cases, the variables are measured by amount (in numeric units).

**Definition 1.8.3 Quantitative Variable** - Variables whose values result from counting or measuring something. For example, height, weight.

**Definition 1.8.4** A **quantitative variable** varies by amount. This variable is measured numerically and is often collected by measuring or counting.

A **qualitative variable**, on the other hand, varies by class. The variables are often labels for the behaviors we observe—so only discrete variables can fall into this category. For example, socioeconomic class (working class, middle class, upper class) is discrete and qualitative; so are many mental disorders such as depression (unipolar, bipolar) or drug use (none, experimental, abusive).

**Definition 1.8.5 Qualitative variable** - Variables that are not measurement variables. Their values do not result from measuring or counting. For example, hair color, religion, political party, profession

**Definition 1.8.6** A **qualitative variable** varies by class. This variable is often represented as a label and describes nonnumeric aspects of phenomena.

**Example 1.8.1 Quantitative data:** This is data that can be expressed numerically for example

1.) Marks of students    2.) Heights                      3.) Weight                      4.) Books in the library

**Example 1.8.2 Qualitative data:** This is data that cannot be expressed numerically for example

1.) Colour                      2.) Shape                      3.) Gender                      4.) Intelligence Quotient (I.Q)

**Example 1.8.3** A list of 20 variables showing how they fit into the three categories that describe them.

Variables	Continuous vs. Discrete	Qualitative vs. Quantitative	Scale of Measurement
Gender (male, female)	Discrete	Qualitative	Nominal
Seasons (spring, summer, fall, winter)	Discrete	Qualitative	Nominal
Number of dreams recalled	Discrete	Quantitative	Ratio
Number of errors	Discrete	Quantitative	Ratio
Duration of drug abuse (in years)	Continuous	Quantitative	Ratio
Ranking of favorite foods	Discrete	Quantitative	Ordinal
Ratings of satisfaction (1 to 7)	Discrete	Quantitative	Interval
Body type (slim, average, heavy)	Discrete	Qualitative	Nominal
Score (from 0 to 100%) on an exam	Continuous	Quantitative	Ratio
Number of students in your class	Discrete	Quantitative	Ratio
Temperature (degrees Fahrenheit)	Continuous	Quantitative	Interval

Time (in seconds) to memorize a list	Continuous	Quantitative	Ratio
The size of a reward (in grams)	Continuous	Quantitative	Ratio
Position standing in line	Discrete	Quantitative	Ordinal
Political Party Affiliation	Discrete	Qualitative	Nominal
Type of distraction (auditory, visual)	Discrete	Qualitative	Nominal
A letter grade (A, B, C, D, F)	Discrete	Qualitative	Ordinal
Weight (in Kg) of an infant	Continuous	Quantitative	Ratio
A college students' MTH score	Discrete	Quantitative	Interval
Number of lever presses per minute	Discrete	Quantitative	Ratio

**Exercise 1.4** What is the difference between descriptive and inferential statistics?

**Exercise 1.5** Distinguish between data and a raw score.

**Solution :** *Data describe a set of measurements (made up of raw scores); a raw score describes individual measurements.* ■

**Exercise 1.6** By definition, how is a sample related to a population?

**Exercise 1.7** State three commonly used research methods in behavioral science.

**Solution :** *Experimental, quasi-experimental, and correlational research methods.* ■

**Exercise 1.8** In an experiment, researchers measure two types of variables: independent and dependent variables.

1.) Which variable is manipulated to create the groups?

2.) Which variable is measured in each group?

**Exercise 1.9** State the four scales of measurement. Which scale of measurement is the most informative?

**Solution :** *The four scales of measurement are nominal, ordinal, interval, and ratio. Ratio scale measurements are the most informative.* ■

**Exercise 1.10** Can a nominal variable be numeric? Explain.

**Exercise 1.11** What is the main distinction between variables on an interval and ratio scale of measurement?



**Exercise 1.12** A quantitative variable varies by .....; a qualitative variable varies by .....

**Exercise 1.13** What are the two types of data that are collected and measured quantitatively?

**Exercise 1.14** State whether each of the following words best describes descriptive statistics or inferential statistics.

- 1.) Describe                                      2.) Infer                                      3.) Summarize

**Exercise 1.15** State whether each of the following is true or false.

- 1.) Graphs, tables, and summary statistics all illustrate the application of inferential statistics.  
*False*
- 2.) Inferential statistics are procedures used to make inferences about a population, given only a limited amount of data. *True*
- 3.) Descriptive statistics can be used to describe populations and samples of data. *True*

**Exercise 1.16** A researcher measured behavior among all individuals in some small population. Are inferential statistics necessary to draw conclusions concerning this population? Explain.

**Exercise 1.17** Appropriately use the terms sample and population to describe the following statement: A statistics class has 25 students enrolled, but only 23 students attended class.

**Exercise 1.18** On occasion, samples can be larger than the population from which it was selected. Explain why this can't be true.

**Exercise 1.19** A researcher demonstrates that eating breakfast in the morning causes increased alertness throughout the day. What research design must the researcher have used in this example? Explain.

**Exercise 1.20** A researcher measures the height and income of participants and finds that taller men tend to earn greater incomes than shorter men. What type of research method did the researcher use in this example? Explain.

### 1.8.3 Recording and Sorting Data

**Example 1.8.4** Sample values (observations, measurements) should be recorded in the order in which they occur. Sorting, that is, ordering the sample values by size, is done as a first step of investigating properties of the sample and graphing it. Sorting is a standard process on the computer

Super alloys is a collective name for alloys used in jet engines and rocket motors, requiring high temperature (typically 1800° F), high strength, and excellent resistance to oxidation. Thirty specimens of Hastelloy C (nickelbased steel, investment cast) had the tensile strength (in 1000lb/sq in), recorded in the order obtained and rounded to integer values,

89	77	88	91	88	93	99	79	87	84	86	82	88	89	78
90	91	81	90	83	83	92	87	89	86	89	81	87	84	89

sorting gives

77	78	79	81	81	82	83	83	84	84	86	86	87	87	87
88	88	88	89	89	89	89	89	90	90	91	91	92	93	99

## 1.9 Methods of Data Presentation

There are four methods of representing statistical data and these are;

- |             |  |
|-------------|--|
| 1.) Text    | 3.) Graphical and                            |
| 2.) Tabular | 4.) Pictorial (diagrammatic) representations |

### 1.9.1 Tabular Representation

A table is a systematic organization of data with rows and columns, a good table should be concise, brief and easy to read. The major considerations in table constructions include;

- 1.) Title which should accompany every table and its always placed above the table it should be complete so as to identify all the data contained in the table. It should be worded as to mention most important consideration first placing towards the end a statement concerning arrangements of items and the period of time covered that is where, how and when.
- 2.) Prefatory note. It's a phrase usually placed just below a title and in lower case of less prominent letters, it provides an explanation concerning the whole table.
- 3.) Foot note. It provides explanations concerning individual figures of column or row figures such explanations are placed below the table. Foot notes referring to the row or column of the figures are usually identified by numbers while foot notes referring to specific figures are identified by symbols such as \*, #, etc.
- 4.) Source note. This gives the source of information contained in the table. If data is from a secondary source then the source should quote the author, title, volume, page, publisher and the date of the publication. A source note is usually placed below the table. Its useful in that it enables the reader to ascertain the reliability of the data and makes it possible for him to refer to the original source to verify the quoted figures or obtain additional information.
- 5.) Stub and caption. The Stub describes data row wise while the caption describes data column wise and this must be described on every table.
- 6.) Units. The units of measurement of the figures in the column or row are sometimes self explanatory when this is not true the nature of the units should be clearly be indicated in the stub or caption. If the units are uniform for all figures in the table their description should be done using a prefatory note.

















### 1.9.2 Diagrammatic (Pictorial) Representation of Data

This includes pictograms and pie charts

#### 1.9.2.1 Pictograms

Here we use small pictures to represent the quantity or variable under study. Pictures of cars are used in representing the population of cars.

**Example 1.9.1** The following pictogram shows the number of computers sold by a company for the months January to March.

January	    
February	      
March	   



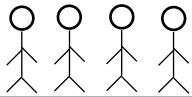

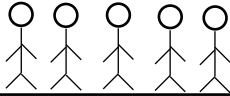
represents 5 computers

**Example 1.9.2** Assume the data below for the numbers of cars produced from a factory in the given years. Here we may choose ♣ to represent 1,000 cars.

Year	Number of cars	Picture
1980	1000	♣
1981	2000	♣♣
1983	3000	♣♣♣

**Definition 1.9.1** A **pictogram** or **pictograph** represents the frequency of data as pictures or symbols. Each picture or symbol may represent one or more units of the data.

**Example 1.9.3** The following pictograph shows the number of students using the various types of transport to go to school.

Walking	
Bus	
Bicycle	
Car	



Represents 4 students

1.) How many students go to school by car?

**Solution :** 20 students ■

2.) If the total number of students involved in the survey is 56 how many symbols must be drawn for the students walking to school?

**Solution :** 56 students should be represented by  $\frac{56}{4} = 14$  symbols.

There are already 11 symbols on the table.

So, the number of symbols to be added for “Walking” is  $14 - 11 = 3$  ■

3.) What is the percentage of students who cycle to school?

**Solution :**

$$\frac{8}{56} \times 100\% = 14.29\%$$
■

4.) What is the difference between students that use a car and those that use bicycles?

**Solution :**

$$5(4) - 2(4) = 12$$
■

5.) How much revenue will the bus driver collect each every other day, if it costs UGX5,00 for each student to get to school.

$$4(4) \times 5,000 = 80,000$$

### 1.9.2.2 Pie Charts

It's a circle which is divided by radial lines into sections or subsections of different angles/sizes so that the area of a particular sector is proportional to the number size of the figures represented.

**Definition 1.9.2** A pie chart is a representation that is used to display the proportion (i.e. percentage or fraction) of the data belonging to different categories.

In order to construct a pie chart, we make use of the fact that the total frequency (population number) corresponds to the total number of degrees ( $360^\circ$ )

**Example 1.9.4** Suppose the table below gives the amount of crops produced in tonnes in a given year, then using the data in the table, draw a pie chart to represent the information.

Crop	Amount(tonnes)
Wheat	3,000
Sorghum	2,500
Maize	1,000

solution

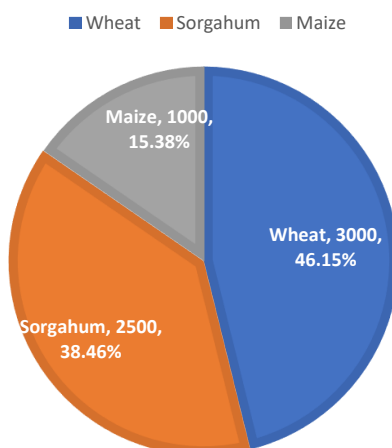
$$1\text{tonne} = \left(\frac{360}{6,500}\right)^\circ$$

$$\text{Wheat} = \left(\frac{3,000 \times 100}{6,500}\right) = 46.15\%$$

$$\text{Sorghum} = \left(\frac{2,500 \times 100}{6,500}\right) = 38.46\%$$

$$\text{Maize} = \left(\frac{1,000 \times 100}{6,500}\right) = 15.3846\%$$

The pie chart showing the crop production in tonnes is



### 1.9.3 Graphical Representation

#### 1.9.3.1 Bar and Column Chart

There are three types of bar graphs namely; simple, compound and multi - bar graphs.

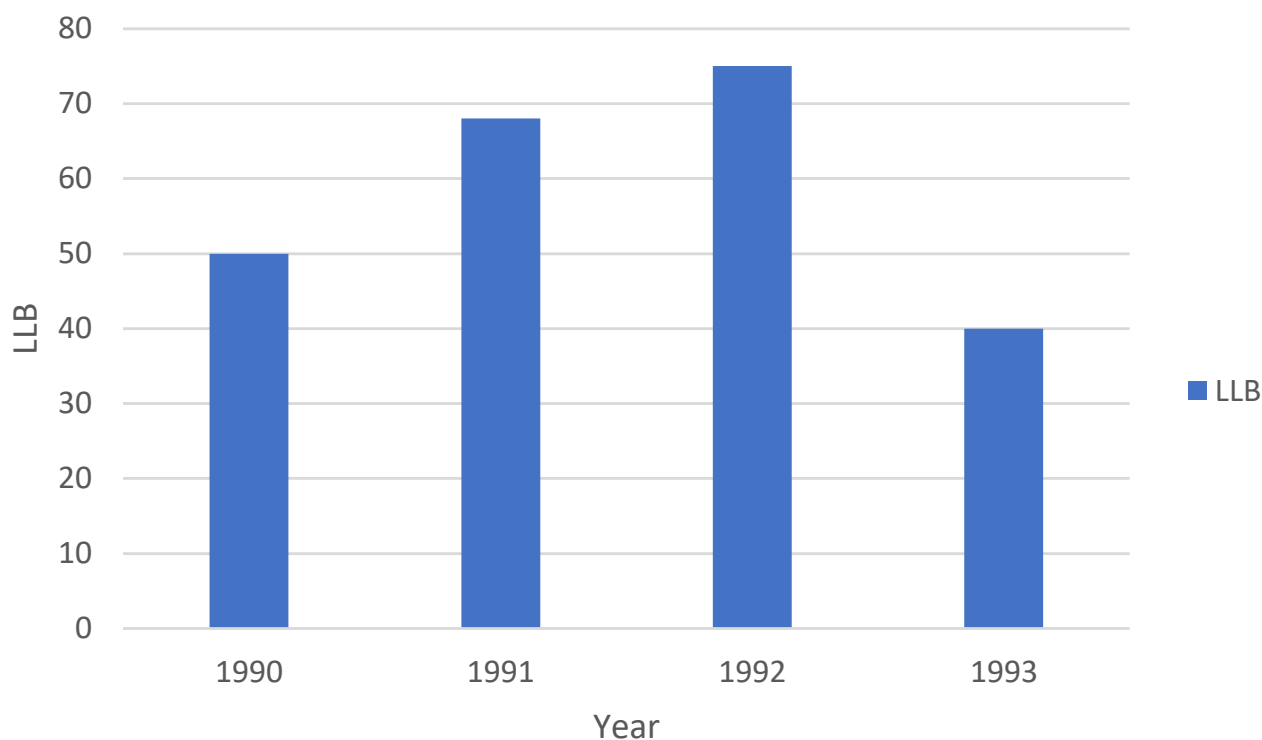
The simple graphs employs the bars or columns to indicate the frequency of occurrences of observations within each category and the height represents the number of the elements of that class. In the compound bar graphs bars are divided into different components each representing a given category of data and in multi - bar graph there are several bars within a given category.

**Example 1.9.5** Given the following table that shows the enrollment at a Uganda university (Hypothetical data)

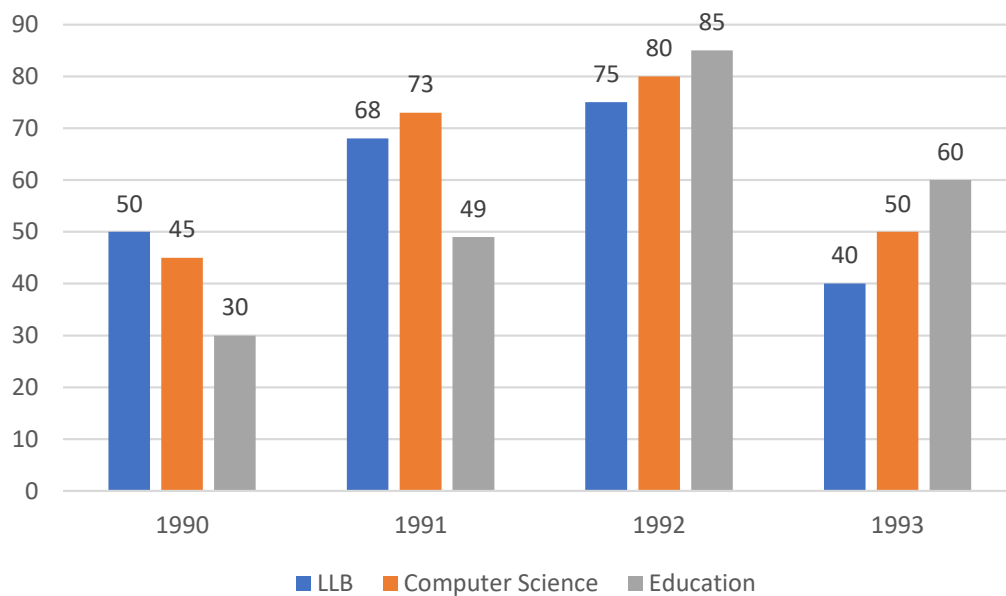
Year	Law(LLB)	Computer	Education	Total
1990	50	45	30	125
1991	68	73	49	190
1992	75	80	85	240
1993	40	50	60	150

We can represent the information in bar and column chart graphs as below,

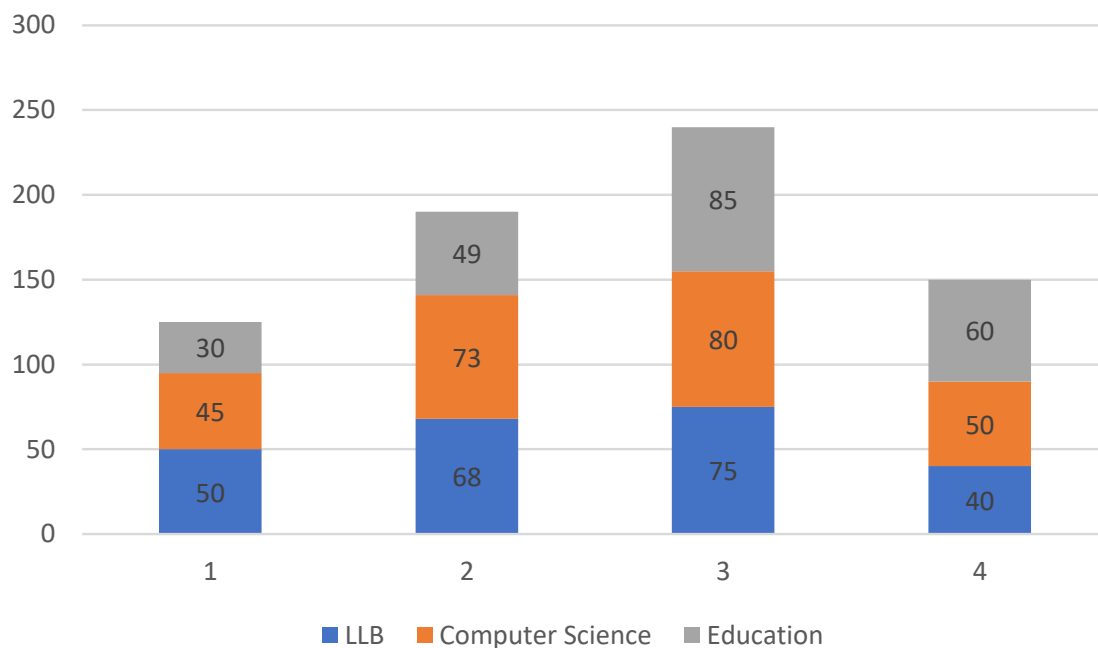
1.) Column chart of LLB intake per year



2.) Another form of the multi-column chart for all courses per year is



3.) The clustered column chart



Both the Bar and the Column charts display data using rectangular bars where the length of the bar is proportional to the data value. Both the charts are used to compare two or more values. However, the difference lies in their orientation. A bar chart is oriented horizontally whereas the column chart is oriented vertically. Although alike, they cannot be always used interchangeably because of the difference in their orientation.

The following are some of the bar chart for the enrollment example

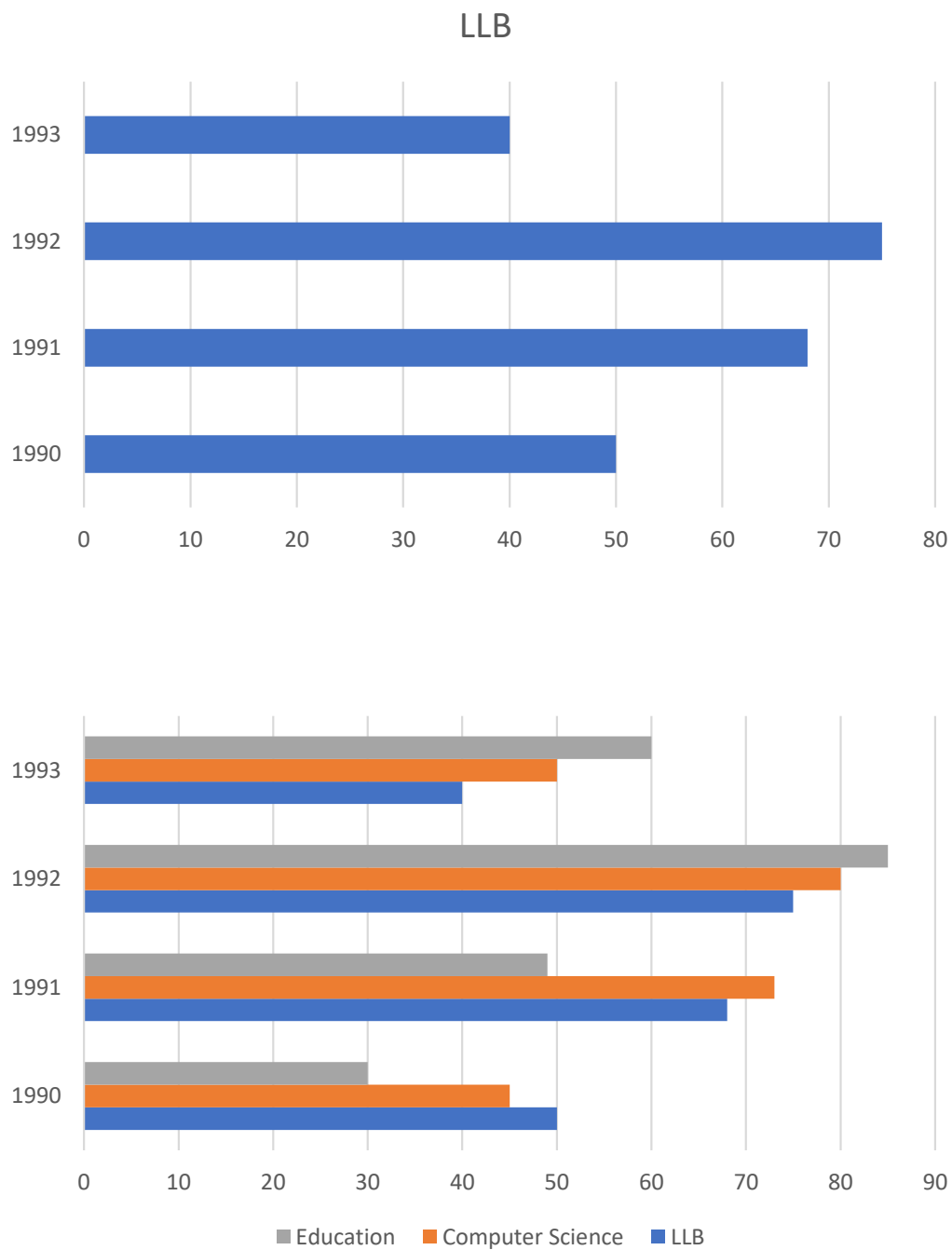


Figure 1.1: Bar graph and Multi-Bar graph respectively



## 1.9.3.2 Line Graphs

**Example 1.9.6** The table below shows the GDP for the East African countries in millions in the different years,

Year	Uganda	Kenya	Tanzania
1990	350	500	250
1991	400	450	430
1992	550	600	500
1993	430	600	450
1994	500	600	480

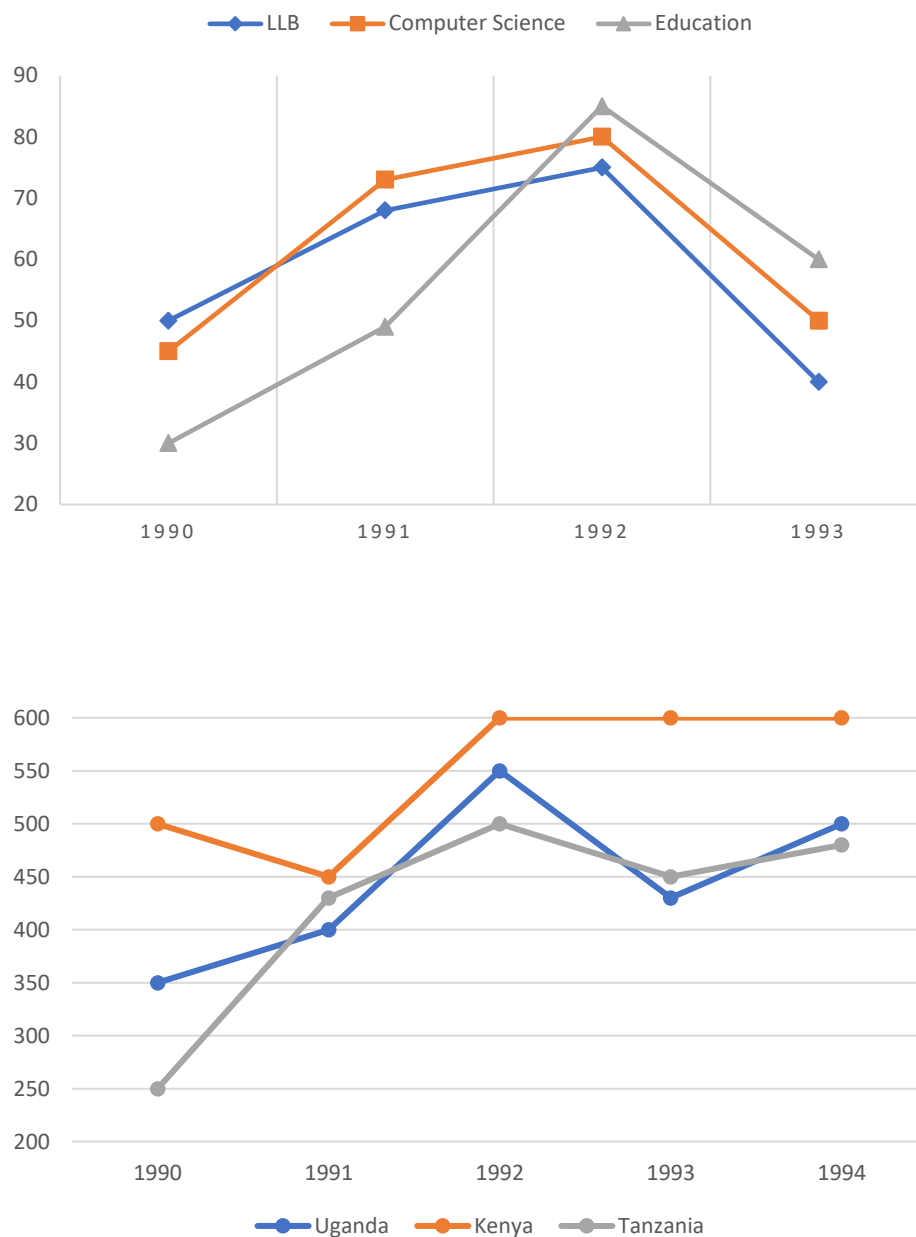


Figure 1.2: The Line graphs for Example 1.9.5 and Example 1.9.6 respectively

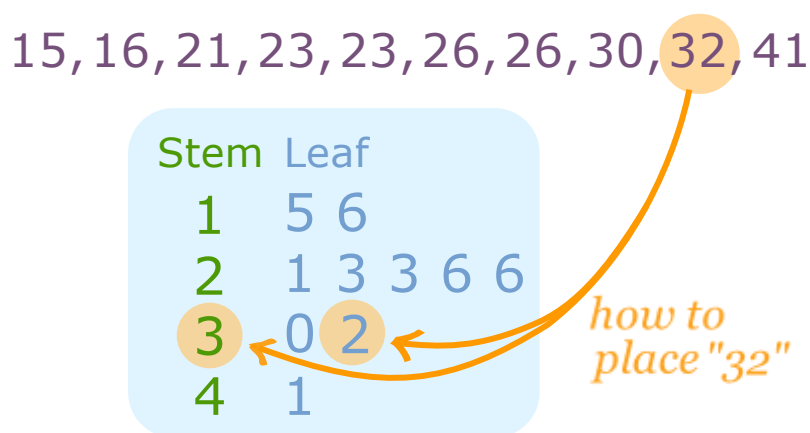
### 1.9.3.3 Stem-and-Leaf Graphs (Stemplots)

**Definition 1.9.3** A Stem and Leaf Plot is a special table where each data value is split into a “stem” (the first digit or digits) and a “leaf” (usually the last digit).

**Example 1.9.7** Given the data

15, 16, 21, 23, 23, 26, 26, 30, 32, 41

can be represented on a Stemplot as follows



1.) Stem “1” Leaf “5” means 15

3.) Stem “2” Leaf “1” means 21

2.) Stem “1” Leaf “6” means 16

When looking at a data set, each observation may be considered as consisting of two parts—a stem and a leaf. To make a stem and leaf plot, each observed value must first be separated into its two parts:

1.) The stem is the first digit or digits,

3.) Each stem can consist of any number of digits,

2.) The leaf is the final digit of a value,

4.) Each leaf can have only a single digit.

**Example 1.9.8** A teacher asked 10 of her students how many books they had read in the last 12 months. Their answers were as follows:

12, 23, 19, 6, 10, 7, 15, 25, 21, 12

Prepare a stem and leaf plot for these data.

Stem	Leaf
0	6 7
1	2 9 0 5 2
2	3 5 1

1|2 represents 12

**Example 1.9.9** A stem-and-leaf plot of the values

20, 30, 32, 35, 41, 41, 43, 47, 48, 51,  
53, 53, 54, 56, 57, 58, 58, 59, 60, 62,  
64, 65, 65, 69, 71, 74, 77, 88 and 102

is given by

Stem	Leaf
2	0
3	0 2 5
4	1 1 3 7 8
5	1 3 3 4 6 7 8 8 9
6	0 2 4 5 5 9
7	1 4 7
8	8
9	
10	2

$n = 29$ , 1|2 represents 12

**Example 1.9.10** Plot the Stem-Leaf graph for the data

44, 46, 47, 49, 63, 64, 66, 68, 68, 72, 72, 75, 76, 81, 84, 88, 106

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

1|2 represents 12

**Remark 1.9.1** For negative numbers, a negative is placed in front of the stem unit

**Example 1.9.11** Plot the Stem-and-Leaf graph for  $-34$ ,  $-23$ , 12, 16, 29

Stem	Leaf
-3	4
-2	3
1	2 6
2	9

1|2 represents 12

**Exercise 1.21** List down the 25 data points used to generate the following Stem-Leaf graph

Stem	Leaf
4	0
5	9
6	0
7	2 5 9 9
8	4 7
9	1 4 4 5 9
10	0 1 3 9
11	1 9 9
12	7
13	1 4
14	1

$n = 25$ , 1|2 represents 1.2

**Exercise 1.22** Each morning, a teacher quizzed his class with 20 geography questions. The class marked them together and everyone kept a record of their personal scores. As the year passed, each student tried to improve his or her quiz marks. Every day, Daniel recorded his quiz marks on a stem and leaf plot. This is what his marks looked like plotted out:

Stem	Leaf
0	3 6 5
1	0 1 4 3 5 6 5 8 9 7 9
2	0 0 0

1|2 represents 12

- 1.) Analyse Daniel's stem and leaf plot.
- 2.) What is his most common score on the geography quizzes?
- 3.) What is his highest score?
- 4.) His lowest score?

Rotate the stem and leaf plot onto its side so that it looks like a bar graph.

Are most of Daniel's scores in the 10s, 20s or under 10?

It is difficult to know from the plot whether Daniel has improved or not because we do not know the order of those scores.

The **main advantage** of a stem and leaf plot is that the data are grouped and all the original data are shown, too

**Exercise 1.23** The following data represent measurements of carbon monoxide content (in mg) for 25 brands of cigarettes:

13.6, 16.6, 23.5, 10.2, 5.4, 15.0, 9.0, 12.3, 16.3, 15.4, 13.0, 14.4, 10.0  
10.2, 9.5, 1.5, 18.5, 12.6, 17.5, 4.9, 15.9, 8.5, 10.6, 13.9, 14.9.

Draw its Stem-and-Leaf graph.

**Example 1.9.12** A random sample of 64 people were selected to take the University Intelligence Test. After each person completed the test, they were assigned an intelligence quotient (IQ) based on their performance on the test. The resulting 64 IQs are as follows:

111	85	83	98	107	101	100	94	101	86	105	122	104	106	90	123
102	107	93	109	141	86	91	88	98	128	93	114	87	116	99	94
94	406	436	402	75	96	78	116	107	106	68	104	91	87	105	97
110	91	107	107	85	117	93	108	91	110	105	99	85	99	99	96

Once the data are obtained, it might be nice to summarize the data. We could, of course, summarize the data using a histogram.

One primary disadvantage of using a **histogram** to summarize data is that the original data aren't preserved in the graph. A **stem-and-leaf plot**, on the other hand, summarizes the data and preserves the data at the same time.

Stem	Leaf
6	8
7	5 8
8	5 3 6 6 8 7 7 5 5
9	8 4 0 3 1 8 3 9 4 4 6 1 7 1 3 1 9 9 9 6
10	7 1 0 1 5 4 6 2 7 9 6 2 7 6 4 5 7 7 8 5
11	1 4 6 6 0 7 0
12	2 3 8
13	6
14	1

1|2 represents 12

Now, rather than looking at a list of 64 unordered IQs, we have a nice picture of the data that quite readily tells us that:

- 1.) the distribution of IQs is bell-shaped
- 2.) most of the IQs are in the 90s and 100s
- 3.) the smallest IQ in the data set is 68, while the largest is 141

**Example 1.9.13** Sam got his friends to do a long jump and got these results

2.3, 2.5, 2.5, 2.7, 2.8, 3.2, 3.6, 3.6, 4.5, 5.0

And here is the stem-and-leaf plot:

Stem	Leaf
2	3 5 5 7 8
3	2 6 6
4	5
5	0

$n = 10$ , 1|2 represents 1.2

Stem “2” Leaf “3” means 2.3

**Example 1.9.14** The results of 41 students' math tests (with a best possible score of 70) are recorded below:

31 49 19 62 50 24 45 23 51 32 48 55 60 40 35  
54 26 57 37 43 65 50 55 18 53 41 50 34 67 56  
44 4 54 57 39 52 45 35 51 63 42

1.) Is the variable discrete or continuous? Explain.

**Solution :** *A test score is a discrete variable. For example, it is not possible to have a test score of  $35.74542341 \dots$*  ■

2.) Prepare an ordered stem and leaf plot for the data and briefly describe what it shows.

**Solution :** *The lowest value is 4 and the highest is 67. Therefore, the stem and leaf plot that covers this range of values looks like this: The stem and leaf plot*

Stem	Leaf
0	4
1	8 9
2	3 4 6
3	1 2 4 5 5 7 9
4	0 1 2 3 4 5 5 8 9
5	0 0 0 1 1 2 3 4 4 5 5 6 7 7
6	0 2 3 5 7

$n = 41$ , 1|2 represents 12

*reveals that most students scored in the interval between 50 and 59. The large number of students who obtained high results could mean that the test was too easy, that most students knew the material well, or a combination of both.* ■

3.) Are there any outliers? If so, which scores?

**Solution :** *The result of 4 could be an outlier, since there is a large gap between this and the next result, 18.* ■

4.) Look at the stem and leaf plot from the side. Describe the distribution's main features such as:

(a) number of peaks

**Solution :** *The distribution has a single peak within the 50–59 interval.* ■

(b) symmetry

**Solution :** *No symmetry. The distribution is skewed to the left or negatively skewed.* ■

(c) value at the centre of the distribution

**Solution :** *Since there are 41 observations, the distribution centre (the median value) will occur at the 21st observation. Counting 21 observations up from the smallest, the centre is 48. (Note that the same value would have been obtained if 21 observations were counted down from the highest observation.)* ■

**Example 1.9.15** Fifteen people were asked how often they drove to work over 10 working days. The number of times each person drove was as follows:

5, 7, 9, 9, 3, 5, 1, 0, 0, 4, 3, 7, 2, 9, 8

Make an ordered stem and leaf plot for this table.

Stem	Leaf
0	0 0 1 2 3 3 4 5 5 7 7 8 9 9 9

$n = 10$ , 1|2 represents 12

**Exercise 1.24** Display data graphically and interpret graphs using Stem-Leaf plots

- 1.) For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):

33 42 49 49 53 55 55 61 63 67 68 68 69 69 72  
73 74 78 80 83 88 88 88 90 92 94 94 94 94 96 100

- 2.) For the University basketball team, scores for the last 30 games were as follows (smallest to largest):

32 32 33 34 38 40 42 42 43 44 46 47 47 48 48  
48 49 50 50 51 52 52 52 53 54 56 57 57 60 61

- 3.) The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data:

1.1 1.5 2.3 2.5 2.7 3.2 3.3 3.3 3.5 3.8 4.0  
4.2 4.5 4.5 4.7 4.8 5.5 5.6 6.5 6.7 12.3

Does the data seem to have any concentration of values?

**Solution :** *The value 12.3 may be an outlier. Values appear to concentrate at three and four kilometers* ■

**Exercise 1.25** Given a Stem-Leaf graph Determine the mode of the data.

Stem	Leaf
3	2 3
4	4 8
5	1 4 5
6	5 6 6 7
7	3 4 5 5 6 9
8	0 1 4 4 4 4 8 8 9
9	0 1 4 7

1|2 represents 12

### 1.9.3.4 Dotplots

A dotplot is an attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values. Each observation is represented by a dot above the corresponding location on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically. As with a stem-and-leaf display, a dotplot gives information about location, spread, extremes, and gaps.

**Example 1.9.16** The tragedy that befell the space shuttle Challenger and its astronauts in 1986 led to a number of studies to investigate the reasons for mission failure. Attention quickly focused on the behavior of the rocket engine's O-rings. Here is data consisting of observations on  $x = 0$ -ring temperature ( $^{\circ}F$ ) for each test firing or actual launch of the shuttle rocket engine (Presidential Commission on the Space Shuttle Challenger Accident, Vol. 1, 1986: 129–131).

84 49 61 40 83 67 45 66 70 69 80 58  
 68 60 67 72 73 70 57 63 70 78 52 67  
 53 67 75 61 70 81 76 79 75 76 58 31

Without any organization, it is difficult to get a sense of what a typical or representative temperature might be, whether the values are highly concentrated about a typical value or quite spread out, whether there are any gaps in the data, what percentage of the values are in the 60s, and so on. Figure 1.2 shows what is called a stem-and-leaf display of the data, as well as a histogram.

Stem	Leaf
3	1
4	0 5 9
5	2 3 7 8 8
6	0 1 1 3 6 7 7 7 7 8 9
7	0 0 0 0 2 3 5 5 6 6 8 9
8	0 1 3 4

Table 1.2:  $N=36$ : 1|2 represents 12

Figure 1.3 shows a dotplot for the O-ring temperature data introduced in Example 1.9.16. The data stretches out more at the lower end than at the upper end, and the smallest observation, 31, can fairly be described as an outlier.

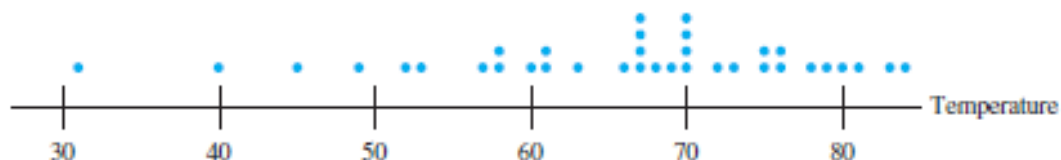


Figure 1.3: A dotplot of the 0-ring temperature data ( $^{\circ}F$ )

If the data set discussed had consisted of 50 or 100 temperature observations, each recorded to a tenth of a degree, it would have been much more cumbersome to construct a dotplot. Histograms technique is well suited to such situations.



## 1.10 Frequency Distributions

A frequency distribution is a tabular representation of data by class together with corresponding class frequencies. There are two types of frequency distributions,

- 1.) The simple or ungrouped frequency distributions
- 2.) The grouped frequency distributions

### 1.10.0.1 Frequency distribution for ungrouped data

Here we arrange numbers in ascending or descending order (array) and tally the scores and the number of tallies made per score represents the number of times that a score occurs in the distribution.

**Example 1.10.1** The ages of ten law students at a certain university are given as,

20, 18, 18, 19, 20, 21, 17, 18, 17, and 18.

Construct a frequency distribution for the above data?.

**Solution :** *The frequency distributions for ages of students is given in the table below,*

<i>Age</i>	<i>Tally</i>	<i>Frequency</i>
<i>17</i>	<i>//</i>	<i>2</i>
<i>18</i>	<i>////</i>	<i>4</i>
<i>19</i>	<i> </i>	<i>1</i>
<i>20</i>	<i>//</i>	<i>2</i>
<i>21</i>	<i> </i>	<i>1</i>
		$\sum f = 10$

■

**Exercise 1.26** Construct a frequency distribution for the following figures of heights obtained from 30 male students in a certain University.

66 70 68 67 71 60

64 70 68 65 64 61

71 66 67 65 68 59

67 65 68 66 69 58

66 65 65 71 70 56

**1.10.0.2 Grouped frequency distribution**

When summarizing large masses of row data its often useful to distribute data into classes or categories and to determine the number of individuals belonging to each category.

Frequency distributions represent data in a relatively compact form that give a good overall picture and contain information that is adequate for many purposes.

However some useful information is lost when such groupings are drawn for instance its impossible to determine the exact size of the lowest or the highest observation of the data.

To construct a frequency distribution table for grouped data involves two stages

- 1.) Choosing the classes
- 2.) Sorting or tallying the data into classes

The following are the steps taken when constructing a frequency table

- 1.) Identify the minimum and maximum observation of the set of data and find their difference (range).
- 2.) Divide the range (difference) into a convenient number of class intervals having the same size choose the appropriate class interval having in mind the following,
  - (a) Class intervals range between 5 and 20 the exact number of classes in a given situation will depend on the nature, magnitude and the range of the data.
  - (b) Ensure that each observation falls into one and only one class. None of the observations should lie/fall into gaps between classes, and the classes should be of the same width.
  - (c) We try to avoid open limit classes because they make it impossible to complete measures interest say the mean and others
- 3.) Determine the number of the observations that fall in each class interval i.e. class frequency and counting here is facilitated by placing a tally mark on the appropriate class and then counting the number of tally marks in each class.

**Example 1.10.2** Draw a grouped frequency distribution to show the heights of students given in the example in the Exercise 1.10.4 above.

Class	Tally	Frequency
56 - 58		2
59 - 61		3
62 - 64		2
65 - 67		12
68 - 70		8
71 - 73		3
		$\sum f = 30$

**Note 1.10.1** An open interval is when one could have a class say  $< 56$ .

Given any frequency distribution for grouped data the following are the basic concepts it involves;

1.) Class limits,

These are the smallest and largest observations that fall in a given class for example the class  $50 - 54$  has its class limits as 50 and 54.

2.) Class boundaries

These are values within which all possible observations which will fall in the given class must lie. Class boundaries are obtained by adding the highest class limit of one class to the lower class limit of the next higher class interval and dividing it by 2 for example for the class  $30 - 35$  in the following classes has the class boundary as,  $24 - 29, 30 - 35, 36 - 41, 42 - 47$  has the class boundary as,  $29.5 - 35.5$  from  $\frac{29 + 30}{2}$  and  $\frac{35 + 36}{2}$  respectively.

3.) Class width

It's the difference between the lower class boundary and the upper class boundary or is the difference between the lower and upper class limits plus 1 or it can be found by listing down all the members in a given interval and their number is the class width.

4.) Class mark

This is the mid point of the class. Its given as the average of the two class limits eg for the class  $30 - 35$  the class mark (mid mark) is

$$\frac{30 + 35}{2} = 32.5$$

5.) Relative frequency

It's the frequency of the class expressed as the percentage of the total frequency. If all frequencies in a frequency distribution are converted into the relative frequency then the resulting distribution is a relative distribution.

$$\text{Relative frequency of a given class} = \left( \frac{f_i}{\sum f_i} \times 100 \right)$$

where  $f_i$  is the frequency of the given class,  $\sum f_i$  is the total frequency or number of observations.

**Example 1.10.3** Given the following frequency distribution for the ages of students in a given faculty, find the relative frequency distribution of the data.

Class	Frequency
60 - 62	25
63 - 65	50
66 - 68	25
	$\sum f = 100$

**Solution :**

<i>Class</i>	<i>Frequency</i>	<i>Relative Frequency</i>
<i>60 - 62</i>	<i>25</i>	$\frac{25}{101} \times 100 = 24.75$
<i>63 - 65</i>	<i>50</i>	$\frac{50}{101} \times 100 = 49.51$
<i>66 - 68</i>	<i>26</i>	$\frac{26}{101} \times 100 = 25.74$
	$\sum f = 101$	

■

**Example 1.10.4** Data below is the number of days lost due to illness of a group of employees:

47 1 55 30 1 3 7 14 7 66 34 6 10 5 12 5 3 9 18 45  
5 8 44 42 46 6 4 24 24 34 11 2 3 13 5 5 3 4 4 1

Represent the data on the frequency distribution table.

**Solution :**

<i>Class</i>	<i>Frequency</i>	<i>Cumulative Frequency</i>	<i>Relative Frequency</i>
<i>0 - 9</i>	<i>22</i>	<i>22</i>	<i>0.55</i>
<i>10 - 19</i>	<i>6</i>	<i>28</i>	<i>0.15</i>
<i>20 - 29</i>	<i>2</i>	<i>30</i>	<i>0.05</i>
<i>30 - 39</i>	<i>3</i>	<i>33</i>	<i>0.075</i>
<i>40 - 49</i>	<i>5</i>	<i>38</i>	<i>0.125</i>
<i>50 - 59</i>	<i>1</i>	<i>39</i>	<i>0.025</i>
<i>60 - 69</i>	<i>1</i>	<i>40</i>	<i>0.025</i>

■

### 1.10.1 Graphical Presentation of Frequency Distribution

The graphical data presentation include the following;

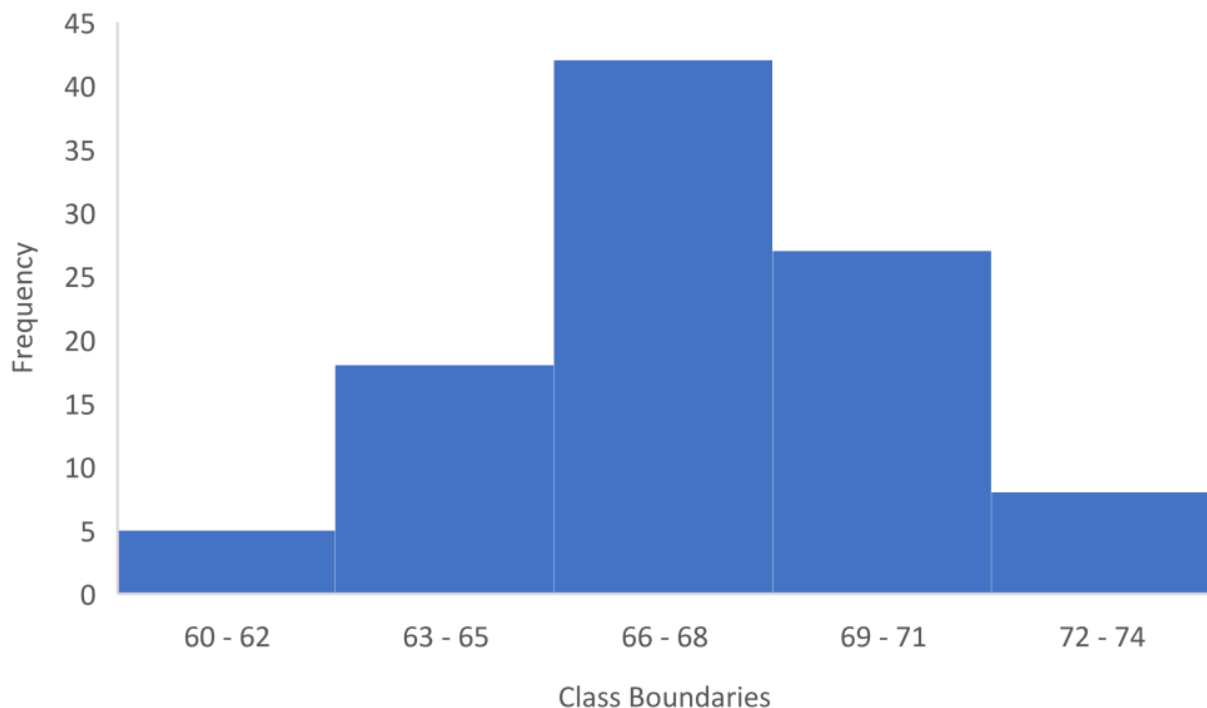
#### 1.10.1.1 Histogram

It's a graph or a plot of class frequencies against class boundaries or a plot of frequencies against classes, unlike the column graph or bar graphs whose bars are not attached together the bars of the histogram should be of the same width and attached to each other.

**Example 1.10.5** Given the following frequency distribution table of marks obtained by students in a certain faculty in a mathematics exam,

Class	Frequency
60 - 62	5
63 - 65	18
66 - 68	42
69 - 71	27
72 - 74	8
	$\sum f = 100$

A histogram to represent the data is given by



Here are some advantages of a histogram:

1.) Spot and track special trends

If you're plotting a time-related data, a histogram can help you spot and track trends.

For example, you noticed that the complaints you received every month seem to have a pattern. So, you started tracking the number of complaints you receive over a 12-month period.

Plotting the distribution frequency of the complaints might reveal specific months when the number of complaints spike.

2.) Checking Distribution Equality

Histograms are a great way to verify the equality of data points distribution.

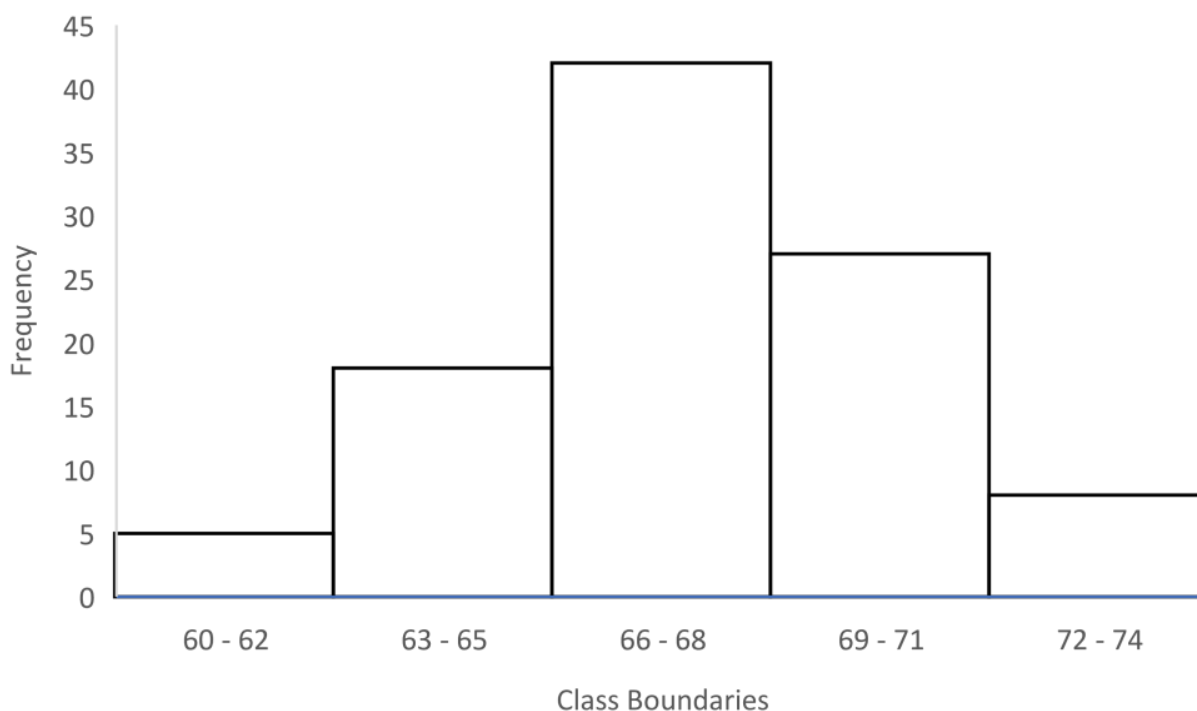
From a glance, you'll immediately see whether you had equal distribution or not.

3.) Data Spread

Basically, the bars in a histogram represents the data points belonging to that range.

One look and you'll know how the data is spread among the ranges.

A histogram could also be with no fill color

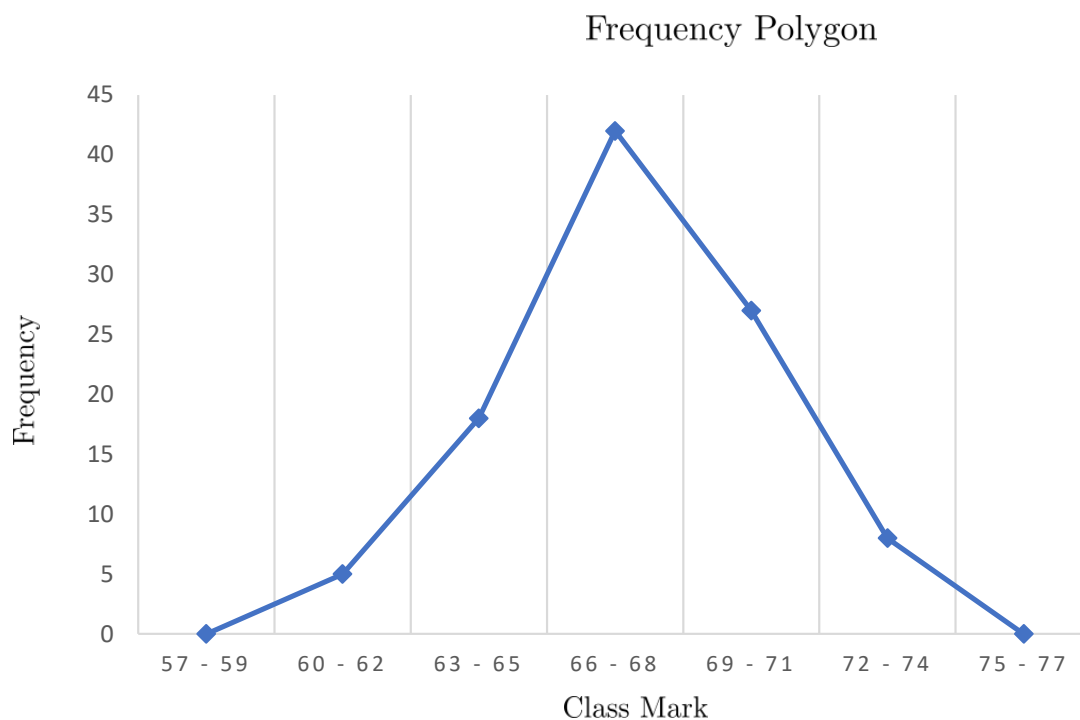


**Exercise 1.27** Draw a histogram for the data in Example 1.10.4 on page (pg. 39).

### 1.10.1.2 Frequency Polygon

It's a plot/graph of class frequencies against the class marks (midpoints) of classes and the successive points are joined by straight lines there by forming a graph similar to a line graph.

A frequency polygon is a line chart variation of a histogram, in which the bars are replaced by lines connecting the midpoints of the tops of the bars. Two class intervals with zero frequency are added to the distribution one at the beginning of the distribution and the other at the other end of the distribution for the purpose closing the polygon.



### 1.10.1.3 Cumulative Frequency Curve (Ogive)

**Definition 1.10.1** Cumulative frequency is the total of all values of frequencies less than the upper boundary of a given class.

The cumulative frequency curve is a plot/graph of cumulative frequency against class boundaries. The curve is useful in estimating the number of units (observations) either falling above or below a given value in the distribution.

The two methods of Ogives are:

#### 1.) Less than Ogive

In the case where cumulative frequency is plotted against upper class boundaries we have the "less than ogive" and such a graph is used to estimate the number of units (observations) falling below a given value of the distribution.

The frequencies of all preceding classes are added to the frequency of a class. This series is called the less than cumulative series. It is constructed by adding the first-class frequency to

the second-class frequency and then to the third class frequency and so on. The downward cumulation results in the less than cumulative series.

2.) Greater than Ogive or more than Ogive

In the case where cumulative frequency is plotted against the lower class boundaries then the resulting graph is called the “more than ogive” and this ogive is used to estimate the number of units (observations) falling above the given value of the distribution.

The frequencies of the succeeding classes are added to the frequency of a class. This series is called the more than or greater than cumulative series. It is constructed by subtracting the first class, second class frequency from the total, third class frequency from that and so on. The upward cumulation result is greater than or more than the cumulative series.

**Remark 1.10.1** The rising curve represents the less than Ogive, and the falling curve represents the greater than Ogive.

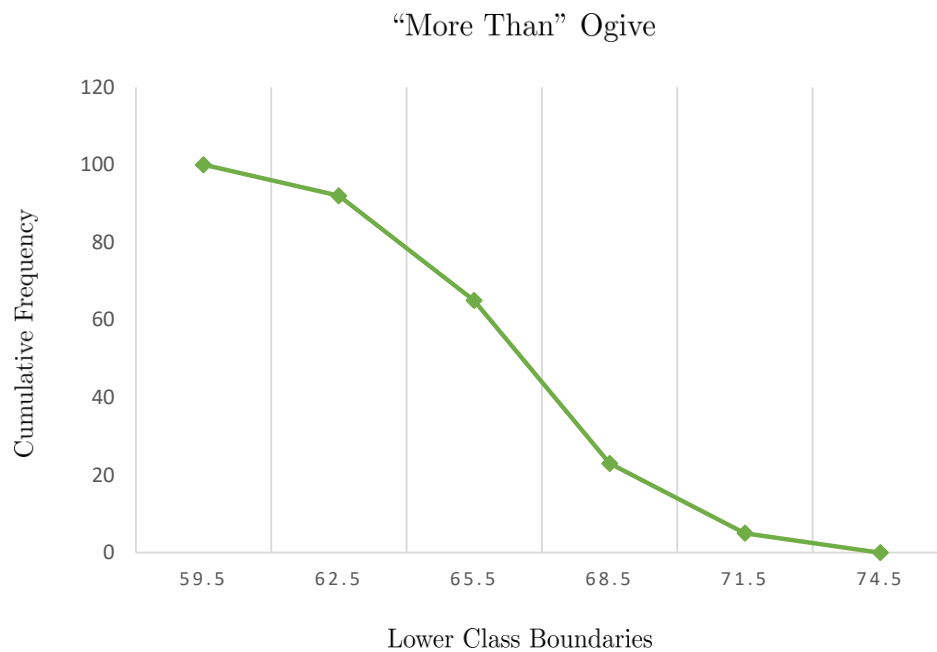
**Example 1.10.6** Cumulative frequencies with class boundaries table for Example 1.10.5

Class	Frequency	Cumulative Frequency (CF)	Class Mark	Class Boundary
60 - 62	5	5	61	59.5 - 62.5
63 - 65	18	23	64	62.5 - 65.5
66 - 68	42	65	67	65.5 - 68.5
69 - 71	27	92	70	68.5 - 71.5
72 - 74	8	100	73	71.5 - 74.5
	$\sum f = 100$			

**Example 1.10.7** “More than” Cumulative Frequency Table for data of Example 1.10.5.

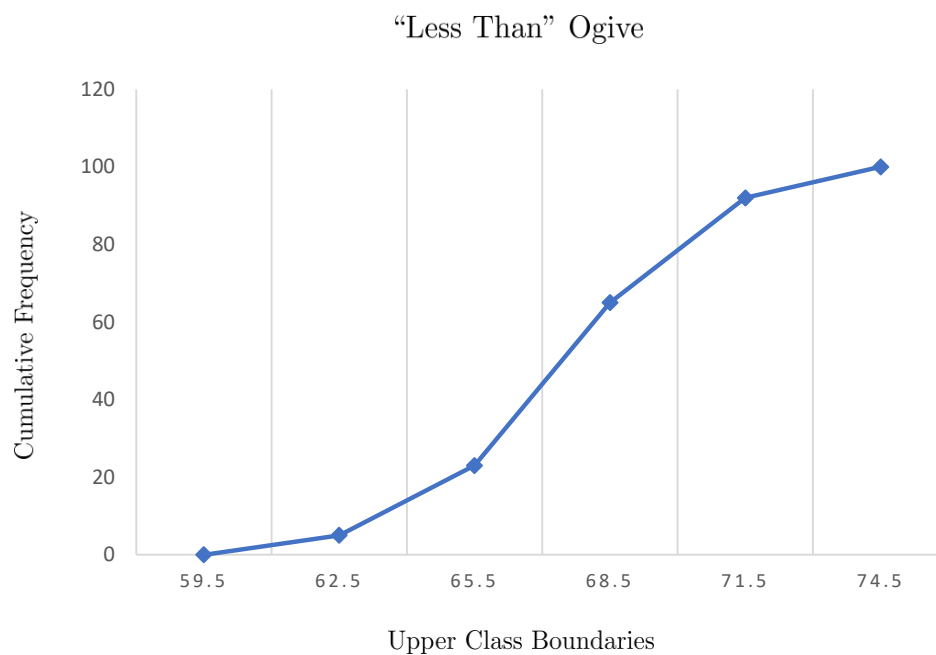
Marks	More than Cumulative Frequency	Upper Class Boundary
Marks more than 60	100	59.5
Marks more than 63	92	62.5
Marks more than 66	65	65.5
Marks more than 69	23	68.5
Marks more than 72	5	71.5
Marks more than 75	0	74.5





**Example 1.10.8** “Less than” Cumulative Frequency Table for data of Example 1.10.5.

Marks	Less than Cumulative Frequency	Lower Class Boundary
Marks less than 60	0	59.5
Marks less than 62	5	62.5
Marks less than 65	23	65.5
Marks less than 68	65	68.5
Marks less than 71	92	71.5
Marks less than 74	100	74.5



## 1.11 Chapter Examples

**Example 1.11.1** The average daily number of sunspots observed from Earth on the sun's disk for the years 1976 to 2015 are as follows:

18.4	39.3	131.0	220.1	218.9	198.9	162.4	91.0
60.5	20.6	14.8	33.9	123.0	211.1	191.8	203.3
133.0	76.1	44.9	25.1	11.6	28.9	88.3	136.3
173.9	170.4	163.6	99.3	65.3	45.8	24.7	12.6
4.2	4.8	24.9	80.8	84.5	94.0	113.3	69.8

- 1.) We shall make a frequency distribution for this data set using nine classes.
- 2.) The minimum data entry is 4.2, and the maximum data entry is 220.1. The range of the data set is thus

$$220.1 - 4.2 = 215.9,$$

and so we determine the class width to be

$$\text{Class width} = \frac{\text{Range}}{\text{Number of classes}} = \frac{215.9}{9} = 23.9888 \approx 24$$

- 3.) The lower limit of the first class we shall set to be 4.2, the minimum data entry. We obtain the lower limits of the other classes by repeatedly adding the class width of 24 to the lower limit of the first class:

4.2	28.2	52.2	76.2	100.2	124.2	148.2	172.2	196.2
-----	------	------	------	-------	-------	-------	-------	-------

Since our data consists of decimals to the tenths place, the upper limit of the first class must be 0.1 less than the lower limit of the second class, which is to say 28.1.

We obtain the upper limits of the other classes by repeatedly adding the class width of 24 to the upper limit of the first class:

28.1	52.1	76.1	100.1	124.1	148.1	172.1	196.1	220.1
------	------	------	-------	-------	-------	-------	-------	-------

- 4.) Tally the number of data entries that fall into each class. A column of tally marks can be included in the frequency distribution table, though it is not required.

The leftmost column of the frequency distribution table must list each of the nine classes, and the rightmost column must list the frequency of each class.

Class	Tally	Frequency
4.2 - 28.1		10
28.2 - 52.1		5
52.2 - 76.1		4
76.2 - 100.1		6
100.2 - 124.1		2
124.2 - 148.1		3
148.2 - 172.1		3
172.2 - 196.1		2
196.2 - 220.1		5

**Example 1.11.2** We extend the frequency distribution constructed in Example 1.11.1 to include class midpoints, relative class frequencies  $f_r$ , and cumulative frequencies  $f_c$ . We also do away with the tally column, which holds the same information as the existing column for frequency  $f$ .

Relative frequencies are given to the thousandths place in this example, since it results in no rounding. In general, if relative frequencies are given to  $k$  decimal places, then all relative frequencies should be written to  $k$  decimal places by inserting zeros as needed. Thus, in this case where relative frequencies are given to 3 decimal places, we write 0.25 as 0.250, and 0.1 as 0.100.

Class	Midpoint	$f$	$f_r$	$f_c$
4.2 - 28.1	16.15	10	0.250	10
28.2 - 52.1	40.15	5	0.125	15
52.2 - 76.1	64.15	4	0.100	19
76.2 - 100.1	88.15	6	0.150	25
100.2 - 124.1	112.15	2	0.050	27
124.2 - 148.1	136.15	3	0.075	30
148.2 - 172.1	160.15	3	0.075	33
172.2 - 196.1	184.15	2	0.050	35
196.2 - 220.1	208.15	5	0.125	40

**Exercise 1.28** Statistics is a body of concepts and methods which deal with data collection, organization, presentation, analysis and interpretation using different phenomena, to draw valid conclusions and making reasonable decisions on the basis of research,

- 1.) Mention any two sources that are used to obtain statistical data.
- 2.) Which of the two sources is most preferred by researchers? Justify your answer.

**Exercise 1.29** One of the methods of data presentation is the graphical method.

- 1.) Mention and give a sketch of each of the three types of bar graphs.
- 2.) The following table shows student enrollment at a certain university in Uganda ( Hypothetical data) in the courses as indicated.

Year	Educ	IT
1990	50	45
1991	68	73
1992	75	80
1993	40	50
1993	48	30

Represent the information using a multi-column and a multi-bar graph.

**Exercise 1.30** Distinguish between the following terms as used in statistics.

- 1.) (a) A parameter and a statistic.  
(b) A sample and a population.
- 2.) Why do you think most researchers choose to use a sample instead of the population in their surveys.

**Exercise 1.31**

- 1.)) Define what is meant by a variable.
- 2.)) Distinguish between the Qualitative and Quantitative classification of data.
- 3.)) What do we mean by a sample in relation to a population.

**Exercise 1.32** With relative examples, briefly explain the scales below.

- 1.) interval scale
- 2.) nominal scale
- 3.) ratio scale
- 4.) ordinal scale

**Exercise 1.33** *Fortune* magazine provides data on how the 500 largest U.S industrial corporations rank in terms of revenues and profits. Data for a sample of *Fortune* 500 companies are given in the table below.

A sample 10 *Fortune* 500 companies

Company	Revenue (\$ millions)	Profit (\$ millions)	Industry Code
US Airways Group	8688.0	538.0	3
International Paper	19500.0	213.0	23
Tyson Foods	7414.1	25.1	20
Hewlett-Packard	47061.1	2945.0	13
Intel	26273.0	6068.0	49
Northrup Grumman	8902.0	214.0	2
Seagate Technology	6819.0	-530.0	11
Unisys	7208.4	387.0	10
Westvaco	2904.7	132.0	23
Campbell Soup	7505.0	660.0	20

(a) How many elements are in this data set?

**Solution :** 10



(b) What is the population?

**Solution :** 500 fortune largest companies



(c) Compute the average revenue for the sample.

**Solution :** \$14,227.59



(d) Using the results in (iii), what is the estimate of the average revenues for the population.

**Solution :** \$14,227.59



**Exercise 1.34** Sketch a “Frequency histogram” and a “Relative frequency histogram” with horizontal axis featuring class boundaries for Example 1.11.2 on page 46.

**Exercise 1.35** Statistics is a body of concepts and methods which deal with data collection, organization, presentation, analysis and interpretation using different phenomena, to draw valid conclusions and making reasonable decisions on the basis of research,

- 1.) Mention any two sources that are used to obtain statistical data.
- 2.) Which of the two sources is most preferred by researchers? Justify your answer.

**Exercise 1.36** One of the methods of data presentation is the graphical method.

- 1.) Mention and give a sketch of each of the three types of bar graphs.
- 2.) The following table shows student enrollment at Department of Distance Education ( Hypothetical data) in the courses as indicated.

Year	Education	Science
1990	50	45
1991	68	73
1992	75	80
1993	40	50
1993	48	30

Represent the information using a multi-column graph.

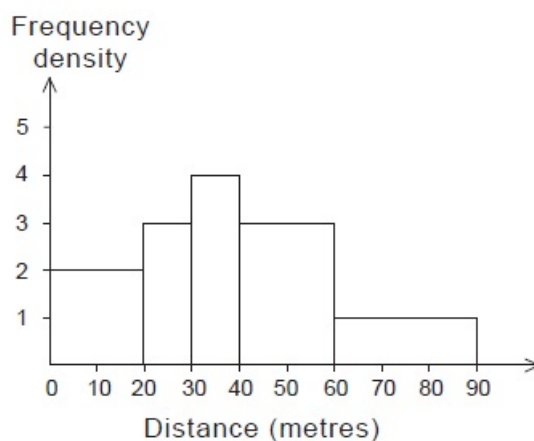
**Exercise 1.37** In descriptive statistics, we learn that it is important to summarize data.

- 1.) Why is it important to summarize data.
- 2.) what do frequency polygon and pie charts represent and how are they constructed.

**Exercise 1.38** “Data can be collected from existing sources or from surveys and experiments designed to obtain a new data”. Describe the two sources of data in the statement above.

**Exercise 1.39** Most of the statistical information in newspapers, magazines, company reports, and other publications consists of data summarized and presented in a form that is easy for the reader to understand. Such summaries of data, which may be tabular, graphical or numerical (descriptive statistics). With an examples, explain the tabular, graphical, or numerical representation of data.

**Example 1.11.3** The histogram below shows the distribution of distances in a throwing competition.



1.) How many competitors threw less than 40 metres?

**Solution :** *Using;*

$$\text{class width} \times \text{frequency density} = \text{frequency}$$

*gives the following table.*

<i>Interval</i>	<i>Class width</i>	<i>Frequency density</i>	<i>Actual frequency</i>
0 – 20	20	2	$2 \times 20 = 40$
20 – 30	10	3	$3 \times 10 = 30$
30 – 40	10	4	$4 \times 10 = 40$
40 – 60	20	3	$3 \times 20 = 60$
60 – 90	30	1	$1 \times 30 = 30$

*Answer is*  $40 + 30 + 40 = 110$ . ■

2.) How many competitors were there in the competition?

**Solution :**  $40 + 30 + 40 + 60 + 30 = 200$ . ■

**Exercise 1.40** Consider the following table:

Age	Men	Women
14 – 16	1637	129
17 – 20	9268	238
21 – 24	7255	235
25 – 29	5847	188
30 – 39	7093	236
40 – 49	3059	132
50 – 59	1128	35
60 and over	262	7

Table 1.5: Age and Sex of Prisoners, England and Wales 1981

Use the information on the ages of sentenced prisoners in the table opposite to draw a composite bar chart. Ignore the uneven group sizes.

**Example 1.11.4** Data are obtained on the topics given below. State whether they are discrete or continuous data.

- 1). The number of days on which rain falls in a month for each month of the year.
- 2). ]The mileage travelled by each of a number of salesmen.
- 3). The time that each of a batch of similar batteries lasts.
- 4). The amount of money spent by each of several families on food.

**Solution :**

- 1). *The number of days on which rain falls in a given month must be an integer value and is obtained by **counting** the number of days. Hence, these data are **discrete**.*
- 2). *A salesman can travel any number of miles (and parts of a mile) between certain limits and these data are **measured**. Hence the data are **continuous**.*
- 3). *The time that a battery lasts is **measured** and can have any value between certain limits. Hence these data are **continuous**.*
- 4). *The amount of money spent on food can only be expressed correct to the nearest pence, the amount being **counted**. Hence, these data are **discrete***

■



**Example 1.11.5** In 1 and 2, state whether data relating to the topics given are discrete or continuous.

- 1). (a) The amount of petrol produced daily, for each of 31 days, by a refinery.  
 (b) The amount of coal produced daily by each of 15 miners.  
 (c) The number of bottles of milk delivered daily by each of 20 milkmen.  
 (d) The size of 10 samples of rivets produced by a machine.

**Solution :**

(a) *continuous*

(c) *discrete*

(b) *continuous*

(d) *continuous*

■

- 2). a). The number of people visiting an exhibition on each of 5 days.  
 b). The time taken by each of 12 athletes to run 100 metres.  
 c). The value of stamps sold in a day by each of 20 post offices.  
 d). The number of defective items produced in each of 10 one-hour periods by a machine.

**Solution :**

a). *discrete*

c). *discrete*

b). *continuous*

d). *discrete*

■

**Exercise 1.41** The distance in miles travelled by four salesmen in a week are as shown below.

Salesmen	<i>P</i>	<i>Q</i>	<i>R</i>	<i>S</i>
Distance travelled (miles)	413	264	597	143

Use a horizontal bar chart to represent these data diagrammatically

**Solution :** *Equally spaced horizontal rectangles of any width, but whose length is proportional to the distance travelled, are used. Thus, the length of the rectangle for salesman P is proportional to 413 miles, and so on. The horizontal bar chart depicting these data is shown in Fig. 1.4*

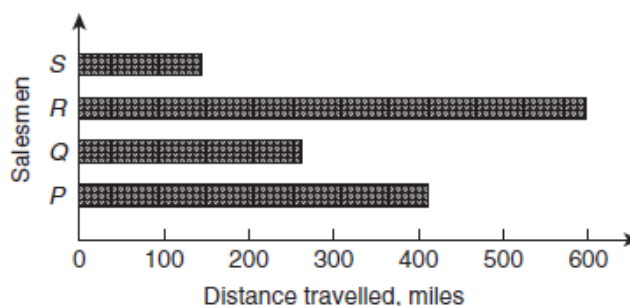


Figure 1.4

■

**Example 1.11.6** The number of issues of tools or materials from a store in a factory is observed for seven, one-hour periods in a day, and the results of the survey are as follows:

Period	1	2	3	4	5	6	7
Number of issues	34	17	9	5	27	13	6

Present these data on a vertical bar chart.

**Solution :** *In a vertical bar chart, equally spaced vertical rectangles of any width, but whose height is proportional to the quantity being represented, are used. Thus the height of the rectangle for period 1 is proportional to 34 units, and so on. The vertical bar chart depicting these data is shown in Fig.1.5*

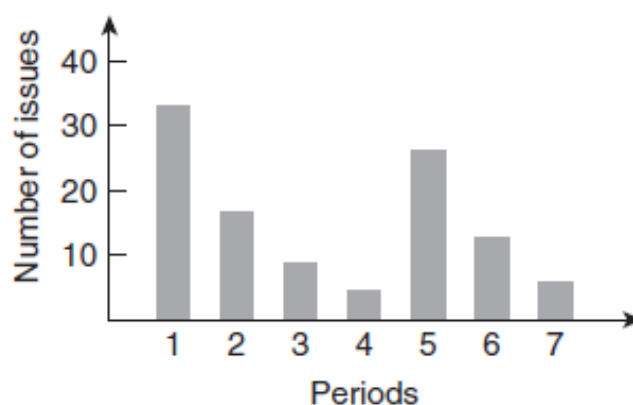


Figure 1.5

■

**Example 1.11.7** The number of various types of dwellings sold by a company annually over a three-year period are as shown below. Draw percentage component bar charts to present these data.

	Year 1	Year 2	Year 3
4-roomed bungalows	24	17	7
5-roomed bungalows	38	71	118
4 roomed houses	44	50	53
5 roomed houses	64	82	147
6 roomed houses	30	30	25

**Solution :** *A table of percentage relative frequency values, correct to the nearest 1%, is the first requirement. Since,*

$$\text{percentage relative frequency} = \frac{\text{frequency of member} \times 100}{\text{total frequency}}$$

then for 4-roomed bungalows in year 1:

$$\text{percentage relative frequency} = \frac{24 \times 100}{24 + 38 + 44 + 64 + 30} = 12\%$$

The percentage relative frequencies of the other types of dwellings for each of the three years are similarly calculated and the results are as shown in the table below.

	Year 1	Year 2	Year 3
4-roomed bungalows	12%	7%	2%
5-roomed bungalows	19%	28%	34%
4 roomed houses	22%	20%	15%
5 roomed houses	32%	33%	42%
6 roomed houses	15%	12%	7%

The percentage component bar chart is produced by constructing three equally spaced rectangles of any width, corresponding to the three years. The heights of the rectangles correspond to 100% relative frequency, and are subdivided into the values in the table of percentages shown above. A key is used (different types of shading or different colour schemes) to indicate corresponding percentage values in the rows of the table of percentages.

The percentage component bar chart is shown in Fig.1.6

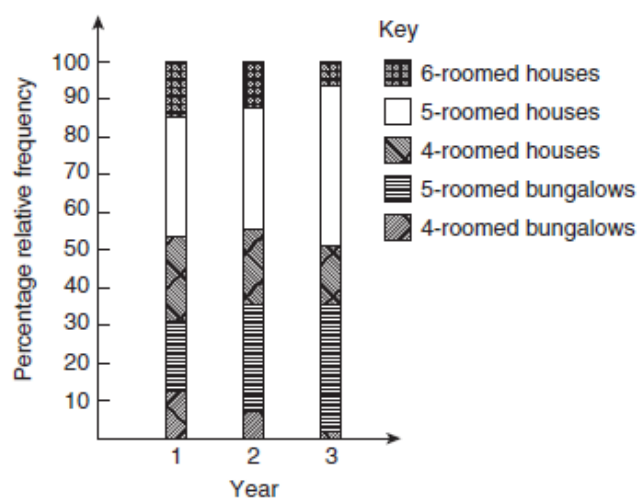


Figure 1.6

■

**Example 1.11.8** The retail price of a product costing £2 is made up as follows: materials 10 p, labour 20 p, research and development 40 p, overheads 70 p, profit 60 p. Present these data on a pie diagram

**Solution :** *A circle of any radius is drawn, and the area of the circle represents the whole, which in this case is £2. The circle is subdivided into sectors so that the areas of the sectors are proportional to the parts, i.e. the parts which make up the total retail price. For the area of a sector to be proportional to a part, the angle at the centre of the circle must be proportional to that part. The whole, £2 or 200 p, corresponds to 360°. Therefore,*

$$10 \text{ p corresponds to } 360 \times \frac{10}{200} \text{ degrees, i.e. } 18^\circ$$

$$20 \text{ p corresponds to } 360 \times \frac{20}{200} \text{ degrees, i.e. } 36^\circ$$

*and so on, giving the angles at the centre of the circle for the parts of the retail price as: 18°, 36°, 72°, 126° and 108°, respectively. The pie diagram is shown in Fig.1.7*

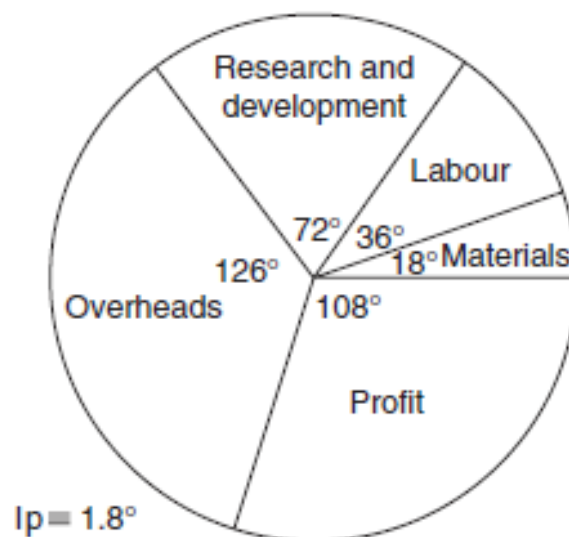


Figure 1.7

■

**Example 1.11.9** A characteristic or measure obtained by using the data values from a sample is called a

- A. quartile      B. parameter      C. statistic      D. percentile      E. quantitative

C

**Example 1.11.10**

- 1). Using the data given in Fig.1.4 only, calculate the amount of money paid to each salesman for travelling expenses, if they are paid an allowance of 37 p per mile.
- 2). Using the data presented in Fig.1.6, comment on the housing trends over the three-year period.
- 3). Determine the profit made by selling 700 units of the product shown in Fig.1.7.

**Solution :**

- 1). *By measuring the length of rectangle P the mileage covered by salesman P is equivalent to 413 miles. Hence salesman P receives a travelling allowance of*

$$\frac{£413 \times 37}{100} \quad i.e \quad £152.81$$

*Similarly, for salesman Q, the miles travelled are 264 and this allowance is*

$$\frac{£264 \times 37}{100} \quad i.e \quad £97.68$$

*Salesman R travels 597 miles and he receives*

$$\frac{£597 \times 37}{100} \quad i.e \quad £220.89$$

*Finally, salesman S receives*

$$\frac{£143 \times 37}{100} \quad i.e \quad £52.91$$

- 2). *An analysis of Fig.1.6 shows that 5-roomed bungalows and 5-roomed houses are becoming more popular, the greatest change in the three years being a 15% increase in the sales of 5-roomed bungalows.*
- 3). *Since  $1.8^\circ$  corresponds to 1 p and the profit occupies  $108^\circ$  of the pie diagram, then the profit per unit is  $\frac{108 \times 1}{1.8}$ , that is, 60 p The profit when selling 700 units of the product is  $£ \frac{700 \times 60}{100}$ , that is, £420*

■

**Example 1.11.11** When all subjects under study are used, the group is called a

- A. study group      B. sample      C. population      D. small group

C

**Exercise 1.42**

- 1). The number of vehicles passing a stationary observer on a road in six ten-minute intervals is as shown. Draw a pictogram to represent these data.

Period of Time	1	2	3	4	5	6
Number of Vehicles	35	44	62	68	49	41

If one symbol is used to represent 10 vehicles, working correct to the nearest 5 vehicles, gives 3.5, 4.5, 6, 7, 5 and 4 symbols respectively.

- 2). The number of components produced by a factory in a week is as shown below:

Day	Mon	Tues	Wed	Thurs	Fri
Number of Components	1580	2190	1840	2385	1280

Show these data on a pictogram. If one symbol represents 200 components, working correct to the nearest 100 components gives: Mon 8, Tues 11, Wed 9, Thurs 12 and Fri 6.5

- 3). For the data given in Problem 1 above, draw a horizontal bar chart.  
6 equally spaced horizontal rectangles, whose lengths are proportional to 35, 44, 62, 68, 49 and 41, respectively.
- 4). Present the data given in Problem 2 above on a horizontal bar chart.  
5 equally spaced horizontal rectangles, whose lengths are proportional to 1580, 2190, 1840, 2385 and 1280 units, respectively.
- 5). For the data given in Problem 1 above, construct a vertical bar chart.  
6 equally spaced vertical rectangles, whose heights are proportional to 35, 44, 62, 68, 49 and 41 units, respectively.
- 6). Depict the data given in Problem 2 above on a vertical bar chart.  
5 equally spaced vertical rectangles, whose heights are proportional to 1580, 2190, 1840, 2385 and 1280 units, respectively.
- 7). A factory produces three different types of components. The percentages of each of these components produced for three, one month periods are as shown below. Show this information on percentage component bar charts and comment on the changing trend in the percentages of the types of component produced.

Month	1	2	3
Component $P$	20	35	40
Component $Q$	45	40	35
Component $R$	35	25	25

Three rectangles of equal height, subdivided in the percentages shown in the columns above.  $P$  increases by 20% at the expense of  $Q$  and  $R$

- 8). A company has five distribution centres and the mass of goods in tonnes sent to each centre during four, one-week periods, is as shown.

Week	1	2	3	4
Centre <i>A</i>	147	160	174	158
Centre <i>B</i>	54	63	77	69
Centre <i>C</i>	283	251	237	211
Centre <i>D</i>	97	104	117	144
Centre <i>E</i>	224	218	203	194

Use a percentage component bar chart to present these data and comment on any trends.

Four rectangles of equal heights, subdivided as follows:

week 1: 18%, 7%, 35%, 12%, 28%

week 3: 22%, 10%, 29%, 14%, 25%

week 2: 20%, 8%, 32%, 13%, 27%

week 4: 20%, 9%, 27%, 19%, 25%.

Little change in centres *A* and *B*, a reduction of about 8% in *C*, an increase of about 7% in *D* and a reduction of about 3% in *E*.

- 9). The employees in a company can be split into the following categories: managerial 3, supervisory 9, craftsmen 21, semi-skilled 67, others 44. Shown these data on a pie diagram.  
A circle of any radius, subdivided into sectors having angles of  $7.5^\circ$ ,  $22.5^\circ$ ,  $52.5^\circ$ ,  $167.5^\circ$  and  $110^\circ$ , respectively.
- 10). The way in which an apprentice spent his time over a one-month period is as follows: drawing office 44 hours, production 64 hours, training 12 hours, at college 28 hours.  
Use a pie diagram to depict this information.  
A circle of any radius, subdivided into sectors having angles of  $107^\circ$ ,  $156^\circ$ ,  $29^\circ$  and  $68^\circ$ , respectively.
- 11). a). With reference to Fig.1.7, determine the amount spent on labour and materials to produce 1650 units of the product.  
b). If in year 2 of Fig.1.6, 1% corresponds to 2.5 dwellings, how many bungalows are sold in that year.  
[(a) £495, (b) 88]
- 12). a). If the company sell 23 500 units per annum of the product depicted in Fig.1.7, determine the cost of their overheads per annum.  
b). If 1% of the dwellings represented in year 1 of Fig.1.6 corresponds to 2 dwellings, find the total number of houses sold in that year.  
[(a) £16450, (b) 138]

**Example 1.11.12** The data given below refer to the gain of each of a batch of 40 transistors, expressed correct to the nearest whole number. Form a frequency distribution for these data having seven classes

81 83 87 74 76 89 82 84  
 86 76 77 71 86 85 87 88  
 84 81 80 81 73 89 82 79  
 81 79 78 80 85 77 84 78  
 83 79 80 83 82 79 80 77

**Solution :** *The range of the data is the value obtained by taking the value of the smallest member from that of the largest member. Inspection of the set of data shows that, range =  $89 - 71 = 18$ . The size of each class is given approximately by range divided by the number of classes. Since 7 classes are required, the size of each class is  $18/7$ , that is, approximately 3. To achieve seven equal classes spanning a range of values from 71 to 89, the class intervals are selected as: 70 – 72, 73 – 75, and so on.*

*To assist with accurately determining the number in each class, a **tally diagram** is produced, as shown in Table 1.6. This is obtained by listing the classes in the left-hand column, and then inspecting each of the 40 members of the set in turn and allocating them to the appropriate classes by putting ‘1s’ in the appropriate rows. Every fifth ‘1’ allocated to a particular row is shown as an oblique line crossing the four previous ‘1s’, to help with final counting.*

*A **frequency distribution** for the data is shown in Table 1.7 and lists classes and*

Table 1.6

Class	Tally
70–72	
73–75	
76–78	
79–81	
82–84	
85–87	
88–90	

Table 1.7

Class	Class mid-point	Frequency
70–72	71	1
73–75	74	2
76–78	77	7
79–81	80	12
82–84	83	9
85–87	86	6
88–90	89	3

*their corresponding frequencies, obtained from the tally diagram. (Class mid-point values are also shown in the table, since they are used for constructing the histogram for these data.* ■



**Example 1.11.13** The amount of money earned weekly by 40 people working part-time in a factory, correct to the nearest £10, is shown below. Form a frequency distribution having 6 classes for these data.

80 90 70 110 90 160 110 80  
140 30 90 50 100 110 60 100  
80 90 110 80 100 90 120 70  
130 170 80 120 100 110 40 110  
50 100 110 90 100 70 110 80

**Solution :** *Inspection of the set given shows that the majority of the members of the set lie between £80 and £110 and*

Table 1.8

Class	Frequency
20–40	2
50–70	6
80–90	12
100–110	14
120–140	4
150–170	2

*that there are a much smaller number of extreme values ranging from £30 to £170. If equal class intervals are selected, the frequency distribution obtained does not give as much information as one with unequal class intervals. Since the majority of members are between £80 and £100, the class intervals in this range are selected to be smaller than those outside of this range. There is no unique solution and one possible solution is shown in Table 1.8.* ■

**Example 1.11.14** Draw a histogram for the data given in Table 1.8

**Solution :** When dealing with unequal class intervals, the histogram must be drawn so that the areas, (and not the heights), of the rectangles are proportional to the frequencies of the classes. The data given are shown in columns 1 and 2 of Table 1.9. Columns 3 and 4 give the upper and lower class boundaries, respectively. In column 5, the class ranges (i.e. upper class boundary minus lower class boundary values) are listed. The heights of the rectangles are proportional to the ratio  $\frac{\text{frequency}}{\text{class range}}$  as shown in column 6. The histogram is shown in Fig. 1.8. ■

Table 1.9

1 Class	2 Frequency	3 U.C.B	4 L.C.B	5 Class range	6 Height of rectangle
20–40	2	45	15	30	$\frac{2}{30} = \frac{1}{15}$
50–70	6	75	45	30	$\frac{6}{30} = \frac{3}{15}$
80–90	12	95	75	20	$\frac{12}{30} = \frac{9}{15}$
100–110	14	115	95	20	$\frac{14}{20} = \frac{10\frac{1}{2}}{15}$
120–140	4	145	115	30	$\frac{4}{30} = \frac{2}{15}$
150–170	2	175	145	30	$\frac{2}{30} = \frac{1}{15}$

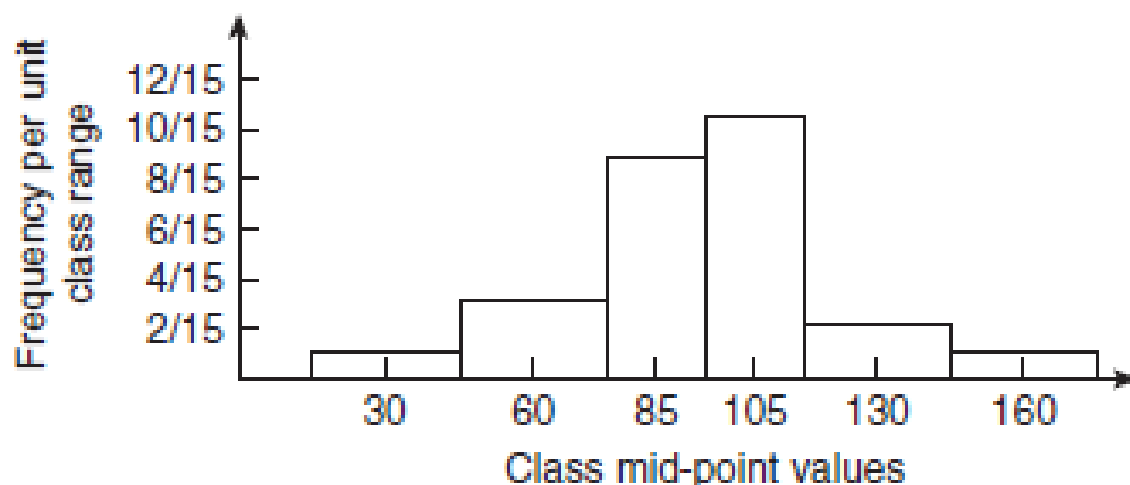


Figure 1.8

**Example 1.11.15** The masses of 50 ingots in kilograms are measured correct to the nearest 0.1 kg and the results are as shown below. Produce a frequency distribution having about 7 classes for these data and then present the grouped data as (a) a frequency polygon and (b) histogram.

8.0 8.6 8.2 7.5 8.0 9.1 8.5 7.6 8.2 7.8  
8.3 7.1 8.1 8.3 8.7 7.8 8.7 8.5 8.4 8.5  
7.7 8.4 7.9 8.8 7.2 8.1 7.8 8.2 7.7 7.5  
8.1 7.4 8.8 8.0 8.4 8.5 8.1 7.3 9.0 8.6  
7.4 8.2 8.4 7.7 8.3 8.2 7.9 8.5 7.9 8.0

**Solution :** *The range of the data is the member having the largest value minus the member having the smallest value. Inspection of the set of data shows that:*

$$\text{range} = 9.1 - 7.1 = 2.0$$

*The size of each class is given approximately by*

$$\frac{\text{range}}{\text{number of classes}}$$

*Since about seven classes are required, the size of each class is  $2.0/7$ , that is approximately 0.3, and thus the class limits are selected as 7.1 to 7.3, 7.4 to 7.6, 7.7 to 7.9, and so on. The **class mid-point** for the 7.1 to 7.3 class is  $\frac{7.35 + 7.05}{2}$  i.e. 7.2, for the 7.4 to 7.6 class is  $\frac{7.65 + 7.35}{2}$ , i.e. 7.5, and so on.*

*To assist with accurately determining the number in each class, a **tally diagram** is produced as shown in Table1.10. This is obtained by listing the classes in the left-hand column and then inspecting each of the 50 members of the set of data in turn and allocating it to the appropriate class by putting a '1' in the appropriate row. Each fifth '1' allocated to a particular row is marked as an oblique line to help with final counting.*

*A **frequency distribution** for the data is shown in Table1.11 and lists classes and their corresponding frequencies. Class mid-points are also shown in this table, since they are used when constructing the frequency polygon and histogram.*

*A **frequency polygon** is shown in Fig.1.9, the co-ordinates corresponding to the class mid-point/ frequency values, given in Table1.11. The co-ordinates are joined by straight lines and the polygon is 'anchored down' at each end by joining to the next class mid-point value and zero frequency.*

*A **histogram** is shown in Fig.1.10, the width of a rectangle corresponding to (upper class boundary*

*value — lower class boundary value) and height corresponding to the class frequency. The easiest way to draw a histogram is to mark class mid-point values on the horizontal scale and to draw the rectangles symmetrically about the appropriate class mid-point values and touching one another. A histogram for the data given in Table1.11 is shown in Fig.1.10. ■*

Table 1.10

Class	Tally
7.1 - 7.3	
7.4 - 7.6	
7.7 - 7.9	
8.0 - 8.2	
8.3 - 8.5	
8.6 - 8.8	
8.9 - 9.1	

Table 1.11

Class	Class mid-point	Frequency
7.1 - 7.3	7.2	3
7.4 - 7.6	7.5	5
7.7 - 7.9	7.8	9
8.0 - 8.2	8.1	14
8.3 - 8.5	8.4	11
8.6 - 8.8	8.7	6
8.9 - 9.1	9.0	2

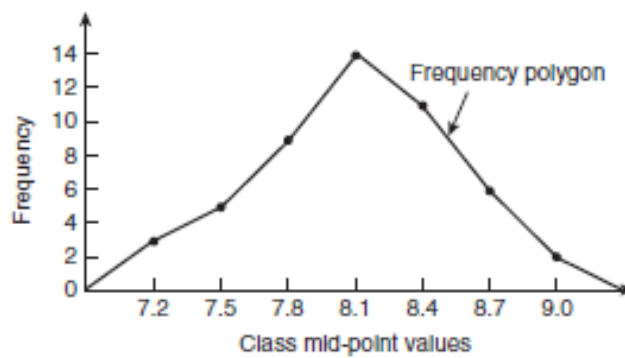


Figure 1.9

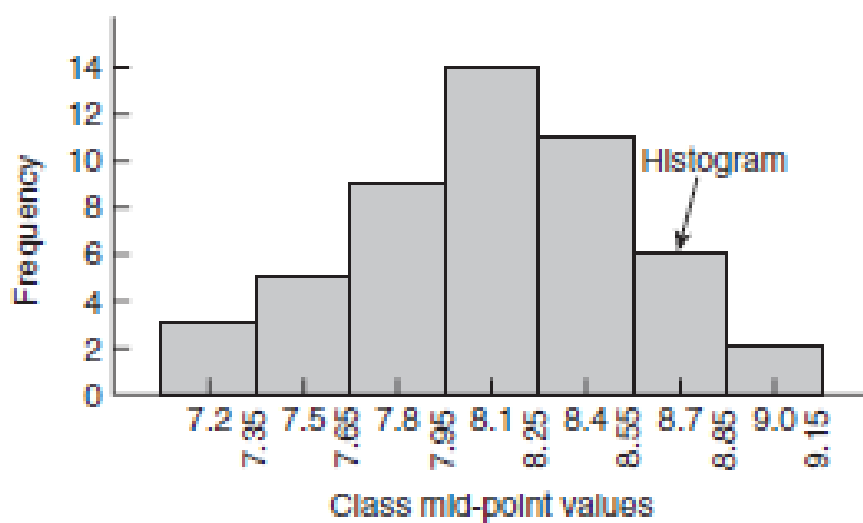


Figure 1.10

**Example 1.11.16** The frequency distribution for the masses in kilograms of 50 ingots is:

7.1 to 7.3	7.4 to 7.6	7.7 to 7.9	8.0 to 8.2	8.3 to 8.5	8.6 to 8.8	8.9 to 9.1
3	5	9	14	11	6	2

Form a cumulative frequency distribution for these data and draw the corresponding ogive

**Solution :** A cumulative frequency distribution is a table giving values of cumulative frequency for the values of upper class boundaries, and is shown in Table 1.12. Columns 1 and 2 show the classes and their frequencies. Column 3 lists the upper class boundary values for the classes given in column 1. Column 4 gives the cumulative frequency values for all frequencies less than the upper class boundary values given in column 3. Thus, for example, for the 7.7 to 7.9 class shown in row 3, the cumulative frequency value is the sum of all frequencies having values of less than 7.95, i.e.  $3 + 5 + 9 = 17$ , and so on. The ogive for the cumulative frequency distribution given in Table 1.12 is shown in Fig. 1.11. The co-ordinates corresponding to each upper class boundary/cumulative frequency value are plotted and the co-ordinates are joined by straight lines (— not the best curve drawn through the co-ordinates as in experimental work). The ogive is ‘anchored’ at its start by adding the co-ordinate (7.05, 0). ■

Table 1.12

1 Class	2 Frequency	3 Upper Class Boundary Less than	4 Cumulative Frequency
7.1–7.3	3	7.35	3
7.4–7.6	5	7.65	8
7.7–7.9	9	7.95	17
8.0–8.2	14	8.25	31
8.3–8.5	11	8.55	42
8.6–8.8	6	8.85	48
8.9–9.1	2	9.15	50

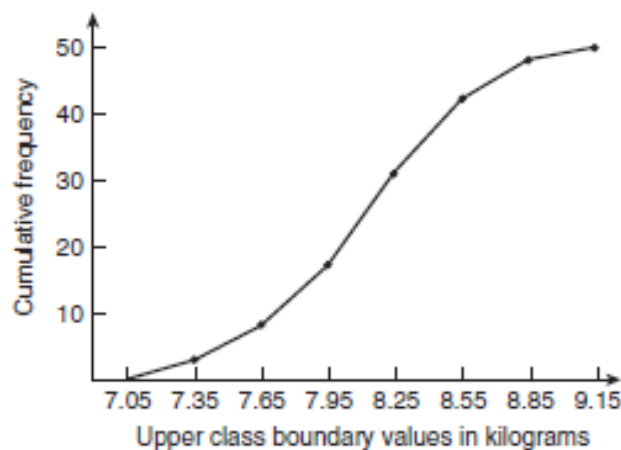


Figure 1.11

**Exercise 1.43**

- 1). The mass in kilograms, correct to the nearest one-tenth of a kilogram, of 60 bars of metal are as shown. Form a frequency distribution of about 8 classes for these data.

39.8 40.3 40.6 40.0 39.6 39.6 40.2 40.3 40.4 39.8  
 40.2 40.3 39.9 39.9 40.0 40.1 40.0 40.1 40.1 40.2  
 39.7 40.4 39.9 40.1 39.9 39.5 40.0 39.8 39.5 39.9  
 40.1 40.0 39.7 40.4 39.3 40.7 39.9 40.2 39.9 40.0  
 40.1 39.7 40.5 40.5 39.9 40.8 40.0 40.2 40.0 39.9  
 39.8 39.7 39.5 40.1 40.2 40.6 40.1 39.7 40.2 40.3

There is no unique solution, but one solution is:

39.3–39.4	39.5–39.6	39.7–39.8	39.9–40.0	40.1–40.2	40.3–40.4	40.5–40.6	40.7–40.8
1	5	9	17	15	7	4	2

- 2). Draw a histogram for the frequency distribution given in the solution of Problem 1.

Rectangles, touching one another, having mid-points of 39.35, 39.55, 39.75, 39.95,  $\dots$  and heights of 1, 5, 9, 17,  $\dots$

- 3). The information given below refers to the value of resistance in ohms of a batch of 48 resistors of similar value. Form a frequency distribution for the data, having about 6 classes

and draw a frequency polygon and histogram to represent these data diagrammatically.

21.0	22.4	22.8	21.5	22.6	21.1	21.6	22.3
22.9	20.5	21.8	22.2	21.0	21.7	22.5	20.7
23.2	22.9	21.7	21.4	22.1	22.2	22.3	21.3
22.1	21.8	22.0	22.7	21.7	21.9	21.1	22.6
21.4	22.4	22.3	20.9	22.8	21.2	22.7	21.6
22.2	21.6	21.3	22.1	21.5	22.0	23.4	21.2

There is no unique solution, but one solution is:

20.5–20.9	21.0–21.4	21.5–21.9	22.0–22.4	22.5–22.9	23.0–23.4
3	10	11	13	9	2

- 4). The time taken in hours to the failure of 50 specimens of a metal subjected to fatigue failure tests are as shown. Form a frequency distribution, having about 8 classes and unequal class intervals, for these data.

28	22	23	20	12	24	37	28	21	25
21	14	30	23	27	13	23	7	26	19
24	22	26	3	21	24	28	40	27	24
20	25	23	26	47	21	29	26	22	33
27	9	13	35	20	16	20	25	18	22

There is no unique solution, but one solution is:

1–10	11–19	20–22	23–25	26–28	29–38	39–48
3	7	12	11	10	5	2

- 5). Form a cumulative frequency distribution and hence draw the ogive for the frequency distribution given in the solution to Problem 3.

20.95	3;	21.45	13;	21.95	24;	22.45	37;	22.95	46;	23.45	48
-------	----	-------	-----	-------	-----	-------	-----	-------	-----	-------	----

- 6). Draw a histogram for the frequency distribution given in the solution to Problem 4.

Rectangles, touching one another, having mid-points of 5.5, 15, 21, 24, 27, 33.5 and 43.5. The heights of the rectangles (frequency per unit class range) are 0.3, 0.78, 4.4.67, 2.33, 0.5 and 0.2

- 7). The frequency distribution for a batch of 50 capacitors of similar value, measured in micro-farads, is:

10.5–10.9   2,   11.0–11.4   7,   11.5–11.9   10,   12.0–12.4   12,   12.5–12.9   11,   13.0–13.48

Form a cumulative frequency distribution for these data.

(10.952), (11.459), (11.9511), (12.4531), (12.9542), (13.4550)

- 8). Draw an ogive for the data given in the solution of Problem 7.
- 9). The diameter in millimetres of a reel of wire is measured in 48 places and the results are as shown.

2.10   2.29   2.32   2.21   2.14   2.22   2.28   2.18   2.17   2.20   2.23   2.13  
2.26   2.10   2.21   2.17   2.28   2.15   2.16   2.25   2.23   2.11   2.27   2.34  
2.24   2.05   2.29   2.18   2.24   2.16   2.15   2.22   2.14   2.27   2.09   2.21  
2.11   2.17   2.22   2.19   2.12   2.20   2.23   2.07   2.13   2.26   2.16   2.12

- (a) Form a frequency distribution of diameters having about 6 classes.

**Solution :**   *There is unique solution, but one solution is:*

2.05–2.09   3;   2.10–2.14   10;   2.15–2.19   11  
2.20–2.24   13;   2.25–2.29   9;   2.30–2.34   2

■

- (b) Draw a histogram depicting the data.

**Solution :**   *Rectangles, touching one another, having mid-points of 2.07, 2.12, ... and heights of 3, 10, ...*

■

- (c) Form a cumulative frequency distribution.

**Solution :**   *Using the frequency distribution given in the solution to part (a) gives:*

2.095   3;   2.145   13;   2.195   24;   2.245   37;   2.295   46;   2.345   48

■

- (d) Draw an ogive for the the data.

**Solution :**   *A graph of cumulative frequency against upper class boundary having the coordinates given in part (c).*

■



- 1.) When should measures of location and dispersion be computed from grouped data rather than from individual data values?
- A. as much as possible since computations are easier.
  - B. whenever computer packages for descriptive statistics are unavailable
  - C. only when the data are from a population
  - D. None of the above answers is correct.
  - E. only when individual data values are unavailable
- 2.) Interviews are conversations with
- A. fun
  - B. purpose
  - C. friendliness
  - D. informality
  - E. none of the above
- 3.) In which of these, more than one candidate is interviewed?
- A. The behavioural interview
  - B. The stress interview
  - C. The group interview
  - D. The audition
  - E. Both C and D.
- 4.) A numerical value used as a summary measure for a sample, such as sample mean, is known as a
- A. population parameter
  - B. sample parameter
  - C. sample statistic
  - D. population mean
  - E. None of the above is correct.
- 5.) The following data show the number of hours worked by 200 statistics students.

Number of hours	Frequency
0 – 9	40
10 – 19	50
20 – 29	70
30 – 39	40

Table 1.13

- (a) Refer to Table1.13. The number of students working 19hours or less is
- A. 40

- B. 50
- C. 90
- D. 110
- E. Can not be determined

(b) Refer to Table 1.13. The relative frequency of students working 9 hours or less is

- A. 2
- B. 45
- C. 40
- D. 0.2
- E. Can not be determined

6.) A tabular summary of a set of data showing the fraction of the total number of items in several classes is a

- A. frequency distribution
- B. relative frequency distribution
- C. frequency
- D. cumulative frequency distribution
- E. None of the above answer is correct

#### Exercise 1.44

1.) Define the term statistics.

2.) (a) Discuss the challenges that are more likely to be met by the interviewee during the personal interview

(b) Data collection requires that the researcher should present the tool before the main interview. why?

(c) Mention any two sources that are used to obtain statistical data.

(d) Which of the two sources is most preferred by researchers? Justify your answer.

3.) Some of the methods of data presentation is the graphical method, tabular, charts etc.

(i) Mention and give a sketch of at least four methods of data presentation.

(ii) The following table shows student enrollment at a Certain university in Uganda (Hypothetical data) in the courses as indicated.

year	Educ	IT
1990	50	45
1991	68	73
1992	75	80
1993	40	50
1993	48	30

Represent the information using a multi-bar graph.

4.) With a clear explanation, describe a real life scenario where you applied statistics. which methods did you use to collect data and why you use them, which challenges you faced, etc. (please be brief and use statistical terms)

5.) Mention any three computer packages used by statisticians to analyze data.

**Exercise 1.45**

- 1.) Differentiate between Primary and Secondary source of data and give two examples on each data source.
- 2.) List the four methods of data presentation.

# Chapter 2

## Descriptive Statistics

### 2.1 Measures of Central Tendency

Measures of central tendency are measures of the location of the middle or the center of a distribution or A measure of central tendency is a sample value ( statistic) around which the distribution is centered.

The most common measures of central tendency are

- 1.) Mean
- 2.) Mode
- 3.) Median

#### 2.1.1 The Mean

There are three categories of mean namely the Arithmetic mean, Geometric mean and the Harmonic mean.

##### 2.1.1.1 The Arithmetic Mean

This is the most commonly used measure of central tendency and is the most widely used of the three categories of means.

##### *Arithmetic Mean for un grouped data*

If observations are in raw form then the mean is computed by summing up the observations and then dividing their sum by the number of observations. That is,

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i \\ &= \frac{\text{Sum of all observations}}{\text{Their total number}}\end{aligned}\tag{2.1}$$

**Example 2.1.1** Given the following observations, 2, 5, 50, 100, 200, 0 and 10.

The mean  $\bar{X}$  is given by

$$\bar{X} = \frac{2 + 5 + 50 + 100 + 200 + 0 + 10}{7} = \frac{367}{7}$$

Arithmetic mean for un-grouped data using the working (assumed) mean method. Here the mean is given by

$$\bar{X} = A + \frac{\sum_{i=1}^n d_i}{n} \quad (2.2)$$

Where

$$d_i = X_i - A$$

With  $A$  the assumed mean.

**Example 2.1.2** Let the observations be 2, 3, 5, and 10 in a given data set. Using an assumed mean of 5. find their arithmetic mean.

A table for the assumed mean

$X_i$	$d_i$
2	-3
3	-2
5	0
10	5
	$\sum d_i = 0$

Thus from,

$$\bar{X} = A + \frac{\sum_{i=1}^n d_i}{n}$$
$$\bar{X} = 5 + \frac{0}{4} = 5$$

**Note 2.1.1** The working( assumed) mean is chosen randomly that is without special consideration from within the range of the observations for easy computations.

***Arithmetic Mean for grouped data***

For grouped data, the arithmetic mean is given by;

$$\mu = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i} \quad (2.3)$$

where  $f_i$  is the frequency of the  $i^{th}$  class mark  $X_i$  - is the class mark of the  $i^{th}$  class.  $n$  is the number of classes.

***Alternatively***, we can compute the grouped mean using the assumed mean. The steps involved using the working mean are

- 1.) Choose the assumed mean  $A$  where  $A$  is preferably a value near or equal to the class mark of a class with highest frequency.
- 2.) Compute the arithmetic mean using the formula.

$$\mu = A + \frac{\sum_{i=1}^n f_i d_i}{\sum f_i} \quad (2.4)$$

**Example 2.1.3** The frequency table below shows the marks obtained by students in a certain examination,

Class	Frequency
60 - 62	5
63 - 65	18
66 - 68	42
69 - 71	27
72 - 74	8

Using the information above, find the arithmetic mean with and without using the working mean.

Class	Class Mark( $X_i$ )	Frequency( $f_i$ )	$f_i X_i$	$d_i = X_i - 67$	$f_i d_i$
60 – 62	61	5	305	–6	–30
63 – 65	64	18	1152	–3	–54
66 – 68	67	42	2814	0	0
69 – 71	70	27	1890	3	81
72 – 74	73	8	584	6	48
		$\sum_{i=1}^5 f_i = 100$	$\sum_{i=1}^5 f_i X_i = 6745$		$\sum_{i=1}^5 f_i d_i = 45$

$$\mu = \frac{\sum_{i=1}^5 f_i X_i}{\sum_{i=1}^5 f_i} = \frac{6745}{100} = 67.45$$

with the working mean  $A = 67$ .

$$\mu = A + \frac{\sum_{i=1}^5 f_i d_i}{\sum_{i=1}^5 f_i} = 67 + \frac{45}{100} = 67.45.$$

**2.1.1.2 Harmonic Mean**

Another measure of central tendency which is only occasionally used is the harmonic mean. It is most frequently employed for averaging speeds where the distances for each section of the journey are equal.

For a given  $n$  observations  $x_i; i = 1, 2, \dots$  then the Harmonic mean is defined as;

$$H = \frac{n}{\sum_{i=1}^n \frac{1}{X_i}} \quad (2.5)$$

**Example 2.1.4** Given the observations 2, 3, 3, 5. Find their Harmonic mean.

$$H = \frac{n}{\sum_{i=1}^n \frac{1}{X_i}} = \frac{4}{\frac{1}{3} + \frac{1}{5} + \frac{1}{2} + \frac{1}{3}} = \frac{120}{41}$$

**2.1.1.3 Geometric Mean**

The geometric mean is seldom used outside of specialist applications. It is appropriate when dealing with a set of data such as that which shows exponential growth. It is sometimes quite difficult to decide where the use of the geometric mean over the arithmetic mean is the best choice.

**Definition 2.1.1** Given any  $n$  observation ie  $x_i; i = 1, 2, \dots$  then the geometric mean is defined as the  $n^{th}$  root of their product;

$$G = \sqrt[n]{x_1 \cdot x_2 \dots} = \text{Antilog} \left( \frac{\sum \log X_i}{n} \right) \quad (2.6)$$

Proof is Example ?? on page (pg. ??).

**Example 2.1.5** Compute the geometric mean of the data in Example 2.1.4 above.

$$G = \sqrt[4]{3 \times 5 \times 3 \times 2} = \sqrt[4]{90}$$

**Example 2.1.6** In 1980 the population of a town is 300,000. In 1990 a new census reveals it has risen to 410,000. Estimate the population in 1985.

**Solution :** *If we assume that was no net immigration or migration then the birth rate will depend on the size of the population (exponential growth) so the geometric mean is appropriate.*

$$G = \sqrt[2]{300,000 \times 410,000} = 350,713$$

■

**Example 2.1.7** An aeroplane travels a distance of 900 miles. If it covers the first third and the last third of the trip at a speed of 250 mph and the middle third at a speed of 300 mph, find the average (harmonic) speed.

**Solution :**

$$H = \frac{n}{\sum_{i=1}^n \frac{1}{X_i}} = \frac{3}{\frac{1}{250} + \frac{1}{300} + \frac{1}{250}} = \frac{3}{0.01133} = 264.7$$

■



**Definition 2.1.2** The Harmonic mean of frequency and grouped data is given by

$$H = \frac{n}{\sum_{i=1}^n \left( \frac{f}{X_i} \right)} \quad (2.7)$$

**Definition 2.1.3** The Geometric mean of frequency and grouped data is given by

$$G = \text{Antilog} \left( \frac{\sum f \log X_i}{n} \right) \quad (2.8)$$

**Example 2.1.8** Calculate Geometric mean, Harmonic mean from the following grouped data

Class	Mark ( $X$ )	$f$	$\log(X)$	$\frac{f}{X}$
2 - 4	3	3	1.4314	1
4 - 6	5	4	2.7959	0.8
6 - 8	7	2	1.6902	0.2857
8 - 10	9	1	0.9542	0.1111
		$\sum f = 10$	$\sum f \log X = 6.8717$	$\sum \left( \frac{f}{X} \right) = 2.1968$

**Solution :**

$$G = \text{Antilog} \left( \frac{\sum f \log X_i}{n} \right) = \text{Antilog} \left( \frac{6.8717}{10} \right) = \text{Antilog} (0.6872) = 4.866$$

$$H = \frac{n}{\sum_{i=1}^n \left( \frac{f}{X_i} \right)} = \frac{10}{2.1968} = 4.552$$

■

**Example 2.1.9** Calculate Geometric mean, Harmonic mean from the following grouped data

Class	0 - 4	5 - 9	10 - 14	15 - 19	20 - 24	25 - 29
Frequency	2	10	7	5	3	8

**Solution :**

$$G = \text{Antilog} \left( \frac{\sum f \log X_i}{n} \right) = \text{Antilog} \left( \frac{3.7387}{35} \right) = \text{Antilog} (1.0925) = 12.3739$$

$$H = \frac{n}{\sum_{i=1}^n \left( \frac{f}{X_i} \right)} = \frac{35}{3.7387} = 9.3616$$

■

**2.1.1.4 Properties of the Mean**

- 1.) It can be calculated for every given data set that it always exists.
- 2.) The set of numerical data has only one mean indicating that the mean is always unique.
- 3.) It takes into account all observations in the data set.
- 4.) The sum of the deviations of a set of observations from their mean is 0 ie

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$

- 5.) The sum of squares of the deviations from the mean are minimal ie the sum of squares of deviation from the mean is less than the sum of sums of squares of deviation from any observation i.e.

$$\sum_{i=1}^n (X_i - \bar{X})^2 < \sum_{i=1}^n (X_i - x^*)^2.$$

Where  $x^*$  is any other observation and  $\bar{X}$  is the mean of the observations

**2.1.2 The Median****2.1.2.1 Median for un-grouped data**

It's a statistical measure that divides a data set into two equal subsets. For an un-grouped data order first arrange the items in either ascending or descending order and the median will then be given by the observation that falls in the middle (for an odd number of observations) but for an even number of observations we get the average of the two middle observations.

**Example 2.1.10** Find the median for the data set,

- 1.) 1, 2, 8, 9, 4, 7, 6.
- 2.) 1, 2, 8, 10, 9, 4, 7, 6.

As a solution, we first arrange in either ascending or descending order

- 1.) 1, 2, 4, 6, 7, 8, 9. thus median = 6.
- 2.) 1, 2, 4, 6, 7, 8, 9, 10., Median =  $\frac{6+7}{2} = 6.5$ .

Generally for a set of  $n$  observation the value of the median is given by  $\left(\frac{n+1}{2}\right)^{th}$  term for  $n$  odd and for  $n$  even the value of the median is given by the average of  $\left(\frac{n}{2}\right)^{th}$  and  $\left(\frac{n+2}{2}\right)^{th}$  terms.

**2.1.2.2 Median for grouped data**

For grouped data the median can be estimated by

1. an orgive
2. linear interpolation

**2.1.2.3 Median by Linear Interpolation method**

The following steps are taken;

- 1.) Compute the cumulative frequency,
- 2.) Divide total frequency by 2 in order to ascertain the median class and locate this class using the cumulative frequency column,
- 3.) Compute the median using the formula,

$$\text{Median} = l_m + \left[ \frac{\frac{N}{2} - cf_b}{f_m} \right] \times c \quad (2.9)$$

Where

$l_m$  - lower class boundary of the median class.

$f_m$  - frequency of the median class.

$cf_b$  - cumulative frequency of the class just before the median class

$c$  - the class width.

*Median by Graphical Method*

1. Compute  $\frac{1}{2} \sum f$  and locate it on the  $y$ -axis (cumulative frequency axis) of the ogive.
2. Draw a perpendicular line from this point and extend it to intersect with the ogive.
3. At the point of intersection with the graph (ogive) draw another perpendicular to the  $x$  - axis (lower class boundaries axis).
4. Read off the value of the median from the  $x$  - axis.

**Example 2.1.11** Given the frequency distribution table as in the Example 2.1.3, find the median mark using,

- 1.) Linear interpolation method
- 2.) The graphical method

Class	Class Mark ( $X_i$ )	Frequency( $f_i$ )	Cumulative Frequency
60 - 62	61	5	5
63 - 65	64	18	23
66 - 68	67	42	65
69 - 71	70	27	92
72 - 74	73	8	100

1.) Using Linear interpolation method.

Median class is

$$66 - 68$$

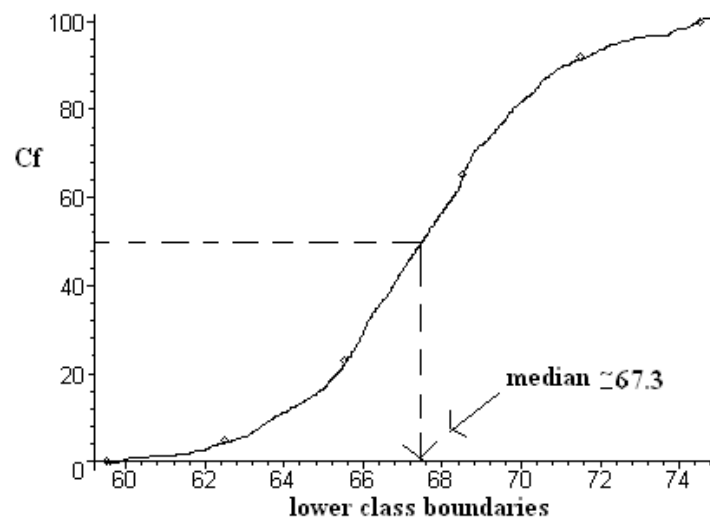
with,

$$l_m = 65.5, \quad c = 3, \quad cf_b = 23, \quad f_m = 42$$

such that

$$\begin{aligned}\text{Median} &= l_m + \left[ \frac{\frac{N}{2} - cf_b}{f_m} \right] \times c \\ &= 65.5 + \left( \frac{50 - 23}{42} \right) \times 3 \\ &= 67.429\end{aligned}$$

2.) The graphical method



## 2.1.3 The Mode

### 2.1.3.1 Mode for un-grouped data

It's the observation with the highest frequency or number of appearances in a given data set for un-grouped data. A given data set may have more than one mode. If a set of data has one mode then it's called unimodal, if it has two modes then its bimodal and with many modes its multi modal.

**Example 2.1.12** Given the data set

$$2, 3, 1, 3, 4, 5, 6, 7, 4$$

then the mode is 3 & 4 (bimodal).

**Example 2.1.13** Following are the margin of victory in the foot ball matches of a league.

$$\begin{array}{cccccccccccccccccccc} 1 & 3 & 2 & 5 & 1 & 4 & 6 & 2 & 5 & 2 & 2 & 2 & 4 & 1 & 2 & 3 & 2 & 3 & 2 & 3 \\ 1 & 1 & 2 & 3 & 2 & 6 & 4 & 3 & 2 & 1 & 1 & 4 & 2 & 1 & 5 & 3 & 4 & 2 & 1 & 2 \end{array}$$

Since 2 has occurred more number of times (14 times), the mode of the given data is 2.

**2.1.3.2 Mode for grouped data**

It's estimated from the class with the highest frequency called the modal class. this is done by the formula below,

$$\text{Mode} = l_m + c \left[ \frac{d_1}{d_1 + d_2} \right]$$

$$\text{Mode} = l_m + \left[ \frac{(f_m - f_b)}{(f_m - f_b) + (f_m - f_a)} \right] \times c \quad (2.10)$$

Where

- $l_m$  : lower class boundary of the modal class
- $f_m$  : frequency of the modal class
- $f_b$  : frequency of the class just before the modal class
- $f_a$  : frequency of the class just after the modal class
- $c$  : the class width

or the mode can be estimated from the histogram.

**Example 2.1.14** Given the frequency distribution table as in Example 2.1.3, find the modal mark of the students.

The modal class with the highest frequency is

$$66 - 68$$

with

$$l_m = 65.5, \quad f_m = 42, \quad f_a = 27, \quad f_b = 18, \quad c = 3$$

Therefore,

$$\begin{aligned} \text{Mode} &= 65.5 + \left( \frac{(42 - 18)}{(42 - 18) + (42 - 27)} \right) \times 3 \\ &= 67.3 \end{aligned}$$

Or It can be estimated from the histogram.

**2.1.3.3 Properties of the Mode**

- 1.) The mode does not always exist
- 2.) It may or may not be unique
- 3.) For grouped data if the modal class happens to be the first or the last class in the distribution then we estimate the mode as mode = 3(median) - 2 (mean).
- 4.) The mode can be estimated practically from the histogram. This is done by drawing lines diagonally from the upper corners of the tallest bar to the upper corner of the adjacent bars and a perpendicular line is drawn from the point of intersection to the  $x$  - axis and the mode is read from the class boundaries axis.

## 2.2 Measures of Position

### 2.2.1 Quartiles

These are measures which divide a given data set into four (4) parts, the first quartile relates to the lower 25% of the observations, the 2<sup>nd</sup> quartile relates to the lower 50% of the observations and the third quartile relates to the lower 75% of the observations.

Interquartile range is the difference between the third quartile and the first quartile.

#### 2.2.1.1 Quartiles from un-grouped data

We arrange the observations in ascending or descending order and locate the quartiles using the formula

$$\left(\frac{n+1}{4}\right)i, \quad \text{for } i = 1, 2, 3,$$

where  $n$  is the total number of observations in the data set, this expression gives the position of the quartile in the data set.

**Example 2.2.1** Consider the monthly salaries of secretaries in a certain organization in dollars as,

441, 430, 515, 420, 490, 438, 435, 447, 445, 500, 510

Find the quartiles together with the interquartile range?.

We arrange the data in an ascending array as below,

420, 430, 435, 438, 441, 445, 447, 490, 500, 510, 515

The positions for the quartiles are,

$$Q_1 = \left(\frac{11+1}{4}\right)(1) = 3^{rd} \text{ observation}$$

$$Q_2 = \left(\frac{11+1}{4}\right)(2) = 6^{th} \text{ observation}$$

$$Q_3 = \left(\frac{11+1}{4}\right)(3) = 9^{th} \text{ observation}$$

That's  $Q_1$  is in the third position,  $Q_2$  is in the sixth position and  $Q_3$  is in the ninth position thus

$$Q_1 = 435, \quad Q_2 = 445 \quad \text{and} \quad Q_3 = 500$$

The interquartile range is given by

$$Q_3 - Q_1 = 500 - 435 = 65$$

**Example 2.2.2** Compute  $Q_3$  for the data 2, 6, 8, 10?

$$Q_3 = \frac{3}{4}(4+1) = 3.75^{th}.obs = 8 + 0.75(10-8) = 9.5$$

**2.2.1.2 Quartiles from grouped data**

Here we locate the quartiles by the help of the formula  $\frac{i}{4}(n+1)$ , for  $i = 1, 2, 3$ , but we locate it with the aid of the cumulative frequencies,

$$Q_i = l_m + \left( \frac{CF_i - cf_b}{f_w} \right) \times c \quad (2.11)$$

Where

- $l_m$  : is the lower class boundary of the  $i^{th}$  quartile class
- $CF_i$  :  $\left( \frac{i}{4} \right) \times n$ , cumulative frequency of the  $i^{th}$  quartile class
- $cf_b$  : cumulative frequency of the class just before the  $i^{th}$  quartile class
- $f_w$  : frequency of the quartile class
- $c$  : the class width

**Example 2.2.3** Given the following frequency distribution table use it to find the quartiles and their interquartile range.

Age of students	Number of students ( $f$ )	Cumulative Frequency
20 - 24	11	11
25 - 29	24	35
30 - 34	30	65
35 - 39	18	83
40 - 44	11	94
45 - 49	5	99
50 - 54	1	100

1.) Lower quartile  $Q_1$

$$Q_1 = \frac{1}{4}(n+1) = \left( \frac{100+1}{4} \right) = 25.25^{th} \text{ position.}$$

this implies that the lower quartile class is = 25 – 29

$$l_m = 24.5, \quad c = 5, \quad cf_b = 11, \quad CF_1 = \left( \frac{i}{4} \right) \times n = \frac{1}{4}(100) = 25, \quad f_w = 24$$

$$Q_1 = 24.5 + \left( \frac{25 - 11}{24} \right) \times 5 = 27.42 \text{ units.}$$

2.) Upper quartile  $Q_3$ :

$$Q_3 = \frac{3}{4}(n+1) = \frac{3}{4}(100+1) = \frac{100+1}{4}(3) = 75.75^{th} \text{ position.}$$

this implies that the Upper quartile class is = 35 – 39

$$l_m = 34.5, \quad c = 5, \quad cf_b = 65, \quad CF_3 = \left( \frac{i}{4} \right) \times n = \frac{3}{4}n = 75, \quad f_w = 18$$

$$Q_3 = 34.5 + \left( \frac{75 - 65}{18} \right) \times 5 = 37.27778 \text{ units.}$$

3.) Interquartile range  $Q_3 - Q_1 = 9.85778$

## 2.3 Measures of Variation

### 2.3.1 The Range

The range gives the distance between the largest and smallest observation in a given set for an un-grouped data set. For an *un-grouped data* set the range is given by the difference between the largest (biggest) and the smallest (least) observation

**Example 2.3.1** For the data set

2, 4, 6, 5, 3, 21, 70

Have a range =  $70 - 2 = 68$

For *grouped data* the range is given by the difference between the class mark of the last class interval and the class mark of the first class interval.

**Note 2.3.1** The greater the range value the wider the dispersion/spread and vice versa.

**Example 2.3.2** Given the following frequency distributions.

Class	Class Mark( $X_i$ )	Frequency( $f_i$ )
60 - 62	61	5
63 - 65	64	18
66 - 68	67	42
69 - 71	70	27
72 - 74	73	8

The range is  $73 - 61 = 12$ .

Though the range has an advantage of easy computation, it has some disadvantages viz;

- 1.) It may be misleading if either of the extreme values are outliers (far smaller or far larger than other observations).
- 2.) The range is silent about the arrangement of the observation that fall in between the two extreme values.



### 2.3.2 The Variance, Standard Deviation and Mean Deviation of un-grouped data

#### 1.) Mean deviation of un-grouped data

For a given data set  $x_i$ ,  $i = 1, 2, \dots, n$  we define the mean deviation as the average of the absolute deviation from the mean given by the formula,

$$\text{Median Deviation} = S_{\bar{X}} = \frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}| \quad (2.12)$$

#### 2.) The Variance is defined as the mean of the squared deviations of individual observations from their arithmetic mean.

(a) For the **ungrouped population**, the population variance, denoted by  $\sigma^2$  defined as;

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \quad (2.13)$$

where  $\mu = \frac{\sum_{i=1}^n X_i}{n}$  for a population with  $n$  as the total number of observations in the population.  $X_i$  is the  $i^{th}$  observation.

(b) For the **ungrouped sample**, the sample variance denoted as  $S^2$  is given by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (2.14)$$

with the sample mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ,  $n$  is the total number of observations in the **sample** with  $X_i$  the  $i^{th}$  observation.

Equation (2.13) gives the population variance and equation (2.14) gives the sample variance.

#### 3.) The standard Deviation is defined as the positive square root of the variance. Equation (2.15) gives the population standard deviation for the un-grouped data and equation (2.16) gives the sample standard deviation for the un-grouped data .

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2} \quad (2.15)$$

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (2.16)$$

#### Note 2.3.2

I. Expression (2.13) can be re-written as,

$$\sigma^2 = \frac{\sum_{i=1}^n X_i^2}{n} - \left( \frac{\sum_{i=1}^n X_i}{n} \right)^2$$

II. Expression (2.14) can be re-written as

$$S^2 = \frac{n \sum_{i=1}^n X_i^2 - \left( \sum_{i=1}^n X_i \right)^2}{n(n-1)}$$

**Example 2.3.3** The figures below show production of a certain product in a Kampala based factory

98, 99, 99, 100, 100, 100, 101, 101, 102

Find

1.) The mean deviation

$$\text{Mean Deviation} = S_{\bar{X}} = \frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}|$$

but

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{900}{9} = 100$$

$$\Rightarrow \text{Mean deviation} = S_{\bar{X}} = \frac{8}{9}$$

2.) The variance and standard deviation

The variance of the sample (less than 30 items)

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = 1.5 \text{ units}$$

and standard deviation is

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} = \sqrt{1.5}$$

### 2.3.3 Variance, Standard Deviation and Mean Deviation of Grouped Data

1.) For a grouped **population**;

(a) Mean Deviation

$$\text{Mean deviation} = \sigma_{\mu} = \frac{1}{n} \sum_{i=1}^n f_i |X_i - \mu| \quad (2.17)$$

(b) The variance  $\sigma^2$

$$\sigma^2 = \frac{1}{n} \left[ \sum_{i=1}^n f_i (X_i - \mu)^2 \right] \quad (2.18)$$

which can be re-written as any of the following formulae,

$$\begin{aligned} \sigma^2 &= \frac{1}{n} \left[ \sum_{i=1}^n f_i X_i^2 - \frac{\left( \sum_{i=1}^n f_i X_i \right)^2}{n} \right] \\ \sigma^2 &= \frac{1}{n^2} \left[ n \sum_{i=1}^n f_i X_i^2 - \left( \sum_{i=1}^n f_i X_i \right)^2 \right] \\ \sigma^2 &= \frac{\sum_{i=1}^n f_i X_i^2}{n} - \frac{\left( \sum_{i=1}^n f_i X_i \right)^2}{n^2} \end{aligned}$$

(c) Standard Deviation: Given by the square root of the Variance.

2.) For a grouped **sample** (subset of the population) of size  $n$ .

(a) Mean deviation

$$\text{Mean deviation} = S_{\bar{X}} = \frac{1}{n} \sum_{i=1}^n f_i |X_i - \bar{X}| \quad (2.19)$$

(b) Variance, denoted by  $S^2$  is defined by the formula,

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n f_i (X_i - \bar{X})^2 \quad (2.20)$$

and can be re-written as,

$$\begin{aligned} S^2 &= \frac{1}{(n-1)} \left[ \sum_{i=1}^n f_i X_i^2 - \frac{\left( \sum_{i=1}^n f_i X_i \right)^2}{n} \right] \\ S^2 &= \frac{1}{n(n-1)} \left[ n \sum_{i=1}^n f_i X_i^2 - \left( \sum_{i=1}^n f_i X_i \right)^2 \right] \\ S^2 &= \frac{\sum_{i=1}^n f_i X_i^2}{(n-1)} - \frac{\left( \sum_{i=1}^n f_i X_i \right)^2}{n(n-1)} \end{aligned}$$

(c) Standard Deviations will respectively be the positive square roots of the Variance.

**Example 2.3.4** Consider the following tiny grouped data set in Table below

Class	Class Boundary	Frequency $f$	Cumulative Frequency cf	Class Mark $X$	$fX$	$fX^2$
1 - 20	0.5 - 20.5	5	5	10.5	52.5	551.25
21 - 40	20.5 - 40.5	25	30	30.5	762.5	23256.25
41 - 60	40.5 - 60.5	37	67	50.5	1868.5	94359.25
61 - 80	60.5 - 80.5	23	90	70.5	1621.5	114315.75
81 - 100	80.5 - 100.5	8	98	90.5	724	65522
		$\sum f = 98$			$\sum fX = 5029$	$\sum fX^2 = 298004.5$

Determine

$$1.) \text{ Mean: } \mu = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i} = \frac{5029}{98} = 51.3163$$

$$2.) \text{ Median: } \frac{98}{2} = 49, \text{ so the median class is } (40.5 - 60.5)$$

$$\text{Median} = l_m + \left[ \frac{\frac{N}{2} - cf_b}{f_m} \right] \times c = 40.5 + \left( \frac{49 - 30}{37} \right) \times 20 = 50.7703$$

$$3.) \text{ Mode: The modal class is the class with the highest frequency, } (40.5 - 60.5)$$

$$\text{Mode} = l_m + \left[ \frac{(f_m - f_b)}{(f_m - f_b) + (f_m - f_a)} \right] \times c = 40.5 + \left[ \frac{(37 - 25)}{(37 - 25) + (37 - 23)} \right] \times 20 = 49.7308$$

$$4.) \text{ Lower quartile: } Q_1 = \left( \frac{n+1}{4} \right) i = \left( \frac{98+1}{4} \right) (1) = 24.75^{th} \text{ value of the observation in } cf \text{ column, so class } (20.5 - 40.5), \text{ and } CF_1 = \left( \frac{i}{4} \right) n = \frac{1}{4} \times 98 = 24.5$$

$$Q_1 = l_m + \left( \frac{CF_1 - cf_b}{f_w} \right) \times c = 20.5 + \left( \frac{24.5 - 5}{25} \right) \times 20 = 36.1$$

$$5.) \text{ Upper quartile: } Q_3 = \left( \frac{n+1}{4} \right) i = \left( \frac{98+1}{4} \right) (3) = 74.25^{th} \text{ value of the observation in } cf \text{ column, so class } (60.5 - 80.5), \text{ and } CF_3 = \left( \frac{i}{4} \right) n = \frac{3}{4} \times 98 = 73.5$$

$$Q_3 = l_m + \left( \frac{CF_3 - cf_b}{f_w} \right) \times c = 60.5 + \left( \frac{73.5 - 67}{23} \right) \times 20 = 66.1522$$

6.) Variance for the sample given

$$S^2 = \frac{1}{(n-1)} \left[ \sum_{i=1}^n f_i X_i^2 - \frac{\left( \sum_{i=1}^n f_i X_i \right)^2}{n} \right] = \frac{1}{(98-1)} \left[ 298004.5 - \frac{5029^2}{98} \right] = 411.6979$$

7.) Standard deviation

$$S = \sqrt{S^2} = \sqrt{411.6979} = 20.2903$$

8.) Show that

(a)  $Q_2 = 50.7703$

(c) Mean deviation  $S_{\bar{X}} = 0.3001$

(b) Range =  $100.5 - 0.5$

(d) Quartile deviation = 15.0261

**Example 2.3.5** Given the following frequency distribution table from a given sample

Class	Class Mark( $X_i$ )	Freq( $f_i$ )	$f_i X_i$	$ X_i - \bar{X} $	$(X_i - \bar{X})^2$	$f_i  X_i - \bar{X} $	$f_i (X_i - \bar{X})^2$
60-62	61	5	305	6.45	41.6025	32.25	208.0125
63-65	64	18	1152	3.45	11.9025	62.1	214.245
66-68	67	42	2814	0.45	.2025	18.9	8.505
69-71	70	27	1890	2.55	6.5025	68.85	175.5675
72-74	73	8	584	5.55	30.8025	44.4	246.42
$\Sigma$		100	6745			226.5	852.75

Find

1.) Mean deviation

$$\text{Mean deviation} = S_{\bar{X}} = \frac{1}{\sum f} \sum_{i=1}^n f_i |X_i - \bar{X}| = \frac{226.5}{100} = 2.265$$

2.) The sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n f_i (X_i - \bar{X})^2 = \frac{852.75}{99} = 8.6136$$

$$\text{and Standard deviation} = \sqrt{8.6136} = 2.9349.$$

**Note 2.3.3** The greater the depression in a given data set the larger will be the value of it's standard deviation and variance. Therefore the standard deviation can be used to compare dispersion of two or more data sets,

However to we at a meaningful conclusion the following conditions should be satisfied;

1.) The data sets should be expressed in the same units

2.) The means of the data sets should be nearly the same in magnitude

Once the two conditions are satisfied then the smaller of the two standard deviations would indicate that the distribution to which it belongs exhibits less dispersion than others under comparison.

**Exercise 2.1** The systolic blood pressure of seven middle aged ladies were as follows:

151, 124, 132, 170, 146, 124 and 113

Compute their

- 1.) Mean systolic blood pressure. [137.14]
- 2.) Median systolic blood pressure [132]
- 3.) Mode systolic blood pressure [124]

**Exercise 2.2** Six men with high cholesterol participated in a study to investigate the effects of diet on cholesterol level. At the beginning of the study, their cholesterol levels (mg/dL) were as follows:

366, 327, 274, 292, 274 and 230

- 1.) Median cholesterol levels [283]
- 2.) Mode cholesterol levels [274]

**Exercise 2.3** Define the median of the random sample, distinguishing between the two cases  $n$  odd and  $n$  even. Show that the median has expected value  $\frac{1}{2}$  if the random sample is drawn from a uniform distribution on  $(0, 1)$ .

Find its variance in the particular case when  $n$  is odd. What is the expected value of the median if the random sample is drawn from a uniform distribution on  $(a, b)$ ?

**Exercise 2.4** Consider the following: Data are Total Patient Care Revenues for a sample of hospitals in Buddu County (Greater Masaka) Note that,

Hospital	Revenue (in millions)
1	414.6
2	358.6
3	439.8
4	64.8
5	159.2
6	130.5
7	395.3

Table 2.1: Hospital Revenues in Buddu in 1996

Hospitals all have different level of revenues

The spread ranges from 64.8 million to 414.6 million

Hospital 4 appears to have unusually low revenues-outlier?

Compute

- 1.) The average revenue in 1996
- 2.) The median revenue in 1996
- 3.) The mode in 1996
- 4.) The third quartile  $Q_3$
- 5.) The range
- 6.) The variance

**Exercise 2.5** Repeat problems in Exercise 2.4 for the grouped data

Heights:	160 - 164	165 - 169	170 - 174	175 - 179	180 -184	185-189
Frequency:	7	11	17	20	16	6

Table 2.2: Height of employees in cm

**Example 2.3.6** The wheat production (in Kg) of 20 acres is given as:

1120 1240 1320 1040 1080 1200 1440 1360 1680 1730  
1785 1342 1960 1880 1755 1720 1600 1470 1750 1885

After arranging the observations in ascending order, we get

1040 1080 1120 1200 1240 1320 1342 1360 1440 1470  
1600 1680 1720 1730 1750 1755 1785 1880 1885 1960

1.)

$$\begin{aligned}Q_1 &= \frac{1}{4}(n+1) = \frac{1}{4}(21) = 5.25^{th} = 5^{th} + 0.25(6^{th} - 5^{th}) \\&= 1240 + 0.25(1320 - 1240) \\&= 1240 + 20 = 1260\end{aligned}$$

2.)

$$\begin{aligned}Q_3 &= \frac{3}{4}(n+1) = \frac{3}{4}(21) = 15.75^{th} = 15^{th} + 0.75(16^{th} - 15^{th}) \\&= 1750 + 0.75(1755 - 1750) \\&= 1750 + 3.75 = 1753.75\end{aligned}$$

3.)

$$\begin{aligned}Q_2 &= \frac{2}{4}(n+1) = \frac{2}{4}(21) = 10.5^{th} = 10^{th} + 0.5(11^{th} - 10^{th}) \\&= 1470 + 0.5(1600 - 1470) \\&= 1470 + 65 = 1533\end{aligned}$$

4.) The *Quartile Deviation* (Q.D)

$$\frac{Q_3 - Q_1}{2} = \frac{1753.75 - 1260}{2} = 246.875$$

5.) *Coefficient of Quartile Deviation*

$$\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{1753.75 - 1260}{1753.75 + 1260} = 0.164$$

**Exercise 2.6** Given the age data of the 12 village members:

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25, 65

- 1.) Find the ages' upper quartile  $Q_3$ . [40.5]
- 2.) Determine the median age. [23.5]
- 3.) Compute the inter quartile range
- 4.) Find the Quartile Deviation (Q.D)
- 5.) Establish the Coefficient of Quartile Deviation

**Exercise 2.7** Based on the grouped data below,

Time to Travel to Work	Frequency
1 - 10	8
11 - 20	14
21 - 30	12
31 - 40	9
41 - 50	7

Find

- 1.) Median [24]
- 2.) Lower Quartile  $Q_1$  [13.7143]
- 3.) Upper Quartile  $Q_3$  [34.3889]
- 4.) Interquartile range [20.6746]
- 5.) Mode [17.5]
- 6.) Variance  $S^2$

**Exercise 2.8** Find the interquartile of the data set:

{1, 3, 4, 5, 5, 6, 9, 14, 21}

$$[Q_3 - Q_1 = 9 - 4 = 5]$$

**Exercise 2.9** In a work study investigation, the times taken by 20 men in a firm to do a particular job were tabulated as follows:

Time Taken (min)	8-10	11-13	14-16	17-19	20-22	23-25
Frequencies	2	4	6	4	3	1

Compute the second quartile  $Q_2$ . [15.50]

**Note 2.3.4** The second quartile  $Q_2$  is usually the median.



**Example 2.3.7** The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value. For example, consider the wages of staff at a factory below:

Staff	1	2	3	4	5	6	7	8	9	10
Salary	5k	18k	16k	14k	15k	15k	12k	17k	90k	95k

The mean salary for these ten staff is \$30.7k. However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers have salaries in the \$12k to \$18k range. The mean is being skewed by the two large salaries. Therefore, in this situation, we would like to have a better measure of central tendency. As we will find out later, taking the median would be a better measure of central tendency in this situation.

**Exercise 2.10** Find the median for the following data

65 , 55 , 89 , 56 , 35 , 14 , 56 , 55 , 87 , 45 , 92

55.5

**Exercise 2.11** Given the data 1, 5, 8, 10, 7 and 5 Calculate

- 1.) mean/average 6
- 2.) range 9
- 3.) median 6
- 4.) and mode 5

**Example 2.3.8** Assume that we have obtained the following 20 observations:

2, 4, 7, -20, 22, -1, 0, -1, 7, 15, 8, 4, -4, 11, 11, 12, 3, 12, 18, 1

In order to calculate the quartiles we first have to sort the observations:

-20, -4, -1, -1, 0, 1, 2, 3, 4, 4, 7, 7, 8, 11, 11, 12, 12, 15, 18, 22

The position of the first quartile is  $x = \text{round}(0.25 \cdot (20+1)) = \text{round}(5.25) = 5$ , which means that  $Q_1$  is the 5th value of the sorted series, namely  $Q_1 = 0$ . The other quartiles are calculated in the same way resulting in  $Q_2 = 5.5$  and  $Q_3 = 12$

**Example 2.3.9** Consider the aptitude test scores of ten children below:

95, 78, 69, 91, 82, 76, 76, 86, 88, 80

Find the mean, median, and mode.

1.) Mean

**Solution :**

$$\bar{X} = \frac{1}{10}(95 + 78 + 69 + 91 + 82 + 76 + 76 + 86 + 88 + 80) = 82.1$$

■

2.) Median

**Solution :** *First, order the data.*

69, 76, 76, 78, 80, 82, 86, 88, 91, 95

With  $n = 10$ , the median position is found by  $\frac{(10 + 1)}{2} = 5.5$ . Thus, the median is the average of the fifth (80) and sixth (82) ordered value and the median = 81

■

3.) Mode

**Solution :** *The most frequent value in this data set is 76. Therefore the mode is 76.*

■

### Exercise 2.12

- 1.) What are measures of central tendency as used in statistics?.
- 2.) Mention any three measures of central tendency you know.
- 3.) Construct a frequency distribution table for the following figures of weights obtained from 36 Elements of mathematics students in a Ugandan university using a class width of 3 and starting with the class 56 – 58.

66 70 68 67 71 60

64 70 68 65 64 61

71 66 67 65 68 59

67 65 68 66 69 58

66 65 65 71 70 56

57 60 62 56 59 72

- 4.) Using the frequency distribution table in (3) above, find
  - (a) the mean weight,
  - (b) the modal weight and
  - (c) the median weight of the students.

## 2.4 Other Statistical Measures

### 2.4.1 Quartile Deviation

**Definition 2.4.1** The quartile deviation is half the difference between the third quartile and the first quartile, and for this reason it is often called the ***semi**-interquartile range*.

$$\text{Quartile Deviation} = \frac{1}{2} [Q_3 - Q_1] \quad (2.21)$$

**Example 2.4.1** Find the quartile deviation of the monthly rainfall for the two towns whose rainfall is given below. The ordered amounts of rainfall are:

*Town A :* 2.1 2.8 3.9 4.2 4.5 4.8 5.2 5.3 5.4 6.5 6.8 7.2

*Town B :* 0.4 2.2 2.8 3.4 4.3 4.8 5.1 5.4 6.2 6.6 6.9 10.6

**Solution :** For Town A,

*The lower quartile*

$$Q_1 = \left( \frac{n+1}{4} \right) i = \left( \frac{12+1}{4} \right) (1) = 3.25^{th}$$

$Q_1$  is a quarter of the way between the 3rd and 4th observations,

$$Q_1 = 3.9 + 0.25(4.2 - 3.9) = 3.975$$

*The upper quartile*

$$Q_3 = \left( \frac{n+1}{4} \right) i = \left( \frac{12+1}{4} \right) (3) = 9.75^{th}$$

$Q_3$  is a three quarter of the way between the 9th and 10th observations,

$$Q_3 = 5.4 + 0.75(6.5 - 5.4) = 6.225$$

*Interquartile range*

$$Q_3 - Q_1 = 6.225 - 3.975 = 2.25$$

*Quartile deviation (Semi interquartile range)*

$$\text{Quartile Deviation} = \frac{1}{2} [Q_3 - Q_1] = \frac{1}{2} [6.225 - 3.975] = 1.125$$

For Town B,

*The lower quartile*

$$Q_1 = \left( \frac{n+1}{4} \right) i = \left( \frac{12+1}{4} \right) (1) = 3.25^{th}$$

$Q_1$  is a quarter of the way between the 3rd and 4th observations,

$$Q_1 = 2.8 + 0.25(3.4 - 2.8) = 2.95$$

*The upper quartile*

$$Q_3 = \left(\frac{n+1}{4}\right)i = \left(\frac{12+1}{4}\right)(3) = 9.75^{th}$$

$Q_3$  is a three quarter of the way between the 9th and 10th observations,

$$Q_3 = 6.2 + 0.75(6.6 - 6.2) = 6.5$$

*Interquartile range*

$$Q_3 - Q_1 = 6.5 - 2.95 = 3.55$$

*Quartile deviation (Semi interquartile range)*

$$\text{Quartile Deviation} = \frac{1}{2}[Q_3 - Q_1] = \frac{1}{2}[6.5 - 2.95] = 1.775$$

■

**Example 2.4.2** Number of days employees are late in a month is given in the frequency table below

Number of Days Late	Number of Employees	Cumulative Frequency
$X$	$f$	$cf$
1	32	32
2	25	57
3	18	75
4	14	89
5	11	100

Determine the quartile deviation.

**Solution :**

*The lower quartile*

$$Q_1 = \left(\frac{n+1}{4}\right)i = \left(\frac{100+1}{4}\right)(1) = 25.25^{th},$$

so  $Q_1$  is a quarter of the way between the 25th and 26th observations; both of these are 1 therefore,  $Q_1 = 1$ .

*The upper quartile*

$$Q_3 = \left(\frac{n+1}{4}\right)i = \left(\frac{100+1}{4}\right)(3) = 75.75^{th}$$

$Q_3$  is three-quarters of the way between the 75th and 76th observations; the 75th observation is 3 and the 76th is 4 so  $Q_3 = 3.75$ .

*Quartile deviation (Semi interquartile range)*

$$\text{Quartile Deviation} = \frac{1}{2}[Q_3 - Q_1] = \frac{1}{2}[3.75 - 1] = 1.375$$

■

**Example 2.4.3** Compute the Quartile deviation for the grouped distribution of the overdraft sizes of 400 bank customers given in the table.

Class Boundaries	Class Mark $X$	Number of Customers $f$	Cumulative Frequency $cf$
0 - 100	50	82	82
100 - 200	150	122	204
200 - 300	250	86	290
300 - 400	350	54	344
400 - 500	450	40	384
500 - 600	550	16	400

**Remark 2.4.1** The class given is not just a class limits, but class boundaries, since upper and lower limits of consecutive classes are the same.

- 1.) Lower quartile:  $Q_1 = \left(\frac{n+1}{4}\right)i = \left(\frac{400+1}{4}\right)(1) = 100.25^{th}$  value of the observation in  $cf$  column, so class boundaries (100 – 200), and  $CF_1 = \left(\frac{i}{4}\right)n = \frac{1}{4} \times 400 = 100$

$$Q_1 = l_m + \left(\frac{CF_1 - cf_b}{f_w}\right) \times c = 100 + \left(\frac{100 - 82}{122}\right) \times 1000 = 114.75$$

- 2.) Upper quartile:  $Q_3 = \left(\frac{n+1}{4}\right)i = \left(\frac{400+1}{4}\right)(3) = 300.75^{th}$  value of the observation in  $cf$  column, so class (300 – 400), and  $CF_3 = \left(\frac{i}{4}\right)n = \frac{3}{4} \times 400 = 300$

$$Q_3 = l_m + \left(\frac{CF_3 - cf_b}{f_w}\right) \times c = 300 + \left(\frac{300 - 290}{54}\right) \times 100 = 318.52$$

- 3.) Quartile deviation

$$\text{Quartile Deviation} = \frac{1}{2} [Q_3 - Q_1] = \frac{1}{2} [318.52 - 114.75] = 101.89$$

**Definition 2.4.2** Inter Quartile range is defined as

$$\text{Inter Quartile range} = Q_3 - Q_1 \quad (2.22)$$

**Definition 2.4.3** Quartile deviation is defined as

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2} \quad (2.23)$$

**Definition 2.4.4** Coefficient of Quartile deviation is defined as

$$\text{Coefficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \quad (2.24)$$

### 2.4.2 Percentiles

These values divide the observations into 100 equal parts and are denoted by

$$P_1, P_2, P_3, \dots, P_{99}$$

e.g.  $P_1$  has 1% below it and 99% above.

**Definition 2.4.5** If a data set contains  $n$  observations, then the  $i$ th percentile is the value

$$P_i = \frac{i}{100}(n+1)^{th} \quad (2.25)$$

in the ordered data set.

**Example 2.4.4** The test score of a sample of 20 students in a class are as follows:

20, 30, 21, 29, 10, 17, 18, 15, 27, 25, 16, 15, 19, 22, 13, 17, 14, 18, 12 and 9

Find the value of  $P_{10}$ ,  $P_{20}$  and  $P_{80}$ .

**Solution :** *The sample size is  $n = 20$ .*

*Arrange the data in ascending order*

9, 10, 12, 13, 14, 15, 15, 16, 17, 17, 18, 18, 19, 20, 21, 22, 25, 27, 29, 30

1.) *The tenth percentile  $P_{10}$  can be computed as follows:*

$$P_{10} = \frac{10}{100}(n+1) = \frac{10}{100}(20+1) = 2.1^{th} \text{ observation,}$$

*which lies a tenth between the second and third entry. Therefore,*

$$P_{10} = 10 + 0.1(12 - 10) = 10.2$$

2.) *The twentieth percentile  $P_{20}$  can be computed as follows:*

$$P_{20} = \frac{20}{100}(n+1) = \frac{20}{100}(20+1) = 4.2^{th} \text{ observation,}$$

*which lies a two tenth between the fourth and fifth entry. Therefore,*

$$P_{20} = 13 + 0.2(14 - 13) = 13.2$$

3.) *The eightieth percentile  $P_{80}$  can be computed as follows:*

$$P_{80} = \frac{80}{100}(n+1) = \frac{80}{100}(20+1) = 16.8^{th} \text{ observation,}$$

*which lies a eighth tenth between the sixteen and seventeen entry. Therefore,*

$$P_{80} = 22 + 0.8(25 - 22) = 24.4$$

■

**Example 2.4.5** Find the 20th percentile of the data represented by the following stem-and-leaf plot.

Stem	Leaf
2	1
3	3 5 8 9
4	1 2 2 3 3 3 4 5 6 7 8 8
5	1 1 2 4 6 7
6	7
7	0 4
8	4 8
9	
10	8

$n = 29$ , 1|2 represents 12

**Solution :**

$$P_{20} = \frac{20}{100}(n+1) = \frac{20}{100}(29+1) = 6^{th} \text{ observation,}$$

which is 41. Therefore,  $P_{20} = 41$ . ■

**Example 2.4.6** The following data gives the hourly wage rates (in dollars) of 25 employees of a company.

20 28 30 18 27 19 22 21 24 25 18 25 20

27 24 20 23 32 20 35 22 26 25 28 31

1.) the upper wage rate for the lowest 15% of the employees,

**Solution :**

$$\begin{aligned}
 P_{15} &= \text{Value of } \left( \frac{15(n+1)}{100} \right)^{th} \text{ obs.} \\
 &= \text{Value of } \left( \frac{15(25+1)}{100} \right)^{th} \text{ obs.} \\
 &= \text{Value of } (3.9)^{th} \text{ obs.} \\
 &= \text{Value of } (3)^{th} \text{ obs.} + 0.9 \left[ \text{Value of } (4)^{th} \text{ obs.} - \text{Value of } (3)^{th} \text{ obs.} \right] \\
 &= 19 + 0.9(20 - 19) \\
 &= 19.9 \text{ dollars}
 \end{aligned}$$

Thus, the upper value of hourly wage rate for the lower 15% of the employees is 19.9 dollars. ■

2.) the upper wage rate for the lowest 45% of the employees,

**Solution :**  $P_{45} = 23.7 \text{ dollars}$  ■

3.) the lower wage rate for the upper 25% of the employees.

**Solution :**  $P_{45} = 27.5 \text{ dollars}$  ■

**Definition 2.4.6** A percentile of a grouped data is defined by

$$P_i = l_m + \left( \frac{\left( \frac{i}{100} \right) n - cf_b}{f_w} \right) \times c \quad (2.26)$$

Where

- $l_m$  : is the lower class boundary of the  $i^{th}$  percentile class
- $\left( \frac{i}{100} \right) \times n$  : The  $i^{th}$  percentile class
- $cf_b$  : cumulative frequency of the class just before the  $i^{th}$  percentile class
- $f_w$  : frequency of the percentile class
- $c$  : the class width, difference between class boundaries

**Example 2.4.7** Determine the 95th percentile  $P_{95}$  of the grouped data of Example 2.4.3 on page (pg. 96).

**Solution :** For  $P_{95}$ , the cumulative frequency class with  $\left( \frac{i}{100} \right) (n + 1) = \left( \frac{95}{100} \right) \times 401 = 380.95^{th}$  observation, a cumulative frequency in the class (400–500).  
Therefore,

$$P_i = l_m + \left( \frac{\left( \frac{i}{100} \right) n - cf_b}{f_w} \right) \times c$$

$$P_{95} = 400 + \left( \frac{\left( \frac{95}{100} \right) (400) - 344}{40} \right) \times 100 = 400 + \left( \frac{380 - 344}{40} \right) \times 100 = 490$$

■

**Definition 2.4.7** Inter Percentile range is defined as

$$\text{Inter Percentile range} = P_{90} - P_{10} \quad (2.27)$$

**Definition 2.4.8** Percentile deviation is defined as

$$\text{Percentile deviation} = \frac{P_{90} - P_{10}}{2} \quad (2.28)$$

**Definition 2.4.9** Coefficient of Percentile deviation is defined as

$$\text{Coefficient of Percentile deviation} = \frac{P_{90} - P_{10}}{P_{90} + P_{10}} \quad (2.29)$$



### 2.4.3 Deciles

These values divide the observations into 10 equal parts and are denoted by

$$D_1, \quad D_2, \quad \dots, \quad D_9$$

e.g.  $D_1$  has 10% below it and 90% above, and  $D_2$  has 20% below it and 80% above.

**Definition 2.4.10** If a data set contains  $n$  observations, then the  $i$ th decile is the value

$$D_i = \frac{i}{10}(n+1)^{th} \quad (2.30)$$

in the ordered data set.

**Example 2.4.8** Calculate Deciles-6 from the following data

85, 96, 76, 108, 85, 80, 100, 85, 70, 95

**Solution :** *Arranging Observations in the ascending order, We get :*

70, 76, 80, 85, 85, 85, 95, 96, 100, 108

*Therefore,*

$$D_5 = \frac{5}{10}(n+1) = \frac{6}{10}(10+1) = 6.6^{th} \text{ observation,}$$

*which lies a sixth between the sixth and seventh entry. Therefore,*

$$D_6 = 85 + 0.6(95 - 85) = 91$$

■

**Example 2.4.9** The Quick oil company has a number of outlets in the metropolitan Hoima area. The numbers of changes at the Hoima Street outlet in the past 20 days are:

51 55 56 59 62 62 63 65 66 70  
71 72 73 79 79 80 85 90 94 98

Calculate 5th deciles.

**Solution :** *Data is already arranged in ascending order.*

$$\begin{aligned} D_5 &= \text{Value of } \left( \frac{5(n+1)}{10} \right)^{th} \text{ obs.} \\ &= \text{Value of } \left( \frac{5(20+1)}{10} \right)^{th} \text{ obs.} \\ &= \text{Value of } (10.5)^{th} \text{ obs.} \\ &= \text{Value of } (10)^{th} \text{ obs.} + 0.5 \left[ \text{Value of } (11)^{th} \text{ obs.} - \text{Value of } (10)^{th} \text{ obs.} \right] \\ &= 70 + 0.5(71 - 70) \\ &= 70.5 \end{aligned}$$

■

**Definition 2.4.11** A decile for grouped data is defined by

$$D_i = l_m + \left( \frac{\left( \frac{i}{10} \right) n - cf_b}{f_w} \right) \times c \quad (2.31)$$

Where

- $l_m$  : is the lower class boundary of the  $i^{th}$  decile class
- $\left( \frac{i}{10} \right) \times n$  : The  $i^{th}$  decile class
- $cf_b$  : cumulative frequency of the class just before the  $i^{th}$  decile class
- $f_w$  : frequency of the decile class
- $c$  : the class width, difference between class boundaries

**Example 2.4.10** Determine the 4th decile  $D_4$  for the grouped data of Example 2.4.3 on page (pg. 96).

**Solution :** For  $D_4$ , the cumulative frequency class with  $\left( \frac{i}{10} \right) (n+1) = \left( \frac{4}{10} \right) \times 401 = 160.4^{th}$  observation, a cumulative frequency in the second class (100 – 200). Therefore,

$$D_i = l_m + \left( \frac{\left( \frac{i}{10} \right) n - cf_b}{f_w} \right) \times c$$

$$D_4 = 100 + \left( \frac{\left( \frac{4}{10} \right) (400) - 82}{122} \right) \times 100 = 100 + \left( \frac{160 - 82}{122} \right) \times 100 = 163.93$$

■

**Definition 2.4.12** Inter Decile range is defined as

$$\text{Inter Decile range} = D_9 - D_1 \quad (2.32)$$

**Definition 2.4.13** Decile deviation is defined as

$$\text{Decile deviation} = \frac{D_9 - D_1}{2} \quad (2.33)$$

**Definition 2.4.14** Coefficient of Decile deviation is defined as

$$\text{Coefficient of Decile deviation} = \frac{D_9 - D_1}{D_9 + D_1} \quad (2.34)$$

**Example 2.4.11** Measure of relative standing of an observation in Grouped Data is given below

Class Boundaries	Class Mark	Frequency	Cumulative Frequency
	$X$	$f$	$cf$
85.5 - 90.5	87	6	6
90.5 - 95.5	93	4	10
95.5 - 100.5	98	10	20
100.5 - 105.5	103	6	26
105.5 - 110.5	108	3	29
110.5 - 115.5	113	1	30

Compute

1.)  $P_{10}$

**Solution :**

$$\begin{aligned}P_{10} &= l_m + \left( \frac{\frac{10n}{100} - cf_b}{f_w} \right) \times c \\&= 85.5 + \left( \frac{3 - 0}{6} \right) \times 5 \\&= 85.5 + 2.5 = 88\end{aligned}$$

■

2.)  $P_{25}$

**Solution :**

$$\begin{aligned}P_{25} &= l_m + \left( \frac{\frac{25n}{100} - cf_b}{f_w} \right) \times c \\&= 90.5 + \left( \frac{7.5 - 6}{4} \right) \times 5 \\&= 90.5 + 1.875 = 92.375\end{aligned}$$

■

3.)  $P_{50}$ **Solution :**

$$\begin{aligned}P_{50} &= l_m + \left( \frac{\frac{50n}{100} - cf_b}{f_w} \right) \times c \\&= 95.5 + \left( \frac{15 - 10}{10} \right) \times 5 \\&= 95.5 + 2.5 = 98\end{aligned}$$

■

4.)  $P_{95}$ **Solution :**

$$\begin{aligned}P_{95} &= l_m + \left( \frac{\frac{95n}{100} - cf_b}{f_w} \right) \times c \\&= 105.5 + \left( \frac{28.5 - 26}{3} \right) \times 5 \\&= 105.5 + 4.1667 = 109.6667\end{aligned}$$

■

5.)  $D_1$ **Solution :**

$$\begin{aligned}D_1 &= l_m + \left( \frac{\left( \frac{1}{10} \right) n - cf_b}{f_w} \right) \times c \\&= 85.5 + \left( \frac{3 - 0}{6} \right) \times 5 = 88\end{aligned}$$

■

6.)  $D_7$ **Solution :**

$$\begin{aligned}D_7 &= l_m + \left( \frac{\left( \frac{7}{10} \right) n - cf_b}{f_w} \right) \times c \\&= 100.5 + \left( \frac{21 - 20}{6} \right) \times 5 = 101.333\end{aligned}$$

■

**Exercise 2.13** Calculate 2nd and 8th deciles of following ordered data

13 13 13 20 26 27 31 34 34 34 35 35 36 37 38  
41 41 41 45 47 47 47 50 51 53 54 56 62 67 82

**Solution :**

$$\begin{aligned}D_2 &= 27 + 0.2(31 - 27) = 27.8 \\D_8 &= 54 + 0.8(53 - 51) = 52.6\end{aligned}$$

■

**Example 2.4.12** For the given data 3, 23, 13, 11, 15, 5, 4, 2

1.)

$$\text{Inter Quartile range} = Q_3 - Q_1 = 14.5 - 3.25 = 11.25$$

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2} = \frac{14.5 - 3.25}{2} = \frac{11.25}{2} = 5.625$$

$$\text{Coefficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{14.5 - 3.25}{14.5 + 3.25} = \frac{11.25}{17.75} = 0.6338$$

2.)

$$\text{Inter Decile range} = D_9 - D_1 = 20.7 - 1.8 = 18.9$$

$$\text{Decile deviation} = \frac{D_9 - D_1}{2} = \frac{20.7 - 1.8}{2} = \frac{18.9}{2} = 9.45$$

$$\text{Coefficient of Decile deviation} = \frac{D_9 - D_1}{D_9 + D_1} = \frac{20.7 - 1.8}{20.7 + 1.8} = \frac{18.9}{22.5} = 0.84$$

3.)

$$\text{Inter Percentile range} = P_{90} - P_{10} = 20.7 - 1.8 = 18.9$$

$$\text{Percentile deviation} = \frac{P_{90} - P_{10}}{2} = \frac{20.7 - 1.8}{2} = \frac{18.9}{2} = 9.45$$

$$\text{Coefficient of Percentile deviation} = \frac{P_{90} - P_{10}}{P_{90} + P_{10}} = \frac{20.7 - 1.8}{20.7 + 1.8} = \frac{18.9}{22.5} = 0.84$$

**Remark 2.4.2** Recall to first rearrange the data in ascending order.

## 2.5 Measures of Shape

These include

- 1.) Skewness
- 2.) Kurtosis

### 2.5.1 Skewness

**Definition 2.5.1** When the items in a distribution are dispersed equally on each side of the mean, we say that the distribution is *symmetrical*.

**Definition 2.5.2** When the items are *not symmetrically* dispersed on each side of the mean, we say that the distribution is **skewed** or **asymmetric**.

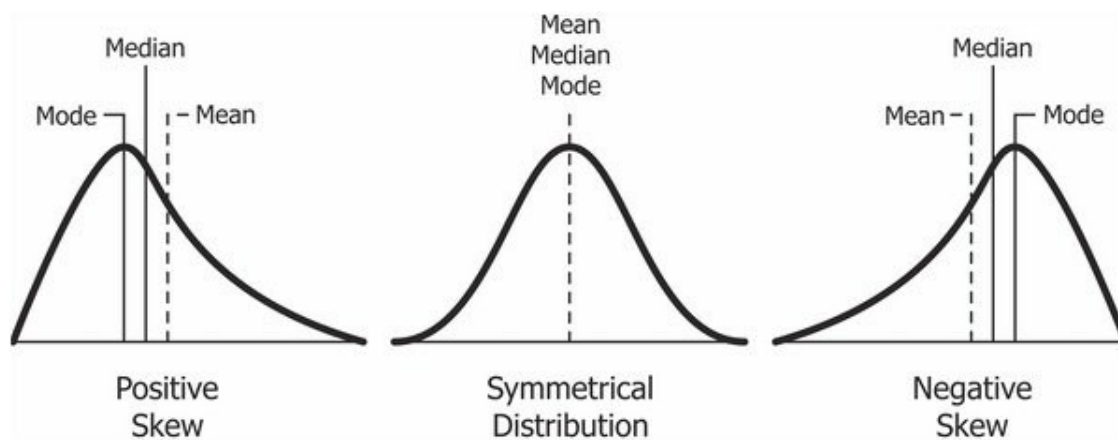
**Definition 2.5.3** **Skewness** is a *measure of asymmetry or distortion of symmetric distribution*.

It measures the deviation of the given distribution of a random variable from a symmetric distribution, such as normal distribution.

**Example 2.5.1** A normal distribution is without any skewness, as it is symmetrical on both sides. Hence, a curve is regarded as skewed if it is shifted towards the right or the left.

Two distributions may have the same mean and the same standard deviation but they may be differently skewed. This will be obvious if you look at one of the skewed distributions and then look at the same one through from the other side of the paper! What, then, does skewness tell us? It tells us that we are to expect a few unusually high values in a positively skewed distribution or a few unusually low values in a negatively skewed distribution.

**Note 2.5.1** If a distribution is **symmetrical**, the mean, mode and the median all occur at the same point, i.e. right in the middle. But in a **skew distribution**, the mean and the median lie somewhere along the side with the “tail”, although the mode is still at the point where the curve is highest.



**Note 2.5.2** The more skew the distribution, the greater the distance from the mode to the mean and the median, but these two are always in the same order; working outwards from the mode, the median comes first and then the mean.

### 2.5.1.1 Positive Skewness

If the given distribution is shifted to the left and with *its tail on the right side*, it is a positively skewed distribution. It is also called the right-skewed distribution. A tail is referred to as the tapering of the curve in a different way from the data points on the other side.

As the name suggests, a positively skewed distribution assumes a skewness value of more than zero. Since the skewness of the given distribution is on the right, the mean value is greater than the median and moves towards the right, and the mode occurs at the highest frequency of the distribution.

### 2.5.1.2 Negative Skewness

If the given distribution is shifted to the right and with *its tail on the left side*, it is a negatively skewed distribution. It is also called a left-skewed distribution. The skewness value of any distribution showing a negative skew is always less than zero. The skewness of the given distribution is on the left; hence, the mean value is less than the median and moves towards the left, and the mode occurs at the highest frequency of the distribution.

### 2.5.1.3 Measuring Skewness

**Definition 2.5.4** Karl Pearson's coefficient of skewness is given by

$$S_K = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} \quad (2.35)$$

or can be written from the relation

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

or equivalently

$$\text{Mode} = 3\text{Median} - 2\text{Mean}$$

For most distributions, except for those with very long tails, the relationship holds approximately.

**Definition 2.5.5** Karl Pearson's coefficient of skewness is given by

$$S_K = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}} \quad (2.36)$$

**Definition 2.5.6** Karl Pearson defined the following  $\beta$  and  $\gamma$  coefficients of skewness, based upon the second and third central moment

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad (2.37)$$

It is used as measure of skewness. For a symmetrical distribution,  $\beta_1$  shall be zero.  $\beta_1$  as a measure of skewness does not tell about the direction of skewness, i.e. positive or negative. Because  $\mu_3$  being the sum of cubes of the deviations from mean may be positive or negative but  $\mu_3^2$  is always positive. Also,  $\mu_2$  being the variance always positive.

Hence,  $\beta_1$  would be always positive. This drawback is removed if we calculate Karl Pearson's Gamma coefficient  $\gamma_1$  which is the square root of  $\beta_1$  i.e

$$\gamma_1 = \pm\sqrt{\beta_1}$$

**Definition 2.5.7** Bowleys's Coefficient of Skewness. This method is based on quartiles. The formula for calculating coefficient of skewness is given by

$$S_K = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \quad (2.38)$$

**Definition 2.5.8** Kelly's Coefficient of Skewness. The coefficient of skewness proposed by Kelly is based on percentiles and deciles. The formula for calculating the coefficient of skewness is given by

$$S_K = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{10})} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} \quad (2.39)$$

or, Kelly's Coefficient of Skewness based on Deciles given by

$$S_K = \frac{(D_9 - D_5) - (D_5 - D_1)}{(D_9 - D_1)} = \frac{D_9 - 2D_5 + D_1}{D_9 - D_1} \quad (2.40)$$

**Definition 2.5.9** Skewness of an ungrouped sample is defined as

$$\text{Skewness} = \frac{\sum (X - \bar{X})^3}{(n - 1)S^3} \quad (2.41)$$

or for grouped sample

$$\text{Skewness} = \frac{\sum f (X - \bar{X})^3}{(n - 1)S^3} \quad (2.42)$$

where  $S$  is the sample standard deviation.

**Definition 2.5.10** Skewness of an ungrouped population is defined as

$$\text{Skewness} = \frac{\sum (X - \mu)^3}{n\sigma^3} \quad (2.43)$$

or for grouped population

$$\text{Skewness} = \frac{\sum f (X - \mu)^3}{n\sigma^3} \quad (2.44)$$

where  $\sigma$  is the population standard deviation.

**Remark 2.5.1** The value of the coefficient of skewness is between  $-3$  and  $+3$ , although values below  $-1$  and above  $+1$  are rare and indicate very skewed distributions.



**Example 2.5.2** Calculate **Sample Skewness** from the following **grouped** data

$X$	Frequency $f$	$fX$	$X - \bar{X}$	$(X - \bar{X})^2$	$f \cdot (X - \bar{X})^2$	$fX^2$	$f \cdot (X - \bar{X})^3$	$f \cdot (X - \bar{X})^4$
0	1	0	-2.2	4.84	4.84	0	-10.648	23.4256
1	5	5	-1.2	1.44	7.2	5	-8.64	10.368
2	10	20	-0.2	0.04	0.4	40	-0.08	0.016
3	6	18	0.8	0.64	3.84	54	3.072	2.4576
4	3	12	1.8	3.24	9.72	48	17.496	31.4928
	<b>25</b>	<b>55</b>			<b>26</b>	<b>147</b>	<b>1.2</b>	<b>67.76</b>

For a grouped sample, the Mean is given by

$$\bar{X} = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i} = \frac{55}{25} = 2.2$$

For a grouped sample, the variance is given by Equation (2.20) on page (pg. 86)

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n f_i (X_i - \bar{X})^2 = \frac{26}{24}$$

$$S = \sqrt{\frac{26}{24}} = 1.0408$$

or by any of its respective forms,

$$S^2 = \frac{1}{(n-1)} \left[ \sum_{i=1}^n f_i X_i^2 - \frac{\left( \sum_{i=1}^n f_i X_i \right)^2}{n} \right]$$

$$S^2 = \frac{1}{n(n-1)} \left[ n \sum_{i=1}^n f_i X_i^2 - \left( \sum_{i=1}^n f_i X_i \right)^2 \right]$$

$$S^2 = \frac{\sum_{i=1}^n f_i X_i^2}{(n-1)} - \frac{\left( \sum_{i=1}^n f_i X_i \right)^2}{n(n-1)}$$

Therefore,

$$\text{Skewness} = \frac{\sum f (X - \bar{X})^3}{(n-1)S^3} = \frac{1.2}{24(1.0408)^3} = 0.0443$$

**Example 2.5.3** For a distribution Karl Pearson's coefficient of skewness is 0.64, standard deviation is 13 and mean is 59.2 Find mode and median.

**Solution :**

$$\begin{aligned}S_K &= \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} \\0.64 &= \frac{59.2 - \text{Mode}}{13} \Rightarrow \\ \text{Mode} &= 50.88\end{aligned}$$

*Applying the Mean-Median-Mode relations*

$$\text{Mode} = 3\text{Median} - 2\text{Mean}$$

*To have*

$$\text{Median} = \frac{1}{3} [\text{Mode} + 2\text{Mean}] = \frac{1}{3} [50.88 + 2(59.2)] = 56.42$$

■

**Example 2.5.4** Karl Pearson's coefficient of population skewness is 1.28, its mean is 164 and mode 100, find the standard deviation.

**Solution :**

$$\begin{aligned}S_K &= \frac{\text{Mean} - \text{Mode}}{\sigma} \\1.28 &= \frac{164 - 100}{\sigma} \Rightarrow \sigma = 50\end{aligned}$$

■

**Exercise 2.14** For a frequency distribution the Bowley's coefficient of skewness is 1.2. If the sum of the 1st and 3rd quarterlies is 200 and median is 76, find the value of third quartile.

**Solution :** *We have that*

$$\begin{aligned}S_K &= 1.2 \\Q_1 + Q_3 &= 200 \\Q_2 &= 76\end{aligned}$$

*From*

$$\begin{aligned}S_K &= 1.2 \\ \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} &= 1.2 \Rightarrow Q_3 - Q_1 = 40\end{aligned}$$

*Solving*

$$\begin{aligned}Q_3 - Q_1 &= 40 \\Q_3 + Q_1 &= 200\end{aligned}$$

*simultaneously, we get  $Q_3 = 120$*

■

**Example 2.5.5** The following are the marks of 150 students (a population, whole class, skewness is always for a distribution not a sample) in an examination. Calculate Karl Pearson's coefficient of skewness.

Marks	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80
No. of Students	10	40	20	0	10	40	16	14

**Solution :** *Class boundaries given not class limits. For the grouped data*

<i>Class Boundary</i>	<i>Class Mark X</i>	<i>Frequency f</i>	<i>Cumulative Frequency cf</i>	<i>fX</i>	<i>fX<sup>2</sup></i>
0 - 10	5	10	10	50	250
10 - 20	15	40	50	600	9000
20 - 30	25	20	70	500	12500
30 - 40	35	0	70	0	0
40 - 50	45	10	80	450	20250
50 - 60	55	40	120	2200	121000
60 - 70	65	16	136	1040	67600
70 - 80	75	14	150	1050	78750
		$\sum f = 150$		$\sum fX = 5890$	$\sum fX^2 = 309350$

*Determine*

$$1.) \text{ Mean: } \mu = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i} = \frac{5890}{150} = 39.27$$

$$2.) \text{ Median: } \frac{150}{2} = 75, \text{ so the median class is } (40 - 50)$$

$$\text{Median} = l_m + \left[ \frac{\frac{N}{2} - cf_b}{f_m} \right] \times c = 40 + \left( \frac{75 - 70}{10} \right) \times 10 = 45$$

3.) *Variance for the population given*

$$\sigma^2 = \frac{1}{n} \left[ \sum_{i=1}^n f_i X_i^2 - \frac{\left( \sum_{i=1}^n f_i X_i \right)^2}{n} \right] = \frac{1}{150} \left[ 309350 - \frac{5890^2}{150} \right] = 520.462$$

4.) *Standard deviation of the population*

$$\sigma = \sqrt{\sigma^2} = \sqrt{533.61} = 22.81$$

such that

$$S_K = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}} = \frac{3(39.27 - 45)}{22.81} = -0.754$$

■

**Example 2.5.6** Calculate **Population Skewness** from the following grouped data

Class	2 - 4	4 - 6	6 - 8	8 - 10
Frequency	3	4	2	1

A more detailed frequency table is illustrated below.

Class	X	Frequency f	fX	X - $\mu$	(X - $\mu$ ) <sup>2</sup>	f · (X - $\mu$ ) <sup>2</sup>	fX <sup>2</sup>	f · (X - $\mu$ ) <sup>3</sup>	f · (X - $\mu$ ) <sup>4</sup>
2 - 4	3	3	9	-2.2	4.84	14.52	27	-31.944	70.2768
4 - 6	5	4	20	-0.2	0.04	0.16	100	-0.032	0.0064
6 - 8	7	2	14	1.8	3.24	6.48	98	11.664	20.9952
8 - 10	9	1	9	3.8	14.44	14.44	81	54.872	208.5136
		<b>10</b>	<b>52</b>			<b>35.6</b>	<b>306</b>	<b>34.56</b>	<b>299.792</b>

For a grouped population, the Mean is given by

$$\mu = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i} = \frac{52}{10} = 5.2$$

For a grouped population, the variance is given by

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n f_i (X_i - \mu)^2 = \frac{35.6}{10}$$

$$\sigma = \sqrt{\frac{35.6}{10}} = 1.8868$$

Therefore,

$$\text{Skewness} = \frac{\sum f (X - \mu)^3}{n\sigma^3} = \frac{34.56}{10(1.8868)^3} = 0.5145$$

**2.5.1.4 How to Interpret**

- 1.) Skewness also includes the extremes of the dataset instead of focusing only on the average. Hence, investors take note of skewness while estimating the distribution of returns on investments. The average of the data set works out in case an investor holds a position for the long term. Therefore, extremes need to be looked at when investors seek short-term and medium-term security positions.
- 2.) Usually, a *standard deviation* is used by investors for prediction of returns, and standard deviation presumes a normal distribution with zero skewness. However, because of skewness risk, it is better to obtain the performance estimations based on skewness. Moreover, the occurrence of return distributions coming close to normal is low.
- 3.) Skewness risk occurs when a symmetric distribution is applied to the skewed data. The financial models seeking to estimate an asset's future performance consider a normal distribution. However, skewed data will increase the accuracy of the financial model.
- 4.) If a return distribution shows a positive skew, investors can expect recurrent small losses and few large returns from investment. Conversely, a negatively skewed distribution implies many small wins and a few large losses on the investment.
- 5.) Hence, a positively skewed investment return distribution should be preferred over a negatively skewed return distribution since the huge gains may cover the frequent – but small – losses. However, investors may prefer investments with a negatively skewed return distribution. It may be because they prefer frequent small wins and a few huge losses over frequent small losses and a few large gains.

**2.5.1.5 Difference between Variance and Skewness**

The following two points of difference between variance and skewness should be carefully noted.

- 1.) Variance tells us about the amount of variability while skewness gives the direction of variability.
- 2.) In business and economic series, measures of variation have greater practical application than measures of skewness. However, in medical and life science field measures of skewness have greater practical applications than the variance.

**2.5.1.6 Remarks about Skewness**

1. If the value of mean, median and mode are same in any distribution, then the skewness does not exist in that distribution. Larger the difference in these values, larger the skewness;
2. If sum of the frequencies are equal on the both sides of mode then skewness does not exist;
3. If the distance of first quartile and third quartile are same from the median then a skewness does not exist. Similarly if deciles (first and ninth) and percentiles (first and ninety nine) are at equal distance from the median. Then there is no asymmetry;
4. If the sums of positive and negative deviations obtained from mean, median or mode are equal then there is no asymmetry; and
5. If a graph of a data become a normal curve and when it is folded at middle and one part overlap fully on the other one then there is no asymmetry.

### 2.5.2 Kurtosis

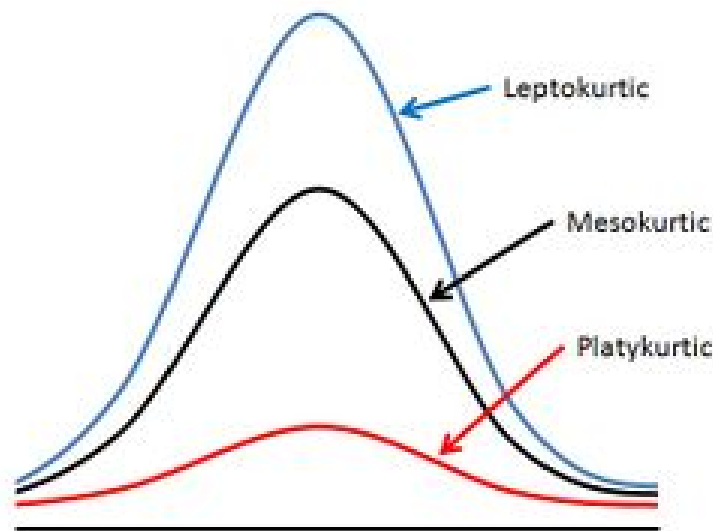
If we have the knowledge of the measures of central tendency, dispersion and skewness, even then we cannot get a complete idea of a distribution. In addition to these measures, we need to know another measure to get the complete *idea about the shape of the distribution* which can be studied with the help of Kurtosis. Prof. Karl Pearson has called it the “Convexity of a Curve”. Kurtosis gives a measure of flatness of distribution.

**Definition 2.5.11** **Kurtosis** is the *degree of peakedness of a distribution*, usually taken relative to normal distribution. We can distinguish between 3 different forms of kurtosis.

**Remark 2.5.2** Kurtosis is actually the measure of outliers present in the distribution.

**Definition 2.5.12** The degree of kurtosis of a distribution is measured relative to that of a normal curve.

- 1.) The curves with greater peakedness than the normal curve are called “**Leptokurtic**”.
- 2.) The curves which are more flat than the normal curve are called “**Platykurtic**”.
- 3.) The normal curve is called “**Mesokurtic**.”



- 1.) Leptokurtic: This is a distribution having a relatively high pick. It has a narrower central portion and higher tails.
- 2.) Platykurtic: a distribution with a flat top. It has a broader central portion and lower tails.
- 3.) Mesokurtic: It is a form exhibited by a normal/bell shaped distribution. It is neither high peaked nor flat topped.

**2.5.2.1 Measure Of Kurtosis****1. Karl Pearson's Measures of Kurtosis**

For calculating the kurtosis, the second and fourth central moments of variable are used. For this, following formula given by Karl Pearson is used:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \quad (2.45)$$

or

$$\gamma_2 = \beta_2 - 3 \quad (2.46)$$

where

$\mu_2$  = Second order central moment of distribution

$\mu_4$  = Fourth order central moment of distribution

Description:

- (a) If  $\beta_2 = 3$  or  $\gamma_2 = 0$ , then curve is said to be mesokurtic;
- (b) If  $\beta_2 < 3$  or  $\gamma_2 < 0$ , then curve is said to be platykurtic;
- (c) If  $\beta_2 > 3$  or  $\gamma_2 > 0$ , then curve is said to be leptokurtic;

**2. Kelly's Measure of Kurtosis**

Kelly has given a measure of kurtosis based on percentiles. The formula is given by

$$\beta_2 = \frac{P_{75} - P_{25}}{P_{90} - P_{10}} \quad (2.47)$$

where  $P_{75}$ ,  $P_{25}$ ,  $P_{90}$  and  $P_{10}$  are the 75<sup>th</sup>, 25<sup>th</sup>, 90<sup>th</sup> and 10<sup>th</sup> percentiles of dispersion respectively.

Description:

- (a) If  $\beta_2 > 0.26315$ , then curve is said to be platykurtic.
- (b) If  $\beta_2 < 0.26315$ , then curve is said to be leptokurtic.

**Definition 2.5.13** Kurtosis of an ungrouped sample is defined as

$$\text{Kurtosis} = \frac{\sum (X - \bar{X})^4}{(n - 1)S^4} \quad (2.48)$$

or for grouped sample

$$\text{Kurtosis} = \frac{\sum f (X - \bar{X})^4}{(n - 1)S^4} \quad (2.49)$$

where  $S$  is the sample standard deviation.

**Definition 2.5.14** Kurtosis of an **ungrouped population** is defined as

$$\text{Kurtosis} = \frac{\sum (X - \mu)^4}{n\sigma^4} \quad (2.50)$$

or for **grouped population**

$$\text{Kurtosis} = \frac{\sum f (X - \mu)^4}{n\sigma^4} \quad (2.51)$$

where  $\sigma$  is the population standard deviation.

**Example 2.5.7** Calculate Sample Kurtosis from the grouped data of Example 2.5.2 on page (pg. 108).

$$Kurtosis = \frac{\sum f (X - \bar{X})^4}{(n - 1)S^4} = \frac{67.76}{24(1.0408)^4} = 2.405986$$

**Example 2.5.8** First four moments about mean of a distribution are 0, 2.5, 0.7 and 18.75. Find coefficient of skewness and kurtosis.

**Solution :** We have  $\mu_1 = 0$ ,  $\mu_2 = 2.5$ ,  $\mu_3 = 0.7$  and  $\mu_4 = 18.7$  Therefore,

$$\begin{aligned} \text{Skewness, } \beta_1 &= \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = 0.031 \\ \text{Kurtosis, } \beta_2 &= \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = 3 \end{aligned}$$

■

**Exercise 2.15** The first four raw moments of a distribution are 2, 136, 320, and 40,000. Find out coefficients of skewness and kurtosis.

**Solution :** We have  $\mu'_1 = 2$ ,  $\mu'_2 = 136$ ,  $\mu'_3 = 320$  and  $\mu'_4 = 40000$ .

First compute the first four **moments about the mean**

$$\begin{aligned} \mu_1 &= \mu'_1 - \mu'_1 = 0 \\ \mu_2 &= \mu'_2 - (\mu'_1)^2 = 136 - 2^2 = 132 \\ \mu_3 &= \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3 \\ &= 320 - 3(132)(2) + 2(2)^3 \\ &= -456 \\ \mu_4 &= \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4 \\ &= 40,000 - 4(2)(320) + 6(136)(2^2) - 3(2^4) \\ &= 40656 \end{aligned}$$

Therefore, the coefficients are given by

$$\begin{aligned} \text{Skewness, } \beta_1 &= \frac{\mu_3^2}{\mu_2^3} = \frac{(-456)^2}{(132)^3} = 0.0904 \\ \text{Kurtosis, } \beta_2 &= \frac{\mu_4}{\mu_2^2} = \frac{40656}{(132)^2} = 2.333 \end{aligned}$$

■



**Example 2.5.9** Calculate Sample Skewness, Sample Kurtosis from the following grouped data

Class	2 - 4	4 - 6	6 - 8	8 - 10
Frequency	3	4	2	1

A more detailed frequency table is illustrated below.

Class	$X$	Frequency $f$	$fX$	$X - \bar{X}$	$(X - \bar{X})^2$	$f \cdot (X - \bar{X})^2$	$fX^2$	$f \cdot (X - \bar{X})^3$	$f \cdot (X - \bar{X})^4$
2 - 4	3	3	9	-2.2	4.84	14.52	27	-31.944	70.2768
4 - 6	5	4	20	-0.2	0.04	0.16	100	-0.032	0.0064
6 - 8	7	2	14	1.8	3.24	6.48	98	11.664	20.9952
8 - 10	9	1	9	3.8	14.44	14.44	81	54.872	208.5136
		<b>10</b>	<b>52</b>			<b>35.6</b>	<b>306</b>	<b>34.56</b>	<b>299.792</b>

For a grouped sample, the Mean is given by

$$\bar{X} = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i} = \frac{52}{10} = 5.2$$

For a grouped sample, the variance is given by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n f_i (X_i - \bar{X})^2 = \frac{35.6}{9}$$
$$S = \sqrt{\frac{35.6}{9}} = 1.9889$$

Therefore,

$$\text{Skewness} = \frac{\sum f (X - \bar{X})^3}{(n-1)S^3} = \frac{34.56}{9(1.9889)^3} = 0.4881$$

$$\text{Kurtosis} = \frac{\sum f (X - \bar{X})^4}{(n-1)S^4} = \frac{299.792}{9(1.9889)^4} = 2.1289$$

**Example 2.5.10** Calculate **Population** Kurtosis from the grouped data of Example 2.5.6 on page (pg. 111).

$$Kurtosis = \frac{\sum f(X - \mu)^4}{n\sigma^4} = \frac{299.792}{10(1.8868)^4} = 2.3655$$

**Example 2.5.11** Calculate Population Skewness, Population Kurtosis from the following grouped data

Class	0 - 2	2 - 4	4 - 6	6 - 8	8 - 10
Frequency	10	20	30	20	10

A more detailed frequency table is illustrated below.

Class	$X$	Frequency $f$	$fX$	$X - \mu$	$f \cdot (X - \mu)^2$	$f \cdot (X - \mu)^3$	$f \cdot (X - \mu)^4$
0 - 2	1	10	10	-4	160	-640	2560
2 - 4	3	20	60	-2	80	-160	320
4 - 6	5	30	150	0	0	0	0
6 - 8	7	20	140	2	80	160	320
8 - 10	9	10	90	4	160	640	2560
		<b>90</b>	<b>450</b>		<b>480</b>	<b>0</b>	<b>5760</b>

For a grouped population, the Mean is given by

$$\mu = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i} = \frac{450}{90} = 5$$

For a grouped population, the variance is given by

$$\begin{aligned}\sigma^2 &= \frac{1}{n} \sum_{i=1}^n f_i (X_i - \mu)^2 = \frac{480}{90} \\ \sigma &= \sqrt{\frac{480}{90}} = 2.3094\end{aligned}$$

Therefore,

$$\begin{aligned}\text{Skewness} &= \frac{\sum f(X - \mu)^3}{n\sigma^3} = \frac{0}{90(2.3094)^3} = 0 \\ \text{Kurtosis} &= \frac{\sum f(X - \mu)^4}{n\sigma^4} = \frac{5760}{90(2.3094)^4} = 2.25\end{aligned}$$

**Example 2.5.12** More figures of Skewness and Kurtosis

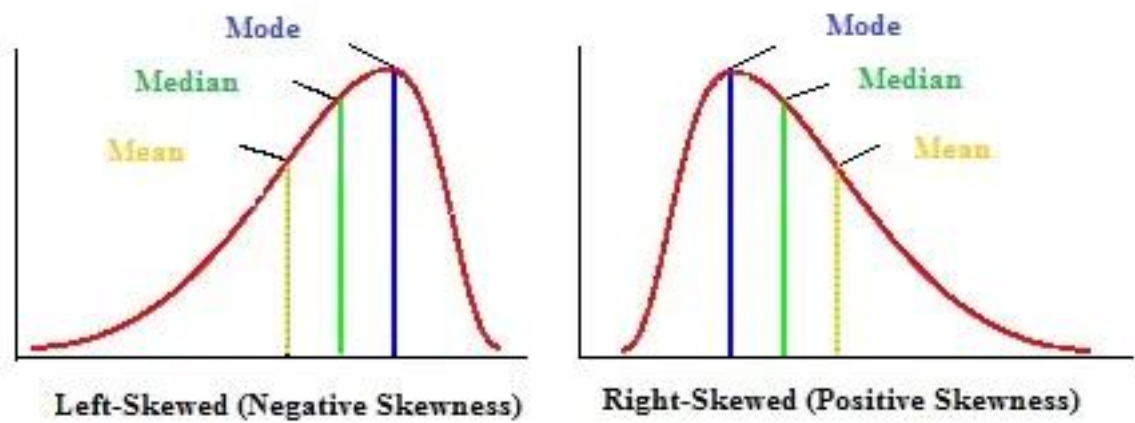


Figure 2.1: Skewness of a distribution

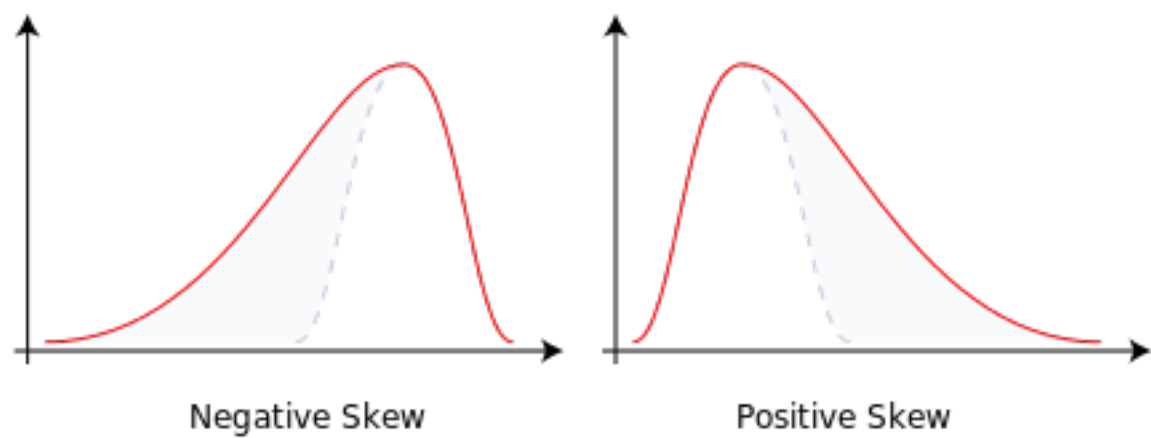


Figure 2.2: Skewness Example

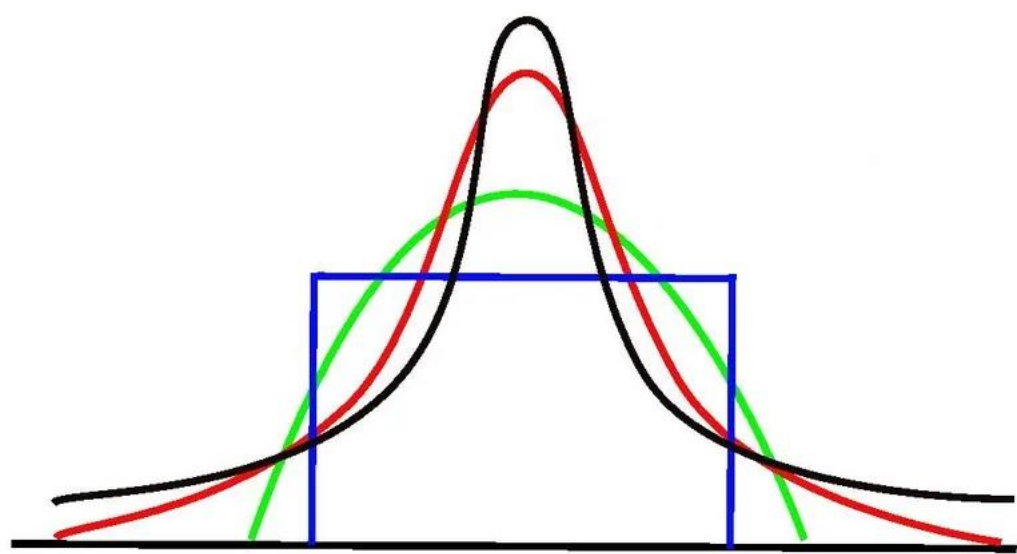


Figure 2.3: Kurtosis of a distribution

## 2.6 Moments

### 2.6.1 Central Moments

The  $r$ -th central moment of a random variable  $X$  with expected value  $\mathbb{E}[X] = \mu$  is defined as:

$$\mu_r = \mathbb{E}[(X - \mu)^r]$$

where:

- $\mu$  is the mean of  $X$
- $r$  is a positive integer
- $\mathbb{E}[\cdot]$  denotes the expectation operator

#### 2.6.1.1 Examples of Central Moments

##### 1. First Central Moment:

$$\mu_1 = \mathbb{E}[(X - \mu)] = 0$$

Since the expectation of deviations from the mean is always zero.

**Proof:**

$$\begin{aligned}\mu_1 &= \mathbb{E}[X - \mu] \\ &= \mathbb{E}(X) - \mu \\ &= \mathbb{E}(X) - \mathbb{E}(X) \\ &= 0.\end{aligned}$$

■

##### 2. Second Central Moment (Variance):

$$\mu_2 = \mathbb{E}[(X - \mu)^2] = \sigma^2$$

which is known as the variance of  $X$ .

##### 3. Third Central Moment (Skewness Measure):

$$\mu_3 = \mathbb{E}[(X - \mu)^3]$$

This moment helps measure the skewness (asymmetry) of the distribution.

##### 4. Fourth Central Moment (Kurtosis Measure):

$$\mu_4 = \mathbb{E}[(X - \mu)^4]$$

This moment is used to measure the peakedness or tailedness of the distribution.

**Remark 2.6.1** The central moments of a random variable  $X$  are given by:

$$\mu'_r = \mu_r = E[(X - \mu)^r], \quad (2.52)$$

where  $\mu = E[X]$  is the mean of  $X$ .

**Example 2.6.1** Consider the following dataset:

$$X = \{2, 4, 6, 8\}$$

with equal probabilities. We will compute the first four central moments.

1. Compute the Mean

$$\mu = \frac{2 + 4 + 6 + 8}{4} = 5.$$

2. Compute Central Moments We compute the central moments using the formula  $\mu_r = E[(X - \mu)^r]$ .

- **First Central Moment ( $\mu_1$ )**

$$\mu_1 = E[(X - 5)] = 0.$$

- **Second Central Moment ( $\mu_2$ ) - Variance**

$$\mu_2 = \frac{(2 - 5)^2 + (4 - 5)^2 + (6 - 5)^2 + (8 - 5)^2}{4} = \frac{9 + 1 + 1 + 9}{4} = 5.$$

- **Third Central Moment ( $\mu_3$ ) - Skewness**

$$\mu_3 = \frac{(2 - 5)^3 + (4 - 5)^3 + (6 - 5)^3 + (8 - 5)^3}{4} = \frac{(-27) + (-1) + (1) + (27)}{4} = 0.$$

The skewness coefficient is:

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = 0,$$

indicating a symmetric distribution.

- **Fourth Central Moment ( $\mu_4$ ) - Kurtosis**

$$\mu_4 = \frac{(2 - 5)^4 + (4 - 5)^4 + (6 - 5)^4 + (8 - 5)^4}{4} = \frac{81 + 1 + 1 + 81}{4} = 41.$$

The kurtosis coefficient is:

$$\beta_2 = \frac{\mu_4}{\sigma^4} = \frac{41}{5^2} = 1.64.$$

Since  $\beta_2 < 3$ , the distribution is platykurtic (flatter than normal).

**Remark 2.6.2** Conclusion

- The variance ( $\mu_2$ ) is 5, indicating the spread of the data.
- The skewness ( $\mu_3$ ) is 0, meaning the distribution is symmetric.
- The kurtosis ( $\beta_2$ ) is 1.64, suggesting a platykurtic distribution.

This example illustrates the computation of central moments and their interpretation in statistics.

**Example 2.6.2** Calculating the Second Central Moment (Variance).

Given the dataset  $X = \{2, 4, 6, 8, 10\}$ , calculate the second central moment.

1. Calculate the mean ( $\mu$ ):

$$\mu = \frac{2 + 4 + 6 + 8 + 10}{5} = \frac{30}{5} = 6$$

2. Compute the squared deviations from the mean:

$$(2 - 6)^2 = 16$$

$$(4 - 6)^2 = 4$$

$$(6 - 6)^2 = 0$$

$$(8 - 6)^2 = 4$$

$$(10 - 6)^2 = 16$$

3. Calculate the second central moment (variance):

$$\mu_2 = \frac{16 + 4 + 0 + 4 + 16}{5} = \frac{40}{5} = 8$$

The second central moment (variance) is 8.

**Example 2.6.3** Calculating the Third Central Moment (Skewness)

Given the dataset  $X = \{1, 2, 2, 3, 12\}$ , calculate the third central moment.

1. Calculate the mean ( $\mu$ ):

$$\mu = \frac{1 + 2 + 2 + 3 + 12}{5} = \frac{20}{5} = 4$$

2. Compute the cubed deviations from the mean:

$$(1 - 4)^3 = -27$$

$$(2 - 4)^3 = -8$$

$$(2 - 4)^3 = -8$$

$$(3 - 4)^3 = -1$$

$$(12 - 4)^3 = 512$$

3. Calculate the third central moment:

$$\mu_3 = \frac{-27 + (-8) + (-8) + (-1) + 512}{5} = \frac{468}{5} = 93.6$$

The third central moment is 93.6.

**Example 2.6.4** Calculating the Fourth Central Moment (Kurtosis)

Given the dataset  $X = \{3, 5, 7, 9, 11\}$ , calculate the fourth central moment.

1. Calculate the mean ( $\mu$ ):

$$\mu = \frac{3 + 5 + 7 + 9 + 11}{5} = \frac{35}{5} = 7$$

2. Compute the fourth power deviations from the mean:

$$(3 - 7)^4 = 256$$

$$(5 - 7)^4 = 16$$

$$(7 - 7)^4 = 0$$

$$(9 - 7)^4 = 16$$

$$(11 - 7)^4 = 256$$

3. Calculate the fourth central moment:

$$\mu_4 = \frac{256 + 16 + 0 + 16 + 256}{5} = \frac{544}{5} = 108.8$$

The fourth central moment is 108.8.

### 2.6.2 Raw Moments

**Definition 2.6.1** The  $r$ -th raw moment of a random variable  $X$  is defined as:

$$\mu'_r = M_r = \mathbb{E}[X^r]$$

where:

- $\mu'_r = M_r$  is the  $r$ -th raw moment.
- $\mathbb{E}[\cdot]$  denotes the expectation operator.
- $r$  is a positive integer representing the order of the moment.

#### 2.6.2.1 Examples of Raw Moments

**1. First Raw Moment (Mean):**

$$\mu'_1 = M_1 = \mathbb{E}[X] = \mu$$

This is simply the mean of the distribution.

**2. Second Raw Moment:**

$$\mu'_2 = M_2 = \mathbb{E}[X^2]$$

This moment is used in variance calculations.

**3. Third Raw Moment:**

$$\mu'_3 = M_3 = \mathbb{E}[X^3]$$

Helps in analyzing skewness.

**4. Fourth Raw Moment:**

$$\mu'_4 = M_4 = \mathbb{E}[X^4]$$

Used in kurtosis calculations.

**Remark 2.6.3** The  $r$ -th raw moment of a random variable  $X$  is defined as:

$$\mu'_r = E[X^r] = \sum_{i=1}^n x_i^r P(x_i) \quad (\text{for discrete case})$$

or

$$\mu'_r = \int_{-\infty}^{\infty} x^r f(x) dx \quad (\text{for continuous case})$$

**Example 2.6.5** Given a discrete probability distribution:

$X$	$P(X)$
1	0.2
2	0.5
3	0.3

We calculate the first and second raw moments:



**• First raw moment ( $\mu'_1$ ):**

- Compute  $E[X]$ :

$$\mu'_1 = (1 \times 0.2) + (2 \times 0.5) + (3 \times 0.3)$$

- Simplify:

$$\mu'_1 = 0.2 + 1.0 + 0.9 = 2.1$$

**• Second raw moment ( $\mu'_2$ ):**

- Compute  $E[X^2]$ :

$$\mu'_2 = (1^2 \times 0.2) + (2^2 \times 0.5) + (3^2 \times 0.3)$$

- Simplify:

$$\mu'_2 = (1 \times 0.2) + (4 \times 0.5) + (9 \times 0.3)$$

$$\mu'_2 = 0.2 + 2.0 + 2.7 = 4.9$$

**Example 2.6.6** Consider a **discrete** random variable  $X$  with the following probability mass function:

$$X : \quad 1, 2, 3$$

$$P(X) : \quad 0.2, 0.5, 0.3$$

**1. First raw moment ( $r = 1$ ):**

$$\mu'_1 = E[X] = (1 \times 0.2) + (2 \times 0.5) + (3 \times 0.3) = 2.1$$

**2. Second raw moment ( $r = 2$ ):**

$$\mu'_2 = E[X^2] = (1^2 \times 0.2) + (2^2 \times 0.5) + (3^2 \times 0.3) = 4.3$$

**3. Third raw moment ( $r = 3$ ):**

$$\mu'_3 = E[X^3] = (1^3 \times 0.2) + (2^3 \times 0.5) + (3^3 \times 0.3) = 10.9$$

**Example 2.6.7** Consider a **continuous** random variable  $X$  with the probability density function (PDF):

$$f(x) = 2x, \quad 0 \leq x \leq 1.$$

**1. First raw moment ( $r = 1$ ):**

$$\mu'_1 = \int_0^1 x(2x) dx = \int_0^1 2x^2 dx = \frac{2}{3}$$

**2. Second raw moment ( $r = 2$ ):**

$$\mu'_2 = \int_0^1 x^2(2x) dx = \int_0^1 2x^3 dx = \frac{1}{2}$$

**3. Third raw moment ( $r = 3$ ):**

$$\mu'_3 = \int_0^1 x^3(2x) dx = \int_0^1 2x^4 dx = \frac{2}{5}$$

## 2.7 The Box And Whisker Plot

What is a box and whisker plot?

A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.

In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum.

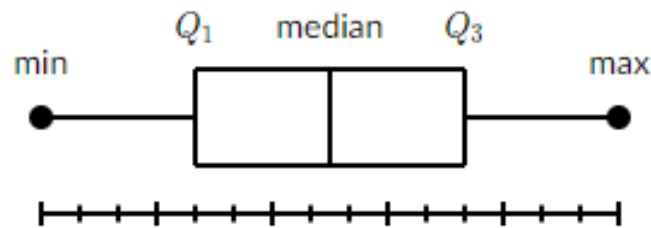


Figure 2.4: Structure of a box plot.

### 2.7.1 Finding The Five-number Summary

**Example 2.7.1** A sample of 10 boxes of raisins has these weights (in grams):

25, 28, 29, 29, 30, 34, 35, 35, 37, 38

. Make a box plot of the data.

**Solution :**

*Step 1: Order the data from smallest to largest.*

*Our data is already in order.*

25, 28, 29, 29, 30, 34, 35, 35, 37, 38

*Step 2: Find the median.*

*The median is the mean of the middle two numbers:*

$$\text{median} = \frac{30 + 34}{2} = 32$$

*Find the quartiles.*

*The first quartile is the median of the data points to the left of the median.*

$$Q_1 = 29$$

*The third quartile is the median of the data points to the right of the median.*

$$Q_3 = 35$$

*Step 4: Complete the five-number summary by finding the min and the max.*

*The min is the smallest data point, which is 25*

*The max is the largest data point, which is 38*

*The five-number summary is 25, 29, 32, 35, 38.*

*Making a box plot*

*Step 1: Scale and label an axis that fits the five-number summary.*

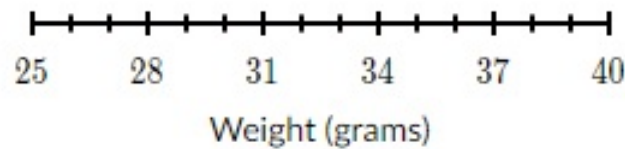


Figure 2.5: Weight(grams).

*Step 2: Draw a box from  $Q_1$  to  $Q_3$  with a vertical line through the median.*

*Recall that  $Q_1 = 29$ , the median is 32, and  $Q_3 = 35$ .*

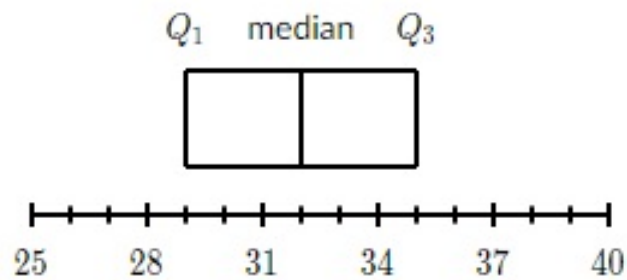


Figure 2.6: Weight(grams).

*Step 3: Draw a whisker from  $Q_1$  to the min and from  $Q_3$  to the maximum.*

*Recall that the min is 25 and the max is 38.*

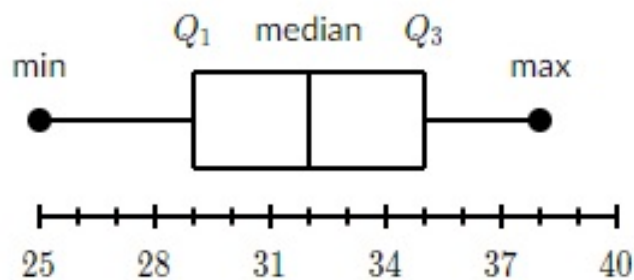


Figure 2.7: Weight(grams).

*We don't need the labels on the final product:*

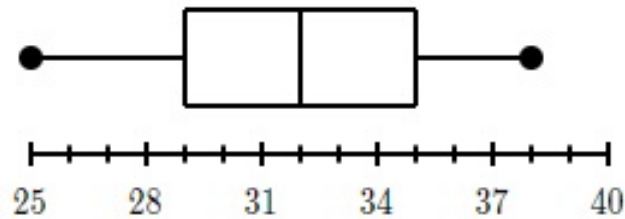
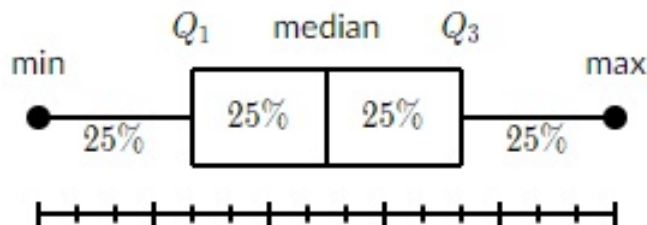


Figure 2.8: Weight(grams).

■

### 2.7.2 Interpreting Quartiles

The five-number summary divides the data into sections that each contain approximately 25% of the data in that set.

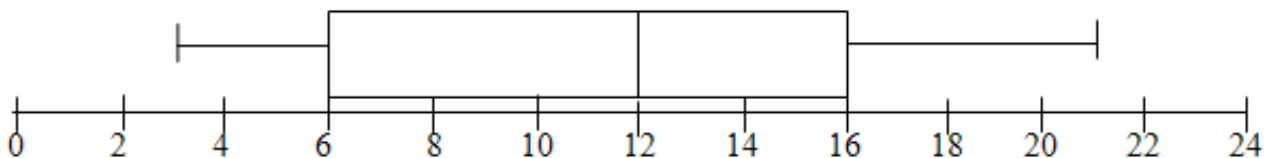


**Example 2.7.2** Draw a box-and-whisker plot for the data set

3, 7, 8, 5, 12, 14, 21, 13, 18

The measures are given by

Minimum : 3,  $Q_1$  : 6, Median : 12,  $Q_3$  : 16, and Maximum : 21.



Notice that in any box-and-whisker plot, the left-side whisker represents where we find approximately the lowest 25% of the data and the right-side whisker represents where we find approximately the highest 25% of the data.

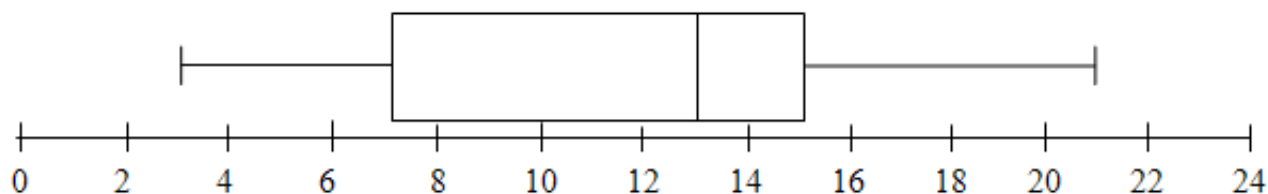
The box part represents the interquartile range and represents approximately the middle 50% of all the data. The data is divided into four regions, which each represent approximately 25% of the data. This gives us a nice visual representation of how the data is spread out across the range.

**Example 2.7.3** Draw a box-and-whisker plot for the data set

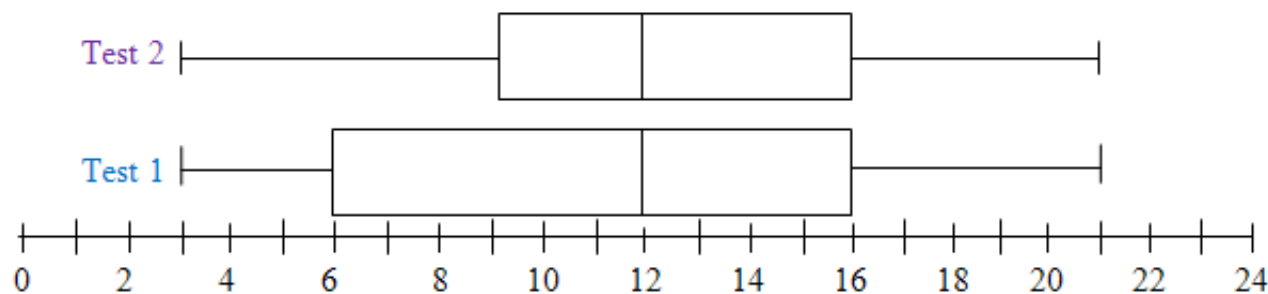
3, 7, 8, 5, 12, 14, 21, 15, 18, 14

The measures are given by

Minimum : 3,  $Q_1$  : 7, Median : 13,  $Q_3$  : 15, and Maximum : 21.



**Example 2.7.4** Suppose that the box-and-whisker plots below represent quiz scores out of 25 points for Quiz 1 and Quiz 2 for the same class. What do these box-and-whisker plots show about how the class did on test 2 compared to test 1?



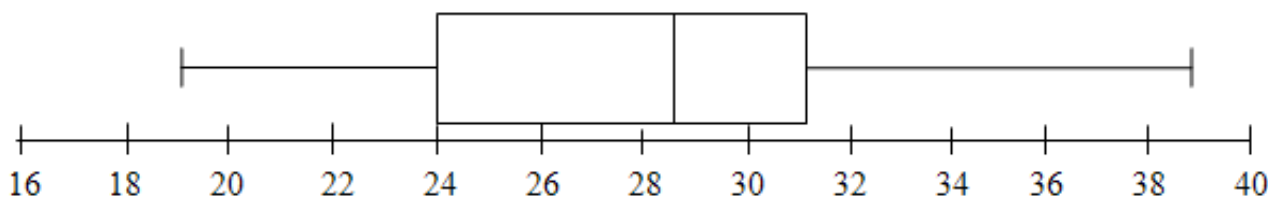
These box-and-whisker plots show that the lowest score, highest score, and  $Q_3$  are all the same for both exams, so performance on the two exams were quite similar. However, the movement  $Q_1$  up from a score of 6 to a score of 9 indicates that there was an overall improvement. On the first test, approximately 75% of the students scored at or above a score of 6. On the second test, the same number of students (75%) scored at or above a score of 9.

**Example 2.7.5** The following dollar amounts were the hourly collections from a Salvation Army kettle at a local store one day in December:

19, 26, 25, 37, 32, 28, 22, 23, 29, 34, 39, and 31.

Construct the box-and-whisker plot for the amount collected.

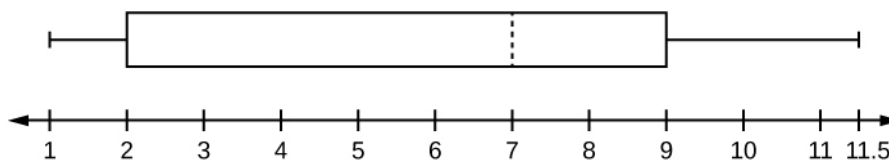
Minimum : 19,  $Q_1$  : 24, Median : 28.5,  $Q_3$  : 33, and Maximum : 39.



**Example 2.7.6** For the given database, draw the Box plots (also called box-and-whisker plots or box-whisker plots)

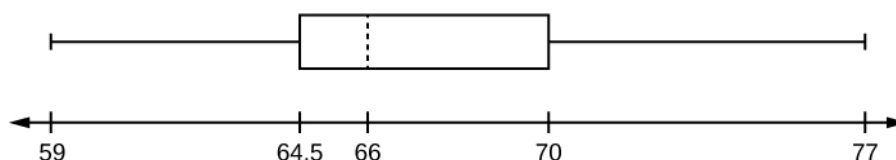
1.)

1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5



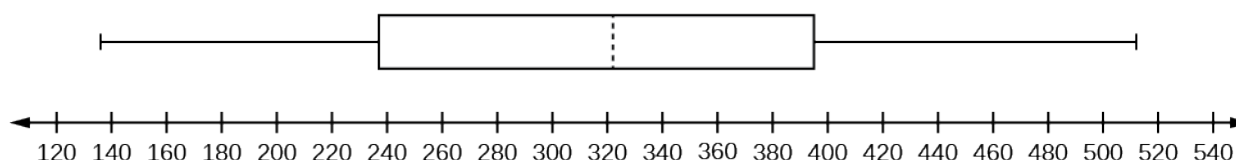
2.)

59 60 61 62 62 63 63 64 64 64 65 65 65 65 65 65 65 65 65 66  
66 67 67 68 68 69 70 70 70 70 70 71 71 72 72 73 74 74 75 77



3.)

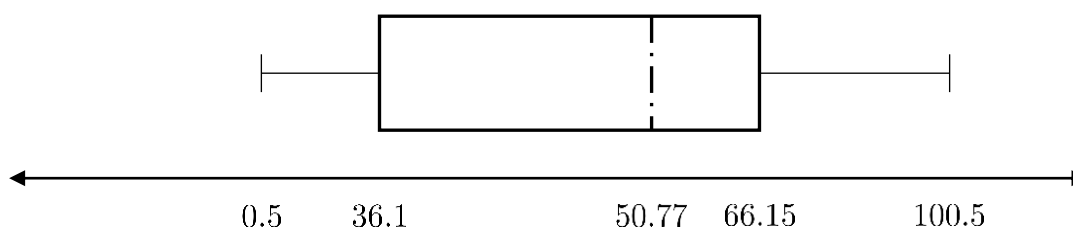
136 140 178 190 205 215 217 218 232 234  
240 255 270 275 290 301 303 315 317 318  
326 333 343 349 360 369 377 388 391 392  
398 400 402 405 408 422 429 450 475 512



**Exercise 2.16** Follow the steps you used to graph a box-and-whisker plot for the data values shown.

0 5 5 15 30 30 45 50 50 60 75 110 140 240 330

**Example 2.7.7** Draw the Boxplot for the grouped data given in Example 2.3.4 on page (87).



**Exercise 2.17** Test scores for a college statistics class held during the day are:

99 56 78 55.5 32 90 80 81 56 59  
45 77 84.5 84 70 72 68 32 79 90

Test scores for a college statistics class held during the evening are:

98 78 68 83 81 89 88 76 65 45 98  
90 80 84.5 85 79 78 98 90 79 81 25.5

- 1.) Find the smallest and largest values, the median, and the first and third quartile for the day class.

**Solution :**  $Min = 32$ ,  $Q_1 = 56$ ,  $M = 74.5$ ,  $Q_3 = 82.5$ ,  $Max = 99$  ■

- 2.) Find the smallest and largest values, the median, and the first and third quartile for the night class.

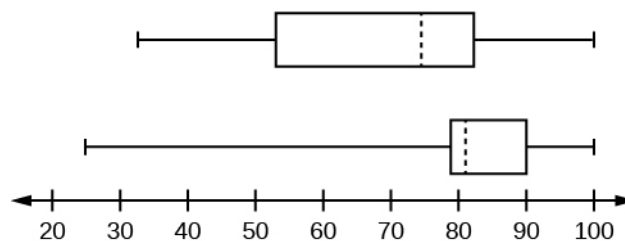
**Solution :**  $Min = 25.5$ ,  $Q_1 = 78$ ,  $M = 81$ ,  $Q_3 = 89$ ,  $Max = 98$  ■

- 3.) For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?

**Solution :** *Day class: There are six data values ranging from 32 to 56 : 30%. There are six data values ranging from 56 to 74.5 : 30%. There are five data values ranging from 74.5 to 82.5 : 25%. There are five data values ranging from 82.5 to 99 : 25%. There are 16 data values between the first quartile, 56, and the largest value, 99 : 75%. Night class:* ■

- 4.) Create a box plot for each set of data. Use one number line for both box plots.

**Solution :**



- 5.) Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

**Solution :** *The first data set has the wider spread for the middle 50% of the data. The IQR for the first data set is greater than the IQR for the second set. This means that there is more variability in the middle 50% of the first data set.* ■

# Chapter 3

## Correlation & Regression

### 3.1 Introduction

There are many statistical investigations in which the main objective is to determine whether a relationship exists between two or more variables. If such a relationship can be expressed by a mathematical formula, we will then be able to use it for the purpose of making predictions.

The reliability of any predictions will, of course, depend on the strength of the relationship between the variables included in the formula.

In dealing with comparisons of two or more sets of data, we tend to develop or establish the kind of knowledge as to whether there is any relationship between these sets of data. We thus assert some conclusions as to whether the relationship is high or no relationship exists, depending on the statistical manipulation taken. For example, one could want to note the extent to which soil erosion affects crop yields; the increase in child births and purchasing of new vehicles, the death rates from cancer and increase in the smoking population. One major reason for this topic is production and control.

### 3.2 Scatter Diagrams

Sometimes we want to find the *relationship*, or *association*, between two variables. This can be done visually with a scatter plot.

The first technique to use when establishing the kind of relationship that exists between sets of data is the use of scatter diagrams. This works best in a bivariate distribution (case of two variables or sets of data). One variable is regarded dependent and the other independent.

A scatter diagram is obtained by plotting points of the independent variable along horizontal axis and the dependent variable along the vertical axis. Points on the scatter graph may not necessarily lie on a straight line. The structure of the relationship is displayed by the pattern of points obtained thereafter.

**Definition 3.2.1** A scatter diagram is a graph that shows the relationship between two variables.

Scatter diagrams can show a relationship between any element of a process, environment, or activity on one axis and a quality defect on the other axis.

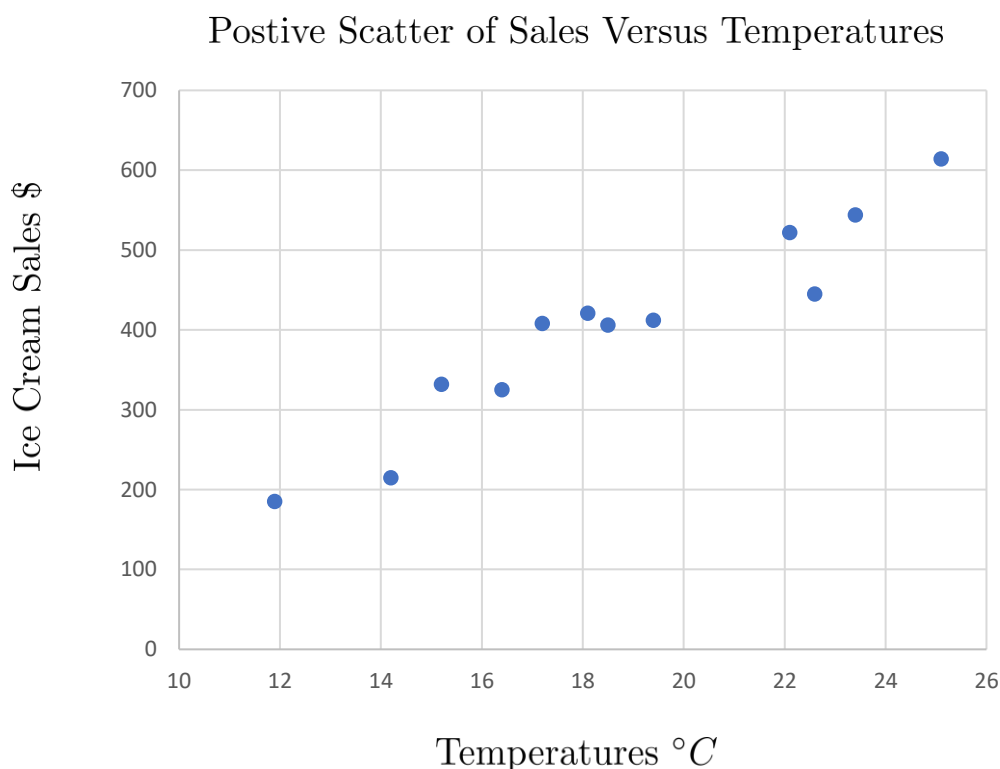


### 3.2.1 Direct Correlation (Positive correlation)

If small values of  $x$  correspond to small values of  $y$  and big values of  $x$  correspond to big values of  $y$  then the pattern of points closely follow from left to right. This is called direct correlation

**Example 3.2.1** The local ice cream shop keeps track of how much ice cream they sell versus the noon temperature on that day. Here are their figures for the last 12 days:

Temp °C	14.2°	16.4°	11.9°	15.2°	18.5°	22.1°	19.4°	25.1°	23.4°	18.1°	22.6°	17.2°
Sales	\$215	\$325	\$185	\$332	\$406	\$522	\$412	\$614	\$544	\$421	\$445	\$408



**Note 3.2.1** It is now easy to see that warmer weather leads to more sales, but the relationship is not perfect.

**Remark 3.2.1** Small values of  $x$  correspond to small values of  $y$ ; large values of  $x$  correspond to large values of  $y$ , so there is a positive co-relation (that is, a positive correlation) between  $x$  and  $y$ .

**Exercise 3.1** In an e-commerce space store we're now visualizing previous year's pageviews and sales given by

Month	Jan	Feb	March	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
Page Views	421	452	496	562	635	681	785	861	998	1187	1357	1521
Sales	33.68	40.68	39.68	44.96	50.80	61.29	70.65	68.88	79.84	94.96	122.13	152.10

Draw it's scatter diagram. Also draw a "Line of Best Fit" (also called a "Trend Line") on your scatter plot.

### 3.2.2 Inverse Correlation (Negative Correlation)

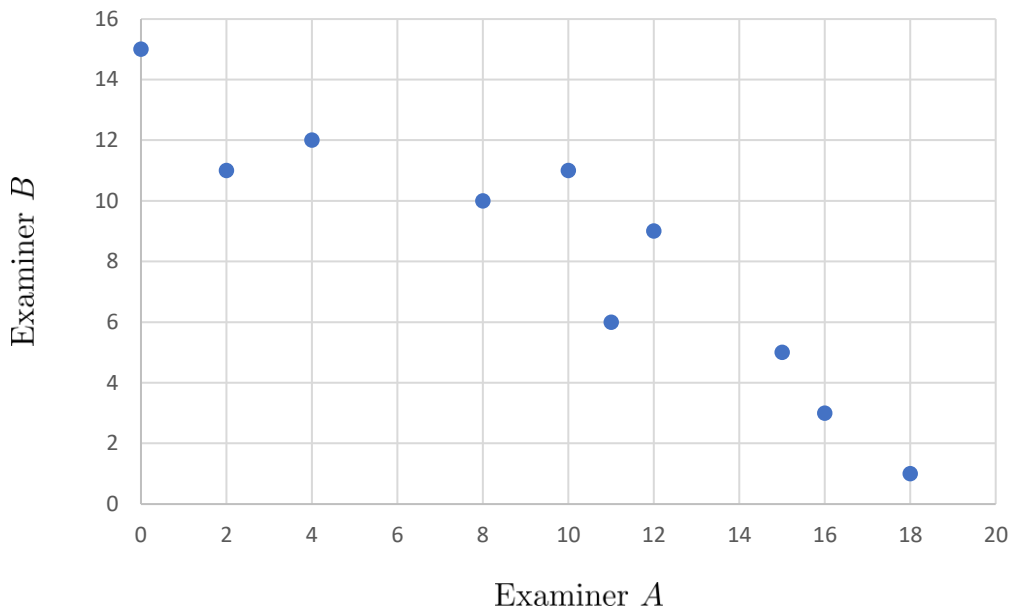
If small values of  $x$  correspond to big values of  $y$  and vice versa, then the pattern of points show a downward slope from left to right.

A negative correlation is a relationship between two variables that move in opposite directions. In other words, when variable  $A$  increases, variable  $B$  decreases. A negative correlation is also known as an inverse correlation.

**Example 3.2.2** Two examiners gave the following marks out of 20 to ten students in a Statistics class after marking the same solutions of each student. Draw their scatter diagram.

Examiner A	18	16	15	11	12	10	8	4	2	0
Examiner B	1	3	5	6	9	11	10	12	11	15

Negative Correlation Scatter Diagram



**Note 3.2.2** The scatter diagram shows quite high negative correlation. There is a very high negative relation on how the two examiners mark. A student one gives a higher mark, the other gives a lower mark. The examiners are negatively correlated.

**Remark 3.2.2** The scatter about the line is quite small, so there is a strong linear relationship. The slope of the line is negative (small values of  $x$  correspond to large values of  $y$ ; large values of  $x$  correspond to small values of  $y$ ), so there is a negative co-relation (that is, a negative correlation) between  $x$  and  $y$ .

**Exercise 3.2** Plot the scatter plot for the following data and determine nature of correlation.

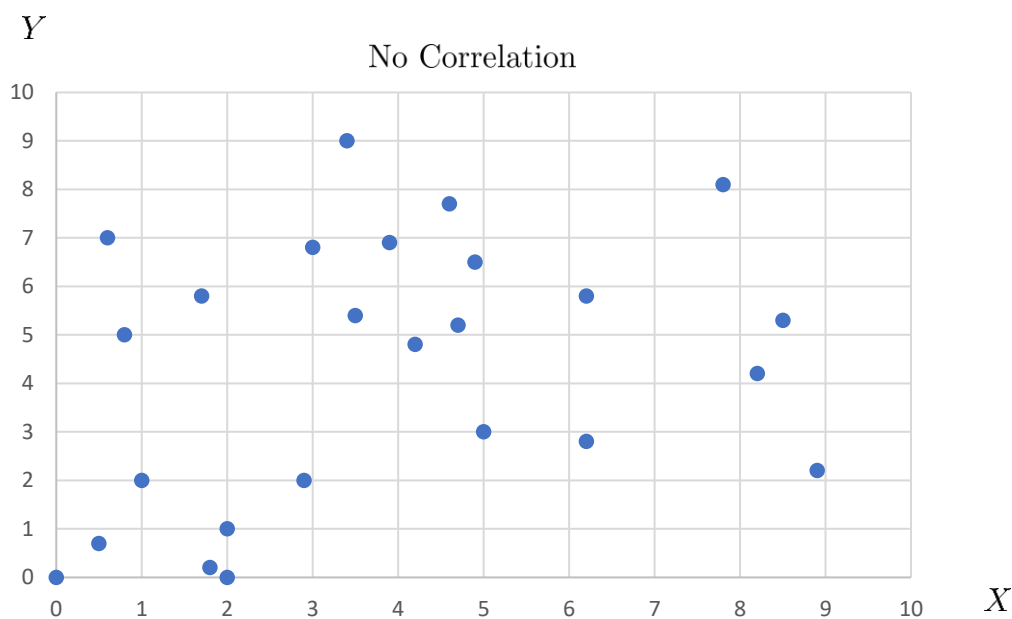
$X$	2	4	5	6	8	11
$Y$	18	12	10	8	7	5

**Exercise 3.3** Plot the scatter diagram for the data of Exercise 3.14 on page (p. 144).

### 3.2.3 Absence of Correlation

If the pattern of points does not follow any of the above trends, i.e., points are scattered all around the graph, then there is no correlation and here we assert that there is zero Correlation or no correlation.

Discussion: The lack of predictability in determining  $Y$  from a given value of  $X$ , and the associated amorphous, non-structured appearance of the scatter plot leads to the summary conclusion: no relationship.



When there is no clear relationship between the two variables, we say there is no correlation between the two variables.

One advantage of scatter diagrams is that they are not affected by time. One needs not to be interested in when the data was obtained to make a prediction.

**Exercise 3.4** Plot the scatter diagram for the following data below, and state the nature of correlation.

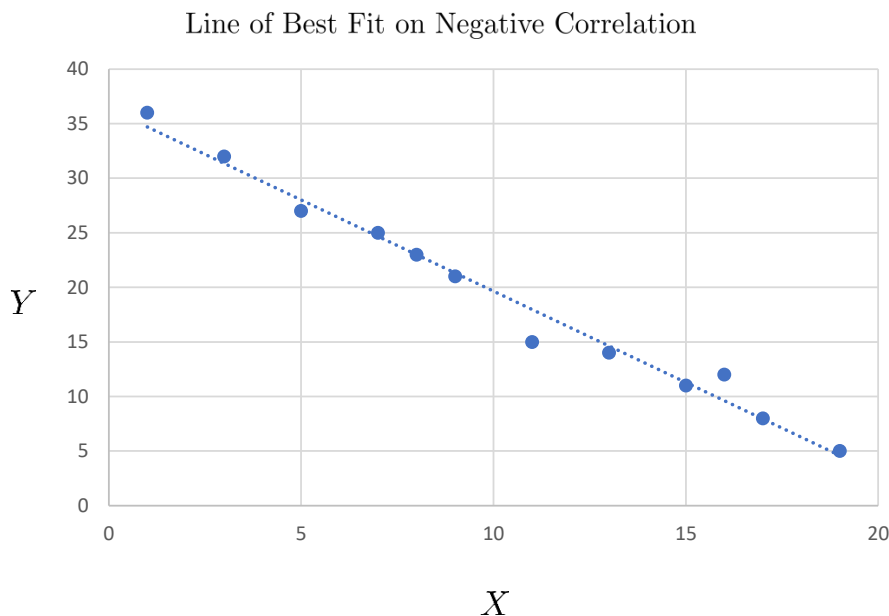
$X$	2	3	4	5	6	7	8
$Y$	4	5	6	12	9	5	4

### 3.3 The Line of Best Fit

In all our cases mentioned, one can try to fit a line through the points plotted. This is regarded as the Line of Best Fit. It is the line designed to suit the pattern of points best.

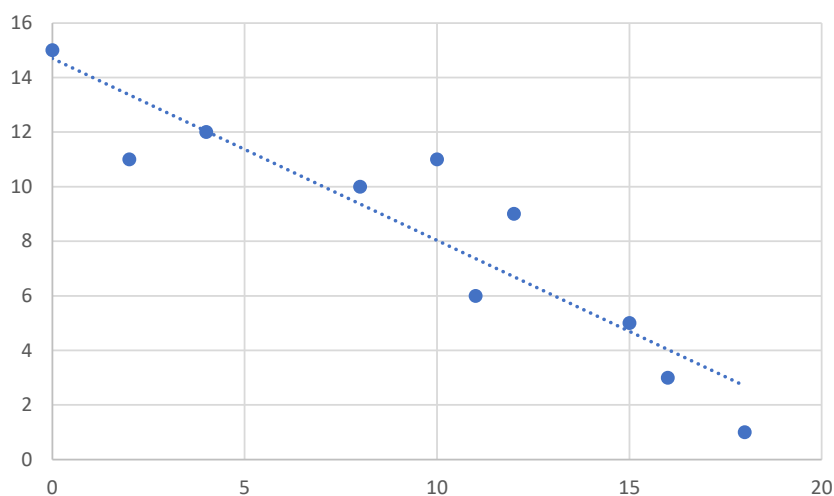
One has to bear in mind that the line to be fitted best, should minimise the divergence of points from this line.

**Example 3.3.1** Draw the line of best fit on a scatter diagram for points  $(x, y)$  given by  $(1, 36), (3, 32), (5, 27), (7, 25), (8, 23), (9, 21), (11, 15), (13, 14), (15, 11), (16, 12), (17, 8), (19, 5)$



Once the line of best fit has been drawn, one easily reads off from the graph, the prediction or the extent of the dependence of  $y$  to the independence of  $x$  and vice versa. Hence the relationship.

**Example 3.3.2** The line of best fit for Example 3.2.2 on page (p. 133) is plotted as



**Exercise 3.5** The following table gives the heights and weights of 10 friends:

Name	Height (cm)	Weight (Kg)
Albert	180	87
Beth	176	55
Cindy	144	52
David	195	94
Emily	159	87
Frank	185	79
Gary	166	59
Helen	173	64
Ida	149	45
Jeremy	168	77

Which one of the following best describes the correlation between their heights and weights?  
(Hint: draw a scatter plot)

- A. High positive correlation                      C. No correlation  
B. Low positive correlation                      D. Low negative correlation

**Example 3.3.3** Plot the scatter diagram for the following data.

$X$	0	1	2	3
$Y$	2	4	6	8

Determine the nature of correlation.

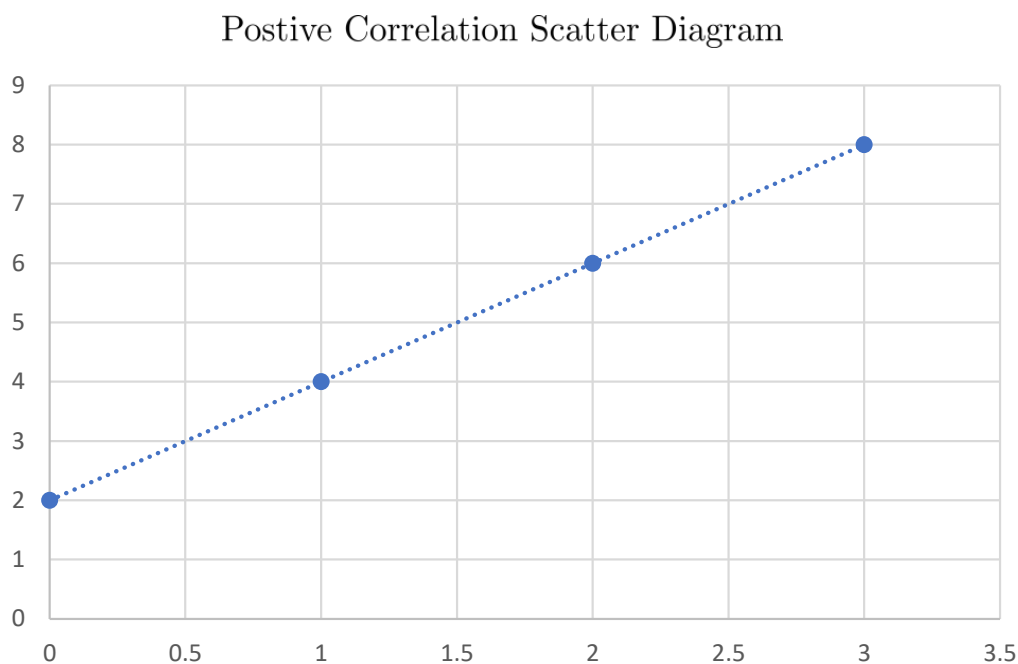
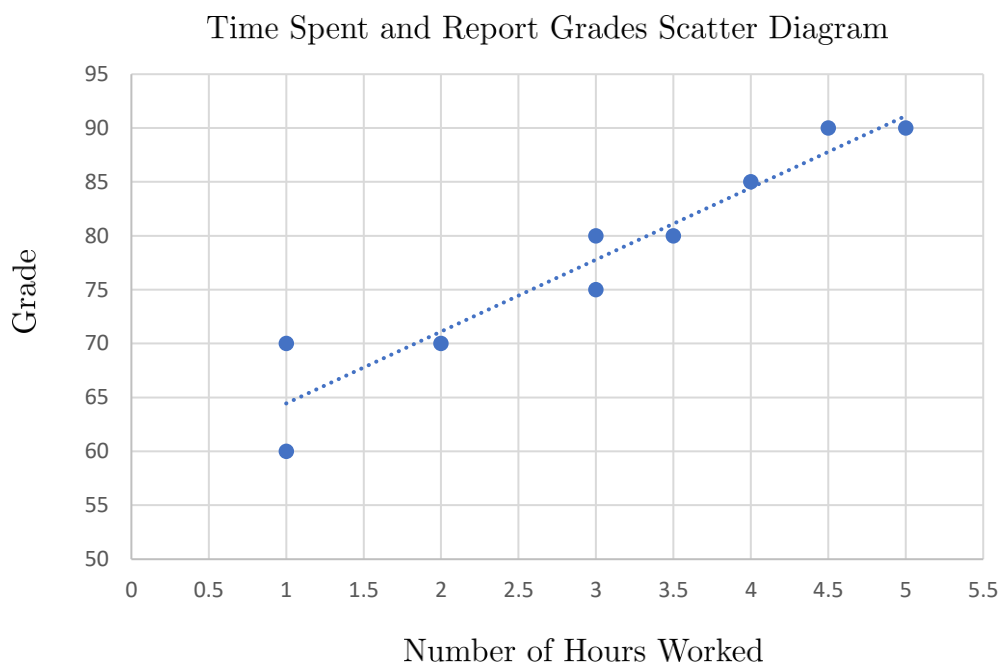


Figure 3.1: Line of best fit on positive scatter diagram

**Example 3.3.4** A teacher wants to know if there is a relationship between the amount of time her students spent working on a social studies report and the grade each student received. She surveyed 10 students and recorded the data below.

Student	Number of Hours Worked	Grade
Ahmad	5	90
Becky	3	80
Ddamulira	3.5	80
Namunana	1	60
Godfrey	4.5	90
Helen	1	70
Kasaija	3	75
Nyakaisiki	4	85
Ogwang	2	70
Zzimula	2.5	75

- 1.) Make a scatter plot based on the data.



- 2.) Determine if there is a relationship between the number of hours worked and the grade received. If so, describe the relationship.

**Solution :** *The answer is that because the line of best fit slants up from left to right, this scatter plot shows a positive relationship between hours worked and grade on the report. This means that in general, the longer a student spent working on the report, the higher the student's grade on the report.* ■

- 3.) Suppose an eleventh student spent 1.5 hours working on her report. Based on the scatter plot, predict the grade you would expect her to receive.

**Solution :** *Because the scatter plot shows a relationship between hours worked on the report and grade on the report, you can use the scatter plot to make predictions for other values within the original range of your data.*

*None of the original ten students worked only 1.5 hours on the report; however, you can assume that this student would fit into the trends seen with the rest of the students.*

*Look at the scatter plot with the line of best fit. Find 1.5 hours on the x-axis and move up until you hit the line of best fit. You are at a height corresponding to a grade of 68 on the report. This means that if a student only worked 1.5 hours on the report, you can predict that they will get a 68 as a grade on the report.*

*The answer is that you predict that a student who only worked 1.5 hours on the report will receive a grade of a 68 on the report.* ■

**Example 3.3.5** Look at the following set of ordered pairs. Would the scatter plot show a positive correlation, a negative correlation or no correlation?

1.)

(1, 12), (8, 9), (3, 10), (2, 4), (5, 2), (6, 6)

**Solution :** *The scatter plot shows no correlation* ■

2.)

(1, 2), (3, 5), (7, 10), (9, 15), (4, 8)

**Solution :** *The scatter plot shows a positive correlation* ■

3.)

(1, 12), (5, 5), (3, 8), (9, 1), (8, 4)

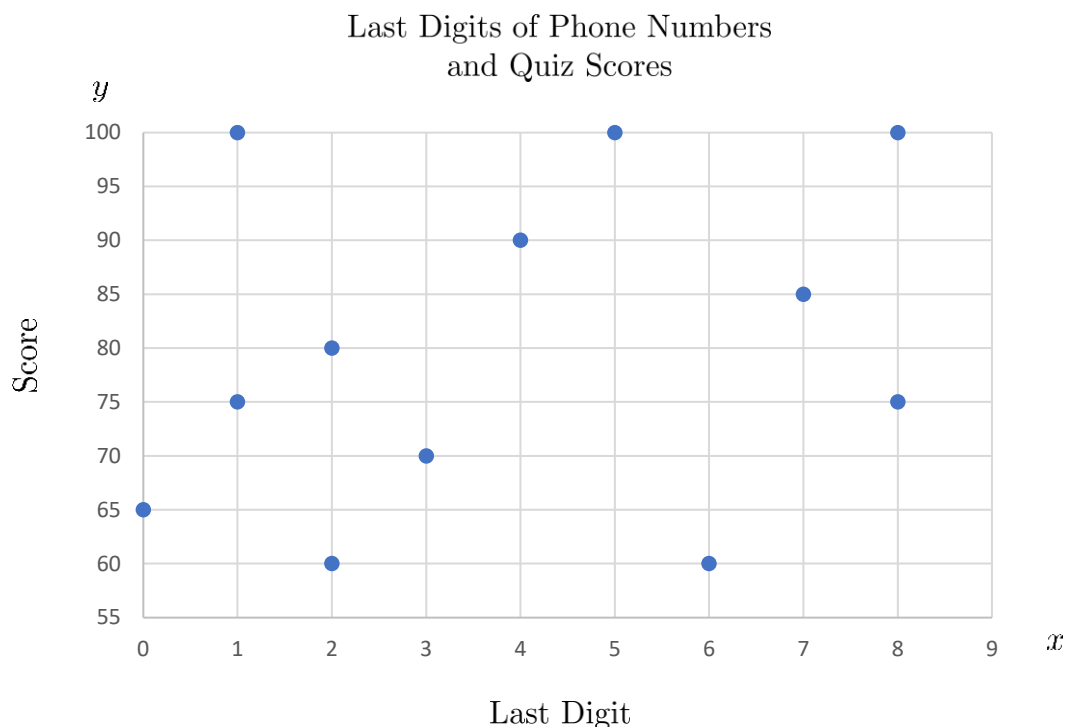
**Solution :** *The scatter plot shows a negative correlation* ■

**Exercise 3.6** The number of umbrellas sold and the amount of rainfall on 9 days is shown on the scatter graph and in the table.

Umbrellas sold	1	10	25	0	1	32	47	8	15
Rainfall (mm)	3	2	4	0	0	5	6	1	1

- 1.) Show this information on the scatter graph.
- 2.) Describe the relationship between the rainfall and umbrellas sold.
- 3.) What is the most number of umbrellas sold in 9 days?
- 4.) What is the greatest amount of rainfall in 9 days?
- 5.) The next rainfall is 4.5mm. Estimate the number of the umbrellas sold.
- 6.) Explain why it may not be appropriate to use your line of best fit to estimate the number of umbrellas sold in a day with 15mm of rainfall.

**Exercise 3.7** This scatter plot shows the relationship between the last digit of ten students' phone numbers and their vocabulary quiz scores.



- 1.) Does this scatter plot show a positive relationship, a negative relationship, or no relationship?
- 2.) How many students have a six as the last digit in their phone number?
- 3.) How many students have an eight as the last digit?
- 4.) How many students received a grade of 70%?
- 5.) How many students received a grade of 60%
- 6.) How many students earned 100%?
- 7.) How many students earned a 75%?
- 8.) *True* or *false*. No one earned below a 60%.
- 9.) *True* or *false*. No one earned an 80%.
- 10.) *True* or *false*. No one earned an 85%.

**Exercise 3.8** Does a line of best fit on a plot with a positive correlation go up or down as you move from left to right on the scatter plot?

**Exercise 3.9** What is a scatterplot?

**Exercise 3.10** Explain the difference between positive correlation, negative correlation, and no correlation.

**Exercise 3.11** What does a weak correlation look like on a scatterplot? Describe a situation where you might find a weak correlation. Explain your thinking.



**Exercise 3.12** Serena wants to know if there is a relationship between a person's age and the number of DVDs they purchased in one year. She surveyed a group of people and recorded the data in the table below.

Person	Age	Number of DVDs Purchased
Person 1	18	14
Person 2	19	13
Person 3	20	13
Person 4	20	12
Person 5	21	11
Person 6	22	12
Person 7	22	11
Person 8	23	10
Person 9	24	9
Person 10	25	9

- 1.) Make a scatter plot for the data in the table.
- 2.) *True or false.* The scatter plot shows a negative correlation.
- 3.) *True or false.* The scatter plot shows no correlation.
- 4.) *True or false.* The scatter plot shows a positive correlation.
- 5.) If the trend in the scatter plot continues, predict the number of DVDs you would expect a 27-year-old person to buy in one year

**Example 3.3.6** What sort of correlation would you expect to find between:

- 1.) a person's age and their house number,

**Solution :** *No correlation, because these two quantities are not linked in any way.* ■

- 2.) a child's age and their height,

**Solution :** *Positive correlation, because children get taller as they get older.* ■

- 3.) an adult's age and their height?

**Solution :** *No correlation, because the height of adults does not change with their age.* ■

**Example 3.3.7** What type of correlation would you expect to find between each of the following quantities:

- 1.) IQ and height,

**Solution :** *No correlation* ■

- 2.) Person's height and shoe size?

**Solution :** *Positive correlation* ■

- 3.) Price of house and number of bedrooms,

**Solution :** *Weak/Moderate positive correlation* ■

## 3.4 Regression Analysis

### 3.4.1 Limitations of Scatter diagram

- 1.) There is uncertainty as to the correct position of the line of best fit. One needs a mathematical formula for consistence.
- 2.) There is lack of measure of closeness of the relationship in case of absence of correlation.

### 3.4.2 Way Forward

With limitations and subjective judgement of the line of best fit, since apparently it depends on one individual's eye, we expect many lines of best fit each depending on the individual involved. As a result, the line will differ from one individual to another. A line of best fit independent of subjectivity or individual's judgement can thus be obtained mathematically. This is known as a Regression Line.

#### *Regression Line*

When drawing a line of best fit, an attempt is made to minimise the total divergence of points from this line. We should also be able to reduce the total divergence of squared deviations from the line if the equation of the line is to be computed mathematically. This mathematical approach is known as the Least Squares Method.

We note that we have two types of deviations, the vertical and horizontal deviations. Vertical deviation is known as the regression of  $y$  on  $x$ . Horizontal deviation is known as regression of  $x$  on  $y$ .

How to calculate the equation of Line of Best Fit using Least Square Method??

**3.4.3 Regression of  $y$  on  $x$** 

Write the general equation of a straight line  $y = mx + c$  as  $y = a + bx$  for regression of  $y$  on  $x$ . Thus, for

$$y = a + bx$$

Multiplying by  $x$

$$yx = ax + bx^2$$

To have the set of equations as

$$y = a + bx \tag{3.1}$$

$$yx = ax + bx^2 \tag{3.2}$$

Taking summation on both sides of both equations,

$$\sum y = \sum a + \sum bx$$

$$\sum yx = \sum ax + \sum bx^2$$

Can be simplified as

$$\sum y = na + b \sum x$$

$$\sum xy = a \sum x + b \sum x^2$$

Where the constants  $a$  and  $b$  can be given as

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \tag{3.3}$$

$$a = \frac{\sum y - b \sum x}{n} \tag{3.4}$$

We obtain the values of  $a$  and  $b$ . The values of  $x$  and  $y$  are usually given in table from which our Equations (3.3) and (3.4) require to get

$$\sum x, \sum y, \sum x^2, \sum y^2, \sum xy$$

**Example 3.4.1** Given the following two data

$x$	4	3	2	6
$y$	2	1	4	5

Find the regression line of  $y$  on  $x$ .

$x$	$y$	$x^2$	$y^2$	$xy$
4	2	16	4	8
3	1	9	1	3
2	4	4	16	8
6	5	36	25	30
$\sum x = 15$	$\sum y = 12$	$\sum x^2 = 65$	$\sum y^2 = 46$	$\sum xy = 49$

From the table or using a calculator

$$\sum x = 15, \quad \sum y = 12, \quad \sum x^2 = 65, \quad \sum y^2 = 46, \quad \sum xy = 49, \quad n = 4$$

$$\begin{aligned} b &= \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \\ &= \frac{4(49) - (15)(12)}{[4(65) - (15)^2]} \\ &= \frac{16}{35} \\ a &= \frac{\sum y - b \sum x}{n} \\ &= \frac{(12) - \left(\frac{16}{35}\right)(15)}{4} \\ &= \frac{9}{7} \end{aligned}$$

the regression of  $y$  on  $x$  is

$$y = a + bx = \frac{9}{7} + \frac{16}{35}x$$

Hence compute  $y(7.5)$ .

$$y = \frac{9}{7} + \frac{16}{35}(7.5) = 4.7143$$

**Note 3.4.1** Students should be able to get the summation answers using a calculator.

**Exercise 3.13** Find the Linear Regression line through  $(3, 1)$ ,  $(5, 6)$ ,  $(7, 8)$ .

**Example 3.4.2** Research showed that the number of worms ( $x$ ) against the fatal sickness is shown as below

$x$	1	2	3	4	5	6
$y$	1	4	2	4	6	5

Find the regression line of  $y$  on  $x$ .

$x$	$y$	$x^2$	$y^2$	$x \cdot y$
1	1	1	1	1
2	4	4	16	8
3	2	9	4	6
4	4	16	16	16
5	6	25	36	30
6	5	36	25	30
$\sum x = 21$	$\sum y = 22$	$\sum x^2 = 91$	$\sum y^2 = 98$	$\sum xy = 9$

**From the calculator**

[MODE REG  $\gg$  LIN  $\gg$  ( $x_i, y_i$ ) M+  $\gg$   $\forall i$   $\gg$  SHIFT 1  $\gg$  SHIFT2  $\gg$  MODE COMP]

$$\sum x = 21, \quad \sum y = 22, \quad \sum x^2 = 91, \quad \sum xy = 91, \quad \sum y^2 = 98, \quad n = 6$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{6(91) - (21)(22)}{6(91) - (21)^2} = \frac{84}{105} = 0.8$$

$$a = \frac{\sum y - b \sum x}{n} = \frac{(22) - \frac{84}{105}(21)}{6} = \frac{13}{15}$$

the regression of  $y$  on  $x$  is

$$y = a + bx = \frac{13}{15} + \frac{4}{5}x$$

Use the line to approximate  $y$  if  $x = 15$

$$y = \frac{13}{15} + \frac{4}{5}x = \frac{13}{15} + \frac{84}{105}(15) = \frac{193}{15}$$

**Exercise 3.14** Determine the regression of line  $y$  on  $x$  for the data set below.

$x$	1	3	4	5	8
$y$	4	2	1	0	0

$$y = -0.575x + 3.81$$

**3.4.4 The Regression of  $x$  on  $y$** 

We now write the general equation  $x = my + c$  to denote the regression of  $x$  on  $y$ . This can be written as  $x = a' + b'y$ . Multiplying by  $y$  on both sides we generate

$$\begin{aligned}x &= a' + b'y \\yx &= a'y + b'y^2\end{aligned}$$

Taking summations will give

$$\sum x = a'n + b'\sum y \quad (3.5)$$

$$\sum xy = a'\sum y + b'\sum y^2 \quad (3.6)$$

to give the values of  $a'$  and  $b'$

$$b' = \frac{n \sum xy - \sum x \sum y}{n \sum y^2 - (\sum y)^2} \quad (3.7)$$

$$a' = \frac{\sum x - b' \sum y}{n} \quad (3.8)$$

and we obtain  $x = a' + b'y$ , the regression of  $x$  on  $y$ . (Interchanging  $x$  and  $y$  in Eqn 3.3, and Eqn 3.4 on page 142)

**Example 3.4.3** Consider the data below

$x$	4	3	2	6
$y$	2	1	4	5

The regression table is given by

$x$	$y$	$x^2$	$y^2$	$xy$
4	2	16	4	8
3	1	9	1	3
2	4	4	16	8
6	5	36	25	30
$\sum x = 15$	$\sum y = 12$	$\sum x^2 = 65$	$\sum y^2 = 46$	$\sum xy = 49$

$$b' = \frac{n \sum xy - \sum x \sum y}{n \sum y^2 - (\sum y)^2} = \frac{(4)(49) - (15)(12)}{(4)(46) - (12)^2} = \frac{16}{40} = \frac{4}{10}$$

$$a' = \frac{\sum x - b' \sum y}{n} = \frac{15 - \frac{4}{10}(12)}{4} = \frac{102}{40} = \frac{51}{20}$$

Therefore, the regression line (line of best fit) of  $x$  on  $y$  written as  $x = a' + b'y$  is given by

$$x = \frac{51}{20} + \frac{4}{10}y$$

**Note 3.4.2** Uses of both regression lines  $y = a + bx$  and  $x = a' + b'y$  are clear and straight forward. For a regression line of  $y$  on  $x$  it implies that for a given value of  $x$  obtains a corresponding value  $y$  and for  $x = a' + b'y$  for a given  $y$  we can get a corresponding value  $x$ .

The choice of which of the lines to use can be noted clearly. For if  $y$  is the dependent variable then one uses the regression line  $y = a + bx$  and vice versa. The value of  $b$  obtained from our simultaneous equations for  $y = a + bx$  is known as the regression coefficient of  $y = a + bx$ . Similarly  $b'$  is the regression coefficient of  $x = a' + b'y$ .

**Example 3.4.4** Find the least square regression line for the following set of data

$$(-1, 0), \quad (0, 2), \quad (1, 4), \quad (2, 5)$$

We use the table as follows

$x$	$y$	$x^2$	$y^2$	$xy$
-1	0	1	0	0
0	2	0	4	0
1	4	1	16	4
2	5	4	25	10
$\sum x = 2$	$\sum y = 11$	$\sum x^2 = 6$	$\sum y^2 = 45$	$\sum xy = 14$

1.) The regression line  $y = a + bx$

**Solution :** Using Equation 3.3, and Equation 3.4 on page 142

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{(4)(14) - (2)(11)}{(4)(6) - (2)^2} = \frac{34}{20} = \frac{17}{10}$$

$$a = \frac{\sum y - b \sum x}{n} = \frac{11 - \frac{17}{10}(2)}{4} = \frac{19}{10}$$

such that

$$y = \frac{19}{10} + \frac{17}{10}x = 1.9 + 1.7x$$

■

2.) The regression line  $x = a' + b'y$

**Solution :** Using Equation 3.7, and Equation 3.8 on page 145

$$b' = \frac{n \sum xy - \sum x \sum y}{n \sum y^2 - (\sum y)^2} = \frac{(4)(14) - (2)(11)}{(4)(45) - (11)^2} = \frac{34}{59}$$

$$a' = \frac{\sum x - b' \sum y}{n} = \frac{2 - \frac{34}{59}(11)}{4} = -\frac{256}{236} = -\frac{64}{59}$$

such that

$$x = -\frac{64}{59} + \frac{34}{59}y = -1.08475 + 0.57627x$$

■

**Example 3.4.5** The values of  $x$  and their corresponding values of  $y$  are shown in the table below

$x$	0	1	2	3	4
$y$	2	3	5	4	6

The regression table is given as

$x$	$y$	$x^2$	$y^2$	$xy$
0	2	0	4	0
1	3	1	9	3
2	5	4	25	10
3	4	9	16	12
4	6	16	36	24
$\sum x = 10$	$\sum y = 20$	$\sum x^2 = 30$	$\sum y^2 = 90$	$\sum xy = 49$

1.) The regression line  $y = a + bx$ . Estimate the value of  $y$  when  $x = 10$ .

**Solution :** Using Equation 3.3, and Equation 3.4 on page 142

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{(5)(49) - (10)(20)}{(5)(30) - (10)^2} = \frac{45}{50} = \frac{9}{10}$$

$$a = \frac{\sum y - b \sum x}{n} = \frac{20 - \frac{9}{10}(10)}{5} = \frac{11}{5}$$

such that

$$y = \frac{11}{5} + \frac{9}{10}x$$

$$\text{such that for } x = 10 \Rightarrow y = \frac{11}{5} + \frac{9}{10}(10) = \frac{56}{5} = 11.2 \quad \blacksquare$$

2.) The regression line  $x = a' + b'y$ . Estimate the value of  $x$  when  $y = -5$

**Solution :** Using Equation 3.7, and Equation 3.8 on page 145

$$b' = \frac{n \sum xy - \sum x \sum y}{n \sum y^2 - (\sum y)^2} = \frac{(5)(49) - (10)(20)}{(5)(90) - (20)^2} = \frac{45}{50} = \frac{9}{10}$$

$$a' = \frac{\sum x - b' \sum y}{n} = \frac{10 - \frac{9}{10}(20)}{5} = -\frac{8}{5}$$

such that

$$x = -\frac{8}{5} + \frac{9}{10}y$$

$$\text{Therefore, for } y = -5 \Rightarrow x = -\frac{8}{5} + \frac{9}{10}(-5) = -\frac{8}{5} - \frac{45}{10} = -\frac{61}{10} = -6.1 \quad \blacksquare$$



## 3.5 Correlation Analysis

With regression, we were establishing the relationship when one variable is independent and the other dependent; and one could certainly get the value of the dependent variable  $y$  predicted from a corresponding value of an independent  $x$ .

With correlation, we now measure rather than predict, the relationship between values where we may not necessarily know the dependence or independence of these values. Thus, an attempt to measure the strength of a relationship in which we may or may not know whether it exists is known as correlation analysis. We draw conclusion after obtaining a computed value known as correlation coefficient.

The main distinction between regression and correlation is that for regression one variable is independent and the other dependent and in correlation analysis dependence or independence of variables on each other is not paramount except that we draw conclusions after getting the correlation coefficient.

## 3.6 Correlation Coefficients

### 3.6.1 Pearson Product - Moment Correlation Coefficient

We define Pearson product - Moment correlation coefficient denoted as  $\gamma$  by

$$\gamma = \frac{n \sum xy - \sum x \sum y}{\sqrt{[(n \sum x^2 - (\sum x)^2)[n \sum y^2 - (\sum y)^2]}} \quad (3.9)$$

**Example 3.6.1** For the experiment carried on two tests; observations got are

Test $x$	2	5	4	6	3
Test $y$	6	10	7	9	8

Find Pearson product - Moment correlation coefficient

$x$	$y$	$x^2$	$y^2$	$xy$
2	6	4	36	12
5	10	25	100	50
4	7	16	49	28
6	9	36	81	54
3	8	9	64	24

$$\sum x = 20, \quad \sum y = 40, \quad \sum x^2 = 90, \quad \sum y^2 = 330, \quad \sum xy = 168, \quad n = 5$$

$$\begin{aligned} \gamma &= \frac{n \sum xy - \sum x \sum y}{\sqrt{[(n \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)]}} \\ &= \frac{(5)(168) - (20)(40)}{\sqrt{[(5)(90) - 20^2][(5)(330) - 40^2]}} \\ &= \frac{840 - 800}{(50)(50)} = \frac{40}{50} = \frac{4}{5} = 0.8 \\ &= +0.8 \end{aligned}$$

**Exercise 3.15** Calculate the Pearson correlation coefficient for the data in Example 3.2.2 on page (p. 133).

**Solution :**  $\gamma = -0.922$ . Shows quite high negative correlation ■

**Exercise 3.16** Calculate the Pearson correlation coefficient for the data in Example 3.4 on page (p. 134).

**Solution :**  $\gamma = 0.077$ . Shows no correlation. ■

**Example 3.6.2** Research showed that the the number of worms ( $x$ ) against the fatal sickness is shown as below

$x$	1	2	3	4	5	6
$y$	1	4	2	4	6	5

Find the correlation coefficient  $\gamma$ .

From the calculator

$$\sum x = 21, \quad \sum y = 22, \quad \sum x^2 = 91, \quad \sum xy = 91, \quad \sum y^2 = 98, \quad n = 6$$

To have the coefficient

$$\gamma = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} = \frac{6(91) - (21)(22)}{\sqrt{[6(91) - (21)^2][6(98) - (22)^2]}}$$

**Exercise 3.17** Is there any correlation between the mathematical achievement test scores ( $x_i$ ) and calculus grades ( $y_i$ ) for 10 college freshmen?

$x$	39	43	21	64	57	47	28	75	34	52
$y$	65	78	52	82	92	89	73	98	56	75

**Exercise 3.18** A sample of 6 children was selected, data about their age in years and weight in kilograms was recorded as shown in the following table .

Age (years)	7	6	8	5	6	9
Weight (Kg)	12	8	12	10	11	13

Compute the Pearson's correlation coefficient  $\gamma$  between the age and weight of children. Interpret the coefficient.

**Exercise 3.19** The relationship between Anxiety ( $x$ ) and Test Scores ( $y$ ) are given as Show

Anxiety ( $x$ )	10	8	2	1	5	6
Test Scores ( $y$ )	2	3	9	7	6	5

that there exists an indirect strong correlation of  $\gamma = -0.94$  between anxiety and test scores.

**Exercise 3.20** Given the following data

$x$	1	2	4	6	7	10
$y$	9	6	4	12	8	3

- 1.) Calculate the Pearson correlation coefficient for the data above.  $\gamma = -0.268$
- 2.) Draw the scatter diagram to show that there is no correlation between the data.

**Exercise 3.21** Given the following pairs of value of the variables  $X$  and  $Y$ :

$X$	2	3	5	6	8	9
$Y$	6	5	7	8	12	11

- 1.) Make a scatter diagram.
- 2.) Do you think that there is any correlation between the variables  $X$  and  $Y$ ?
- 3.) Is it positive or negative?
- 4.) Is it high or low?
- 5.) By graphic inspection draw an estimated line.

**Exercise 3.22** The lengths and weights of a sample of six articles manufactured by a factor are given here. Find the Pearson's correlation coefficient.

Length $X$	3	5	6	7	10	11
Weight $Y$	8	12	11	14	16	17

$$\gamma = 0.97$$

**Exercise 3.23** Find Karl Pearson's coefficient of correlation between the values of  $X$  and  $Y$  given here under:

$X$	46	68	72	75	80	70	93	100
$Y$	64	50	39	48	12	52	46	30

$$\gamma = 0.86$$

**Exercise 3.24** Determine the Pearson's correlation coefficient for Exercise 3.14 on page (p. 144).

**Exercise 3.25** For a regression line  $y = a + bx$ , show that

$$a = \frac{\sum y - b \sum x}{n}$$
$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

**Exercise 3.26** A new computer circuit was tested and the times (in micro seconds) required to carryout different subroutines were recorded as follows:

$x$	1	2	3	4
$y$	1	5	8	13

- 1.) Calculate the values of  $a$  and  $b$  for the given data for the regression line  $y = a + bx$
- 2.) Hence estimate  $y$  when  $x = 2.5$
- 3.) Sketch the scatter diagram for the figures above
- 4.) Is there a linear relationship between the variables  $x$  and  $y$ ?

**Exercise 3.27**

- 1.) With relevant diagrams (sketches), distinguish between the positive, negative and zero correlation.
- 2.) Explain the major difference between correlation and regression.

**Exercise 3.28** The head of Statistical data in the mathematics department found the following data after quizzing seven students in the BED(T) first year students

Individual	No. of Sodas Consumed ( $x$ )	No. of Bathroom Trips ( $y$ )
Rick	1	2
Janice	2	1
Paul	3	3
Susan	3	4
Cindy	4	6
John	5	5
Donald	6	5

Compute the Pearson correlation coefficient between the number of sodas and the visits to bathrooms as a result of the drinks.

$$\text{Hint : } \gamma = \frac{n \sum xy - \sum x \sum y}{\sqrt{[(n \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)]}}$$

**Exercise 3.29** Given the following data

$x$	12	8	5	3	2	0
$y$	1	7	4	6	4	2

- 1.) Calculate the Pearson correlation coefficient for the data above.  $\gamma = -0.120$
- 2.) Draw the scatter diagram to show that there is no correlation between the data.

## 3.7 Correlation by Ranks

At times dealing with exact figures may be laborious and tiresome and an alternative is to rank the values by either from the highest to lowest or lowest to highest giving positions  $n(\text{ranks})$  say from  $1, 2, \dots, n$ . The procedure is that get the values of variables  $x$  and  $y$  ranked in order. Sum the product of the corresponding ranks of  $x$  and  $y$ . We obtain the value known as the coefficient of rank correlation.

### 3.7.1 Spearman Rank Correlation Coefficient

We define Spearman correlation coefficient denoted as  $\rho$  by

$$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} \quad (3.10)$$

where  $D$  is the difference between corresponding ranks of  $x$  and  $y$ .

**Example 3.7.1** The data given is observations obtained by measuring weights of children and their corresponding foods.

Weight(kg) $x$	2.75	2.15	4.41	5.52	3.21	4.32	2.31	4.30	3.71
Foods (gm) $y$	29.5	26.5	32.2	36.5	27.2	27.7	28.3	30.3	27.7

Ranking in order:

$x$	$y$	Rank $x$	Rank $y$	$D$	$D^2$
2.75	29.5	7	4	3	9
2.15	26.5	9	9	0	0
4.41	32.2	2	2	0	0
5.52	36.5	1	1	0	0
3.21	27.2	6	8	-2	4
4.32	27.7	3	6.5	-3.5	12.25
2.31	28.3	8	5	3	9
4.30	30.3	4	3	1	1
3.71	27.7	5	6.5	-1.5	2.25
					$\sum D^2 = 37.5$

$$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6(37.5)}{9(80)} = +0.6875$$

**Example 3.7.2** The data given provides scores of a certain class in Mathematics and History tests in percentage:

Mathematics	84	72	67	81
History	46	48	53	19

Calculate Spearman's rank correlation coefficient.

Maths $x$	History $y$	Rank $x$	Rank $y$	$D$	$D^2$
84	46	1	3	-2	4
72	48	3	2	1	1
67	53	4	1	3	9
81	19	2	4	-2	4
					$\sum D^2 = 18$

Spearman's rank correlation coefficient:

$$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6(18)}{4(15)} = 1 - \frac{9}{5} = -0.8$$

**Note 3.7.1** Some cases may happen when the values take on some rank, i.e in an array there are two or more same values. Here if such two values happen to be similar e.g, the data

10.1    10.2    10.2    11

Let the rank of 10.1 = 1, rank of 10.2 = 2. The rank of 10.2 should be 3, but since we have two 10.2 then we add the ranks 2 and 3 and get the average.

i.e

$$\frac{2 + 3}{2} = 2.5$$

and our new ranking becomes

10.1    10.2    10.2    11  
1        2.5    2.5    4

Note that the next value after those with similar ranks gets a rank that it should have got if there were not same values, in our case the value 11 takes rank 4. If there are three such values similar add their proposed ranks and divide by 3 and so on.

**Example 3.7.3** For the following data, obtain the ranks accordingly

$x$	Proposed Rank $x$	New Rank
9.8	2	2
7.2	4	4
7.9	3	3
3.1	9	9.5
3.1	10	9.5
5.0	8	8
6.4	5	6
6.4	6	6
6.4	7	6
10.1	1	1
2.0	11	11

Adding  $\frac{5+6+7}{3} = 6$  and adding  $\frac{9+10}{2} = 9.5$

And the computation for Spearman's rank correlation coefficient continues as before.

**Example 3.7.4** Given the data for rainfall records for two stations on specific days in mm

Station $x$	Station $y$	Rank $x$	Rank $y$	$D$	$D^2$
10	17	6	3	3	9
11	4	5	10	-5	25
9	18	7.5	2	5.5	30.25
9	6	7.5	7	0.5	0.25
8	6	9	7	2	4
13	6	2	7	-5	25
12	14	3.5	4	-0.5	0.25
12	20	3.5	1	2.5	6.25
14	5	1	9	-8	64
4	11	10	5	5	25
					$\sum D^2 = 189$

$$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6(189)}{10(99)} = 1 - \frac{63}{55} = -\frac{8}{55} = -0.1455$$



**Example 3.7.5** Following were marks obtained by 8 pupils in maths and physics. Calculate the Spearman's coefficients of rank correlation of rank correlation.

Math $x$	Physics $y$	Rank $x$	Rank $y$	$D$	$D^2$
67	70	5	4	1	1
42	59	7	6	1	1
85	71	2	3	-1	1
51	38	6	8	-2	4
39	55	8	7	1	1
97	62	1	5	-4	16
81	80	3	1	2	4
70	76	4	2	2	4
					$\sum D^2 = 32$

$$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6(32)}{8(63)} = 1 - \frac{192}{504} = +0.6190$$

**Example 3.7.6** Marks of 10 students in French and German tests were as follows:

French $x$	German $y$	Rank $x$	Rank $y$	$D$	$D^2$
12	6	4.5	8	-3.5	12.25
8	5	9	9	0	0
16	7	2	6.5	4.5	20.25
11	7	6	6.5	-0.5	0.25
7	4	10	10	0	0
10	9	7	4	3	9
13	8	3	5	-2	4
17	13	1	1	0	0
12	10	4.5	3	1.5	2.25
9	11	8	2	6	36
					$\sum D^2 = 84$

$$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)} = 1 - \frac{6(84)}{10(99)} = 1 - \frac{504}{990} = +0.4909$$

### 3.7.2 Kendall's Rank Correlation Coefficient

**Definition 3.7.1** Then Kendall's rank correlation coefficient denoted as  $\tau$  is given by

$$\tau = \frac{S}{\frac{1}{2}n(n-1)} \quad (3.11)$$

where  $S$  is the total scores **on the right**.

**The first method** has got a procedure that ranks  $x$  in ascending order then rank  $y$  correspondingly. Then each pair  $(R_x, R_y)$  is given order  $A, B, C, \dots$ , consider every pair of letters that can be got from these e.g  $AB, BC, BC$ , etc for each  $R_x$  or  $R_y$  then these are  $\frac{1}{2}n(n-1)$  such pairs expected. Each pair is attached on a  $(+1)$  or  $(-1)$  score depending on whether  $(R_x, R_y)$  is in right or reverse order. Eg. If  $(R_x, R_y) = (3, 4)$  then it is right order and thus  $+1$  if  $(R_x, R_y) = (6, 3)$  then it is a reverse (descending) order hence  $-1$  score. Then the product of say  $AB$  for rank  $x$  and  $AB$  for rank  $y$  is obtained as either  $+1$  or  $-1$  depending on the sign of  $AB$  for  $R_x$  and  $AB$  for  $R_y$ . Add the room scores got in each line pair. Sum total scores on the right.

**The second method** for Kendall requires also to arrange Rank  $x$  ( $R_x$ ) in ascending order. Then fit in rank  $y$  ( $R_y$ ) correspondingly, on the line of Rank  $y$ , start from first value. Note of how many other values are greater or less than that value. Values greater than that in consideration are counted positive and those less, counted negative. Add corresponding results vertically and get a row of sums. Add those values in the new row to get  $S$ , the total sum. Then Kendall's  $\tau$  is given by equation (3.11).

**Example 3.7.7** Given ranks of  $x$ ,  $R_x$  in ascending order and corresponding ranks  $R_y$  of  $y$ . Obtain Kendall's correlation coefficient  $\tau$ .

	$G$	$D$	$A$	$E$	$H$	$B$	$C$	$F$
$R_x$	1	2	3	4	5	6	7	8
$R_y$	2	1	5	3	6	7	4	8
$P$	6	6	3	4	2	1	1	0
$Q$	1	0	2	0	1	1	0	0

Consider a 2 in  $R_y$ . There 6 values (5,3,6,7,4,8) **greater than** a 2 on its right, hence record 6 in row of  $P$  and record 1 in row of  $Q$  since only one value (1) is less than a 2 on its right.

Consider a 5 in the  $R_y$  row, there are 3 values (6,7,8) greater than a 5 on its right and there are 2 values (3,4) **less than** a 5 on its right. Hence record a 3 in  $P$  row and a 2 in the  $Q$  row.

$$\begin{aligned} S &= \sum P - \sum Q = 23 - 5 = 18 \Rightarrow \\ \tau &= \frac{S}{\frac{1}{2}n(n-1)} = \frac{18}{(4)(7)} = +0.6429 \end{aligned}$$

**Example 3.7.8** For the ranked data below, compute the Kendalls' correlation coefficient

	<i>E</i>	<i>C</i>	<i>I</i>	<i>F</i>	<i>G</i>	<i>A</i>	<i>D</i>	<i>K</i>	<i>H</i>	<i>B</i>
$R_x$	1	2	3	4	5	6	7	8	9	10
$R_y$	6	8	4	1	9	2	10	3	5	7
$P$	4	2	4	6	1	4	0	2	1	0
$Q$	5	6	3	0	4	0	3	0	0	0
$S$	-1	-4	+1	6	-3	4	-3	2	1	0

$$S = \sum P - \sum Q = 24 - 21 = 3, \Rightarrow \tau = \frac{S}{\frac{1}{2}n(n-1)} = \frac{3}{\frac{1}{2}(10)(9)} = \frac{1}{15} = +0.0667$$

**Note 3.7.2** Note that if there is no value greater or less than the one in consideration record zero.

**Example 3.7.9** Below are numbers of hours studied by 10 candidates for an exam and their grades:

Number of hours( <i>x</i> )	10	7	13	15	12	7	20	17	4	10
Grades( <i>y</i> )	58	46	51	74	72	56	96	87	35	67

calculate Kendall's coefficient of rank correlation using method I.

In this method, we need to first arrange the first row in Descending order, then use **row two** to determine whether what is on right is higher(*P*) or lower (*Q*) than the column rank.

<i>x</i>	20	17	15	13	12	10	10	7	7	4
<i>y</i>	96	87	74	51	72	58	67	46	56	35
$R_x$	1	2	3	4	5	6.5	6.5	8.5	8.5	10
$R_y$	1	2	3	8	4	6	5	9	7	10
P	9	8	7	2	5	3	3	1	1	0
Q	0	0	0	4	0	1	0	1	0	0

$$\sum P = 39, \sum Q = 6, S = \sum P - \sum Q = 39 - 6 = 33 \Rightarrow \tau = \frac{S}{\frac{1}{2}n(n-1)} = \frac{33}{5(9)} = +0.7333$$

Alternatively

	A	B	C	D	E	F	G	H	I	J	
$R_x$	1	2	3	4	5	6.5	6.5	8.5	8.5	10	
$R_y$	1	2	3	8	4	6	5	9	7	10	
		AB	AC	AD	AE	AF	AG	AH	AI	AJ	
		+1	+1	+1	+1	+1	+1	+1	+1	+1	+9
			BC	BD	BE	BF	BG	BH	BI	BJ	
			+1	+1	+1	+1	+1	+1	+1	+1	+8
				CD	CE	CF	CG	CH	CI	CJ	
				+1	+1	+1	+1	+1	+1	+1	+7
					DE	DF	DG	DH	DI	DJ	
					-1	-1	-1	-1	+1	+1	-2
						EF	EG	EH	EI	EJ	
						+1	+1	+1	+1	+1	+5
							FG	FH	FI	FJ	
							-1	+1	+1	+1	+2
								GH	GI	GJ	
								+1	+1	+1	+3
									HI	HJ	
									-1	+1	0
										IJ	
										+1	+1

Total Score

$$S = 9 + 8 + 7 - 2 + 5 + 2 + 3 + 0 + 1 = 33$$

to have

$$\tau = \frac{S}{\frac{1}{2}n(n-1)} = \frac{33}{5(9)} = +0.7333$$

**Note 3.7.3** To determine whether +1 or -1, we only concentrate on the second row of ranks.

**Example 3.7.10** The crop yield for two firms in the period 1980-1985 are given in the table below:

Year	1980	1981	1982	1983	1984	1985
Farm $x$	100	101	103	140	151	106
Farm $y$	98	87	100	101	140	121

Calculate Kendall's rank correlation coefficient.

$x$	100	101	103	140	151	106
$y$	98	87	100	101	140	121
$R_x$	1	2	3	4	5	6
$R_y$	1	3	2	4	6	5
$P$	5	3	3	2	0	0
$Q$	0	1	0	0	1	0

$$S = \sum P - \sum Q = 13 - 2 = 11 \Rightarrow \tau = \frac{S}{\frac{1}{2}n(n-1)} = \frac{11}{\frac{1}{2}(6)(5)} = \frac{11}{15} = +0.7333$$

Alternatively,

Year	1984	1983	1985	1982	1981	1980	
	$A$	$B$	$C$	$D$	$E$	$F$	
$R_x$	1	2	3	4	5	6	
$R_y$	1	3	2	4	6	5	
		$AB$	$AC$	$AD$	$AE$	$AF$	
		+1	+1	+1	+1	+1	+5
			$BC$	$BD$	$BE$	$BF$	
			-1	+1	+1	+1	+2
				$CD$	$CE$	$CF$	
				+1	+1	+1	+3
					$DE$	$DF$	
					+1	+1	+2
						$EF$	
						-1	-1
							Total Score $S = 11$

And Kendall's rank correlation coefficient for  $n = 6$

$$\tau = \frac{S}{\frac{1}{2}n(n-1)} = \frac{11}{\frac{1}{2}(6)(5)} = \frac{11}{15} = +0.7333$$

**Example 3.7.11** Ranks of commodities  $x$  and  $y$  are given below

Rank  $x$    6   5   7.5   7.5   9   2   3.5   3.5   1   10

Rank  $y$    3   10   2   7   7   7   4   1   9   5

Compute the Kendall's rank correlation coefficient.

$R_x$    6   5   7.5   7.5   9   2   3.5   3.5   1   10

$R_y$    3   10   2   7   7   7   4   1   9   5

$P$    7   0   6   1   1   1   2   2   0   0

$Q$    2   8   1   3   3   3   1   0   1   0

$$S = \sum P - \sum Q = 20 - 22 = -2, \Rightarrow \tau = \frac{S}{\frac{1}{2}n(n-1)} = \frac{-2}{\frac{1}{2}(10)(9)} = -\frac{2}{45} = -0.0444$$

Alternatively,

	$A$	$B$	$C$	$D$	$E$	$F$	$G$	$H$	$I$	$J$	
$R_x$	6	5	7.5	7.5	9	2	3.5	3.5	1	10	
$R_y$	3	10	2	7	7	7	4	1	9	5	
	$AB$	$AC$	$AD$	$AE$	$AF$	$AG$	$AH$	$AI$	$AJ$		
	+1	-1	+1	+1	+1	+1	-1	+1	+1	+5	
		$BC$	$BD$	$BE$	$BF$	$BG$	$BH$	$BI$	$BJ$		
		-1	-1	-1	-1	-1	-1	-1	-1	-8	
			$CD$	$CE$	$CF$	$CG$	$CH$	$CI$	$CJ$		
			+1	+1	+1	+1	-1	+1	+1	+5	
				$DE$	$DF$	$DG$	$DH$	$DI$	$DJ$		
				0	0	-1	-1	+1	-1	-2	
					$EF$	$EG$	$EH$	$EI$	$EJ$		
					0	-1	-1	+1	-1	-2	
						$FG$	$FH$	$FI$	$FJ$		
						-1	-1	+1	-1	-2	
							$GH$	$GI$	$GJ$		
							-1	+1	+1	+1	
								$HI$	$HJ$		
								+1	+1	+2	
									$IJ$		
									-1	-1	

Then total Score  $S = -2$  to have

$$\tau = \frac{S}{\frac{1}{2}n(n-1)} = \frac{-2}{\frac{1}{2}(10)(9)} = -\frac{2}{45} = -0.0444$$

## 3.8 Interpretation of the Coefficient

We note in all examples done that the value of  $\gamma, \rho, \tau$  can be any value positive or negative. They specifically take on any value between  $-1$  and  $1$ .

We have to note that the correlation is positive correlation or variables are positively correlated if an increase in one variable is associated with an increase in the other, the other side of the statement being a negative correlation.

If the value computed is close to  $+1$  say  $+0.8, +0.78, +0.9$  etc then we say there is a high positive correlation or the variables are highly positively correlated.

If the value computed is close to  $0$  say  $+0.12, +0.07, +0.3$ , then this is a weak positive correlation. If the computed value is  $0$  then no correlation, if it is  $+1$  then there is a perfectly positive correlation or variables are perfectly correlated.

If the value is  $-1$  then perfect negative correlation.

Perfect correlation occurs when all points lie on the line in the scatter diagram. Make sure that the computed value is accompanied by an explanation to show whether perfectly correlated, high, weak, positive or negative correlation.

### 3.8.1 Covariance

**Definition 3.8.1** Let  $g(X, Y) = [X - \mathbb{E}(X)][Y - \mathbb{E}(Y)]$  then  $\mathbb{E}[g(X, Y)]$  is called the covariance of  $X$  and  $Y$ , that is,

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \quad (3.12)$$

It is easy to show that Equation (3.12) can be rewritten as

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \quad (3.13)$$

the computational form of  $\text{cov}(X, Y)$ .

**Proof :**

$$\begin{aligned} \text{cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\ &= \mathbb{E}[XY - X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y)] \\ &= \mathbb{E}[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y] \\ &= \mathbb{E}(XY) - \mu_Y\mathbb{E}(X) - \mu_X\mathbb{E}(Y) + \mathbb{E}(\mu_X\mu_Y) \\ &= \mathbb{E}(XY) - \mu_Y\mathbb{E}(X) - \mu_X\mathbb{E}(Y) + \mu_X\mu_Y \\ &= \mathbb{E}(XY) - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y \\ &= \mathbb{E}(XY) - \mu_X\mu_Y \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \end{aligned}$$

■

**Exercise 3.30** If  $X_1 = aX + b$  and  $X_2 = cY + d$ , prove that

$$\text{cov}(X_1, X_2) = ac \text{cov}(X, Y)$$

# Chapter 4

## Probability Spaces

### Introduction

Every busy mind at the onset of a day starts with an activity that involves some element of “chance”. The most probable words that would be involved are ‘either’ and ‘or’. These tend to originate from probabilistic situations that hang round every aspect of life. Think of a pregnant woman. Either she would produce a baby boy or a baby girl (leave alone the rare cases of “both”). Consider a football match in which the verdict must be decided on. Either one team will win or lose. Initially, the probability (chance) of doing something or not doing it would seem to be  $\frac{1}{2}$  to most of you. On a cloudy day, it would even be obvious that the chance or raining or not raining is  $\frac{1}{2}$ .

The relevance of probability theory in life’s problems was identified as early as the 16<sup>th</sup> century by great Mathematicians such as Abraham de Moivre (1776 – 1754), Reverend Thomas Baye (1702 – 1761), and Joseph Lagrange (1736 – 1813). This was later strengthened in the 19<sup>th</sup> century by researchers such as Laplace who unified all these early ideas.

**Definition 4.0.1** Probability is a measure of uncertainty of events.



## Set Theory, Events and Probability of Events

### 4.1 Sample Space, Events and Experiment

Any probability pertains to the results of a situation which we call an experiment. An experiment is any process by which data is obtained (either through controlled or uncontrolled procedures).

**Example 4.1.1** Examples of an experiment could be:

- 1.) Tossing a coin
- 2.) Rolling a die
- 3.) Measuring heights of students
- 4.) Giving a test to a class
- 5.) Administering a drug to patients

**Definition 4.1.1** The results of an experiment are called *outcomes*. For instance, in tossing a coin, a head or a tail appearing is an outcome. An experiment can result in various outcomes.

**Definition 4.1.2** A set of all possible outcomes that may occur in a particular experiment is called a *sample space*. We denote this by  $\Omega$ .

**Example 4.1.2** For the examples above

- 1.) Tossing a coin once

$$\Omega = \{H, T\}$$

- 2.) Rolling a dice once

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

- 3.) Sex of a newly born

$$\Omega = \{b, g\}$$

**Definition 4.1.3** Any subset of a sample space  $\Omega$  is called an *event*.

An event is a set consisting of possible outcomes of an experiment with desired qualities.

**Example 4.1.3** Examples of events could be:

- 1.) Toss a die, then  $\Omega = \{1, 2, 3, 4, 5, 6\}$ . If the desired quality is that an “odd” number shows up then Odd is the event  $E$  and thus  $E = \{1, 3, 5\}$ .
- 2.) If two coins are flipped then

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}.$$

If the desired quality is that a tail  $T$  appears first then

$$E = \{(T, H), (T, T)\}.$$

## 4.2 Operations with Events

You will notice that operation with probability events is similar to familiar set theory and it will be helpful to you to try and use Venn diagrams to verify some of the results much as you apply them in sets. Take an experiment in which multiple events are allowed to occur. Then the following could result:

### 4.2.1 Intersection of Events

For any two events  $A$  and  $B$ , we define a new event  $A \cap B$  called the Intersection of  $A$  and  $B$  and consists of all outcomes that are in both  $A$  and  $B$ .

**Example 4.2.1**  $A = \{1, 2, 3, 4, 5\}$ ,  $B = \{3, 4, 6\}$  then  $A \cap B = \{3, 4\}$ .

### 4.2.2 Union of Events

For any two events  $A$  and  $B$ , we define a new event  $A \cup B$  called the Union of  $A$  and  $B$  and consists of all outcomes that are either in  $A$  or in  $B$  or both. Thus event  $A \cup B$  will occur if either  $A$  or  $B$  occurs.

**Example 4.2.2**  $A = \{1, 2, 3, 4\}$ ,  $B = \{2, 4, 5, 6\}$  then  $A \cup B = \{1, 2, 3, 4, 5, 6\}$

### 4.2.3 Complementary Events

To each event  $A$ , we define a new event  $A'$  or  $A^c$  known as the complement of  $A$  and consists of all outcomes in a sample space not in  $A$ .

**Example 4.2.3** For  $\xi = \Omega = \{1, 2, 3, 4, 5\}$ ,  $A = \{2, 4, 5\}$  then  $A^c = A' = \{1, 3\}$ .

### 4.2.4 Null Event

If an event  $A$  does not contain any outcomes and hence could not occur, we call it a null event and we denote it by  $A = \emptyset$  or  $\{\}$  but not  $\{\emptyset\}$

**Example 4.2.4** Given

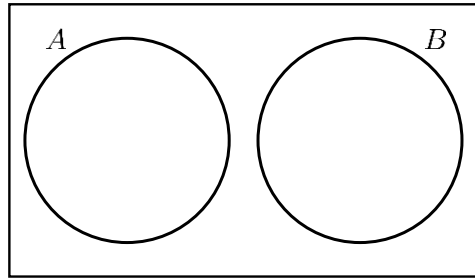
$$A = \{\text{mathematics, chemistry, biology}\}$$

$$B = \{\text{math, chem, bio}\}$$

$$A \cap B = \emptyset$$

### 4.2.5 Disjoint Events

For any two events  $A$  and  $B$ , if  $A \cap B = \phi$  then  $A$  and  $B$  are said to be disjoint. We shall refer to  $A$  and  $B$  in this case as being mutually exclusive events - they cannot occur together.



Showing disjoint (Mutually exclusive) events  $A$  and  $B$ .

**Note 4.2.1** It should be noted that an event may be simple or compound. For instance if in tossing a coin the event  $E$  is to obtain an odd number then this is a simple event and if in shuffling a pack of cards an event is drawing a spade or heart then this is compound. The number of outcomes in an event or a sample space need not be finite or countable.

### 4.2.6 Subsets

For any two events  $A$  and  $B$ , if all the outcomes in  $A$  are also in  $B$  then we say that  $A$  is contained in  $B$  and we write  $A \subset B$  (or  $B \supset A$ ) then the occurrence of  $A$  necessarily implies the occurrence of  $B$ .

If  $A \subset B$  and  $B \subset A$  then we say that  $A$  and  $B$  are equal and we write  $A = B$ .

### 4.2.7 Other Resulting Operations of Events

The operations of forming unions, intersections and complements of events obey other rules similar to those of sets.

1.) Commutative Law

$$\left. \begin{aligned} A \cup B &= B \cup A \\ A \cap B &= B \cap A \end{aligned} \right\}$$

2.) Associative Law

$$\left. \begin{aligned} (A \cup B) \cup C &= A \cup (B \cup C) \\ (A \cap B) \cap C &= A \cap (B \cap C) \end{aligned} \right\}$$

3.) Distributive Law

$$\left. \begin{aligned} (A \cup B) \cap C &= (A \cap C) \cup (B \cap C) \\ (A \cap B) \cup C &= (A \cup C) \cap (B \cup C) \end{aligned} \right\}$$

4.) Complementary Law

$$\left. \begin{aligned} A \cup A^c &= \Omega \\ A \cap A^c &= \phi \end{aligned} \right\}$$

5.) De Morgan's Law

$$\left. \begin{aligned} (A \cap B)^c &= A^c \cup B^c \\ (A \cup B)^c &= A^c \cap B^c \end{aligned} \right\}$$

At times we generalize these laws especially the De Morgan Laws:

$$\begin{aligned} \left[ \bigcup_{i=1}^n E_i \right]^c &= \bigcap_{i=1}^n E_i^c \\ \left[ \bigcap_{i=1}^n E_i \right]^c &= \bigcup_{i=1}^n E_i^c \end{aligned}$$

for events  $E_1, E_2, \dots, E_n$  or for  $E_i, i = 1, 2, \dots, n$

**Example 4.2.5** An ordinary die is thrown. Let  $A$  denote the event that the score obtained will be even,  $B$  the event that the score will be less than 3 and  $C$  the event that the score will be a multiple of 3.

Write down the sample space for the possible scores and elements of events

- |                |                       |                                  |
|----------------|-----------------------|----------------------------------|
| 1.) $A'$       | 3.) $A \cup C'$       | 5.) $A \cap B \cap C$            |
| 2.) $A \cap B$ | 4.) $A \cup B \cup C$ | 6.) $(A \cap C) \cup (B \cap C)$ |

The sets are given by

$$\begin{aligned} \xi = U = \Omega &= \{1, 2, 3, 4, 5, 6\} \\ A &= \{2, 4, 6\} \\ B &= \{1, 2\} \\ C &= \{3, 6\} \end{aligned}$$

- |                                     |   |
|-------------------------------------|---|
| 1.) $A' = \{1, 3, 5\}$              | 4.) $A \cup B \cup C = \{1, 2, 3, 4, 6\}$   |
| 2.) $A \cap B = \{2\}$              | 5.) $A \cap B \cap C = \emptyset$ or $\{\}$ |
| 3.) $A \cup C' = \{1, 2, 4, 5, 6\}$ | 6.) $(A \cap C) \cup (B \cap C) = \{6\}$    |

## 4.3 Probabilities of Events

By the probability of an event, we mean a measure of how likely it is that an event will occur in a sample space. Thus, probabilities are numbers that are assigned to each of the elements of the sample space.

The number assigned to a particular outcome of a sample space is the proportion of times that specific outcome occurs over long run if an experiment is performed.

For instance, if a die is rolled once then a 3 can only occur once in the six possible outcomes and so one out of six is the number we can assign to a 3 occurring.

We shall denote the probability of an event  $A$  in a sample large number of times  $N$  and if the outcome of an event  $A$  occurs  $n$  times then the probability of  $A$  is

$$P(A) = \frac{n}{N}; N \neq 0 \quad (4.1)$$

Thus the probability of a 3 occurring since it occurs once out of six possible outcomes is

$$P(3) = \frac{1}{6}$$

**Example 4.3.1** What is the probability of getting a similar values when a pair of dice is tossed?

$$\frac{6}{36}$$

**Example 4.3.2** What is the probability of getting a head and an even number when a die and a coin are thrown together?

$$\frac{3}{12}$$

**Example 4.3.3** What is the probability of getting two heads when a pair of coins is tossed once?

$$\frac{1}{4}$$

**Exercise 4.1** What is the probability of getting two heads when three coins are tossed once?

## 4.4 Axioms of Probability

Basing on our definition of probability of an event, we can thus state or attach important conditions on the probability function of any event or sample space.

Consider an event  $A$  of a sample space  $\Omega$ . Then we shall define

$n(A)$  as the number of members or elements in  $A$

$n(\Omega)$  would mean the number of members in  $\Omega$ .

Thus the probability of an event  $A$  of  $\Omega$  occurring

$$P(A) = \frac{n(A)}{n(\Omega)}; \quad n(\Omega) \neq 0 \quad (4.2)$$

If we assume that for each event  $A_i$  of a sample space  $\Omega$ , a number  $P(A_i)$  is defined then it will satisfy the following **Axioms**.

### Axiom 1

$$0 \leq P(A_i) \leq 1 \quad (4.3)$$

### Axiom 2

$$P(\Omega) = \sum_{i=1}^n P(A_i) = 1 \quad (4.4)$$

### Axiom 3 For *disjoint events* $A_i$ s

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) \quad (4.5)$$

## 4.5 Theorems of Probability

Some probability theorems include

1.) Since  $A \cup \phi = A$

$$P(\phi) = 0 \quad (4.6)$$

2.) For any two events  $A$  and  $B$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (4.7)$$

This is called the additive rule of probability.

3.)

$$P(A') = 1 - P(A) \quad (4.8)$$

**Proof :**

$$\begin{aligned} P(A \cup A') &= P(\Omega) = P(A) + P(A') - P(A \cap A') \\ P(\Omega) &= P(A) + P(A') - P(\emptyset) \\ 1 &= P(A) + P(A') - 0 \\ \Rightarrow P(A') &= 1 - P(A) \end{aligned}$$

■

### 4.5.1 Other Related Theorems of Probability

4a.) If  $A \subseteq B$  then

$$P(B) \geq P(A)$$

4b.) For any events  $A$  and  $B$ ,

$$P(A \cap B) \geq P(A) + P(B) - 1$$

**Example 4.5.1** *Proof of Axiom 1, Equation (4.3)*

**Proof :**

$$\begin{aligned} \phi &\subset A \subset \Omega \\ P(\phi) &\leq P(A) \leq P(\Omega) \\ 0 &\leq P(A) \leq 1 \end{aligned}$$

■

**Example 4.5.2** From the additive rule of probability theorem, Equation (4.7)

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

We may generalize the additive rule to more than two events. For instance if for two events  $A$  and  $B$ ,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

then for three events  $A, B$  and  $C$

$$\begin{aligned} P(A \cup B \cup C) &= P(A \cup (B \cup C)) \\ &= P(A) + P(B \cup C) - P(A \cap (B \cup C)) \\ &= P(A) + P(B) + P(C) - P(B \cap C) - P(A \cap (B \cup C)) \end{aligned}$$

and

$$\begin{aligned} P(A \cap (B \cup C)) &= P((A \cap B) \cup (A \cap C)) \\ &= P(A \cap B) + P(A \cap C) - P(A \cap B \cap C) \end{aligned}$$

To have

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

## 4.6 Contingency Table

The table below will help you compute probabilities of complementary events that are associated with union and intersection of complements.

From a Venn-Diagram

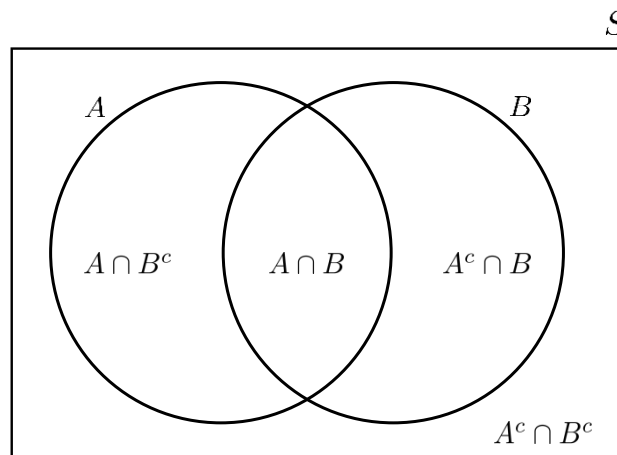


Figure 4.1: Two sets Venn-Diagram



The Contingency table is a tabular summary of the venn diagram as shown in Table 4.1.

	$A$	$A^c$	
$B$	$A \cap B$	$A^c \cap B$	$P(B)$
$B^c$	$A \cap B^c$	$A^c \cap B^c$	$P(B^c)$
	$P(A)$	$P(A^c)$	1

Table 4.1: Contingency table showing operation with complementary events

From the table we note that:

$$P(A) = P(A \cap B) + P(A \cap B^c) \quad (4.9)$$

$$P(A^c) = P(A^c \cap B) + P(A^c \cap B^c) \quad (4.10)$$

$$P(B) = P(A \cap B) + P(A^c \cap B) \quad (4.11)$$

$$P(B^c) = P(A \cap B^c) + P(A^c \cap B^c) \quad (4.12)$$

Similarly the Contingency table shows that

$$P(A) + P(A^c) = 1$$

$$P(B) + P(B^c) = 1$$

## 4.7 De Morgan's Laws

$$P(A^c \cap B^c) = P(A \cup B)^c \quad (4.13)$$

$$P(A^c \cup B^c) = P(A \cap B)^c \quad (4.14)$$

**Example 4.7.1** We also note from the table that

$$\begin{aligned}
 P(B^c) &= P(A \cap B^c) + P(A^c \cap B^c) \\
 &= P(A) - P(A \cap B) + 1 - P(A \cup B) \\
 &= P(A) - P(A \cap B) + 1 - (P(A) + P(B)) - P(A \cap B) \\
 &= P(A) - P(A \cap B) + 1 - P(A) - P(B) + P(A \cap B) \\
 &= 1 - P(B)
 \end{aligned}$$

Hence justifying the complementary laws Theorem Equation (4.8).

**Example 4.7.2** Events  $A$  and  $B$  are such that  $P(A \cap B) = 0.4$  and  $P(A \cup B) = 0.7$ , Given that  $P(A) = P(B) = x$ , find  $x$ .

From the additive rule

$$\begin{aligned}
 P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\
 \Rightarrow 0.7 &= x + x - 0.4 \\
 \Rightarrow 2x &= 0.7 + 0.4 = 1.1 \\
 \Rightarrow x &= 0.55
 \end{aligned}$$

**Example 4.7.3** Given that for events  $A$  and  $B$

$$P(A \cup B) = \frac{7}{8}, \quad P(A \cap B) = \frac{1}{4} \text{ and } P(A^c) = \frac{5}{8}.$$

Find the values of

- |              |                         |   |
|--------------|-------------------------|---|
| 1.) $P(A)$ , | 3.) $P(A \cap B^c)$ ,   | 5.) $P((A \cap B^c) \cup (A^c \cap B))$ |
| 2.) $P(B)$ , | 4.) $P(A^c \cup B^c)$ , |   |

The solutions are

- 1.) By the Axiom of probability (Axiom 1)

$$\begin{aligned} P(A) &= 1 - P(A^c) \\ &= 1 - \frac{5}{8} = \frac{3}{8} \end{aligned}$$

- 2.) Using the theorem of probability

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ \Rightarrow P(B) &= P(A \cup B) + P(A \cap B) - P(A) \\ &= \frac{7}{8} + \frac{1}{4} - \frac{3}{8} = \frac{3}{4} \end{aligned}$$

- 3.) From the contingency table

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= \frac{3}{8} - \frac{1}{4} = \frac{1}{8} \end{aligned}$$

- 4.) By De'morgan's rule

$$\begin{aligned} P(A^c \cup B^c) &= P(A \cap B)^c \\ &= 1 - P(A \cap B) \\ &= 1 - \frac{1}{4} = \frac{3}{4} \end{aligned}$$

- 5.) By the Axiom of probability (Axiom 3) since  $(A \cap B^c)$ ,  $(A^c \cap B)$  are disjoint sets

$$\begin{aligned} P[(A \cap B^c) \cup (A^c \cap B)] &= P(A \cap B^c) + P(A^c \cap B) \\ &= \{P(A) - P(A \cap B)\} - \{P(B) - P(A \cap B)\} \\ &= \frac{1}{8} + \frac{1}{2} = \frac{5}{8} \end{aligned}$$

**Example 4.7.4** Events  $A$  and  $B$  are such that  $P(A) = \frac{19}{30}$ ,  $P(B) = \frac{2}{5}$  and  $P(A \cup B) = \frac{4}{5}$ .  
Find  $P(A \cap B)$ . [7/30]

**Example 4.7.5** A coin and a die are tossed together, draw the possibility sample diagram and find the probability of getting

- 1.) a head. [1/2]
- 2.) number greater than 4. [1/3]
- 3.) a head and a number greater than 4. [1/6]
- 4.) a head or a number greater than 4. [2/3]

**Example 4.7.6** A card is drawn from an ordinary pack of 52 playing cards. Find the probability that a card selected is a club or a diamond. [1/2]

**Exercise 4.2** An ordinary die is thrown once. Let  $E$  denote the event that the score obtained will be even,  $F$  the event that the score obtained will be less than 3 and  $G$  the event that the score will be multiple of 3. Give verbal descriptions and set representation of the events.

- |                  |                           |                                  |
|------------------|---------------------------|----------------------------------|
| 1.) $E^c$        | 4.) $E \cap F \cap G$     | 7.) $(E \cap F) \cup (F \cap G)$ |
| 2.) $E \cap F$   | 5.) $E \cap F^c \cap G^c$ |                                  |
| 3.) $E \cup F^c$ | 6.) $(E \cup F) \cap G$   |                                  |

**Exercise 4.3** A green die and a blue die are thrown simultaneously. Write down the sample space of this experiment expressing each element as an ordered pair. Determine the set representation of each of these verbal events given

- $A$  = the sum of the score is divisible by 4  
 $B$  = the scores are equal  
 $C$  = both scores are even  
 $D$  = the scores differ by at least 4

- |                |                    |
|----------------|--------------------|
| 1.) $A \cap C$ | 3.) $B \cap A^c$   |
| 2.) $B \cup D$ | 4.) $(A \cup B^c)$ |

**Exercise 4.4** Given that events  $A$  and  $B$  are such that

$$P(A) = 0.5, P(A \cup B) = 0.8, P(A \cap B) = 0.2$$

Compute

- |                |                       |                         |
|----------------|-----------------------|-------------------------|
| 1.) $P(A^c)$ , | 3.) $P(A \cap B^c)$   | 5.) $P(A^c \cup B^c)$ . |
| 2.) $P(B)$     | 4.) $P(A^c \cap B)$ , |                         |

**Exercise 4.5** Events  $A$  and  $B$  are such that

$$P(A) = 0.36, P(B) = 0.25, P(A \cap B^c) = 0.24$$

Find the values of

$$1.) P(A^c), \quad 2.) P(A \cap B) \quad 3.) P(A^c \cap B), \quad 4.) P(A^c \cup B^c).$$

Also find the probability that exactly one of  $A, B$  occurs.

**Exercise 4.6** Given

$$P(A) = 0.59, P(B) = 0.30, P(A \cap B) = 0.21$$

Find the probabilities

$$1.) P(A \cup B) \quad 2.) P(A \cap B^c) \quad 3.) P(A^c \cup B^c)$$

**Exercise 4.7** For a married couple, the probability that the husband will vote is 0.21 and that the wife will vote is 0.28 and that both will vote is 0.15. What is the probability that at least one of them will vote?

**Exercise 4.8** At Nile College, it is known that  $\frac{1}{4}$  of the students live off campus,  $\frac{5}{9}$  of the students come from Busoga and  $\frac{3}{4}$  of the students are out-of-Busoga or live on campus. What is the probability that a student selected at random from Nile College is from Out-of-Busoga and lives on campus? [4/9]

**Exercise 4.9** Prove the Demorgan's Law.

## 4.8 Special Events

When applying the five special ways of computing probabilities, some events have special properties such as

1.) Mutually exclusive - disjoint events

2.) Conditional events

which result into Total probabilities and Baye's theorem.

### 4.8.1 Mutually Exclusive Events

Two events  $A$  and  $B$  are said to be mutually exclusive if they cannot occur together. Thus if two events are mutually exclusive, the probability that they both occur is zero, that is

$$P(A \cap B) = 0 \quad (4.15)$$

Then from the additive rule of events, if two events  $A$  and  $B$  are mutually exclusive

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ P(A \cup B) &= P(A) + P(B) - 0 \\ P(A \cup B) &= P(A) + P(B) \end{aligned} \quad (4.16)$$

**Example 4.8.1** In the example of selecting a number from a set of integers, events

$A$  = odd number selected

$B$  = even number selected

are mutually exclusive since we cannot select even and odd numbers at the same time picking.

In general, the additive rule, for mutually exclusive events  $A_1, A_2, \dots, A_n$

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$

We define for three events  $A, B, C$  (or more) that if they are mutually exclusive pairwise, that is

$$P(A \cap B) = P(B \cap C) = P(A \cap C) = 0$$

and in addition

$$A \cup B \cup C = \xi \quad (4.17)$$

the universal set, then they are mutually exhaustive.

**Example 4.8.2**  $A$  and  $B$  are mutually exclusive events such that  $P(A) = 0.5$  and  $P(B) = 0.2$ . Find

1.)  $P(A \cup B)$

2.)  $P(A^c \cap B^c)$

Since mutually exclusive,  $P(A \cap B) = 0$

1.)

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.5 + 0.2 - 0 = 0.7 \end{aligned}$$

2.)

$$\begin{aligned} P(A^c \cap B^c) &= P(A \cup B)^c \text{ (by De Morgan).} \\ &= 1 - P(A \cup B) \text{ (by Complementary)} \\ &= 1 - 0.7 = 0.3 \end{aligned}$$

**Example 4.8.3** In a race, the probability that John wins is 0.3; the probability that Paul wins is 0.2 and the probability that Mark wins is 0.4. Find the probability that

1.) John or Mark wins

2.) Neither John nor Paul wins.

3.) Either John or Paul or Mark to win

Assume that there are no dead heats. But even if not said, winning a race is mutually exclusive since both people cannot win a race = mutually exclusive.

Define Events

$$J \equiv \text{John wins}, \quad M \equiv \text{Mark wins}, \quad P \equiv \text{Paul wins}$$

Then

$$P(J) = 0.3 \quad P(M) = 0.4 \quad P(P) = 0.2$$

(a)

$$\begin{aligned} P(J \cup M) &= P(J) + P(M) - P(J \cap M) \\ &= 0.3 + 0.4 - 0 = 0.7 \end{aligned}$$

(b)  $P(\text{neither John nor Paul})$ 

$$\begin{aligned} P(J^c \cap P^c) &= P(J \cup P)^c = 1 - P(J \cup P) \\ &= 1 - [P(J) + P(P) - P(J \cap P)] \\ &= 1 - (0.3 + 0.2 - 0) = 0.5 \end{aligned}$$

### 4.8.2 Conditional Events

Consider two events  $A$  and  $B$ . At times it is impossible for any one of the two events to occur without another. In such cases, a preconceived idea about the one which has occurred would help you greatly to ascertain the occurrences of the other.

**Definition 4.8.1** Conditional probability: A probability computed under the assumption that some condition holds.

Then the probability that an event  $A$  will occur given that another event  $B$  has occurred (or must occur) is known as the conditional probability of  $A$  given  $B$  and is written as  $P(A | B)$ .

**Definition 4.8.2** Regardless of whatever they stand for, the conditional probability of event  $A$  given that  $B$  has occurred denoted by  $P(A | B)$ , is

$$P(A | B) = \frac{P(B \cap A)}{P(B)}; \quad P(B) \neq 0 \quad (4.18)$$

And the conditional probability of  $B$  given  $A$  is

$$P(B | A) = \frac{P(A \cap B)}{P(A)}; \quad P(A) \neq 0 \quad (4.19)$$

### 4.8.3 Multiplicative Rule

From (4.18) and (4.19), we clearly define the multiplicative rule of two events or joint probability that both events  $A$  and  $B$  will occur as

$$P(A \cap B \cap C) = P(C | A \cap B) \cdot P(B | A) \cdot P(A) \quad (4.20)$$

Consequences:

$$1.) \quad P(\Omega | B) = \frac{P(B \cap \Omega)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

$$2.) \quad P(\phi | B) = \frac{P(B \cap \phi)}{P(B)} = \frac{P(\phi)}{P(B)} = \frac{0}{P(B)} = 0$$

3.) If  $A_1, A_2, \dots, A_n$  are mutually exclusive

$$P(A_1 \cup A_2 \cup \dots \cup A_n | B) = P(A_1 | B) + P(A_2 | B) + \dots + P(A_n | B) = \sum_{i=1}^n P(A_i | B)$$

$$4.) \quad a) \quad P(A^c | B) = 1 - P(A | B)$$

**Proof :**

$$P(A^c | B) = \frac{P(B \cap A^c)}{P(B)} = \frac{P(B) - P(A \cap B)}{P(B)} = 1 - P(A | B)$$

■

$$b) \quad P(A^c | B^c) = \frac{P(B^c \cap A^c)}{P(B^c)} = \frac{1 - P(A \cup B)}{1 - P(B)}$$

5.) For any event  $A_1$  and  $A_2$

$$(i) P(A_1 | B) = P(A_1 \cap A_2 | B) + P(A_1 \cap A_2^c | B)$$

$$(ii) P(A_1 \cup A_2 | B) = P(A_1 | B) + P(A_2 | B) - P(A_1 \cap A_2 | B)$$

6.) If events  $A_1$  and  $A_2$  are such that  $A_1 \subset A_2$  then  $P(A_1 | B) \leq P(A_2 | B)$

7.) a) If  $A, B, C$  are any three events, such that  $P(A \cap B) \neq 0$  then

$$P(A \cap B \cap C) = P(C | A \cap B) \cdot P(B | A) \cdot P(A)$$

For, writing  $P(A \cap B \cap C)$  as

$$P((A \cap B) \cap C) = P(C | A \cap B) \cdot P(A \cap B) = P(C | A \cap B) \cdot P(B | A) \cdot P(A)$$

b) If  $A_1, A_2, \dots, A_n$  are  $n$  events, the joint probability of these events

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_n | A_1 \cap A_2 \cap \dots \cap A_{n-1}) \cdots P(A_3 | A_1 \cap A_2) \cdot P(A_2 | A_1) \cdot P(A_1)$$

**Example 4.8.4** Given that events  $A$  and  $B$  are such that  $P(A) = 0.6, P(B) = 0.2$  and  $P(B | A) = 0.1$ . Find

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

2.)  $P(A | B)$

3.)  $P(A^c | B^c)$

The solutions are given by

1.)  $P(A \cap B) = P(B | A) \cdot P(A) = (0.1)(0.6) = 0.06$

2.)

$$P(A | B) = \frac{P(B \cap A)}{P(B)} = \frac{0.06}{0.2} = 0.3$$

$$P(B \cap A) = P(A \cap B), P(B \cup A) = P(A \cup B) \text{ for such events with no order.}$$

3.)

$$\begin{aligned} P(A^c | B^c) &= \frac{P(B^c \cap A^c)}{P(B^c)} = \frac{P(A \cup B)^c}{P(B^c)} = \frac{1 - P(A \cup B)}{1 - P(B)} \\ &= \frac{1 - [P(A) + P(B) - P(A \cap B)]}{1 - P(B)} \\ &= \frac{1 - [0.6 + 0.2 - 0.06]}{[1 - 0.2]} \\ &= \frac{0.26}{0.8} \\ &= 0.325 \end{aligned}$$

**Note 4.8.1**  $A \cap B = B \cap A$  only if the occurrence of events do not matter, however they usually do.



**Example 4.8.5** An urn contains 20 mangos of which 5 are bad. If 2 mangos are selected at random in succession without replacement of the first one what is the probability that both are bad.

$$\begin{aligned}P(B_1 \cap B_2) &= P(B_2 | B_1) \cdot P(B_1) \\&= \left(\frac{4}{19}\right) \left(\frac{1}{4}\right) = \frac{1}{19}\end{aligned}$$

**Example 4.8.6** The probability that a student wakes up early is 0.83 and the probability that he arrives at school on time is 0.92. The probability that he wakes up early and arrives on time is 0.78. Find the probability that the student

1.) arrives on time given that he wakes up early

Define events

$$\begin{aligned}E &= \text{wakes up early} \\A &= \text{arrives on time}\end{aligned}$$

$$\begin{aligned}P(E) = 0.83 &\Rightarrow P(E') = 0.17 \\P(A) = 0.92 &\Rightarrow P(A') = 0.08 \\P(E \cap A) = 0.78 &\Rightarrow P(E \cap A)' = P(E' \cup A') = 0.22\end{aligned}$$

Thus the solution is

$$P(A | E) = \frac{P(E \cap A)}{P(E)} = \frac{0.78}{0.83} = 0.94$$

2.) wakes up early given that he arrives on time.

$$P(E | A) = \frac{P(A \cap E)}{P(A)} = \frac{???}{0.92}$$

**Note 4.8.2** “and” is usually not commutative, (waking up early “and” arrive on time, might not be, arriving on time “and” then wake up early) that is  $P(A \cap E) \neq P(E \cap A)$

**Example 4.8.7** A coin is flipped twice. If we assume that all four elements in the sample space  $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$  are equally likely to occur, what is the probability that both flips result in heads given that the first flip resulted in a head?

$$P(H_2 | H_1) = \frac{P(H_1 \cap H_2)}{P(H_1)} = \frac{P(H, H)}{P((H, H) \text{ or } (H, T))} = \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{4}} = \frac{1}{2}$$

**Example 4.8.8** An Urn contains 10 white, 5 yellow and 10 black marbles. A marble is chosen at random from the urn and it is noted that it is not black. What is the probability that it was yellow?

$$\begin{aligned}P(Y | B') &= \frac{P(B' \cap Y)}{P(B')} \\&= \frac{P(B' \cap Y)}{1 - P(B)} \\&= \frac{P(Y)}{1 - P(B)} \\&= \frac{5/25}{1 - 10/25} = \frac{1}{3}\end{aligned}$$

#### 4.8.4 Independent events

Consider a situation where you are at a party and another where it will rain. You shall note that your being at a party does not stop rain. We would say that the two are independent. That is, the occurrence of one does not affect the occurrences of the other.

**Example 4.8.9** Using a pen when it is raining or not raining are independent events.

**Definition 4.8.3** Two events  $A$  and  $B$  are said to be Independent if the occurrence of one does not affect or hinder the occurrence of the other. Thus, events  $A$  and  $B$  are said to be independent if

$$P(A | B) = P(A) \quad (4.21)$$

$$P(B | A) = P(B) \quad (4.22)$$

Then from (4.21) and (4.22), that two events are independent if

$$\begin{aligned} P(A \cap B) &= P(B | A) \cdot P(A) \\ &= P(B) \cdot P(A) \\ &= P(A) \cdot P(B) \end{aligned}$$

That is

$$P(A \cap B) = P(A) \cdot P(B) \quad (4.23)$$

This is a special multiplicative rule of probability of events.

**Definition 4.8.4** Three events  $A, B, C$  are said to be pairwise independent if and only if each pair of events chosen from  $A, B, C$  are independent (that is,  $A$  and  $B$  are independent,  $B$  and  $C$  are independent,  $A$  and  $C$  are independent).

**Definition 4.8.5** If in addition to being pairwise independent

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$$

then they are said to be totally independent.

Consequences:

If  $A, B, C$  are independent events then

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$$

In general, if events  $A_1, A_2, \dots, A_n$  are independent.

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \dots P(A_n)$$

**Example 4.8.10** If  $A, B, C$  are independent events, so are  $A$  and  $B \cup C$ .

Note that, to prove independence, is to prove the *and* to be the *product*.

**Proof :**

$$\begin{aligned}
 P(A \cap (B \cup C)) &= P[(A \cap B) \cup (A \cap C)] \\
 &= P(A \cap B) + P(A \cap C) - P(A \cap B \cap C) \\
 &= P(A) \cdot P(B) + P(A) \cdot P(C) - P(A)P(B \cap C) \\
 &= P(A) [P(B) + P(C) - P(B \cap C)] \\
 &= P(A) \cdot P(B \cup C).
 \end{aligned}$$

■

**Example 4.8.11** If  $A, B$  are independent events then so are  $A$  and  $B^c$ .

Since  $A$  and  $B$  are independent,  $P(A \cap B) = P(A) \cdot P(B)$

**Proof :**

$$\begin{aligned}
 P(A \cap B^c) &= P(A) - P(A \cap B) \\
 &= P(A) - [P(A) \cdot P(B)] \\
 &= P(A) \cdot [1 - P(B)] \\
 &= P(A) \cdot P(B^c)
 \end{aligned}$$

■

**Example 4.8.12** If  $A, B$  are independent so are  $A^c$  and  $B^c$ .

**Proof :**

$$\begin{aligned}
 P(A^c \cap B^c) = P(A \cup B)^c &= 1 - P(A \cup B) \\
 &= 1 - [P(A) + P(B) - P(A \cap B)] \\
 &= 1 - P(A) - P(B) + P(A) \cdot P(B), \quad A, B \text{ are independent} \\
 &= 1 - P(A) - P(B)(1 - P(A)) \\
 &= [1 - P(A)][1 - P(B)] \\
 &= P(A^c) \cdot P(B^c)
 \end{aligned}$$

■

Alternatively,

**Proof :**

$$\begin{aligned}
 P(A^c \cap B^c) &= P(B^c) - P(A \cap B^c) \\
 &= P(B^c) - P(A) \cdot P(B^c), \quad \text{Example 4.8.11} \\
 &= P(B^c)[1 - P(A)] \\
 &= P(A^c) \cdot P(B^c)
 \end{aligned}$$

■

**Example 4.8.13** A fair coin is tossed twice. Let  $A$  denote the event that a head is obtained on the first toss,  $B$  the event that a head is obtained at the second toss and  $C$  the event that exactly one head is obtained. Discuss the independence of  $A, B, C$

Define events

$$\begin{aligned}S &= \{(H, H), (H, T), (T, H), (T, T)\} \\A &= \{(H, H), (H, T)\} \\B &= \{(H, H), (T, H)\} \\C &= \{(H, T), (T, H)\}\end{aligned}$$

Then

$$P(A) = \frac{1}{2}, P(B) = \frac{1}{2}, P(C) = \frac{1}{2}$$

Since  $A \cap B = \{H, H\}$  then

$$P(A \cap B) = \frac{1}{4} = P(A) \cdot P(B)$$

Hence  $A$  and  $B$  are independent.

Similarly

$$\begin{aligned}P(A \cap C) &= P(A) \cdot P(C) \\P(B \cap C) &= P(B) \cdot P(C)\end{aligned}$$

so  $A$  and  $C$  are independent and  $B$  and  $C$  are independent.

Hence  $A, B, C$  are pairwise independent.

**Note 4.8.3** We note that  $A \cap B \cap C = \phi$  thus

$$P(A \cap B \cap C) = 0 \neq P(A) \cdot P(B) \cdot P(C) = \frac{1}{8}$$

Hence  $A, B, C$  are not totally independent. Thus pairwise independence does not imply independence.

**Example 4.8.14** The probability that a student passes Mathematics is 0.98 and the probability that he passes English is 0.92. Find the probability that he passes both.

To pass one subject is *independent* of passing or failing another subject.

$$\begin{aligned}P(M \cap E) &= P(E | M) \cdot P(M) \\&= P(E) \cdot P(M) \\&= (0.92) \cdot (0.98) \\&= 0.9016\end{aligned}$$

**Example 4.8.15** Two events  $A$  and  $B$  are independent. If

$$P(A \cap B) = \frac{1}{3}, \quad P(A \cup B) = \frac{5}{6}$$

Find possible values of

1.)  $P(A)$

2.)  $P(B)$

Let  $P(A) = x$  and  $P(B) = y$ .

Then since  $A$  and  $B$  are independent  $\Rightarrow$

$$\begin{aligned} P(A \cap B) &= P(B | A) \cdot P(A) = P(B) \cdot P(A) = P(A) \cdot P(B) \\ \frac{1}{3} &= xy \end{aligned} \tag{4.24}$$

Also

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ \frac{5}{6} &= x + y - \frac{1}{3} \end{aligned} \tag{4.25}$$

Solving (4.24) and (4.25) simultaneously, either

$$\begin{aligned} P(A) = \frac{1}{2} \quad \text{and} \quad P(B) = \frac{2}{3} \quad \text{or} \\ P(A) = \frac{2}{3} \quad \text{and} \quad P(B) = \frac{1}{2} \end{aligned}$$

**Example 4.8.16** A die is thrown four times. Find the probability that a 5 is obtained each throw.

Define events (these events are also independent)

$$\begin{aligned} A \equiv 5 \text{ obtained at first throw} &\Rightarrow P(A) = \frac{1}{6} \Rightarrow P(A') = \frac{5}{6} \\ B \equiv 5 \text{ obtained at second throw} &\Rightarrow P(B) = \frac{1}{6} \Rightarrow P(B') = \frac{5}{6} \\ C \equiv 5 \text{ obtained at third throw} &\Rightarrow P(C) = \frac{1}{6} \Rightarrow P(C') = \frac{5}{6} \\ D \equiv 5 \text{ obtained at fourth throw} &\Rightarrow P(D) = \frac{1}{6} \Rightarrow P(D') = \frac{5}{6} \end{aligned}$$

Result from one throw is independent of the other, and to obtain a 5 on each throw

$$P(A \cap B \cap C \cap D) = P(D) \cdot P(C) \cdot P(B) \cdot P(A) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{1296}$$

**Exercise 4.10** A coin is tossed twice, find the probability of getting at least a head.

[3/4]

**Exercise 4.11** Professor Janet loves to give  $F$ 's. If the probability that Jones will make an  $F$  is  $\frac{2}{3}$  and the probability that Smith will make an  $F$  is  $\frac{1}{2}$ , what is the probability that at least one of them will make an  $F$ .

**Example 4.8.17** Three people  $A, B$  and  $C$  decided to enter a marathon race. Their respective probabilities that they will complete the marathon are 0.9, 0.7 and 0.6. Find the probability that

- |                                     |   |
|-------------------------------------|---|
| 1.) only $A$ completes the marathon | 4.) only two complete the marathon          |
| 2.) all complete the marathon       | 5.) at least two will complete the marathon |
| 3.) only one complete the marathon  | 6.) no one completes the marathon           |

Assume that the performances of each other are independent.

Define events

$A \equiv$  First person completes the marathon  
 $B \equiv$  Second person completes the marathon  
 $C \equiv$  Third person completes the marathon.

$$P(A) = 0.9 \Rightarrow P(A') = 0.1$$

$$P(B) = 0.7 \Rightarrow P(B') = 0.3$$

$$P(C) = 0.6 \Rightarrow P(C') = 0.4$$

- 1.) Only  $A$  completes the marathon

$$\begin{aligned} P(A \cap B' \cap C') &= P(C' | (A \cap B')) \cdot P(B' | A) \cdot P(A) \\ &= P(C') \cdot P(B') \cdot P(A), \text{ Independence} \\ &= (0.4) \cdot (0.3) \cdot (0.9) \\ &= 0.108 \end{aligned}$$

- 2.) All complete the marathon

$$\begin{aligned} P(A \cap B \cap C) &= P(C | (A \cap B)) \cdot P(B | A) \cdot P(A) \\ &= P(C) \cdot P(B) \cdot P(A), \text{ Independence} \\ &= (0.6) \cdot (0.7) \cdot (0.9) \\ &= 0.378 \end{aligned}$$

- 3.) Only one complete the marathon

$$\begin{aligned} \text{Only one} &\equiv A \cap B' \cap C' \text{ or } A' \cap B \cap C' \text{ or } A' \cap B' \cap C \\ P(\text{Only one}) &= P(A \cap B' \cap C') + P(A' \cap B \cap C') + P(A' \cap B' \cap C) \\ P(\text{Only one}) &= P(C' | (A \cap B')) \cdot P(B' | A) \cdot P(A) + P(C' | (A' \cap B)) \cdot P(B | A') \cdot P(A') \\ &\quad + P(C | (A' \cap B')) \cdot P(B' | A') \cdot P(A') \\ &= P(C') \cdot P(B') \cdot P(A) + P(C') \cdot P(B) \cdot P(A') \\ &\quad + P(C) \cdot P(B') \cdot P(A'), \text{ Independence} \\ &= (0.4) \cdot (0.3) \cdot (0.9) + (0.4) \cdot (0.7) \cdot (0.1) + (0.6) \cdot (0.3) \cdot (0.1) \\ &= 0.154 \end{aligned}$$

4.) Only two complete the marathon

$$\begin{aligned}
 \text{Only two} &\equiv A \cap B \cap C' \text{ or } A \cap B' \cap C \text{ or } A' \cap B \cap C \\
 P(\text{Only two}) &= P(A \cap B \cap C') + P(A \cap B' \cap C) + P(A' \cap B \cap C) \\
 &= P(C' \mid (A \cap B)) \cdot P(B \mid A) \cdot P(A) + P(C \mid (A \cap B')) \cdot P(B' \mid A) \cdot P(A) \\
 &\quad + P(C \mid (A' \cap B)) \cdot P(B \mid A') \cdot P(A') \\
 &= P(C') \cdot P(B) \cdot P(A) + P(C) \cdot P(B') \cdot P(A) \\
 &\quad + P(C) \cdot P(B) \cdot P(A'), \quad \text{Independency} \\
 &= (0.4) \cdot (0.7) \cdot (0.9) + (0.6) \cdot (0.3) \cdot (0.9) + (0.6) \cdot (0.7) \cdot (0.1) \\
 &= 0.456
 \end{aligned}$$

5.) At least two will complete the marathon

$$\begin{aligned}
 \text{At least two} &\equiv \text{Two or all three} \\
 P(\text{At least two}) &= P(\text{Exactly two}) + P(\text{All three}) \\
 P(\text{At least two}) &= 0.456 + 0.378 \\
 &= 0.834
 \end{aligned}$$

6.) No one completes the marathon

$$\begin{aligned}
 P(A' \cap B' \cap C') &= P(C' \mid (A' \cap B')) \cdot P(B' \mid A') \cdot P(A') \\
 &= P(C') \cdot P(B') \cdot P(A'), \quad \text{Independency} \\
 &= (0.4) \cdot (0.3) \cdot (0.1) = 0.012
 \end{aligned}$$

**Example 4.8.18** If  $A$  and  $B$  are independent events, with  $P(A) = \frac{1}{3}$  and  $P(B) = \frac{1}{4}$ , find the following:

1.)  $P(A^c \cap B^c)$ .

*Solution 1.* Since  $A$  and  $B$  are independent, then  $A^c$  and  $B^c$  are also independent [See Example 4.8.12, (pg. 182)]. So

$$P(A^c \cap B^c) = P(A^c) \cdot P(B^c) = [1 - P(A)][1 - P(B)] = \left[1 - \frac{1}{3}\right] \left[1 - \frac{1}{4}\right] = \frac{1}{2}$$

*Solution 2.*  $P(A \cap B) = P(A) \cdot P(B) = \frac{1}{12}$ .

$$P(A^c \cap B^c) = 1 - P(A \cup B) = 1 - (P(A) + P(B) - P(A \cap B)) = 1 - \left[\frac{1}{3} + \frac{1}{4} - \frac{1}{12}\right] = \frac{1}{2}$$

2.)  $P(A^c|B)$ .

Since  $A$  and  $B$  are independent,  $A^c$  and  $B$  are also independent. So

$$P(A^c|B) = P(A^c) = 1 - P(A) = \frac{2}{3}$$

**Note 4.8.4** If two events are independent, one does not depend on the occurrence of the other. But if Mutually exclusive, one does not influence the other.

**Example 4.8.19** In a race the probability that John wins is  $\frac{1}{3}$ , Paul wins is  $\frac{1}{4}$ , and Mark to win is  $\frac{1}{5}$ . Find the probability that

1.) John or Mark wins [8/15]

$$P(J \cup M) = P(J) + P(M) - P(J \cap M) = \frac{1}{3} + \frac{1}{5} - 0 = \frac{8}{15}$$

Since *winning*, its mutually exclusive, both cannot win, only one wins, its different from *completing* race.

2.) Neither John nor Paul wins [5/12]

$$1 - P(J \cup P) = 1 - [P(J) + P(P) - P(J \cap P)] = 1 - \left[ \frac{1}{3} + \frac{1}{4} - 0 \right] = \frac{5}{12}$$

Neither = 1- Either



## 4.9 Total Probability Theorem

Consider mutually exclusive and exhaustive events  $B_1, B_2, \dots, B_n$  that constitute a proportion of a sample space  $S$  such that

$$\Omega = B_1 \cup B_2 \cup \dots \cup B_n$$

Then the event  $A$  of  $S$  is seen to be the union of events.

$$\begin{aligned} A &= B_1 \cap A, B_2 \cap A, \dots, B_n \cap A \\ \Rightarrow A &= (B_1 \cap A) \cup (B_2 \cap A) \cup \dots \cup (B_n \cap A) \\ \Rightarrow P(A) &= P(B_1 \cap A) + P(B_2 \cap A) + \dots + P(B_n \cap A) \end{aligned} \quad (4.26)$$

Using definition of conditional probabilities then

$$P(A) = P(A | B_1) \cdot P(B_1) + P(A | B_2) \cdot P(B_2) + \dots + P(A | B_n) \cdot P(B_n) \quad (4.27)$$

Equation (4.26) or (4.27) is called the theorem of Total probability.

## 4.10 Bayes' Theorem

Consider events  $B_1, B_2, \dots, B_k, \dots, B_n$  that constitute a partition of a sample space  $S$  where  $P(B_k) \neq 0$ . For  $i = 1, 2, \dots, n$ , then for any event  $A$  in  $S$  such that  $P(A) \neq 0$

$$P(B_k | A) = \frac{P(A | B_k) \cdot P(B_k)}{\sum_{i=1}^n P(A | B_i) \cdot P(B_i)} \quad (4.28)$$

**Proof :** *By definition of conditional probability*

$$P(B_k | A) = \frac{P(A \cap B_k)}{P(A)} \quad (4.29)$$

*and since*

$$P(B_k \cap A) = P(A | B_k) \cdot P(B_k) \text{ by multiplicative rule, and} \quad (4.30)$$

$$P(A) = \sum_{i=1}^n P(A | B_i) \cdot P(B_i) \text{ by total probability} \quad (4.31)$$

*then equation (4.29) becomes*

$$P(B_i | A) = \frac{P(A | B_k) \cdot P(B_k)}{P(A | B_1) \cdot P(B_1) + P(A | B_2) \cdot P(B_2) + \dots + P(A | B_n) \cdot P(B_n)} \quad (4.32)$$

■

**Example 4.10.1** Using the Law of total Probability,

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B | A) \cdot P(A) + P(B | A') \cdot P(A')}$$

where:

$P(A)$  = probability that event A occurs

$P(B)$  = probability that event B occurs

$P(A')$  = probability that event A does not occur

$P(A | B)$  = probability that event A occurs given that event B has occurred already

$P(B | A)$  = probability that event B occurs given that event A has occurred already

$P(B | A')$  = probability that event B occurs given that event A has not occurred already

**Example 4.10.2** Three professors  $A, B, C$  are nominated for the post of Dean of Science. Their respective probabilities of being elected are 0.3, 0.5 and 0.2. If Professor  $A$  is elected the probability that the mathematics department gets a new computer is 0.8 and if  $B$  is elected is 0.1 and if  $C$  is elected is 0.3. The Head of Mathematics goes away to London and learns that the department has received a new computer.

1.) Find the probability that a Mathematics department gets a new computer,

2.) The probability that professors

a)  $A$

b)  $B$

c)  $C$

is elected if Mathematics department got new computer.

3.) if the department did not get a new computer, what is the probability that professor  $B$  was elected?

Define events

$A$  = professor A is elected

$B$  = Professor B is elected

$C$  = Professor C is elected

$N$  = Maths department gets a new computer.

Then

$$P(A) = 0.3 \Rightarrow P(A') = 0.7$$

$$P(B) = 0.5 \Rightarrow P(B') = 0.5$$

$$P(C) = 0.2 \Rightarrow P(C') = 0.8$$

$$P(N | A) = 0.8 \Rightarrow P(N' | A) = 0.2$$

$$P(N | B) = 0.1 \Rightarrow P(N' | B) = 0.9$$

$$P(N | C) = 0.3 \Rightarrow P(N' | C) = 0.7$$

1.) Then by Theorem of total probability

$$\begin{aligned}N &\equiv (A \cap N) \text{ or } (B \cap N) \text{ or } (C \cap N) \\P(N) &= P(A \cap N) + P(B \cap N) + P(C \cap N) \\P(N) &= P(N | A) \cdot P(A) + P(N | B) \cdot P(B) + P(N | C) \cdot P(C) \\&= (0.8)(0.3) + (0.1)(0.5) + (0.2)(0.3) \\&= 0.24 + 0.05 + 0.06 = 0.35\end{aligned}$$

2.) By Bayes' Theorem

$$\begin{aligned}P(A | N) &= \frac{P(N | A) \cdot P(A)}{P(N | A) \cdot P(A) + P(N | B) \cdot P(B) + P(N | C) \cdot P(C)} \\&= \frac{P(N | A) \cdot P(A)}{P(N)} \\&= \frac{(0.8)(0.3)}{0.35} = \frac{24}{35} \\P(B | N) &= \frac{P(N | B) \cdot P(B)}{P(N | A) \cdot P(A) + P(N | B) \cdot P(B) + P(N | C) \cdot P(C)} \\&= \frac{P(N | B) \cdot P(B)}{P(N)} \\&= \frac{(0.1)(0.5)}{0.35} = \frac{1}{7} \\P(C | N) &= \frac{P(N | C) \cdot P(C)}{P(N | A) \cdot P(A) + P(N | B) \cdot P(B) + P(N | C) \cdot P(C)} \\&= \frac{P(N | C) \cdot P(C)}{P(N)} = \frac{(0.3)(0.2)}{0.35} = \frac{6}{35}\end{aligned}$$

3.) We need

$$P(B | N')$$

$$\begin{aligned}P(B | N') &= \frac{P(N' | B) \cdot P(B)}{P(N' | A) \cdot P(A) + P(N' | B) \cdot P(B) + P(N' | C) \cdot P(C)} \\&= \frac{(0.9)(0.5)}{(0.2)(0.3) + (0.9)(0.5) + (0.7)(0.2)} \\&= 0.692\end{aligned}$$

**Exercise 4.12** For the problem in Example 4.10.3, find the probability that if he did not complete the journey in under three hours calculate the probability that he traveled by the first of the four routes.

**Example 4.10.3** A motorist travels regularly from one town to another. On each occasion he chooses a route at random from four possible routes. From his experience the probabilities of completing a journey under three hours via those routes are 0.5, 0.8, 0.9 and 0.9 respectively. Given that on a certain occasion he completed the journey in under three hours calculate the probability that he traveled by the first of the four routes.

Define events

- $A_1$  = he traveled by first route
- $A_2$  = he traveled by second route
- $A_3$  = he traveled by third route
- $A_4$  = he traveled by fourth route
- $E$  = he completed the journey in under three hours

$$\begin{aligned}P(A_1) &= 0.25 \Rightarrow P(A'_1) = 0.75 \\P(A_2) &= 0.25 \Rightarrow P(A'_2) = 0.75 \\P(A_3) &= 0.25 \Rightarrow P(A'_3) = 0.75 \\P(A_4) &= 0.25 \Rightarrow P(A'_4) = 0.75 \\P(E | A_1) &= 0.5 \Rightarrow P(E' | A_1) = 0.5 \\P(E | A_2) &= 0.8 \Rightarrow P(E' | A_2) = 0.2 \\P(E | A_3) &= 0.9 \Rightarrow P(E' | A_3) = 0.1 \\P(E | A_4) &= 0.9 \Rightarrow P(E' | A_4) = 0.1\end{aligned}$$

$$\begin{aligned}P(A_1 | E) &= \frac{P(E | A_1) \cdot P(A_1)}{P(E | A_1) \cdot P(A_1) + P(E | A_2) \cdot P(A_2) + P(E | A_3) \cdot P(A_3) + P(E | A_4) \cdot P(A_4)} \\&= \frac{(0.5)(0.25)}{(0.5)(0.25) + (0.8)(0.25) + (0.9)(0.25) + (0.9)(0.25)} = \frac{0.125}{0.775}\end{aligned}$$

**Example 4.10.4** There are five urns, and are numbered 1 to 5. Each urn contains 10 balls. Urn  $i$  has  $i$  defective balls and  $10 - i$  non defective balls,  $i = 1, 2, \dots, 5$ . For example urn 4 has 4 defective balls and 6 non defective balls. Consider the following experiment: First, an urn is selected at random and then a ball is selected at random from the selected urn. (The experimenter does not know the urn selected). Lets us ask two questions:

- 1.) What is the probability that a defective ball will be selected?
- 2.) If we have already selected a ball and noted it is defective what is the probability that it came from urn 5?

Define event

$$\begin{aligned} D &\equiv \text{defective ball is selected} \\ B_i &\equiv \text{urn } i \text{ is selected, } i = 1, 2, \dots, 5 \end{aligned}$$

then

$$P(B_i) = \frac{1}{5}, \quad i = 1, 2, \dots, 5$$

and also

$$\begin{aligned} P(D | B_1) &= \frac{1}{10} \Rightarrow P(D' | B_1) = \frac{9}{10} \\ P(D | B_2) &= \frac{2}{10} \Rightarrow P(D' | B_2) = \frac{8}{10} \\ P(D | B_3) &= \frac{3}{10} \Rightarrow P(D' | B_3) = \frac{7}{10} \\ P(D | B_4) &= \frac{4}{10} \Rightarrow P(D' | B_4) = \frac{6}{10} \\ P(D | B_5) &= \frac{5}{10} \Rightarrow P(D' | B_5) = \frac{5}{10} \end{aligned}$$

- 1.) From Theorem of total probability

$$\begin{aligned} P(D) &= P(B_1 \cap D) + P(B_2 \cap D) + P(B_3 \cap D) + P(B_4 \cap D) + P(B_5 \cap D) \\ &= P(D | B_1) \cdot P(B_1) + P(D | B_2) \cdot P(B_2) + P(D | B_3) \cdot P(B_3) \\ &\quad + P(D | B_4) \cdot P(B_4) + P(D | B_5) \cdot P(B_5) \\ &= \frac{1}{10} \cdot \frac{1}{5} + \frac{2}{10} \cdot \frac{1}{5} + \frac{3}{10} \cdot \frac{1}{5} + \frac{4}{10} \cdot \frac{1}{5} + \frac{5}{10} \cdot \frac{1}{5} \\ &= \frac{15}{50} = \frac{3}{10} \end{aligned}$$

- 2.) To find the probability that if a defective ball is selected, came from urn 5

$$\begin{aligned} &P(B_5 | D) \\ &= \frac{P(D | B_5) \cdot P(B_5)}{P(D | B_1) \cdot P(B_1) + P(D | B_2) \cdot P(B_2) + P(D | B_3) \cdot P(B_3) + P(D | B_4) \cdot P(B_4) + P(D | B_5) \cdot P(B_5)} \\ &= \frac{P(D | B_5) \cdot P(B_5)}{P(D)} \\ &= \frac{\frac{5}{10} \cdot \frac{1}{5}}{\frac{3}{10}} = \frac{1}{3} \end{aligned}$$

And generally

$$P(B_k | D) = \frac{\frac{k}{10} \cdot \frac{1}{5}}{\frac{3}{10}} = \frac{k}{15}, \quad k = 1, 2, \dots, 5.$$

**Example 4.10.5** A box contains 4 red and 6 black balls. If two balls are picked without replacement from the box, what is the probability that

1.) the first ball is black and the second is red?

$$\begin{aligned}\text{Black \& then Red} &\equiv B_1 \cap R_2 \\ P(\text{Black \& then Red}) &= P(B_1 \cap R_2) \\ &= P(R_2 | B_1) \cdot P(B_1) \\ &= \frac{4}{9} \cdot \frac{6}{10}\end{aligned}$$

2.) both balls are of same color

$$\begin{aligned}\text{Same color} &\equiv B_1 \cap B_2 \text{ or } R_1 \cap R_2 \\ P(\text{Same color}) &= P(B_1 \cap B_2) + P(R_1 \cap R_2) \\ &= P(B_2 | B_1) \cdot P(B_1) + P(R_2 | R_1) \cdot P(R_1) \\ &= \frac{5}{9} \cdot \frac{6}{10} + \frac{3}{9} \cdot \frac{4}{10}\end{aligned}$$

3.) the two balls are of different colors

$$\begin{aligned}\text{Different colors} &\equiv B_1 \cap R_2 \text{ or } R_1 \cap B_2 \\ P(\text{Different colors}) &= P(B_1 \cap R_2) + P(R_1 \cap B_2) \\ &= P(R_2 | B_1) \cdot P(B_1) + P(B_2 | R_1) \cdot P(R_1) \\ &= \frac{4}{9} \cdot \frac{6}{10} + \frac{6}{9} \cdot \frac{4}{10}\end{aligned}$$

**Example 4.10.6** Repeat the problem above in Example 4.10.5 with replacement.

1.)

$$\begin{aligned}\text{Black \& then Red} &\equiv B_1 \cap R_2 \\ P(\text{Black \& then Red}) &= P(B_1 \cap R_2) \\ &= P(R_2 | B_1) \cdot P(B_1) \\ &= \frac{4}{10} \cdot \frac{6}{10}\end{aligned}$$

2.)

$$\begin{aligned}\text{Same color} &\equiv B_1 \cap B_2 \text{ or } R_1 \cap R_2 \\ P(\text{Same color}) &= P(B_1 \cap B_2) + P(R_1 \cap R_2) \\ &= P(B_2 | B_1) \cdot P(B_1) + P(R_2 | R_1) \cdot P(R_1) \\ &= \frac{6}{10} \cdot \frac{6}{10} + \frac{4}{10} \cdot \frac{4}{10}\end{aligned}$$

3.) Different colors

$$\frac{4}{10} \cdot \frac{6}{10} + \frac{6}{10} \cdot \frac{4}{10}$$

**Example 4.10.7** Two boxes have balls. The first box (Box  $A$ ) contains 9 yellow and 7 purple balls. The second box (Box  $B$ ) contains 8 yellow and 10 purple balls. A ball is picked from box  $A$  and placed in box  $B$  without noticing its color. Then after thorough shaking of box  $B$ , a ball is picked from box  $B$ . What is the probability

1.) that the ball now picked from box  $B$  is purple?

$$\begin{aligned} P_2 &= Y_1 \cap P_2 \text{ or } P_1 \cap P_2 \\ P(P_2) &= P(Y_1 \cap P_2) + P(P_1 \cap P_2) \\ &= P(P_2 | Y_1) \cdot P(Y_1) + P(P_2 | P_1) \cdot P(P_1) \\ &= \frac{10}{19} \cdot \frac{9}{16} + \frac{11}{19} \cdot \frac{7}{16} \end{aligned}$$

2.) the second ball is yellow given that the ball brought was yellow

$$P(Y_2 | Y_1) = \frac{9}{19}$$

3.) the first ball picked is purple

$$P(P_1) = \frac{7}{16}$$

4.) the two balls are of the same colors

$$\begin{aligned} \text{Same color} &\equiv Y_1 \cap Y_2 \text{ or } P_1 \cap P_2 \\ &= P(Y_1 \cap Y_2) + P(P_1 \cap P_2) \\ &= P(Y_2 | Y_1) \cdot P(Y_1) + P(P_2 | P_1) \cdot P(P_1) \\ &= \frac{9}{19} \cdot \frac{9}{16} + \frac{11}{19} \cdot \frac{7}{16} \end{aligned}$$

**Example 4.10.8** Suppose that  $A$  and  $B$  are independent events such that the probability that neither occurs is  $\frac{1}{2}$  and the probability of  $B$  occurring is  $\frac{1}{3}$ . Determine the probability of  $A$  occurring.

$$P((A \cup B)^c) = 1 - P(A \cup B) = 1/2$$

To have

$$P(A \cup B) = 1/2$$

Since  $A$  and  $B$  are independent

$$P(A \cap B) = P(A)P(B) = P(A)/3$$

By substituting terms into  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ , we have

$$\frac{1}{2} = P(A) + \frac{1}{3} - \frac{1}{3}P(A)$$

Hence  $P(A) = 1/4$ .

## 4.11 Chapter Examples

**Example 4.11.1** Suppose that  $A$  and  $B$  are mutually exclusive events for which

$$P(A) = 0.3, \quad P(B) = 0.5$$

What is the probability

- |                               |     |
|-------------------------------|-----|
| 1.) either $A$ or $B$ occurs? | 0.8 |
| 2.) $A$ occurs but not $B$ ?  | 0.3 |
| 3.) both $A$ and $B$ occurs?  | 0   |

**Example 4.11.2** An elementary school is offering 3 language classes; one in Spanish, one in French, and one in German. These classes are open to any of the 100 students in the school. There are 28 students in the Spanish class, 26 in the French class, and 16 in the German class. There are 12 students that are in both Spanish and French, 4 in both Spanish and German, 6 in both French and German, and 2 taking all three.

- 1.) If a student is chosen randomly, what is the probability that he or she is not in any of these classes?

$$P(S \cup F \cup G)' = 1 - P(S \cup F \cup G) = 0.5$$

- 2.) If a student is chosen randomly, what is the probability that he or she is taking exactly one language course?

$$\frac{32}{100} = 0.32$$

- 3.) If 2 students are chosen randomly, what is the probability that at least 1 is taking a language class?

$$\frac{149}{198}$$

**Example 4.11.3** Prove

$$P(A \cap B) \geq P(A) + P(B) - 1$$

when  $A, B$  are two events in the same sample space.

We have shown in the previous sections

$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \text{ and} \quad (4.33)$$

$$P(A \cup B) \leq 1 \quad (4.34)$$

So by rearranging the terms,  $P(A \cap B) \geq P(A) + P(B) - 1$ . This inequality is called *Bonferroni inequality*.

**Example 4.11.4** Show  $P(A_1 \cup A_2 \cup A_3) \leq P(A_1) + P(A_2) + P(A_3)$ .

$P(A_1 \cup B) = P(A_1) + P(B) - P(A_1 \cap B) \leq P(A_1) + P(B)$ . Now let  $B = A_2 \cup A_3$ . Then similarly  $P(B) \leq P(A_2) + P(A_3)$ . So

$$P(A_1 \cup A_2 \cup A_3) \leq P(A_1) + P(B) \leq P(A_1) + P(A_2) + P(A_3)$$

**Example 4.11.5** Find the sample space for the gender of the children if a family has three children. Use  $B$  for boy and  $G$  for girl.

$$\{BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG\}$$



**Example 4.11.6** If two dice are rolled one time, find the probability of getting a sum of 7 or 11.

$$\frac{6 + 2}{36} = \frac{2}{9}$$

**Example 4.11.7** In a sample of 50 people, 21 had type *O* blood, 22 had type *A* blood, 5 had type *B* blood, and 2 had type *AB* blood. Find the probability that a person has neither type *A* nor type *O* blood.

$$\frac{5}{50} + \frac{2}{50} = \frac{7}{50}$$

**Example 4.11.8** Determine which events are mutually exclusive and which are not, when a single die is rolled.

- 1.) Getting an odd number and getting an even number

Mutually Exclusive

- 2.) Getting a 3 and getting an odd number

Not Mutually Exclusive

- 3.) Getting an odd number and getting a number less than 4

Not Mutually Exclusive

- 4.) Getting a number greater than 4 and getting a number less than 4

Mutually Exclusive

**Example 4.11.9** At a political rally, there are 20 NRM supporters, 13 Democrats, and 6 Independents. If a person is selected at random, find the probability that he or she is either a Democrat or an NRM.

$$\text{Mutually Exclusive, } \frac{13}{39} + \frac{20}{39} - 0 = \frac{33}{39}$$

**Example 4.11.10** In a hospital unit there are 8 nurses and 5 physicians; 7 nurses and 3 physicians are females. If a staff person is selected, find the probability that the subject is a nurse or a male.

$$\begin{aligned} P(\text{Nurse or Male}) &= P(\text{Nurse}) + P(\text{Male}) - P(\text{Male Nurse}) \\ &= \frac{8}{13} + \frac{3}{13} - \frac{1}{13} = \frac{10}{13} \end{aligned}$$

**Example 4.11.11** When a die is thrown, an odd number occurs, what is the probability that a number is prime.  $\left[ \frac{1}{3} \right]$

**Example 4.11.12** A coin is flipped and a die is rolled. Find the probability of getting a head on the coin and a 4 on the die.

$$\begin{aligned} P(\text{Head and 4}) &= P(4 \mid \text{Head}) \cdot P(\text{Head}) \\ &= P(4) \cdot P(\text{Head}), \text{ Independent events} \\ &= \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12} \end{aligned}$$

Results by tossing a tossing a coin and a die are independent.

**Example 4.11.13** A Kiwa poll found that 46% of Makerere students say they suffer great stress at least once a week. If three people are selected at random, find the probability that all three will say that they suffer great stress at least once a week.

Independent Events

$$\begin{aligned}P(S \cap S \cap S) &= P(S) \cdot P(S) \cdot P(S) \\&= (0.46)(0.46)(0.46) \\&= 0.097\end{aligned}$$

**Example 4.11.14** The probability that Sam parks in a no-parking zone and gets a parking ticket is 0.06, and the probability that Sam cannot find a legal parking space and has to park in the no-parking zone is 0.20. On Tuesday, Sam arrives at school and has to park in a no-parking zone. Find the probability that he will get a parking ticket.

$N$  = parking in a no-parking zone,  $T$  = getting a ticket

$$\begin{aligned}P(T | N) &= \frac{P(N \text{ and } T)}{P(N)} \\&= \frac{0.06}{0.20} = 0.30\end{aligned}$$

**Example 4.11.15** A recent survey asked 100 people if they thought women in the armed forces should be permitted to participate in combat. The results of the survey are shown.

Gender	Yes	No	Total
Male	32	18	50
Female	8	42	50
<b>Total</b>	<b>40</b>	<b>60</b>	<b>100</b>

- 1.) Find the probability that the respondent answered yes (Y), given that the respondent was a female (F).

$$P(Y | F) = \frac{P(F \text{ and } Y)}{P(F)} = \frac{P(F \cap Y)}{P(F)} = \frac{8/100}{50/100} = \frac{8}{50} = \frac{4}{25}$$

- 2.) Find the probability that the respondent was a male (M), given that the respondent answered no (N)

$$P(M | N) = \frac{P(N \cap M)}{P(N)} = \frac{18/100}{60/100} = \frac{18}{60} = \frac{3}{10}$$

**Example 4.11.16** The Neckware Association of Masaka reported that 3% of ties sold in Buddu are bow ties ( $B$ ). If 4 customers who purchased a tie are randomly selected, find the probability that at least 1 purchased a bow tie.

$$P(B) = 0.03 \Rightarrow P(B') = 0.97$$

$$\begin{aligned}P(\text{no bow ties}) &= P(B' \cap B' \cap B' \cap B') \\&= P(B') \cdot P(B') \cdot P(B') \cdot P(B') \text{ Independent Events} \\&= (0.97) \cdot (0.97) \cdot (0.97) \cdot (0.97) = 0.885\end{aligned}$$

$$P(\text{at least 1 bow tie}) = 1 - P(\text{no bow ties}) = 1 - 0.885 = 0.115$$

**Example 4.11.17** A bucket contains 3 black balls and 7 green balls. We draw a ball from the bucket, replace it, and draw a second ball.

- 1.) A black ball drawn on first draw

$$P(B) = 0.30$$

- 2.) Two green balls drawn

$$P(G_1 \cap G_2) = P(G_2 | G_1) \cdot P(G_1) = (0.7)(0.7) = .49$$

- 3.) A black ball drawn on second draw if the first draw is green

$$P(B_2 | G_1) = P(B_2) = 0.30$$

a conditional probability but equal to the marginal because the two draws are independent events since with replacement.

- 4.) A green ball is drawn on the second if the first draw was green

$$P(G_2 | G_1) = P(G_2) = 0.70$$

a conditional probability but equal to the marginal because the two draws are independent events since with replacement.

**Example 4.11.18** If we have  $E = \{(H, T)\}$  and  $F = \{(T, H)\}$  then  $E \cup F = \{(H, T), (T, H)\}$  is the event that one coin is head and the other is tail.

**Example 4.11.19** (horse race) If we have

$$\begin{aligned} E &= \{\text{all outcomes in } S \text{ starting with a 7}\} \quad \text{and} \\ F &= \{\text{all outcomes in } S \text{ finishing with a 3}\} \end{aligned}$$

then  $E \cup F$  is the event that the race was won by horse 7 or/and the last horse was horse 3.

**Example 4.11.20** If  $E = \{x : 0 \leq x \leq 5\}$  and  $F = \{x : 10 \leq x < \infty\}$  then  $E \cup F$  is the event that your pet will die before 5 or will die after 10.

**Example 4.11.21** (coins): If we have  $E = \{(H, H), (H, T), (T, H)\}$  (event that one H at least occurs) and  $F = \{(H, T), (T, H), (T, T)\}$  (event that one T at least occurs) then  $E \cap F = \{(H, T), (T, H)\}$  is the event that one H and one T occur.

**Example 4.11.22** (horse race): If we have  $E = \{\text{all outcomes in } S \text{ starting with a 7}\}$  and  $F = \{\text{all outcomes in } S \text{ starting with a 8}\}$  then  $E \cap F$  does not contain any outcome and is denoted by  $\emptyset$ .

**Example 4.11.23** (lifetime): If we have  $E = \{x : 0 \leq x \leq 5\}$  and  $F = \{x : 3 \leq x < 7\}$  then  $E \cap F = \{x : 3 \leq x \leq 5\}$  is the event that your pet will die between 3 and 5.

**Example 4.11.24** Suppose we've tossed a fair die, and we know only that an even number has come up. We find the probability that a two has come up using conditional probability:

$$\mathbb{P}(\{2\} | \{2, 4, 6\}) = \frac{\mathbb{P}(\{2\}, \{2, 4, 6\})}{\mathbb{P}(\{2, 4, 6\})} = \frac{\mathbb{P}(\{2\} \cap \{2, 4, 6\})}{\mathbb{P}(\{2, 4, 6\})} = \frac{\mathbb{P}(\{2\})}{\mathbb{P}(\{2, 4, 6\})} = \frac{1/6}{1/2} = 1/3.$$

**Example 4.11.25** A die is rolled and we assume

$$P(\{1\}) = P(\{2\}) = \cdots = P(\{6\}) = 1/6$$

Hence as a consequence from Axiom 3, the probability of having an even or odd number is equal to

$$\begin{aligned}P(\{1, 3, 5\}) &= P(\{1\}) + P(\{3\}) + P(\{5\}) = 1/2, \\P(\{2, 4, 6\}) &= P(\{2\}) + P(\{4\}) + P(\{6\}) = 1/2.\end{aligned}$$

**Example 4.11.26** If  $E \subset F$  then  $P(E) \leq P(F)$ .

We have  $F = E \cup (E^c \cap F)$  and  $E \cap (E^c \cap F) = \emptyset$  so

$$P(F) = P(E) + \underbrace{P(E^c \cap F)}_{\geq 0 \text{ by Axiom 1}} \geq P(E).$$

**Example 4.11.27** You are in a restaurant and ordered 2 dishes. With probability 0.6, you will like the first dish; with probability 0.4, you will like the second dish. With probability 0.3, you will like both of them. What is the probability that you like neither dish?

Let  $A_i$  the event: "You like dish  $i$ ". Then the probability you like at least one is

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) = 0.6 + 0.4 - 0.3 = 0.7.$$

The event that you like neither dish is the complement of liking at least one, so

$$\begin{aligned}P(\text{"you will like neither dish"}) &= P(A_1 \cup A_2)^c \\&= 1 - P(A_1 \cup A_2) \\&= 0.3\end{aligned}$$

**Example 4.11.28** A dice is thrown twice and the number on each throw is recorded. Assuming the dice is fair, what is the probability of obtaining at least one 6?

There are clearly 6 possible outcomes for the first throw and 6 for the second throw. By the counting principle, there are 36 possible outcomes for the two throws. Let  $A_i$  the event "I have obtained a 6 for throw  $i$ ". The probability we are interested in is

$$\begin{aligned}P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\&= \frac{1}{6} + \frac{1}{6} - \frac{1}{36} \\&= \frac{11}{36}\end{aligned}$$

**Example 4.11.29** A day of the week is selected at random. Find the probability that it is a weekend day.

$$P(\text{Saturday or Sunday}) = P(\text{Saturday}) + P(\text{Sunday}) = \frac{1}{7} + \frac{1}{7} = \frac{2}{7}$$

**Example 4.11.30** An urn contains 3 red balls , 2 blue balls and 5 white balls. A ball is selected and its color noted. Then it is replaced. A second ball is selected and its color noted . Find the probability of each of these.

- 1.) Selecting 2 blue balls 1/25
- 2.) Selecting 1 blue ball and then 1 white ball 1/10
- 3.) Selecting 1 red ball and then 1 blue ball 3/50

**Example 4.11.31** If two dice are rolled one time, find the probability of getting these results.

- 1.) A sum of 6
- 2.) Doubles
- 3.) A sum of 7 or 11
- 4.) A sum greater than 9
- 5.) A sum less than or equal to 4

**Example 4.11.32** If

$$P(A) = \frac{1}{3}, P(B) = \frac{1}{4}, P(A | B) = \frac{2}{5}$$

Compute the following probabilities

- (i)  $P(B | A)$  [3/10]                      (ii)  $P(A \cap B)$  [1/10]

**Example 4.11.33** A bag contains 10 balls, of which 7 are green, 3 are black. A ball is picked at random from the bag and its color noted. The ball is not replaced. A second ball is then picked out. Find the probability that

- 1.) the balls are of different colors [7/15]
- 2.) the first is black [3/10]
- 3.) the second is black [ $G_1B_2$  or  $B_1B_2$ ]
- 4.) both balls are of the same color [ $G_1G_2$  or  $B_1B_2$ ]

**Example 4.11.34** At a local university 54.3% of incoming first-year students have computers. If three students are selected at random, find the probability that at least one has a computer.

$$P(\text{at least one}) = 1 - P(\text{none of the three}) = 1 - 0.0954 = 0.9046$$

**Example 4.11.35** In Kadomola Housing Plan, 42% of the houses have a deck and a garage; 60% have a deck. Find the probability that a home has a garage, given that it has a deck.

$$P(\text{garage} | \text{deck}) = \frac{0.42}{0.60} = 0.70$$

**Example 4.11.36** A die is thrown twice. Find the probability of obtaining a 4 on the first throw and an odd number on the second throw. [ $\frac{1}{12}$ ]

**Example 4.11.37** A fair die is thrown twice, find the probability that

- 1.) neither throw results in a 4 [25/36]
- 2.) at least one throw results in a 4 [11/36]

**Example 4.11.38** If events  $A$  and  $B$  are independent such that

$$P(A) = \frac{1}{3}, P(A \cap B) = \frac{1}{12}$$

Find

1.)  $P(B)$   $[1/4]$

2.)  $P(A \cup B)$   $[1/2]$

**Example 4.11.39** Two events  $A$  and  $B$  are such that  $P(A) = \frac{1}{4}$ ,  $P(A | B) = \frac{1}{2}$  and  $P(B | A) = \frac{2}{3}$ . Find

1.)  $P(A \cap B)$   $[1/6]$

2.)  $P(B)$   $[1/3]$

**Example 4.11.40** A bag contains 10 balls, of which 4 are white, 6 are red. Two balls are picked without replacement, Find the probability that

1.) both balls are of the same color  $[42/90]$

2.) the balls are of different color  $[24/45]$

3.) the second is red  $[54/90]$

**Example 4.11.41** In a shooting competition 3 men  $A$ ,  $B$ , and  $C$  participate. Their respective chances of hitting the target are  $\frac{1}{3}$ ,  $\frac{1}{4}$  and  $\frac{1}{2}$  respectively. Find the probability that the target will be hit by

1.) only one bullet if all the three fire.  $[11/24]$

2.) only two hitting.  $[1/4]$

3.) at least by one of the candidates.

**Example 4.11.42** A box contains 3 Black and 5 red beads. If three beads are picked without replacement, Find the probability that

1.) no Black bead will be selected  $[5/28]$

2.) exactly one Red bead would be selected  $[15/56]$

3.) at least one Red bead should be selected  $[55/56]$

**Example 4.11.43** If

$$P(A) = \frac{1}{3}, P(B | A) = \frac{1}{4}, P(B' | A') = \frac{4}{5}$$

Find

1.)  $P(A \cap B)$   $[1/12]$

3.)  $P(B)$   $[13/60]$

2.)  $P(B' | A)$   $[3/4]$

4.)  $P(A \cup B)$   $[7/15]$

**Example 4.11.44** Two bags  $A$  and  $B$ . Bag  $A$  contains 8 White and 4 Black balls, while Bag  $B$  contains 5 White and 10 Black balls. A ball is picked from bag  $A$  and transferred into bag  $B$ . A ball is the picked from bag  $B$ . Find the probability that its white?  $[17/48]$

**Example 4.11.45** Two boxes  $A$  and  $B$ . Box  $A$  contains 2 White and 8 Black balls, while Box  $B$  contains 7 White and 3 Black balls. One box is chosen at random, and a ball is picked from the chosen box, what is the probability of getting a black ball? [11/20]

**Example 4.11.46** If  $A$  and  $B$  are independent events and  $P(A) = \frac{2}{5}$ ,  $P(A \cup B) = \frac{4}{5}$ . Find

- 1.)  $P(B)$  [2/3]                      2.)  $P(A' \cup B')$  [11/15]

**Example 4.11.47** Two boxes  $A$  and  $B$ . Box  $A$  contains 6 Green and 4 Red balls, while Box  $B$  contains 2 Green and 7 Red balls. One box is chosen at random. If the ball is red, what is the probability that it comes from box  $A$ . [18/53]

**Example 4.11.48** PitaPata travels to work by a bicycle or a bus. The probability that he travels by a bicycle is  $\frac{1}{4}$ . The probability that he is late for work if he travels by a bicycle is  $\frac{2}{3}$ , and if by bus is  $\frac{1}{3}$ . Find

- 1.) the probability that he is late [5/12]
- 2.) if he is late, what is the probability that he traveled by bus [3/5]

**Example 4.11.49** The probability that Amos will be safe is  $\frac{5}{8}$  and the probability that Kayanja will be alive is  $\frac{5}{6}$ . Determine the probability that

- 1.) both are alive     $[25/48]$                       2.) at least one of them is alive     $[15/18]$

**Example 4.11.50** In a committee of 5 men and 7 women, determine the probability that when 2 people are selected,

- 1.) both are women  $[7/22]$                       2.) will have one man and one woman  $[35/66]$

**Example 4.11.51** In a country, 60% of the cars are privately owned. Of them 70% are small. Of those not privately owned, 40% are small. Find the probability that a car chosen at random,

- 1.) is small [0.58]                  2.) its privately owned given its large [3/7]

**Example 4.11.52** In a shooting contest 3 marksmen  $A$ ,  $B$ , and  $C$  participate. Their respective chances of hitting the target are  $\frac{1}{3}$ ,  $\frac{1}{7}$  and  $\frac{1}{9}$ . Find the probability that the target

- |                             |          |
|-----------------------------|----------|
| 1.) will be hit by one man. | [76/189] |
| 2.) will be hit.            | [31/63]  |
| 3.) all hit.                | [1/189]  |

**Example 4.11.53** Two girls  $A$  and  $B$  take part in a competition in which each of them throws a ball at a target. The probability that  $A$  will hit a target on any throw is  $\frac{1}{3}$  and for  $B$  is  $\frac{1}{4}$ . The competition is terminated when any one of the girls hits the target and therefore becomes the winner. Suppose that  $B$  throws first and take two turns throwing, determine the probability that  $A$  will win on the third throw?  $\left[\frac{1}{16}\right] B' \cdot A' \cdot B' \cdot A' \cdot B' \cdot A$

**Example 4.11.54** A box contains 10 defective and 30 non-defective mangos. If 3 mangos are picked, find the probability that  $\frac{2}{3}$  (two out of the three picked) are defective,

1.) picked with replacement. [9/64]

2.) without replacement. [135/988]

**Exercise 4.13** The probability that a student passes his Maths exam is 0.7. Given that he does not pass his Math exam, the probability he joins the University is 0.2. Find the probability that he passes his Math exam and joins the University, if the probability of joining the University is 0.8. [0.74]

**Example 4.11.55** A box  $P$  contains 3 Red and 5 Black balls, while another box  $Q$  contains 6 Red and 4 Black balls. A box is chosen at random and from it a ball is picked at random and put it into another box. A ball is then randomly drawn from the later box, Find the probability that

1.) the both balls are Red [0.2527]

2.) the first ball drawn is Black [0.5125]

**Exercise 4.14** A box  $A$  contains 8 White and 4 Blue balls, while another box  $B$  contains 5 White and 10 Blue balls. A ball is randomly chosen from box  $A$  and placed into box  $B$ . After mixing the balls, a ball is now randomly selected from box  $B$  and placed in box  $A$ . If now a ball is picked from box  $A$ , find the probability that

1.) its White [41/64]

2.) there are 5 White balls in basket  $B$  at the end? [92/192]

**Example 4.11.56** Its known that in Nnyendo city, the probability of selecting a person with cancer is 0.02. If the probability of a doctor correctly diagnosing a person with cancer as having the disease is 0.78, and the probability of incorrectly diagnosing a person with cancer is 0.06. What is the probability that a person is diagnosed as having cancer?

[0.0744]

**Exercise 4.15** Three coins are tossed. Two of these are fair and one is biased so that a head is three times as likely to occur as a tail. Use a tree diagram to list down all possible outcomes and hence obtain the probability of obtaining two heads and a tail.

[7/16]

**Exercise 4.16** A Jar  $A$  contains 5 White, 8 Yellow and 10 Black beads, while another Jar  $B$  contains 6 White, 13 Yellow and 7 Black beads. A Jar is chosen at random and from it two beads are randomly drawn without replacement, find the probability that the beads are of

1.) same colors [0.3394]

2.) different colors [0.6606]

**Example 4.11.57** An A.D.A selected at random a village from among villages  $A$ ,  $B$ , and  $C$  for a visit. The probability that it rains in  $A$  is  $\frac{1}{3}$ , in  $B$  is  $\frac{1}{4}$  and in  $C$  is  $\frac{1}{4}$ . He came back with mud on his car, what is the probability that he visited village  $C$ . [3/10]



**Example 4.11.58** A bag  $A$  contains 5 White and 7 Yellow marbles, while another bag  $B$  contains 6 White and 8 Yellow marbles. One marble is taken from the second bag and placed into the first bag. After thorough mixing, one marble is taken from the first bag and placed into the second. What is the probability that there are 6 White marbles in the second bag. [50/91]

**Example 4.11.59** Three candidates have been nominated for the post of head teacher in a certain school. The probability of candidate  $A$  will be selected is 0.1, the probability that candidate  $B$  will be selected is 0.2 and probability of candidate  $C$  will be elected is 0.3. Its expected that school fees will increase if any one of these is elected as head teacher. The probability of increase of school fees if  $A$  is elected is 0.5, the corresponding probabilities for  $B$  and  $C$  are 0.6 and 0.4 respectively.

- 1.) Find the probability that there will be no increase in school fees. [0.71]
- 2.) Given that there was increase in school fees, calculate the probability that candidate  $A$  was elected [5/29]

**Exercise 4.17** Two friends Whitney and Mariah often go to hair dressers on Friday afternoons. If Mariah goes, the probability that Whitney goes is 0.96. If Whitney goes, the probability that Mariah goes is 0.8. The probability that neither goes is 0.07. Find the probability that both go to hair dresser? [0.72]

**Exercise 4.18** Events  $A$  and  $B$  are such that

$$P(A) = \frac{1}{2}, \quad P(B) = \frac{1}{3}, \quad P(A \cap B) = \frac{1}{4}$$

Compute

- 1.)  $P(A | B)$
- 2.)  $P(B | A)$
- 3.)  $P(A^c | B)$
- 4.)  $P(A^c | B^c)$

**Example 4.11.60** A bag contains 7 black and 3 white balls. If they are drawn one by one from the bag, find the probability of drawing first a black ball then a white ball and so on alternating until only black balls remain. [1/120]

**Example 4.11.61** The probability that Mirumbe passes Maths is  $\frac{2}{3}$  and that he passes Physics is  $\frac{4}{9}$ . And the probability that he passes at least one subject is  $\frac{4}{5}$ . Find Mirumbe's probability of passing both papers. [14/45]

**Exercise 4.19** Two balls were drawn at random one after the other without replacement from a box which contained ten balls, of which six are black. Calculate the conditional probability that the second ball was also black. [Ans 5/9]

**Example 4.11.62** Three people throw a coin. A game becomes successful when one gets a result different from the others.

- 1.) Find the probability of a success for the first throw. [6/8 = 3/4]
- 2.) Find the probability that at least one success will occur in two throws.  
Hint: Apply Binomial [15/16 = 0.94]



**Exercise 4.27** If  $A$  and  $B$  are mutually exclusive events such that

$$P(A) = 0.5, P(B) = 0.7, \text{ and } P(A \cup B) = 0.8,$$

find

$$1.) P(A^c \cap B^c) \qquad 2.) P(A \cap B^c)$$

**Exercise 4.28** A bag contains 3 black and 5 white balls. Two balls are drawn at random one at a time without replacement. Find

$$1.) \text{ the probability that the second ball is white} \qquad [5/8]$$

$$2.) \text{ the probability that the first ball is white given that the second ball is white.} \qquad [4/7]$$

**Exercise 4.29** If  $A$  and  $B$  are independent events prove or disprove that

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

Also prove that if  $A$  and  $B$  are independent,  $P(B \mid A^c) = P(B \mid A)$ .

**Exercise 4.30** Given that  $P(A) = 0.5$  and  $P(A \cup B) = 0.6$  find  $P(B)$  if

1.)  $A$  and  $B$  are mutually exclusive

2.)  $A$  and  $B$  are independent

$$3.) P(A \mid B) = 0.4$$

**Exercise 4.31** Prove that if events  $A_1, A_2, \dots, A_n$  are independent then

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = 1 - [1 - P(A_1)][1 - P(A_2)] \times \dots \times [(1 - P(A_n))]$$

**Exercise 4.32** Let each of the mutually disjoint sets  $E_1, E_2, \dots, E_n$  have nonzero probabilities. If the set  $B_i$  subset of the union of  $E_i, i = 1, 2, \dots, n$  show that

$$P(B) = P(B \mid E_1) \cdot P(E_1) + P(B \mid E_2) \cdot P(E_2) \dots + P(B \mid E_n) \cdot P(E_n)$$

If  $P(B) > 0$  prove Bayes' formula

$$P(E_i \mid B) = \frac{P(B \mid E_i) \cdot P(E_i)}{\sum_{i=1}^n P(B \mid E_i) \cdot P(E_i)}$$

**Exercise 4.33** Three teachers have been nominated for the office of Headmaster. the probability that Mr. Apuli will be elected is 0.3, the probability that Mr. Babi will be elected is 0.5 and the probability that Mr. Courts will be elected is 0.2. Should Mr. Apuli be elected the probability that there will be an increase in school fees is 0.8 and for Mr Babi and Mr. Courts are 0.3 and 0.1 respectively.

1.) What is the probability that there will be an increase in School fees?

2.) If School fees increased, what is the probability that Mr. Babi was elected?

**Exercise 4.34** The probability that a house girl is in the kitchen when a visitor arrives is 0.6. Given that she is in the kitchen the probability that she breaks a plate is 0.4. Find the probability that the house girl is in the kitchen and breaks a plate when a visitor arrives.

**Exercise 4.35** That the probability that Cheeye will be alive in 20 years time is 0.7 and the probability that Alice will be alive in 20 years is 0.9. What is the probability that neither will be alive in 20 years [Ans:0.03]

**Example 4.11.63** Toss two fair coins. There are four equally probable outcomes:

$$HH, HT, TH, TT$$

Let  $X$  equal 1 if first coin is heads, 0 if first coin is tails. Let  $Y$  equal 1 if second coin is heads, 0 if second coin is tails. Then  $X$  and  $Y$  are independent because, for example,

$$\text{Prob}(X = 1 \text{ and } Y = 0) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \text{Prob}(X = 1) \cdot \text{Prob}(Y = 0),$$

and similarly, for all other possible values,

$$\text{Prob}(X = x_i \text{ and } Y = y_j) = \text{Prob}(X = x_i) \cdot \text{Prob}(Y = y_j)$$

In contrast, if we define  $Y$  to be 0 if outcome is  $TT$  and 1 otherwise, then  $X$  and  $Y$  are not independent because  $\text{Prob}(X = 1 \text{ and } Y = 0) = 0$ , yet  $\text{Prob}(X = 1) = 1/2$  and  $\text{Prob}(Y = 0) = 1/4$ .

**Exercise 4.36** Give a possible sample space  $\Omega$  for each of the following experiments:

- 1.) An election decides between two candidates A and B.
- 2.) A two-sided coin is tossed.
- 3.) A student is asked for the month of the year and the day of the week on which her birthday falls.
- 4.) A student is chosen at random from a class of ten students.

**Exercise 4.37** Describe in words the events specified by the following subsets of

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

- 1.)  $E = \{HHH, HHT, HTH, HTT\}$ .
- 2.)  $E = \{HHH, TTT\}$ .
- 3.)  $E = \{HHT, HTH, THH\}$ .
- 4.)  $E = \{HHT, HTH, HTT, THH, THT, TTH, TTT\}$ .

**Exercise 4.38** A die is loaded in such a way that the probability of each face turning up is proportional to the number of dots on that face. (For example, a six is three times as probable as a two.) What is the probability of getting an even number in one throw?

**Example 4.11.64** The tree diagram when a coin is tossed three times (three coins tossed once)

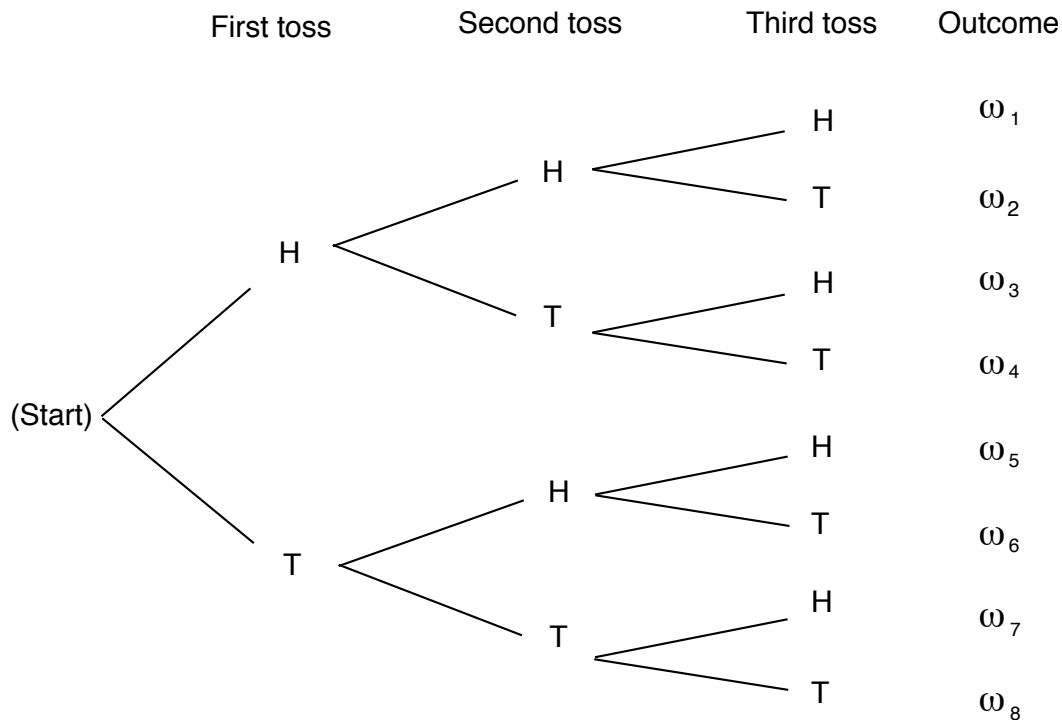


Figure 4.2: Tree diagram for three tosses of a coin

**Exercise 4.39** Let  $A$  and  $B$  be events such that  $P(A \cap B) = 1/4$ ,  $P(A') = 1/3$ , and  $P(B) = 1/2$ . What is  $P(A \cup B)$ ?

**Exercise 4.40** A student must choose one of the subjects, art, geology, or psychology, as an elective. She is equally likely to choose art or psychology and twice as likely to choose geology. What are the respective probabilities that she chooses art, geology, and psychology?

**Exercise 4.41** A student must choose exactly two out of three electives: Fine Art, French, and Mathematics. He chooses Fine Art with probability  $5/8$ , French with probability  $5/8$ , and Fine Art and French together with probability  $1/4$ . What is the probability that he chooses mathematics? What is the probability that he chooses either art or French?

**Exercise 4.42** What odds should a person give in favor of the following events?

- 1.) A card chosen at random from a 52-card deck is an ace.
- 2.) Two heads will turn up when a coin is tossed twice.
- 3.) Boxcars (two sixes) will turn up when two dice are rolled.

**Exercise 4.43** John and Mary are taking a mathematics course. The course has only three grades:  $A$ ,  $B$ , and  $C$ . The probability that John gets a  $B$  is 0.3. The probability that Mary gets a  $B$  is 0.4. The probability that neither gets an  $A$  but at least one gets a  $B$  is 0.1. What is the probability that at least one gets a  $B$  but neither gets a  $C$ ?

**Exercise 4.44** If  $A$ ,  $B$ , and  $C$  are any three events, show that

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(B \cap C) - P(A \cap C) \\ &\quad + P(A \cap B \cap C) . \end{aligned}$$

**Example 4.11.65** We have two urns, I and II. Urn I contains 2 black balls and 3 white balls. Urn II contains 1 black ball and 1 white ball. An urn is drawn at random and a ball is chosen at random from it.

Let  $B$  be the event “a black ball is drawn,” and  $I$  the event “urn I is chosen.”

Suppose we wish to calculate  $P(I|B)$ . Using the formula, we obtain

$$\begin{aligned} P(I|B) &= \frac{P(B \cap I)}{P(B)} \\ &= \frac{P(I \cap B)}{P(I \cap B) + P(II \cap B)} \\ &= \frac{P(B | I) \cdot P(I)}{P(B | I) \cdot P(I) + P(B | II) \cdot P(II)} \\ &= \frac{\frac{2}{5} \cdot \frac{1}{2}}{\frac{2}{5} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}} = \frac{1/5}{1/5 + 1/4} = \frac{4}{9} . \end{aligned}$$

**Example 4.11.66** We consider now a problem called the *Monty Hall* problem. This has long been a favorite problem

Suppose you’re on Monty Hall’s *Let’s Make a Deal!* You are given the choice of three doors, behind one door is a car, the others, goats. You pick a door, say 1, Monty opens another door, say 3, which has a goat. Monty says to you “Do you want to pick door 2?” Is it to your advantage to switch your choice of doors?

Marilyn gave a solution concluding that you should switch, and if you do, your probability of winning is  $2/3$ . Several irate readers, some of whom identified themselves as having a PhD in mathematics, said that this is absurd since after Monty has ruled out one door there are only two possible doors and they should still each have the same probability  $1/2$  so there is no advantage to switching. Marilyn stuck to her solution and encouraged her readers to simulate the game and draw their own conclusions from this. We also encourage the reader to do this.

Other readers complained that Marilyn had not described the problem completely. In particular, the way in which certain decisions were made during a play of the game were not specified. We will assume that the car was put behind a door by rolling a three-sided die which made all three choices equally likely. Monty knows where the car is, and always opens a door with a goat behind it. Finally, we assume that if Monty has a choice of doors (i.e., the contestant has picked the door with the car behind it), he chooses each door with probability  $1/2$ . Marilyn clearly expected her readers to assume that the game was played in this manner.

As is the case with most apparent paradoxes, this one can be resolved through careful analysis. We begin by describing a simpler, related question. We say that a contestant is using the “stay” strategy if he picks a door, and, if offered a chance to switch to another door, declines to do so (i.e., he stays with his original choice). Similarly, we say that the contestant is using the “switch” strategy if he picks a door, and, if offered a chance to switch to another door, takes the offer. Now suppose that a contestant decides in advance to play the “stay” strategy. His only action in this case is to pick a door (and decline an invitation to switch, if one is offered). What is the probability that he wins a car? The same question can be asked about the “switch”

strategy.

Using the “stay” strategy, a contestant will win the car with probability  $1/3$ , since  $1/3$  of the time the door he picks will have the car behind it. On the other hand, if a contestant plays the “switch” strategy, then he will win whenever the door he originally picked does not have the car behind it, which happens  $2/3$  of the time.

This very simple analysis, though correct, does not quite solve the problem that Craig posed. Craig asked for the conditional probability that you win if you switch, given that you have chosen door 1 and that Monty has chosen door 3. To solve this problem, we set up the problem before getting this information and then compute the conditional probability given this information. This is a process that takes place in several stages; the car is put behind a door, the contestant picks a door, and finally Monty opens a door. Thus it is natural to analyze this using a tree measure. Here we make an additional assumption that if Monty has a choice of doors (i.e., the contestant has picked the door with the car behind it) then he picks each door with probability  $1/2$ . The assumptions we have made determine the branch probabilities and these in turn determine the tree measure. The resulting tree and tree measure are shown in Figure 4.3. It is tempting to reduce the tree’s size by making certain assumptions such as: “Without loss of generality, we will assume that the contestant always picks door 1.” We have chosen not to make any such assumptions, in the interest of clarity.

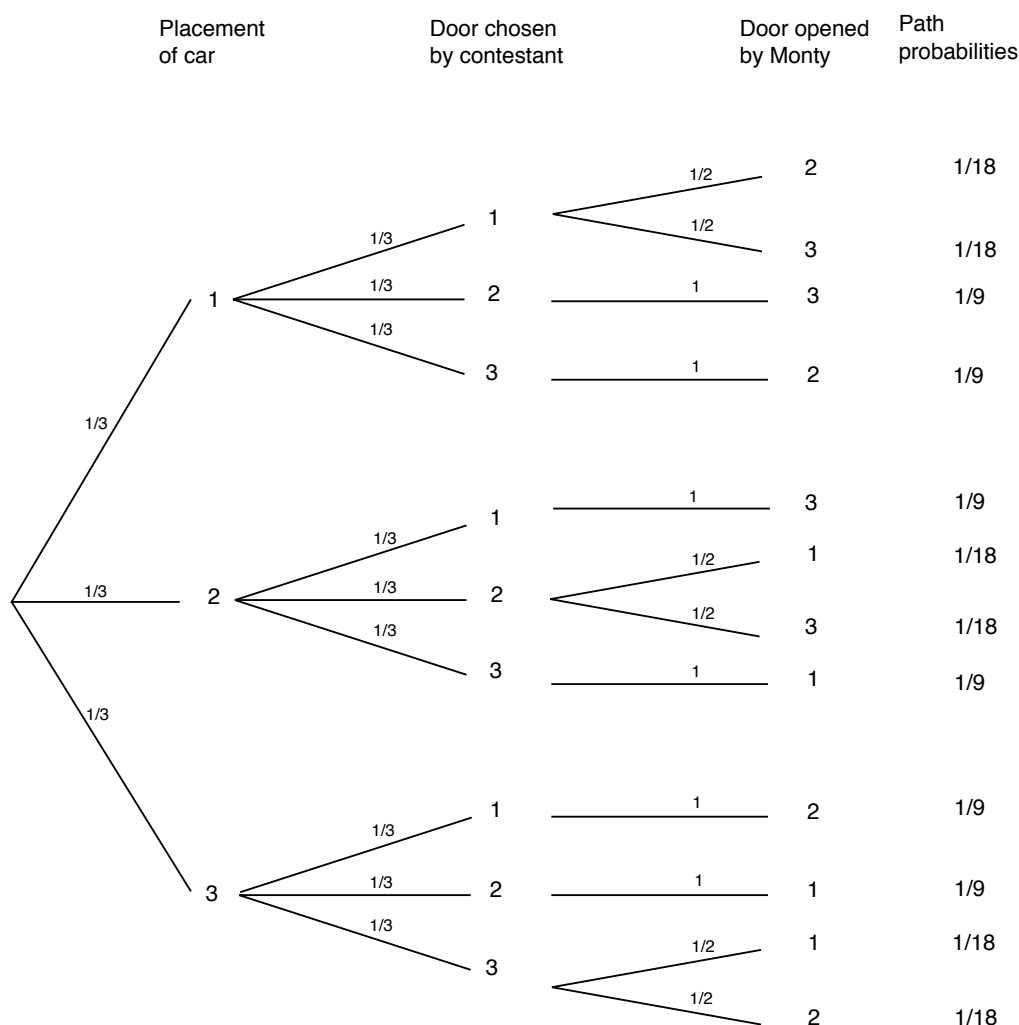


Figure 4.3: The Monty Hall problem.

Now the given information, namely that the contestant chose door 1 and Monty chose door 3, means only two paths through the tree are possible (see Figure 4.4).

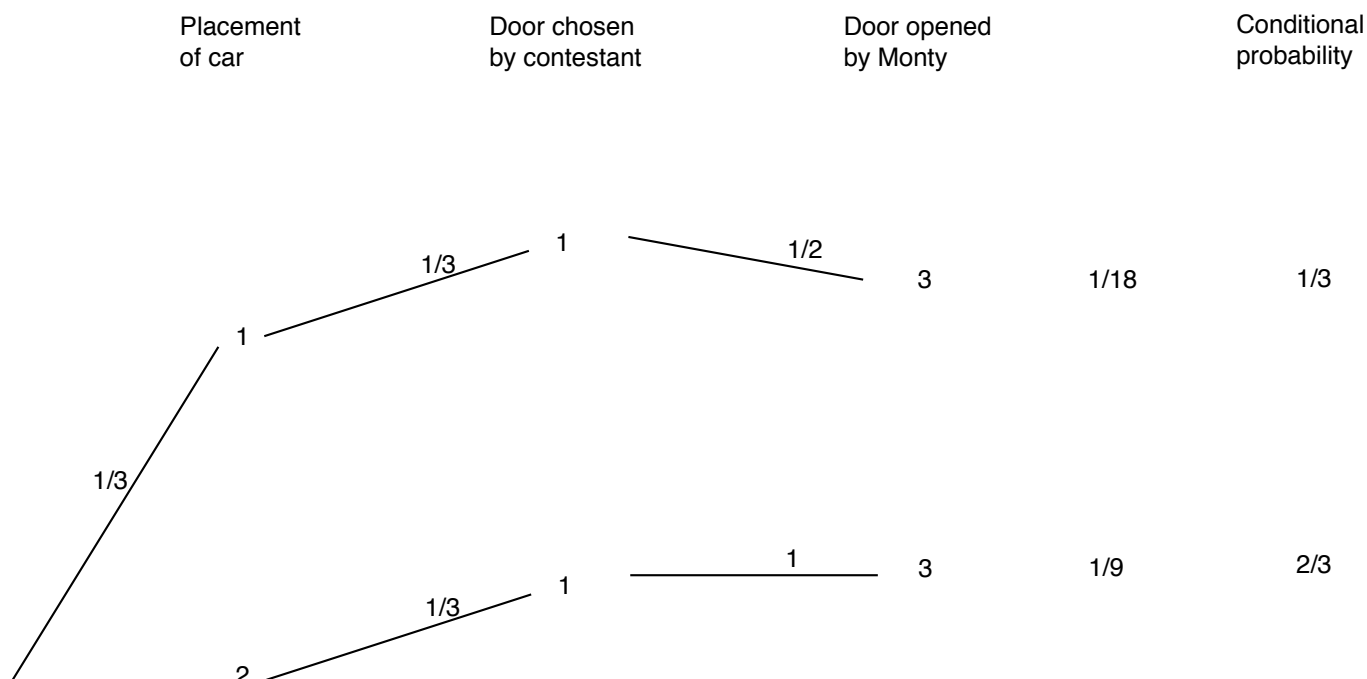


Figure 4.4: Conditional probabilities for the Monty Hall problem.

For one of these paths, the car is behind door 1 and for the other it is behind door 2. The path with the car behind door 2 is twice as likely as the one with the car behind door 1. Thus the conditional probability is  $2/3$  that the car is behind door 2 and  $1/3$  that it is behind door 1, so if you switch you have a  $2/3$  chance of winning the car, as Marilyn claimed.

At this point, the reader may think that the two problems above are the same, since they have the same answers. Recall that we assumed in the original problem if the contestant chooses the door with the car, so that Monty has a choice of two doors, he chooses each of them with probability  $1/2$ . Now suppose instead that in the case that he has a choice, he chooses the door with the larger number with probability  $3/4$ . In the “switch” vs. “stay” problem, the probability of winning with the “switch” strategy is still  $2/3$ . However, in the original problem, if the contestant switches, he wins with probability  $4/7$ . The reader can check this by noting that the same two paths as before are the only two possible paths in the tree. The path leading to a win, if the contestant switches, has probability  $1/3$ , while the path which leads to a loss, if the contestant switches, has probability  $1/4$ .



**Exercise 4.45** A coin is tossed three times. What is the probability that exactly two heads occur, given that

- 1.) the first outcome was a head?
- 2.) the first outcome was a tail?
- 3.) the first two outcomes were heads?
- 4.) the first two outcomes were tails?
- 5.) the first outcome was a head and the third outcome was a head?

**Exercise 4.46** A die is rolled twice. What is the probability that the sum of the faces is greater than 7, given that

- 1.) the first outcome was a 4?
- 2.) the first outcome was greater than 3?
- 3.) the first outcome was a 1?
- 4.) the first outcome was less than 5?

**Exercise 4.47** A card is drawn at random from a deck of cards. What is the probability that

- 1.) it is a heart, given that it is red?
- 2.) it is higher than a 10, given that it is a heart? (Interpret J, Q, K, A as 11, 12, 13, 14.)
- 3.) it is a jack, given that it is red?

**Exercise 4.48** A coin is tossed three times. Consider the following events

*A*: Heads on the first toss.

*B*: Tails on the second.

*C*: Heads on the third toss.

*D*: All three outcomes the same (HHH or TTT).

*E*: Exactly one head turns up.

- 1.) Which of the following pairs of these events are independent?
  - (1) *A*, *B*
  - (2) *A*, *D*
  - (3) *A*, *E*
  - (4) *D*, *E*
- 2.) Which of the following triples of these events are independent?
  - (1) *A*, *B*, *C*
  - (2) *A*, *B*, *D*
  - (3) *C*, *D*, *E*

**Exercise 4.49** From a deck of five cards numbered 2, 4, 6, 8, and 10, respectively, a card is drawn at random and replaced. This is done three times. What is the probability that the card numbered 2 was drawn exactly two times, given that the sum of the numbers on the three draws is 12?

**Exercise 4.50** A coin is tossed twice. Consider the following events.

$A$ : Heads on the first toss.

$B$ : Heads on the second toss.

$C$ : The two tosses come out the same.

1.) Show that  $A$ ,  $B$ ,  $C$  are pairwise independent but not independent.

2.) Show that  $C$  is independent of  $A$  and  $B$  but not of  $A \cap B$ .

**Exercise 4.51** Let  $\Omega = \{a, b, c, d, e, f\}$ . Assume that

$$\rho(a) = \rho(b) = 1/8 \text{ and } \rho(c) = \rho(d) = \rho(e) = \rho(f) = 3/16$$

Let  $A$ ,  $B$ , and  $C$  be the events  $A = \{d, e, a\}$ ,  $B = \{c, e, a\}$ ,  $C = \{c, d, a\}$ . Show that  $P(A \cap B \cap C) = P(A)P(B)P(C)$  but no two of these events are independent.

**Exercise 4.52** What is the probability that a family of two children has

1.) two boys given that it has at least one boy?

2.) two boys given that the first child is a boy?

**Example 4.11.67** A family has two children. At least one of them is a boy. What is the probability that both are boys?

Let  $X$  be the number of boys. Then

$$\begin{aligned} P(X = 2 | X \geq 1) &= \frac{P(X \geq 1, X = 2)}{P(X \geq 1)} \\ &= \frac{P(X = 2)}{1 - P(X = 0)} = \frac{1/4}{1 - 1/4}. \end{aligned}$$

**Example 4.11.68** 1/10 of men and 1/7 of women are color-blind. A person is chosen at random and that person is color-blind. What is the probability that the person is male. Assume males and females to be in equal numbers.

Let  $M$ =male,  $F$ =female,  $C$ =color-blind. Then by Baye's theorem

$$\begin{aligned} P(M | C) &= \frac{P(C | M)P(M)}{P(C | M)P(M) + P(C | F)P(F)} \\ &= \frac{\frac{1}{10} \cdot \frac{1}{2}}{\frac{1}{10} \cdot \frac{1}{2} + \frac{1}{7} \cdot \frac{1}{2}}. \end{aligned}$$

**Example 4.11.69** A box contains  $w$  white balls,  $b$  black balls or  $r$  red balls. A ball is chosen at random and if it is either black or red then it is replaced by a white ball and if it is white then it is replaced by a red ball. Now again draw a ball. What is the probability that the second ball drawn is red when the first ball drawn is red? What is the probability that the second ball drawn is white?

Let  $W_i, B_i, R_i$  be the event that the  $i$ -th draw is a white, black and red ball respectively. The sample space is given by

$$S = W_1 \cup B_1 \cup C_1 = W_2 \cup B_2 \cup C_2.$$

$$P(R_2|R_1) = \frac{r-1}{w+b+r}.$$

$$\begin{aligned} P(W_2) &= P(W_2|W_1)P(W_1) + P(W_2|B_1)P(B_1) + P(W_2|R_1)P(R_1) \\ &= \frac{w-1}{w+b+r} \frac{w}{w+b+r} + \frac{w+1}{w+b+r} \frac{b}{w+b+r} + \frac{w+1}{w+b+r} \frac{r}{w+b+r}. \end{aligned}$$

**Example 4.11.70** Suppose that  $A \subset B$  and  $P(A) > 0$  and  $P(B) > 0$ . Are two events  $A$  and  $B$  independent?

Since  $A \subset B$ ,  $P(A \cap B) = P(A)$ . The condition for the independence is

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

But

$$P(A) = P(A) \cdot P(B)$$

iff.  $P(B) = 1$ . Hence, if  $P(B) = 1$ ,  $A$  and  $B$  are independent but if  $P(B) < 1$ ,  $A$  and  $B$  are not independent.

**Example 4.11.71** In a bolt factory, machines 1, 2 and 3 respectively produce 20 %, 30% and 50% of the total output. Of their output, 5%, 3% and 2% are defective. A bolt is selected random.

- 1.) What is the probability that it is defective? Let  $D$  be the event that the bolt is defective and  $M_1, M_2, M_3$  be the events that the selected bolt comes from machines 1, 2 and 3 respectively.  $P(M_1) = 0.2, P(M_2) = 0.3, P(M_3) = 0.5$ . From the law of total probability,

$$P(D) = \sum_{i=1}^3 P(D | M_i)P(M_i),$$

where  $P(D | M_1) = 0.05, P(D | M_2) = 0.03, P(D | M_3) = 0.02$ . Hence,

$$P(D) = 0.029$$

- 2.) Given that it is defective, what is the probability that it was made by machine 1?

$$P(M_1 | D) = \frac{P(D | M_1)P(M_1)}{P(D)} = 0.3448.$$

# Chapter 5

## Random Variables

### 5.1 Random Variables & Sample Spaces

It is commonly the case when an experiment is performed that we are mainly interested in some function of the outcome as opposed to the actual outcome itself. For instance in tossing a pair of dice we are often interested in the sum of the two dice and not really about the separate values of the dice. That is, we may be interested in knowing that the sum is 9 and not concerned about whether the actual outcome are (1, 8), (2, 7), (3, 6), (4, 5), (6, 3), (7, 2) or (8, 1). Here then we are interested in the quantities not special qualities involved. These quantities of interest or more formally, these real valued functions defined on the sample space shall be referred to as random variables.

**Definition 5.1.1** A random variable  $f(x)$  or  $P(X = x) = \rho(x)$  is a real-valued function representing the outcome of a random experiment. It takes from a sample space  $\Omega$  to  $[0, 1]$

$$f(x) : \Omega \rightarrow [0, 1] \quad (5.1)$$

**Remark 5.1.1** We use capital  $X$  to denote the random variables, and a small  $x$  to denote its outcome.

State the random variables for the experiments below:

**Example 5.1.1** Toss a coin 2 times:  $X$  = no. of tails obtained

The sample space is

$$HH, HT, TH, TT$$

This can be summarized (written as a random variable) as

$x$	0	1	2
$f(x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

alternatively the random variable  $P(X = x)$  or  $f(x)$  can be written as

$$f(x) = \begin{cases} \frac{1}{4} & ; x = 0, 2, \\ \frac{1}{2} & ; x = 1. \end{cases}$$

**Example 5.1.2** Roll a pair of dice:  $X = \text{sum of face values}$

If we toss two dice simultaneously and note the sum on the two, we have the following table of possible outcomes.

	Die 1					
Die 2	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

with a random variable

With a corresponding table of sums

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

$$f(x) = \begin{cases} \frac{1}{36} & ; x = 2, 12, \\ \frac{2}{36} & ; x = 3, 11, \\ \frac{3}{36} & ; x = 4, 10, \\ \frac{4}{36} & ; x = 5, 9, \\ \frac{5}{36} & ; x = 6, 8, \\ \frac{6}{36} & ; x = 7, \\ 0 & ; \text{otherwise.} \end{cases}$$

Alternatively in tabular form, the random variable is given by

$x$	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

**Note 5.1.1** A random variable is a summary of a sample space. The advantage of a random variable to sample spaces when computing probabilities is that for  $n \geq 3$ , a sample space becomes more complicated.

**Exercise 5.1** Classify each random variable as either discrete or continuous.

- 1.) The number of arrivals at an emergency room between midnight and 6 : 00 *a.m.*
- 2.) The weight of a box of cereal labeled “18 ounces.”
- 3.) The duration of the next outgoing telephone call from a business office.
- 4.) The number of kernels of popcorn in a 1-pound container.
- 5.) The number of applicants for a job.

**Exercise 5.2** Classify each random variable as either discrete or continuous.

- 1.) The time between customers entering a checkout lane at a retail store.
- 2.) The weight of refuse on a truck arriving at a landfill.
- 3.) The number of passengers in a passenger vehicle on a highway at rush hour.
- 4.) The number of clerical errors on a medical chart.
- 5.) The number of accident-free days in one month at a factory.

**Exercise 5.3** Classify each random variable as either discrete or continuous.

- 1.) The number of boys in a randomly selected three-child family.
- 2.) The temperature of a cup of coffee served at a restaurant.
- 3.) The number of no-shows for every 100 reservations made with a commercial airline.
- 4.) The number of vehicles owned by a randomly selected household.
- 5.) The average amount spent on electricity each July by a randomly selected household in a certain state.

**Exercise 5.4** Classify each random variable as either discrete or continuous.

- 1.) The number of patrons arriving at a restaurant between 5 : 00 *p.m.* and 6 : 00 *p.m.*
- 2.) The number of new cases of influenza in a particular county in a coming month.
- 3.) The air pressure of a tire on an automobile.
- 4.) The amount of rain recorded at an airport one day.
- 5.) The number of students who actually register for classes at a university next semester.

**Exercise 5.5** Identify the set of possible values for each random variable. (Make a reasonable estimate based on experience, where necessary.)

- 1.) The number of heads in two tosses of a coin.
- 2.) The average weight of newborn babies born in a particular county one month.
- 3.) The amount of liquid in a 12-ounce can of soft drink.

- 4.) The number of games in the next World Series (best of up to seven games).
- 5.) The number of coins that match when three coins are tossed at once.

**Exercise 5.6** Identify the set of possible values for each random variable. (Make a reasonable estimate based on experience, where necessary.)

- 1.) The number of hearts in a five-card hand drawn from a deck of 52 cards that contains 13 hearts in all.
- 2.) The number of pitches made by a starting pitcher in a major league baseball game.
- 3.) The number of breakdowns of city buses in a large city in one week.
- 4.) The distance a rental car rented on a daily rate is driven each day.
- 5.) The amount of rainfall at an airport next month.

## 5.2 Discrete Random Variables

If a random experiment can result into countable (finite) outcomes then the random variables  $X$  representing that random experiment will be discrete.

**Definition 5.2.1** A discrete random variable  $X$  is one that takes on a countable random experiment and their corresponding set of possible outcomes (countable)

### Example 5.2.1

- 1.) The score when a die is tossed,  $X = 1, 2, 3, 4, 5, 6$
- 2.) The number of heads when two coins are tossed simultaneously,  $X = 0, 1, 2$
- 3.) The number of boys in a family of five,  $X = 0, 1, 2, 3, 4, 5$

**Definition 5.2.2** Let  $X$  be a discrete random variable with values  $x_1, x_2, \dots, x_n$ . Let the associated probabilities be  $\rho_1, \rho_2, \dots, \rho_n$  : where

$$P(X = x_1) = \rho_1, P(X = x_2) = \rho_2, \dots, P(X = x_n) = \rho_n$$

Then  $X$  is a discrete random variable if  $\sum_{i=1}^n \rho_i = 1$

### 5.2.1 Discrete Probability Distributions

**Definition 5.2.3** A table of formula listing all possible values that a discrete random variable  $X$  can take on, together with associated probabilities is called a discrete probability distribution.

**Definition 5.2.4** Let  $\rho(x) = f(x) = P(X = x)$  be a probability mass function of a discrete random variable  $X$ . Then the following properties are true

Property 1:  $P(c) = P(X = c)$

Property 2:  $\rho(x) \geq 0$

Property 3:  $\sum \rho(x) = 1$  [or  $\sum P(X = x) = 1$ ]

Examples of a probability mass function of a discrete random variable  $X$  are tossing coins, die, coin and dice etc

**Example 5.2.2** Toss a die once. Then  $X$  would take up values 1, 2, 3, 4, 5, 6 with probability 1/6, of occurring. Thus in tabular form we have:

$x$	1	2	3	4	5	6
$f(x)$	1/6	1/6	1/6	1/6	1/6	1/6

Observe that

- 1.)  $P(1) = P(X = 1) = P(2) = P(X = 2) = \dots = P(6) = P(X = 6) = 1/6$ .
- 2.)  $\sum \rho(x) = 1$ .
- 3.)  $\rho(x) \geq 0$  hence a probability mass function.



**Example 5.2.3** Toss three coins simultaneously and note the number of heads that occur. Then we have a sample space

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

so that either we have no head, one, two, or three heads with respective probabilities  $1/8, 3/8, 3/8, 1/8$ . We tabulate this as:

$x$	0	1	2	3
$P(X = x)$	$1/8$	$3/8$	$3/8$	$1/8$

Observe that

1.)  $P(X = 0) = 1/8, P(X = 1) = 3/8, P(X = 2) = 3/8$  and  $P(X = 3) = 1/8$ .

2.)  $\sum_{x=0}^{x=3} f(x) = 1/8 + 3/8 + 3/8 + 1/8 = 1$ .

**Definition 5.2.5** The probability that a discrete random variable  $X$  takes on values between two specified outcomes  $a$  and  $b$

$$P(a \leq X \leq b) = \sum_{x=a}^b f(x).$$

**Example 5.2.4** When tossing a die once, as in Example (5.2.2)

1.)

$$\begin{aligned} P(2 \leq X \leq 5) &= \sum_{x=2}^5 f(x) \\ &= P(X = 2) + P(x = 3) + P(X = 4) + P(x = 5) \\ &= 1/6 + 1/6 + 1/6 + 1/6 = 2/3 \end{aligned}$$

2.)

$$\begin{aligned} P(2 < X < 5) &= \sum_{x=3}^4 \rho(x) = \sum_{x=3}^4 f(x) \\ &= P(x = 3) + P(x = 4) \\ &= 1/6 + 1/6 = 2/6 \\ &= 1/3 \end{aligned}$$

**Example 5.2.5** When tossing three coins simultaneously, as in Example 5.2.3, compute

1.)

$$\begin{aligned} P(X > 1) &= P(X = 2) + P(X = 3) \\ &= 3/8 + 1/8 = 4/8 = 1/2 \end{aligned}$$

2.)

$$\begin{aligned} P(X < 3) &= P(x = 0) + P(X = 1) + (X = 2) \\ &= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8} \end{aligned}$$

**Example 5.2.6** When tossing two dice, as in Example 5.1.2

$$\begin{aligned}P(4 < X \leq 7) &= P(X = 5) + P(X = 6) + P(X = 7) \\&= \frac{4}{36} + \frac{5}{36} + \frac{6}{36} = \frac{15}{36} = \frac{5}{12}\end{aligned}$$

**Example 5.2.7** A fair coin is tossed twice. Let  $X$  be the number of heads that are observed.

1.) Construct the probability distribution of  $X$ .

**Solution :** *The possible values that  $X$  can take are 0, 1, and 2. Each of these numbers corresponds to an event in the sample space  $S = \{HH, HT, TH, TT\}$  of equally likely outcomes for this experiment:*

$X = 0$  to  $\{TT\}$ ,  $X = 1$  to  $\{HT, TH\}$ , and  $X = 2$  to  $HH$ .

*The probability of each of these events, hence of the corresponding value of  $X$ , can be found simply by counting, to give*

$x$	0	1	2
$P(x)$	0.25	0.50	0.25

*This table is the probability distribution of  $X$ .* ■

2.) Find the probability that at least one head is observed.

**Solution :** *“At least one head” is the event  $X \geq 1$ , which is the union of the mutually exclusive events  $X = 1$  and  $X = 2$ . Thus*

$$\begin{aligned}P(X \geq 1) &= P(1) + P(2) = 0.50 + 0.25 \\&= 0.75\end{aligned}$$

■

**Example 5.2.8** A pair of fair dice is rolled. Let  $X$  denote the sum of the number of dots on the top faces.

1.) Construct the probability distribution of  $X$  for a pair of fair dice.

**Solution :** *The sample space of equally likely outcomes is*

11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56
61	62	63	64	65	66

*where the first digit is die 1 and the second number is die 2.*

*The possible values for  $X$  are the numbers 2 through 12.  $X = 2$  is the event  $\{11\}$ , so  $P(2) = 1/36$ .  $X = 3$  is the event  $\{12, 21\}$ , so  $P(3) = 2/36$ . Continuing this way we obtain the following table*

$x$	2	3	4	5	6	7	8	9	10	11	12
$P(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

*This table is the probability distribution of  $X$ .* ■

2.) Find  $P(X \geq 9)$ .

**Solution :** *The event  $X \geq 9$  is the union of the mutually exclusive events  $X = 9, X = 10, X = 11$ , and  $X = 12$ . Thus*

$$\begin{aligned}P(X \geq 9) &= P(9) + P(10) + P(11) + P(12) \\&= \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} \\&= \frac{10}{36} \\&= 0.2\bar{7}\end{aligned}$$

■

3.) Find the probability that  $X$  takes an even value.

**Solution :** *Before we immediately jump to the conclusion that the probability that  $X$  takes an even value must be 0.5, note that  $X$  takes six different even values but only five different odd values. We compute*

$$\begin{aligned}P(X \text{ is even}) &= P(2) + P(4) + P(6) + P(8) + P(10) + P(12) \\&= \frac{1}{36} + \frac{3}{36} + \frac{5}{36} + \frac{5}{36} + \frac{3}{36} + \frac{1}{36} \\&= \frac{18}{36} \\&= 0.5\end{aligned}$$

■

**Note 5.2.1** For a discrete random variable  $\leq \not\equiv <$  and  $\geq \not\equiv >$

$$1.) P(a \leq X \leq b) = \sum_{x=a}^b \rho(x) = \sum_{x=a}^b f(x)$$

$$2.) P(a < X \leq b) = \sum_{x=a+1}^b \rho(x) = \sum_{x=a+1}^b f(x)$$

$$3.) P(a \leq X < b) = \sum_{x=a}^{b-1} \rho(x) = \sum_{x=a}^{b-1} f(x)$$

$$4.) P(X > a) = \sum_{x=a+1}^b \rho(x) = \sum_{x=a+1}^b f(x)$$

### 5.2.2 Cumulative Distribution Function (CDF)

**Definition 5.2.6** We define the cumulative distribution function  $CDF$  of a discrete random variable  $X$ , denoted as  $F(x)$  by

$$\begin{aligned} F(X) &= P(X \leq x) \\ &= \sum_{t \leq x} f(t) \end{aligned}$$

You must bear in mind that while it is called the distribution function of  $x$  it is not a function of  $x$ , that is, its argument is not a random variable  $X$ , but the real number  $x$ .

We also define the c.d.f of a discrete random variable as the sum of all values of the probability function of  $x$  at points to the left of  $x$ .

**Example 5.2.9** Consider the probability mass function of discrete random variable defined by

$x$	0	1	2	3
$f(x) = P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Then the cumulative distribution function of the above example is:

$x$	0	1	2	3
$P(X \leq x)$	$\frac{1}{8}$	$\frac{4}{8}$	$\frac{7}{8}$	1

**Example 5.2.10** Consider a probability mass function

$x$	2	3	4	5	6	6	8	9	10	11	12
$f(x) = P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Then the cumulative distribution function is

$x$	2	3	4	5	6	6	8	9	10	11	12
$P(X \leq x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

**Example 5.2.11** Consider a probability mass function

$x$	1	2	3	4	5	6
$f(x) = P(X = x) = \rho(x) = \mathbb{P}(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

The cumulative distribution function is

$x$	1	2	3	4	5	6
$P(X \leq x)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

**Example 5.2.12** A discrete random variable has a probability mass function defined by

$$f(x) = \begin{cases} \frac{x}{c} & ; x = 1, 2, 3, 4 \\ 0 & ; \text{elsewhere} \end{cases}$$

- 1.) Determine the value of  $c$
- 2.) Find the cumulative distribution function  $F(x)$ .

Since  $\sum f(x) = 1$  then  $\frac{1}{c} + \frac{2}{c} + \frac{3}{c} + \frac{4}{c} = 1 \Rightarrow 10 = c$  or  $c = 10$ . Then we tabulate our outcomes as:

$x$	1	2	3	4
$f(x) = P(X = x)$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$

So that our cumulative distribution function is

$x$	1	2	3	4
$P(X \leq x)$	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{6}{10}$	1

**Example 5.2.13** A discrete random  $X$  is defined by

$$f(x) = \begin{cases} \frac{x}{15k} & ; x = 1, 2, 3, 4, 5 \\ 0 & ; \text{elsewhere} \end{cases}$$

1.) Determine  $k$

$$\begin{aligned} \sum_{\forall i} f(x_i) &= 1 \\ \frac{1}{15k} + \frac{2}{15k} + \frac{3}{15k} + \frac{4}{15k} + \frac{5}{15k} &= 1 \\ \Rightarrow \frac{15}{15k} &= 1 \\ \Rightarrow k &= 1 \end{aligned}$$

2.) Find the distribution function  $F(x)$ .

In tabular form, the probability function is

$x$	1	2	3	4	5
$P(X = x)$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{5}{15}$

To have the cumulative distribution function as

$x$	1	2	3	4	5
$P(X \leq x)$	$\frac{1}{15}$	$\frac{3}{15}$	$\frac{6}{15}$	$\frac{10}{15}$	1

**Note 5.2.2** For the cumulative distribution function  $F(x)$ , we note that

1.)

$$0 \leq F(x) \leq 1$$

since  $F(x)$  is a probability

2.)  $F(x)$  is a nondecreasing function of  $x$ .

3.) If  $F(y) = 1$  then  $y$  is any value greater or equal to the largest value in the range  $R$  given and if for any  $z$ ,  $F(z) = 0$  then  $z$  is any value less than the small value in  $R$ .

4.) If  $X$  is a random variable of a discrete type then  $F(x)$  is a step function and the height of set  $P$  at  $x$ , ( $x \in R$ ), equals the probability  $P(X = x)$ .

### 5.2.3 Expectation of a Discrete Random Variable $X$

**Definition 5.2.7** For a discrete random variable  $X$ , we define the mean  $\mu$  of  $X$  or expectation of  $X$  or expected value of  $X$  denoted by  $E(X)$  as

$$E(X) = \sum_{\forall i} x_i f(x_i) = \sum x f(x) \quad (5.2)$$

The difference between the expected value and the actual mean already seen in descriptive statistics is that in a practical approach, we get results in a frequency distribution and obtain the mean as  $\frac{\sum fx}{\sum f}$ . Whereas in a theoretical approach, we get results in a probability distribution and an expected value known as the expectation, by multiplying each score by its corresponding probability.

#### 5.2.3.1 Properties of Expectation for a Discrete Random Variable

For a discrete random variable  $X$  with  $E(X) = \sum x f(x)$

Property 1: If  $g(X)$  is any function of the discrete random variable  $X$ , then

$$E[g(X)] = \sum g(x) f(x) \quad (5.3)$$

For example,

$$\begin{aligned} E(5X) &= \sum 5x f(x), \\ E(X^2) &= \sum x^2 f(x) \\ E(2X + 3) &= \sum (2x + 3) f(x) \end{aligned}$$

Property 2:

$$E[a] = a \quad (5.4)$$

where  $a$  is any constant

**Proof :**

$$\begin{aligned} E[a] &= \sum a f(x) = a \sum f(x) \\ &= (a)(1) \\ &= a \end{aligned}$$

■

Property 3: For  $a$ , a constant

$$E[aX] = aE[X] \quad (5.5)$$

**Proof :**

$$\begin{aligned} E[aX] &= \sum a x f(x) \\ &= a \sum x f(x) \\ &= aE[X] \end{aligned}$$

■

Property 4: For  $a$  and  $b$  constants,

$$E[aX + b] = aE[x] + b \quad (5.6)$$

**Proof :**

$$\begin{aligned} E[aX + b] &= \sum (ax + b)f(x) \\ &= \sum axf(x) + \sum bf(x) \\ &= a \sum xf(x) + b \sum f(x) \\ &= aE[X] + b \end{aligned}$$

■

Property 5: The expectation of the sum, is the sum of the expectation.

$$E \left[ \sum_{i=1}^n f_i \right] = \sum_{i=1}^n E[f_i] \quad (5.7)$$

For any two functions  $g(X), h(X)$

$$E[g(X) + h(X)] = E[g(x)] + E[h(X)] \quad (5.8)$$

**Proof :**

$$\begin{aligned} E[g(X) + h(X)] &= \sum [g(x) + h(x)] f(x) \\ &= \sum g(x)f(x) + \sum h(x)f(x) \\ &= E[g(X)] + E[h(X)] \end{aligned}$$

■

**Example 5.2.14** In Example 5.2.15, if  $g(x) = x^2$ , then we have

$x$	0	1	2	3
$x^2$	0	1	4	9
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Then using

$$\begin{aligned} E[g(X)] &= \sum g(x)f(x) \\ &= \sum x^2 f(x) \\ &= (0)\frac{1}{8} + (1)\frac{3}{8} + (4)\frac{3}{8} + (9)\frac{1}{8} \\ &= \frac{24}{8} \\ &= 3 \end{aligned}$$



**Example 5.2.15** Find the expected number of heads that will be obtained when a fair coin is tossed three times.

Let  $X$  denote the number of heads in three tosses, and from the random variable in Example 5.2.3, (pg. 220)

$$\begin{aligned} E[X] &= \sum_{\forall x} xf(x) \\ &= (0)\frac{1}{8} + (1)\frac{3}{8} + (2)\frac{3}{8} + (3)\frac{1}{8} \\ &= \frac{3}{2} \end{aligned}$$

**Example 5.2.16** A discrete random variable  $X$  has probability mass function defined by

$$f(x) = \frac{x}{10}, \quad x = 1, 2, 3, 4.$$

Find  $E[X]$ , the expectation of  $X$ .

We can express our problem in tabular form as

$x$	0	1	2	3	4
$P(X = x)$	$\frac{0}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$

Then from definition,

$$\begin{aligned} E[X] &= \sum xf(x) \\ &= (0)\frac{0}{10} + (1)\frac{1}{10} + (2)\frac{2}{10} + (3)\frac{3}{10} + (4)\frac{4}{10} \\ &= 3 \end{aligned}$$

**Example 5.2.17** In Example 5.2.16, Find  $E[g(x)]$  if  $g(x) = 2x - 1$ . Then we have

$x$	1	2	3	4
$2x - 1$	1	3	5	7
$P(X = x)$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$

And so

$$\begin{aligned} E[2X - 1] &= \sum (2x - 1)f(x) \\ &= (1)\frac{1}{10} + (3)\frac{2}{10} + (5)\frac{3}{10} + (7)\frac{4}{10} \\ &= \frac{50}{10} \\ &= 5 \end{aligned}$$

**Example 5.2.18** Find the mean of the discrete random variable  $X$  whose probability distribution is

$x$	-2	1	2	3.5
$P(x)$	0.21	0.34	0.24	0.21

**Solution :** *Using the definition of mean gives*

$$\begin{aligned}\mu &= \sum xP(x) \\ &= (-2)(0.21) + (1)(0.34) + (2)(0.24) + (3.5)(0.21) \\ &= 1.135\end{aligned}$$

■

**Example 5.2.19** A service organization in a large town organizes a raffle each month. One thousand raffle tickets are sold for \$1 each. Each has an equal chance of winning. First prize is \$300, second prize is \$200, and third prize is \$100. Let  $X$  denote the net gain from the purchase of one ticket.

1.) Construct the probability distribution of  $X$ .

**Solution :** *If a ticket is selected as the first prize winner, the net gain to the purchaser is the \$300 prize less the \$1 that was paid for the ticket, hence  $X = 300 - 1 = 299$ . There is one such ticket, so  $P(299) = 0.001$ . Applying the same “income minus outgo” principle to the second and third prize winners and to the 997 losing tickets yields the probability distribution:*

$x$	299	199	99	-1
$P(x)$	0.001	0.001	0.001	0.997

■

2.) Find the probability of winning any money in the purchase of one ticket.

**Solution :** *Let  $W$  denote the event that a ticket is selected to win one of the prizes. Using the table*

$$\begin{aligned}P(W) &= P(299) + P(199) + P(99) = 0.001 + 0.001 + 0.001 \\ &= 0.003\end{aligned}$$

■

3.) Find the expected value of  $X$ , and interpret its meaning.

**Solution :** *Using the definition of expected value,*

$$\begin{aligned}E(X) &= (299) \cdot (0.001) + (199) \cdot (0.001) + (99) \cdot (0.001) + (-1) \cdot (0.997) \\ &= -0.4\end{aligned}$$

*The negative value means that one loses money on the average. In particular, if someone were to buy tickets repeatedly, then although he would win now and then, on average he would lose 40 cents per ticket purchased. The concept of expected value is also basic to the insurance industry, as the following simplified example illustrates.*

■

**Example 5.2.20** A life insurance company will sell a \$200,000 one-year term life insurance policy to an individual in a particular risk group for a premium of \$195. Find the expected value to the company of a single policy if a person in this risk group has a 99.97% chance of surviving one year.

**Solution :** Let  $X$  denote the net gain to the company from the sale of one such policy. There are two possibilities: the insured person lives the whole year or the insured person dies before the year is up. Applying the “income minus outgo” principle, in the former case the value of  $X$  is  $195 - 0$ ; in the latter case it is  $195 - 200,000 = -199,805$ . Since the probability in the first case is 0.9997 and in the second case is  $1 - 0.9997 = 0.0003$ , the probability distribution for  $X$  is:

$x$	195	-199,805
$P(x)$	0.9997	0.0003

Therefore

$$\begin{aligned} E(X) &= \sum xP(x) \\ &= (195) \cdot (0.9997) + (-199,805) \cdot (0.0003) \\ &= 135 \end{aligned}$$

Occasionally (in fact, 3 times in 10,000) the company loses a large amount of money on a policy, but typically it gains \$195, which by our computation of  $E(X)$  works out to a net gain of \$135 per policy sold, on average. ■

### 5.2.4 Variance of a Discrete Random Variable

Consider these three sets of quiz scores:

Section A:

$$5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5$$

Section B:

$$0, 0, 0, 0, 0, 10, 10, 10, 10, 10$$

Section C:

$$4, 4, 4, 5, 5, 5, 5, 6, 6, 6$$

All three of these sets of data have a mean of 5 and a median of 5, yet the sets of scores are clearly quite different. In section A, everyone had the same score; in section B, half the class got no points and the other half got a perfect score, assuming this was a 10-point quiz. Section C was not as consistent as section A, but not as widely varied as section B.

**Remark 5.2.1** In addition to the mean and median, which are measures of the “typical” or “middle” value, we also need a measure of how “spread out” or varied each data set is.

There are several ways to measure this “spread” of the data. The measures of variation such as range, standard deviation, variance and quartiles.

**Definition 5.2.8** The **variance** of random variable  $X$  is the measure of the dispersion or spread of a distribution. It is denoted by  $\sigma^2$  or  $\text{Var}(X)$  and defined by

$$\sigma^2 = E[(X - \mu)^2] \quad (5.9)$$

or by definition of expectation,

$$\sigma^2 = \sum (x - \mu)^2 f(x) \quad (5.10)$$

It is important to note that computationally, the variance of  $X$

$$\begin{aligned} \text{Var}(X) &= E[(X - \mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

Thus the variance given by Equation (5.9) is given by

$$\text{Var}(X) = E(X^2) - [E(X)]^2 \quad (5.11)$$

The variance of a discrete random variable  $X$  is defined as

$$\begin{aligned} \text{Var}(X) &= E(X^2) - [E(X)]^2 \\ &= \sum x^2 f(x) - \left[ \sum x f(x) \right]^2 \end{aligned}$$

**Definition 5.2.9** A **standard deviation** denoted by  $\sigma$  is the square root of variance.

**5.2.4.1 Properties of Variance**

For a discrete random variable  $X$  with

$$\begin{aligned} E(X) &= \sum xf(x) \\ \text{Var}(X) &= E(X^2) - [E(X)]^2 \end{aligned}$$

Property 1: With  $a$ , a constant.

$$\text{Var}(a) = 0 \quad (5.12)$$

**Proof :**

$$\begin{aligned} \text{Var}(a) &= E(a^2) - [E(a)]^2 \\ &= a^2 - a^2 \\ &= 0 \end{aligned}$$

■

*Thus the variance of a constant is zero.*

Property 2: For a constant value  $a$

$$\text{Var}(aX) = a^2 \text{Var}(X) \quad (5.13)$$

**Proof :**

$$\begin{aligned} \text{Var}(aX) &= E(aX)^2 - [E(aX)]^2 \\ &= a^2 E(X)^2 - a^2 [E(X)]^2 \\ &= a^2 (E(X^2) - [E(X)]^2) \\ &= a^2 \text{Var}(X) \end{aligned}$$

■

Property 3: For constants,  $a$  and  $b$ ,

$$\text{Var}(aX + b) = a^2 \text{Var}(X) \quad (5.14)$$

**Proof :**

$$\begin{aligned} \text{Var}(aX + b) &= E(aX + b)^2 - [E(aX + b)]^2 \\ &= E(a^2 X^2 + 2abX + b^2) - [aE(X) + b]^2 \\ &= a^2 E(X^2) + 2abE(X) + b^2 - a^2 [E(X)]^2 - 2abE(X) - b^2 \\ &= a^2 E(X^2) - a^2 [E(X)]^2 \\ &= a^2 (E(X^2) - [E(X)]^2) \\ &= a^2 \text{Var}(X) \end{aligned}$$

■

**Example 5.2.21** A discrete random variable  $X$  has a probability distribution

$x$	1	2	3	4
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

Find

- |              |               |                   |
|--------------|---------------|-------------------|
| 1.) $E(X)$   | 3.) $Var(X)$  | 5.) $Var(3X - 2)$ |
| 2.) $E(X^2)$ | 4.) $Var(2X)$ | 6.) $E[2X + 5]$   |

We need to compute  $E(X)$  and  $E(X^2)$ . Thus the table becomes

$x$	1	2	3	4
$x^2$	1	4	9	16
$P(X = x)$	1/2	1/4	1/8	1/8

$$\begin{aligned}E(X) &= \sum xf(x) \\&= (1)\frac{1}{2} + (2)\frac{1}{4} + (3)\frac{1}{8} + (4)\frac{1}{8} \\&= \frac{15}{8} \\E(X^2) &= \sum x^2f(x) \\&= (1)\frac{1}{2} + (4)\frac{1}{4} + (9)\frac{1}{8} + (16)\frac{1}{8} \\&= \frac{37}{8}\end{aligned}$$

Then from definition

$$\begin{aligned}Var(X) &= E(X^2) - [E(X)]^2 \\&= \frac{37}{8} - \left[\frac{15}{8}\right]^2 \\&= \frac{37}{8} - \frac{225}{64} = \frac{71}{64} \\Var(2X) &= 2^2 Var(X) \\&= 4Var(X) \\&= (4)\frac{71}{64} \\&= \frac{71}{16} \\Var(3X - 2) &= 3^2 Var(X) \\&= 9Var(X) \\&= (9)\frac{71}{64} \\&= \frac{639}{64}\end{aligned}$$


**Example 5.2.22** A discrete random variable  $X$  has the following probability distribution:

$x$	$-1$	$0$	$1$	$4$
$P(x)$	$0.2$	$0.5$	$a$	$0.1$

Compute each of the following quantities.


1.)  $a$ .

**Solution :** *Since all probabilities must add up to 1,*

$$a = 1 - (0.2 + 0.5 + 0.1) = 0.2$$



2.)  $P(0)$ .

**Solution :** *Directly from the table,  $P(0)=0.5$*

$$P(0) = 0.5$$



3.)  $P(X > 0)$ .

**Solution :** *From probability distribution table,*

$$P(X > 0) = P(1) + P(4) = 0.2 + 0.1 = 0.3$$



4.)  $P(X \geq 0)$ .

**Solution :** *From probability distribution table,*

$$P(X \geq 0) = P(0) + P(1) + P(4) = 0.5 + 0.2 + 0.1 = 0.8$$



5.)  $P(X \leq -2)$ .

**Solution :** *Since none of the numbers listed as possible values for  $X$  is less than or equal to  $-2$ , the event  $X \leq -2$  is impossible, so*

$$P(X \leq -2) = 0$$


6.) The mean  $\mu$  of  $X$ .

**Solution :** *Using the formula in the definition of  $\mu$*

$$\begin{aligned}\mu &= \sum xP(x) \\ &= (-1) \cdot (0.2) + (0) \cdot (0.5) + (1) \cdot (0.2) + (4) \cdot (0.1) \\ &= 0.4\end{aligned}$$


7.) The variance  $\sigma^2$  of  $X$ .

**Solution :** *Using the formula in the definition of  $\sigma^2$  (Equation 5.10) and the value of  $\mu$  that was just computed,*

$$\begin{aligned}\sigma^2 &= \sum (x - \mu)^2 f(x) \\ &= (-1 - 0.4)^2 \cdot (0.2) + (0 - 0.4)^2 \cdot (0.5) + (1 - 0.4)^2 \cdot (0.2) + (4 - 0.4)^2 \cdot (0.1) \\ &= 1.84\end{aligned}$$

■

8.) The standard deviation  $\sigma$  of  $X$ .

**Solution :** *Using the result of the above part,*

$$\sigma = \sqrt{1.84} = 1.3565$$

■



## 5.3 Discrete Random Variables Chapter Examples

**Example 5.3.1** Determine whether or not the table is a valid probability distribution of a discrete random variable. Explain fully.

1.)

$x$	$-2$	$0$	$2$	$4$
$f(x)$	$0.3$	$0.5$	$0.2$	$0.1$

**Solution :** *no: the sum of the probabilities exceeds 1* ■

2.)

$x$	$0.5$	$0.25$	$0.25$
$f(x)$	$-0.4$	$0.6$	$0.8$

**Solution :** *no: a negative probability* ■

3.)

$x$	$1.1$	$2.5$	$4.1$	$4.6$	$5.3$
$f(x)$	$0.16$	$0.14$	$0.11$	$0.27$	$0.22$

**Solution :** *no: the sum of the probabilities is less than 1* ■

**Example 5.3.2** A discrete random variable  $X$  has the following probability distribution:

$x$	$77$	$78$	$79$	$80$	$81$
$f(x)$	$0.15$	$0.15$	$0.20$	$0.40$	$0.10$

Compute each of the following quantities.

1.)  $P(80)$ .

**Solution :**  $0.4$  ■

2.)  $P(X > 80)$ .

**Solution :**  $0.1$  ■

3.)  $P(X \leq 80)$ .

**Solution :**  $0.9$  ■

4.) The mean  $\mu$  of  $X$ .

**Solution :**  $79.15$  ■

5.) The variance  $\sigma^2$  of  $X$ .

**Solution :**  $\sigma^2 = 1.5275$  ■

6.) The standard deviation  $\sigma$  of  $X$ .

**Solution :**  $\sigma = 1.2359$  ■

**Example 5.3.3** If each die in a pair is “loaded” so that one comes up half as often as it should, six comes up half again as often as it should, and the probabilities of the other faces are unaltered, then the probability distribution for the sum  $X$  of the number of dots on the top faces when the two are rolled is

$x$	2	3	4	5	6	7
$f(x)$	$\frac{1}{144}$	$\frac{4}{144}$	$\frac{8}{144}$	$\frac{12}{144}$	$\frac{16}{144}$	$\frac{22}{144}$

$x$	8	9	10	11	12
$f(x)$	$\frac{24}{144}$	$\frac{20}{144}$	$\frac{16}{144}$	$\frac{12}{144}$	$\frac{9}{144}$

Compute each of the following.

1.)  $P(5 \leq X \leq 9)$ .

**Solution :** 0.6528



2.)  $P(X \geq 7)$ .

**Solution :** 0.7153



3.) The mean  $\mu$  of  $X$ . (For fair dice this number is 7).

**Solution :**  $\mu = 7.8333$



4.) The standard deviation  $\sigma$  of  $X$ . (For fair dice this number is about 2.415).

**Solution :**  $\sigma^2 = 5.4866, \sigma = 2.3424$



**Example 5.3.4** In a hamster breeder’s experience the number  $X$  of live pups in a litter of a female not over twelve months in age who has not borne a litter in the past six weeks has the probability distribution

$x$	3	4	5	6	7	8	9
$f(x)$	0.04	0.10	0.26	0.31	0.22	0.05	0.02

1.) Find the probability that the next litter will produce five to seven live pups.

**Solution :** 0.79



2.) Find the probability that the next litter will produce at least six live pups.

**Solution :** 0.60



3.) Compute the mean and standard deviation of  $X$ . Interpret the mean in the context of the problem.

**Solution :**  $\mu = 5.8, \sigma = 1.2570$



**Example 5.3.5** Let  $X$  denote the number of boys in a randomly selected three-child family. Assuming that boys and girls are equally likely, construct the probability distribution of  $X$ .

**Solution :**

$x$	0	1	2	3
$f(x)$	$1/8$	$3/8$	$3/8$	$1/8$



**Example 5.3.6** Five thousand lottery tickets are sold for \$1 each. One ticket will win \$1,000, two tickets will win \$500 each, and ten tickets will win \$100 each. Let  $X$  denote the net gain from the purchase of a randomly selected ticket.

- 1.) Construct the probability distribution of  $X$ .

**Solution :**

$x$	$-1$	$999$	$499$	$99$
$f(x)$	$\frac{4987}{5000}$	$\frac{1}{5000}$	$\frac{2}{5000}$	$\frac{10}{5000}$

■

- 2.) Compute the expected value  $E(X)$  of  $X$ . Interpret its meaning.

**Solution :**  $-0.4$

■

- 3.) Compute the standard deviation  $\sigma$  of  $X$ .

**Solution :**  $17.8785$

■

**Example 5.3.7** An insurance company will sell a \$90,000 one-year term life insurance policy to an individual in a particular risk group for a premium of \$478. Find the expected value to the company of a single policy if a person in this risk group has a 99.62% chance of surviving one year.

**Solution :**  $136$

■

**Example 5.3.8** An insurance company estimates that the probability that an individual in a particular risk group will survive one year is 0.9825. Such a person wishes to buy a \$150,000 one-year term life insurance policy. Let  $C$  denote how much the insurance company charges such a person for such a policy.

- 1.) Construct the probability distribution of  $X$ . (Two entries in the table will contain  $C$ ).

**Solution :**

$x$	$C$	$C$	$-150,000$
$f(x)$	$0.9825$		$0.0175$

■

- 2.) Compute the expected value  $E(X)$  of  $X$ .

**Solution :**  $C - 2625$

■

- 3.) Determine the value  $C$  must have in order for the company to break even on all such policies (that is, to average a net gain of zero per policy on such policies).

**Solution :**  $C \geq 2625$

■

- 4.) Determine the value  $C$  must have in order for the company to average a net gain of \$250 per policy on all such policies.

**Solution :**  $C \geq 2875$

■

**Example 5.3.9** A roulette wheel has 38 slots. Thirty-six slots are numbered from 1 to 36; half of them are red and half are black. The remaining two slots are numbered 0 and 00 and are green. In a \$1 bet on red, the bettor pays \$1 to play. If the ball lands in a red slot, he receives back the dollar he bet plus an additional dollar. If the ball does not land on red he loses his dollar. Let  $X$  denote the net gain to the bettor on one play of the game.

1.) Construct the probability distribution of  $X$ .

**Solution :**

$x$	$-1$	$1$
$f(x)$	$\frac{20}{38}$	$\frac{18}{38}$

■

2.) Compute the expected value  $E(X)$  of  $X$ , and interpret its meaning in the context of the problem.

**Solution :**  $E(X) = -0.0526$ . *In many bets the bettor sustains an average loss of about 5.25 cents per bet.*

■

3.) Compute the standard deviation of  $X$ .

**Solution :** 0.9986

■

**Example 5.3.10** The time, to the nearest whole minute, that a city bus takes to go from one end of its route to the other has the probability distribution shown. As sometimes happens with probabilities computed as empirical relative frequencies, probabilities in the table add up only to a value other than 1.00 because of round-off error.

$x$	42	43	44	45	46	47
$f(x)$	0.10	0.23	0.34	0.25	0.05	0.02

1.) Find the average time the bus takes to drive the length of its route.

**Solution :** 43.54

■

2.) Find the standard deviation of the length of time the bus takes to drive the length of its route.

**Solution :** 1.2046

■

**Example 5.3.11** The number  $X$  of nails in a randomly selected 1-pound box has the probability distribution shown. Find the average number of nails per pound.

$x$	100	101	102
$f(x)$	0.01	0.96	0.03

**Solution :** 101.02

■

**Example 5.3.12** Two fair dice are rolled at once. Let  $X$  denote the difference in the number of dots that appear on the top faces of the two dice. Thus for example if a one and a five are rolled,  $X = 4$ , and if two sixes are rolled,  $X = 0$ .

1.) Construct the probability distribution for  $X$ .

**Solution :**

$x$	0	1	2	3	4	5
$f(x)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

■

- 2.) Compute the mean  $\mu$  of  $X$ .

**Solution :** 1.9444

■

- 3.) Compute the standard deviation  $\sigma$  of  $X$ .

**Solution :** 1.4326

■

**Example 5.3.13** A manufacturer receives a certain component from a supplier in shipments of 100 units. Two units in each shipment are selected at random and tested. If either one of the units is defective the shipment is rejected. Suppose a shipment has 5 defective units.

- 1.) Construct the probability distribution for the number  $X$  of defective units in such a sample. (A tree diagram is helpful).

**Solution :**

$x$	0	1	2
$f(x)$	0.902	0.096	0.002

■

- 2.) Find the probability that such a shipment will be accepted.

**Solution :** 0.902

■

**Example 5.3.14** The owner of a proposed outdoor theater must decide whether to include a cover that will allow shows to be performed in all weather conditions. Based on projected audience sizes and weather conditions, the probability distribution for the revenue  $X$  per night if the cover is not installed is

<i>Weather</i>	$x$	$f(x)$
<i>Clear</i>	\$3,000	0.61
<i>Threatening</i>	\$2,800	0.17
<i>Light Rain</i>	\$1,975	0.11
<i>Show – cancelling rain</i>	\$0	0.11

The additional cost of the cover is \$410,000. The owner will have it built if this cost can be recovered from the increased revenue the cover affords in the first ten 90-night seasons.

- 1.) Compute the mean revenue per night if the cover is not installed.

**Solution :** 2523.25

■

- 2.) Use the answer to (1.) to compute the projected total revenue per 90-night season if the cover is not installed.

**Solution :** 227,092.5

■

- 3.) Compute the projected total revenue per season when the cover is in place. To do so assume that if the cover were in place the revenue each night of the season would be the same as the revenue on a clear night.

**Solution :** 270,000



- 4.) Using the answers to (2.) and (3.), decide whether or not the additional cost of the installation of the cover will be recovered from the increased revenue over the first ten years. Will the owner have the cover installed?

**Solution :** *The owner will install the cover.*



## 5.4 Continuous Random Variables

In the continuous situation we shall extend our consideration to random variables whose range space elements cannot be listed individually but can be defined within an interval of values.

**Definition 5.4.1** A continuous random variable  $X$  is one where the random outcomes can take uncountable (infinite) values.

**Definition 5.4.2** Random variable  $X$  is a continuous random variable if there is a function  $f : \mathbb{R} \rightarrow [0, \infty >$  such that

$$P(X \leq a) = \int_{-\infty}^a f(t)dt, \quad \forall a \in \mathbb{R}$$

Function is called the **probability density function** or PDF.

**Definition 5.4.3** Let  $X$  be such a random variable (of continuous type). Then the non negative function  $f(x)$  defined on an interval will have a probability given by

$$\rho(x) = \int_{-\infty}^{\infty} f(x)dx.$$

The function  $f(x)$  is called the probability density function (p.d.f) of the random variable  $X$ .

**Remark 5.4.1** Since the continuous random variable cannot take on individual values, we consequently cannot express its outcomes in tabular form as we did to the discrete type.

**Remark 5.4.2** A continuous random variable  $X$  is specified by its probability density function  $f(x)$ , where  $f(x) \geq 0$ ; through the range of values for which  $x$  is valid. This probability density function can be represented by a curve and the area under this curve represents the probabilities.

### 5.4.1 Properties of Continuous Random Variables

Let  $x$  be a continuous random variable and  $f(x)$  be the corresponding *pdf* then we have the following definition.

**Definition 5.4.4** For a continuous random variable with a probability density function  $f(x)$

$$P(a \leq X \leq b) = \int_a^b f(x)dx \quad (5.15)$$

**Note 5.4.1** Consequently, for continuous random variable  $X$ , we have  $\leq \equiv <$  and  $\geq \equiv >$

$$\begin{aligned} P(X \leq a) = P(X < a) &= \int_{-\infty}^a f(x)dx \\ P(X \geq b) = P(X > b) &= \int_b^{\infty} f(x)dx \\ P(X = a) &= \int_a^a f(x)dx = 0 \end{aligned}$$

**Example 5.4.1** Given a random variable  $X$  with a probability density function

$$f(x) = c(4x - 2x^2), \quad 0 \leq x < 2$$

- 1.) Find the value of  $c$
- 2.) Verify that  $f(x)$  is indeed a probability density function
- 3.) Find  $P(X > 1)$ .

So for the probability density function given in Example 5.4.1 above

- 1.) Since  $f(x)$  is a probability density function, then

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)dx &= 1 \\ \int_0^2 c(4x - 2x^2)dx &= 1 \\ c \left[ 2x^2 - \frac{2}{3}x^3 \right]_0^2 &= 1 \\ \Rightarrow c &= \frac{3}{8}\end{aligned}$$

- 2.) To verify that  $f(x)$  is indeed a *pdf* we need to show that within  $0 < x < 2$

- a)  $f(x) \geq 0$

$$f(x) = \frac{3}{8}(4x - 2x^2), \quad f(0) = 0, \quad f(2) = 0 \Rightarrow f(x) \geq 0$$

- b)  $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\int_0^2 \frac{3}{8}(4x - 2x^2)dx = \frac{3}{8} \left[ 2x^2 - \frac{2}{3}x^3 \right]_0^2 = 1$$

Hence  $f(x)$  is a *pdf*

- 3.)

$$P(X > 1) = \frac{3}{8} \int_1^2 (4x - 2x^2)dx = \frac{1}{2}$$



**Example 5.4.2** A continuous random variable  $X$  has a *pdf* defined by

$$f(x) = \begin{cases} kx(16 - x^2) & ; 0 < x < 4 \\ 0 & ; \text{elsewhere} \end{cases}$$

1.) Find the constant  $k$

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) dx &= 1 \\ \int_0^4 kx(16 - x^2) dx &= 1 \\ k \left[ 8x^2 - \frac{1}{4}x^4 \right]_0^4 &= 1 \\ \Rightarrow k &= \frac{1}{64} \\ \Rightarrow f(x) &= \begin{cases} \frac{1}{64} (16x - x^3) & ; 0 < x < 4 \\ 0 & ; \text{elsewhere} \end{cases} \end{aligned}$$

2.) Evaluate

(a)  $P(1 < X < 2)$

$$\begin{aligned} P(1 < X < 2) &= \int_1^2 f(x) dx \\ &= \frac{1}{64} \int_1^2 (16x - x^3) dx \\ &= \frac{1}{64} \left[ 8x^2 - \frac{1}{4}x^4 \right]_1^2 \\ &= \frac{81}{256} \end{aligned}$$

(b)  $P(X \geq 3)$

$$\begin{aligned} P(X \geq 3) &= \int_3^4 f(x) dx \\ &= \frac{1}{64} \int_3^4 (16x - x^3) dx \\ &= \frac{1}{64} \left[ 8x^2 - \frac{1}{4}x^4 \right]_3^4 \\ &= \frac{1}{64} \left[ (128 - 64) - \left( 72 - \frac{81}{4} \right) \right] \\ &= \frac{49}{456} \end{aligned}$$

(c)  $P(X < 2 \mid 1 < X < 3)$ 

$$\begin{aligned}P(X < 2 \mid 1 < X < 3) &= \frac{P\{(1 < X < 3) \cap (X < 2)\}}{P(1 < X < 3)} \\&= \frac{P(1 < X < 2)}{P(1 < X < 3)} \\&= \frac{\frac{1}{64} \int_1^2 (16x - x^3) dx}{\frac{1}{64} \int_1^3 (16x - x^3) dx} \\&= \frac{\frac{1}{64} \left[ 8x^2 - \frac{1}{4}x^4 \right]_1^2}{\frac{1}{64} \left[ 8x^2 - \frac{1}{4}x^4 \right]_1^3} \\&= \frac{81}{176}\end{aligned}$$

### 5.4.2 The Cumulative Distribution Function

The cumulative distribution function *cdf* of a continuous random variable  $X$  with a *pdf*  $f(x)$  is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt$$

As an immediate consequence, we note that for a continuous random variable  $X$

1.)

$$P(a < X < b) = F(b) - F(a),$$

and

2.)

$$f(x) = \frac{d}{dx}[F(x)],$$

if the derivative exists.

**Example 5.4.3** A Random variable  $X$  has a p.d.f defined by

$$f(x) = \begin{cases} \frac{1}{3}x^2 & ; -1 < x < 2 \\ 0 & ; \text{elsewhere} \end{cases}$$

Find the cumulative distribution function  $F(x)$

$$F(x) = \int_{-\infty}^x \frac{t^2}{3}dt = \int_{-1}^x \frac{t^2}{3}dt = \left[ \frac{t^3}{9} \right]_{-1}^x = \frac{x^3 + 1}{9}$$

Thus

$$F(x) = \begin{cases} 0 & ; x < -1 \\ \frac{x^3 + 1}{9} & ; -1 \leq x < 2 \\ 1 & ; x \geq 2 \end{cases}$$

For example,

$$\begin{aligned} P(0 < X < 2) &= P(X < 2) - P(X \leq 0) \\ &= P(X \leq 2) - P(X \leq 0) \quad \text{continuous random variable} \\ &= F(2) - F(0) \\ &= 1 - \frac{1}{9} \\ &= \frac{8}{9} \end{aligned}$$

which is the same as

$$\int_0^2 \frac{x^2}{3}dx = \frac{8}{9}$$

**Example 5.4.4** A continuous random variable  $X$  has a *pdf* defined by

$$f(x) = \begin{cases} x - 2 & ; 2 \leq x \leq 3 \\ 4 - x & ; 3 \leq x \leq 4 \\ 0 & ; \text{otherwise} \end{cases}$$

obtain the cumulative distribution function  $F(x)$

Case 1: For  $2 \leq x < 3$

$$F_1(x) = \int_2^x (t - 2)dt = \left[ \frac{1}{2}t^2 - 2t \right]_2^x = \frac{1}{2}x^2 - 2x + 2$$

Case 2: At  $x = 3$ ,

$$F_1(3) = \frac{1}{2}(9) - 2(3) + 2 = \frac{1}{2}$$

Case 3: For  $3 \leq x < 4$ ,  $f(x) = 4 - x$ , and we have

$$\begin{aligned} F_2(x) &= F_1(3) + \int_3^x (4 - t)dt. \\ &= \frac{1}{2} + \left[ 4t - \frac{1}{2}t^2 \right]_3^x \\ &= \frac{1}{2} + 4x - \frac{x^2}{2} - 12 + \frac{9}{2} \\ &= 4x - \frac{x^2}{2} - 7 \end{aligned}$$

Case 4: At  $x = 4$ ,

$$F_2(4) = 4(4) - \frac{16}{2} - 7 = 1$$

**Thus**, the CDF,  $F(x)$  for the  $f(x)$  is given by

$$F(x) = \begin{cases} 0 & ; x < 2 \\ \frac{1}{2}x^2 - 2x + 2 & ; 2 \leq x < 3 \\ 4x - \frac{x^2}{2} - 7 & ; 3 \leq x < 4 \\ 1 & ; x \geq 4 \end{cases}$$

**Example 5.4.5** The cumulative distribution function of a random variable  $X$  is given by

$$F(x) = 1 - \frac{1}{\pi} \arccos x, -1 < x < 1$$

Find the probability density function  $f(x)$ .

$$\begin{aligned} f(x) &= \frac{d}{dx}[F(x)] = \frac{d}{dx} \left[ 1 - \frac{1}{\pi} \arccos x \right] = \frac{1}{\pi \sqrt{1-x^2}} \\ &= \begin{cases} \frac{1}{\pi \sqrt{1-x^2}} & ; -1 < x < 1 \\ 0 & ; \text{otherwise} \end{cases} \end{aligned}$$

**Exercise 5.7** Let  $X$  represent the outcome when a honest die is tossed. Find  $E(Y)$  where  $Y = 2X^2 - 5$ .

**Exercise 5.8** A discrete random variable  $X$  has a probability distribution defined by

$x$	0.5	1	1.5	2
$P(X = x)$	$p$	$p^2$	$2p^2$	$p$

- 1.) Find the value of  $p$
- 2.) Find  $P(X > 1)$
- 3.) Find  $P(1 \leq X < 2)$

**Exercise 5.9** A box contains four red and six black balls. If a random sample of three balls is to be drawn from the box without replacement find the expected number of black balls in the sample. [Ans:1.8]

**Exercise 5.10** From a box, containing five coins, four of which are marked 1 and one 24 a person selects two coins at random without replacement. If the person is to receive the sum of the two respective amounts, compute the expectation of payment. [Ans. 11.20]

**Exercise 5.11** A discrete random variable has a probability distribution defined by

$$f(x) = \frac{x}{k}, \quad x = 1, 2, \dots, n$$

Given that the mean is  $7/3$ , find the values of  $k$  and  $n$ . Find the variance, cumulative function of  $X$  and  $P(X = 2 \mid X \geq 2)$

**Exercise 5.12** A discrete random variable  $X$  has a probability distribution given by

$$P(X = x) = \begin{cases} 0.1x & ; x = 1, 2, 3, 4 \\ 0 & ; \text{elsewhere} \end{cases}$$

Find  $E(X)$  and  $P(X < 3 \mid 2 \leq X \leq 4)$ .

**Exercise 5.13** A continuous random variable  $X$  has a p.d.f given by

$$f(x) = \begin{cases} ax & ; 0 \leq x \leq 1 \\ a & ; 1 \leq x \leq 2 \\ 3a - ax & ; 2 \leq x \leq 3 \\ 0 & ; \text{otherwise} \end{cases}$$

- 1.) Determine the value of  $a$ .
- 2.) Determine  $P(X \leq 1.5)$
- 3.) Find  $F(x)$ . Sketch  $f(x)$  and  $F(x)$ .

**Exercise 5.14** A random variable  $X$  has a density function given by

$$f(x) = \begin{cases} k(x^2 - 2x + 1) & ; 0 \leq x \leq 1 \\ 0 & ; \text{elsewhere} \end{cases}$$

- 1.) Determine  $k$
- 2.) Graph  $f(x)$

**Exercise 5.15** Let a ball be picked from a bowl that contains six white, three red and one blue balls. Let a random variable  $X = 1$  if the outcome is white,  $X = 5$  if the outcome is red and  $X = 10$  if the outcome is blue.

- 1.) Find the  $df$  of  $X$
- 2.) Graph the  $pdf$
- 3.) Find  $F(x)$  and graph it.

**Exercise 5.16** The probability density function of  $f(x)$ , the lifetime of a certain type of electronic device (measured in hours), is given by

$$f(x) = \begin{cases} \frac{10}{x^2} & ; x > 10 \\ 0 & ; x \leq 10 \end{cases}$$

- 1.) Find  $P(X > 20)$
- 2.) What is the cumulative distribution function of  $X$ ?

### 5.4.3 Expectation of a Continuous Random Variable

**Definition 5.4.5** For a continuous random variable  $X$ , we define the mean of  $X$  or expectation of  $X$  denoted by  $E(X)$  as

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \quad (5.16)$$

In there is a new random variable  $g(X)$  defined by a function  $g(X)$ , then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

### 5.4.4 Variance of a Continuous Random Variable

**Definition 5.4.6** The variance of a continuous random variable  $X$  is defined by

$$\begin{aligned} \text{Var}(X) &= E(X^2) - [E(X)]^2 \\ &= \int_{-\infty}^{\infty} x^2 f(x)dx - \left[ \int_{-\infty}^{\infty} x f(x) dx \right]^2 \end{aligned}$$

### 5.4.5 Median of A continuous Random Variable

**Definition 5.4.7** Let  $m$  be the median of a random variable  $X$  and  $m \in [-\infty, \infty]$ , then

$$\int_{-\infty}^m f(x)dx = \frac{1}{2} \quad (5.17)$$

Observe that  $m$  must lie within the interval defined.

**Exercise 5.17** A random variable  $Y$  has *pdf* defined by

$$f(y) = \begin{cases} \frac{1}{8}(y+1) & ; 2 < y < 4 \\ 0 & ; \text{elsewhere} \end{cases}$$

Find

- 1.)  $P(Y < 3.2)$  and  $P(2.9 < Y < 3.2)$
- 2.) the cumulative distribution function of  $Y$
- 3.)  $E(Y)$  and variance.

**Example 5.4.6** A random variable  $X$  has a *pdf*

$$f(x) = \begin{cases} x & ; 0 < x < 1 \\ 2-x & ; 1 \leq x \leq 2 \\ 0 & ; \text{elsewhere} \end{cases}$$

$$P(0.8 < X < 1.2) = \int_{0.8}^1 x dx + \int_1^{1.2} (2-x) dx$$

**Example 5.4.7** A continuous random variable  $X$  has a p.d.f defined by

$$f(x) = \begin{cases} 2x & ; 0 < x < 1 \\ 0 & ; \text{otherwise} \end{cases}$$

1.)  $E(X)$                       2.)  $Var(X)$                       3.)  $F(X)$                       4.) Median

1.)

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_0^1 2x^2 dx = \left[ \frac{2x^3}{3} \right]_0^1 = \frac{2}{3} \\ E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx \\ &= \int_0^1 2x^3 dx = \left[ \frac{1}{2} x^4 \right]_0^1 = \frac{1}{2} \end{aligned}$$

2.)

$$\begin{aligned} Var(X) &= E(X^2) - [E(X)]^2 \\ &= \frac{1}{2} - \frac{4}{9} \\ &= \frac{1}{18} \end{aligned}$$

3.)

$$\begin{aligned} F(X) &= \int_{-\infty}^x f(t) dt \\ &= \int_0^x 2t dt = [t^2]_0^x = x^2 \end{aligned}$$

therefore,

$$F(X) = \begin{cases} 0 & ; x \leq 0 \\ x^2 & ; 0 < x < 1 \\ 1 & ; x \geq 1 \end{cases}$$

4.) For Median  $m$

$$\begin{aligned} \int_{-\infty}^m f(x) dx &= \frac{1}{2} \\ \int_0^m 2x dx &= \frac{1}{2} \\ [x^2]_0^m &= \frac{1}{2} \\ \Rightarrow m &= \pm \frac{1}{\sqrt{2}} \end{aligned}$$



Since  $0 < x < 1$  then  $m = 1/\sqrt{2}$ . If the density function  $f(x)$  has more than one interval, we first integrate  $f(x)$  within the first interval. If the answer to this is less than  $\frac{1}{2}$  then this interval does not contain the median. We next add the answer from the integral of the first function to the integral of the second function up to  $m$ . the sum which should be  $\frac{1}{2}$ .

**Note 5.4.2** Observe that, always the value to  $\text{Var}(X)$  should be positive.

**Example 5.4.8** Determine the median of the distribution given a *pdf* defined by

$$f(x) = \begin{cases} \frac{1}{2}x & ; 0 \leq x \leq 1 \\ \frac{3}{2} - \frac{1}{2}x & ; 1 \leq x \leq 2 \\ 0 & ; \text{elsewhere} \end{cases}$$

We realize that

$$\int_0^1 \frac{1}{2}x dx = \left[ \frac{x^2}{4} \right]_0^1 = \frac{1}{4}$$

Since the answer is less than  $\frac{1}{2}$ , we move to the next interval and check. Thus

$$\begin{aligned} \int_{-\infty}^m f(x) dx &= \frac{1}{2} \\ \int_0^1 \frac{1}{2}x dx + \int_1^m \left( \frac{3}{2} - \frac{1}{2}x \right) dx &= \frac{1}{2} \\ \frac{1}{4} + \frac{3}{2}m - \frac{1}{4}m^2 - \frac{3}{2} + \frac{1}{4} &= \frac{1}{2} \\ m^2 - 6m + 6 &= 0 \end{aligned}$$

Then the value of  $m$  that lies within the interval is computed from the equation.

**Example 5.4.9** Given

$$f(x) = \begin{cases} x & ; 0 \leq x \leq 1 \\ k - x & ; 1 \leq x \leq 2 \\ 0 & ; \text{elsewhere} \end{cases}$$

- 1.) Find the constant  $k$  [Ans:  $k = 2$ ]
- 2.)  $P\left(\frac{1}{2} \leq X \leq \frac{3}{2}\right)$  [Ans:  $3/4$ ]
- 3.)  $P\left(X \leq \frac{1}{4}\right)$  [Ans:  $1/32$ ]
- 4.)  $E(X)$  [Ans:  $1$ ]
- 5.)  $\text{Var}(X)$  [Ans:  $1/6$ ]

## 5.5 Continuous Random Variables Chapter Examples

**Proposition 5.5.1** Let  $X$  be a continuous random variable with density  $f$ .

1.  $\forall a, b \in \mathbb{R}, a \leq b$  we have

$$P(a < X \leq b) = P(\{X \leq b\} \setminus \{X \leq a\}) = P(X \leq b) - P(X \leq a) = \int_a^b f(t)dt$$

2. All probabilities in measured space and up to 1. In continuous variables integral is interpreted as sum.

$$\int_{-\infty}^{+\infty} f(t)dt = P(X \in \mathbb{R}) = 1$$

3. Notice that for all  $a \in \mathbb{R}$

$$P(X = a) = \int_a^a f(t)dt = 0$$

Consequently,

$$P(X \leq a) = P(X < a) + P(X = a) = P(X < a)$$

Moreover,

$$P(a < X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a \leq X \leq b)$$

i.e. it's not important if the edge is included or not.

**Example 5.5.1** Let  $X$  be continuous random variable with PDF:

$$f(t) = \begin{cases} c \cdot t^4, & \text{if } 0 < t < 1. \\ 0, & \text{otherwise.} \end{cases}$$

- 1.) Find constant  $c$

**Solution :** *e know that  $\int_{-\infty}^{+\infty} f(t)dt = 1$ . Since the function has value 0 for every  $t$  outside of the interval  $(0, 1)$  the integral is equal to*

$$\int_0^1 c \cdot t^4 dt = 1$$

*Further, lets solve this integral by using the polynomial rule*

$$\frac{c}{5} \cdot t^5 \Big|_0^1 = 1 \Rightarrow c = 5$$

*Consequently, PDF looks like this*

$$f(t) = \begin{cases} 5 \cdot t^4, & \text{if } 0 < t < 1. \\ 0, & \text{otherwise.} \end{cases}$$

■

2.) Calculate  $P\left(X > \frac{1}{2}\right)$

**Solution :** Probability of  $X > \frac{1}{2}$  can be calculated in two ways. First way,

$$\begin{aligned} P\left(X > \frac{1}{2}\right) &= \int_{-\infty}^{+\infty} 5 \cdot t^4 dt - \int_{-\infty}^{\frac{1}{2}} 5 \cdot t^4 dt \\ &= 1 - \int_{-\infty}^{\frac{1}{2}} 5 \cdot t^4 dt \\ &= 1 - \left. \frac{5}{5} \cdot t^5 \right|_0^{\frac{1}{2}} = 1 - \left(\frac{1}{2}\right)^5 = 0.96875 \end{aligned}$$

Second way,

$$P\left(X > \frac{1}{2}\right) = P\left(\frac{1}{2} < X < +\infty\right)$$

In this case  $a = \frac{1}{2}$  and  $b = +\infty$ . Consequently, integral is equal to

$$\int_{\frac{1}{2}}^{+\infty} 5 \cdot t^4 dt$$

Moreover, our function is zero outside of the interval  $< 0, 1 >$ . Because of that we can rewrite our integral as

$$\int_{\frac{1}{2}}^1 5 \cdot t^4 dt$$

Finally, we solve the given integral

$$\int_{\frac{1}{2}}^1 5 \cdot t^4 dt = \left. \frac{5}{5} \cdot t^5 \right|_{\frac{1}{2}}^1 = 1 - \left(\frac{1}{2}\right)^5 = 0.96875$$

■

3.)  $E[X]$  and  $\text{Var}(X)$ .

**Solution :** We've learned before that expected value is equal to

$$E[X] := \int_{-\infty}^{+\infty} t f(t) dt$$

We simply put our PDF and it's edges into the given formula and solve the integral. It looks like this:

$$E[X] := \int_0^1 t \cdot 5 \cdot t^4 dt = 5 \cdot \int_0^1 t^5 dt = \frac{5}{6}$$

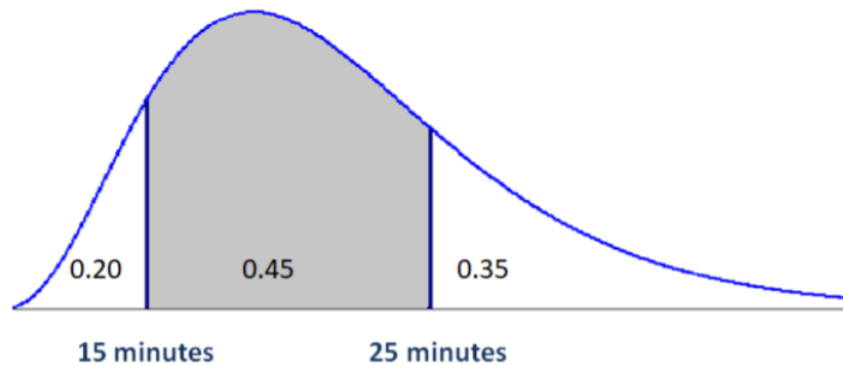
$$E[X^2] := 5 \int_0^1 t^2 \cdot t^4 dt = 5 \cdot \int_0^1 t^6 dt = \frac{5}{7}$$

Finally

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{5}{7} - \left(\frac{5}{6}\right)^2 = 0.0198413$$

■

**Example 5.5.2** The time to drive to school for a community college student is an example of a continuous random variable. The probability density function and areas of regions created by the points 15 and 25 minutes are shown in the graph.



- 1.) Find the probability that a student takes less than 15 minutes to drive to school.

**Solution :**

$$P(X < 15) = 0.20$$

■

- 2.) Find the probability that a student takes no more than 15 minutes to drive to school.

**Solution :** *This answer is the same as the prior question, because points have no probability with continuous random variables.*

$$P(X \leq 15) = 0.20$$

■

- 3.) Find the probability that a student takes more than 15 minutes to drive to school.

**Solution :**

$$P(X > 15) = 0.45 + 0.35 = 0.80$$

■

- 4.) Find the probability that a student takes between 15 and 25 minutes to drive to school.

**Solution :**

$$P(15 \leq X \leq 25) = 0.45$$

■

## 5.6 Probability Distribution Functions

Yes now we think we can almost find the probabilities of all events, but maybe not; for example we can draw sample space for throwing a die twice, but what of throwing it 3 or 9 or 1000 times??, Or what if selecting 25 balls from a bag containing 20 different types/colors of balls??, these types of sample spaces are very very very hard if not impossible using sample spaces. What to do then????!!! There are some functions which can act as the sample spaces, the Distribution functions.

# Chapter 6

## Common Discrete Distributions

### 6.1 The Binomial Distribution

#### Introduction

Often an experiment may consist of repeated trials each with two possible outcomes which we may label success or failure. This is true in testing items, say, of an assembly line, where each test or trial may include a defective or non defective item.. We may decide to choose any of these as success. Also for a deck of cards if five cards are drawn in succession and each trial is labeled either a success or failure depending on whether the card is red or black and this is repeated several times and then note that each of these trials are independent of one another and the probability of success remains the same. We shall refer to these types of experiments as binomial.

#### 6.1.1 Binomial Experiment

**Definition 6.1.1** If an experiment has the following properties

- 1.) consists of  $n$  repeated trials
- 2.) each trial results in an outcome that may be classified as a success or failure.
- 3.) the probability of success, denoted by  $p$  remains constant from trial to trial (or at each trial)
- 4.) the repeated trials are independent then it is known as a binomial experiment.

**Definition 6.1.2** The number  $X$ , of successes in  $n$  independent trials of a binomial experiment is called a binomial random-variable. this is the simplest and most important of the special discrete random variables. The distribution of the random variable  $X$  that can take up the stated properties is known as the binomial distribution. If  $X$  is the number of successes in  $n$  independent trials of a binomial experiment with  $n$  trials and  $p$  the probability of success, then we say that it is a binomial experiment with parameters  $n$  and  $p$ , and is written as  $b(x, n, p)$  or  $b(n, p)$ .

**Definition 6.1.3** Bernoulli trial: If the experiment is conducted only once, that is, if  $n = 1$  then  $X$  is often called a Bernoulli trial (after a Swiss Mathematician, James Bernoulli(1654 – 1705)). A Bernoulli trial also must satisfy the properties of a binomial experiment. Thus a Bernoulli random variable is a binomial random variable with parameters  $(1, p)$ . Observe that if  $p$  is the probability of success then  $q = 1 - p$  is the probability of failure.

### 6.1.2 Binomial Distribution

We have already noted that the probability distribution of a binomial random variable  $X$  is called a binomial distribution and is denoted by  $b(x, n, p)$ .

**Definition 6.1.4** If a binomial experiment can result in a success with probability  $p$  and a failure with probability  $q = 1 - p$ , then the probability distribution of a binomial random variable  $X$ , with  $x$  number of success in  $n$  independent trials, is

$$P(X = x) = b(x, n, p) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \dots, n \quad (6.1)$$

where the combination  $\binom{n}{x}$  is given by

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} = {}^nC_x$$

**Note 6.1.1** Whenever any experiment is done more than once, i.e  $n \geq 2$ , then we apply Binomial. E.g a coin/die thrown 5 times

- 1.)  $p, q$  are for a single throw/ toss/ experiment.
- 2.)  $n$  is the number of times it's repeated (total number of the experiment)
- 3.)  $x$  is the question (the number of the event)

**Definition 6.1.5** The probability that a binomial random variable  $X$  takes up a value, say  $a$  is given by

$$P(X = a) = b(a, n, p) = {}^nC_a \cdot p^a \cdot q^{n-a} \quad (6.2)$$

**Example 6.1.1** If  $p$  is the probability of success and  $q = 1 - p$  is the probability of failure, find the probability of 0, 1, 2, ..., 5 successes in 5 independent trials of an experiment.

$$\begin{aligned} P(X = x) &= {}^nC_x p^x q^{n-x}, \quad x = 0, 1, 2, \dots, 5. \quad \text{So } n = 5 \text{ gives} \\ P(X = 0) &= {}^5C_0 p^0 q^5 = q^5 \\ P(X = 1) &= {}^5C_1 p^1 q^4 = 5q^4 p \\ P(X = 2) &= {}^5C_2 p^2 q^3 = 10q^3 p^2 \\ P(X = 3) &= {}^5C_3 p^3 q^2 = 10q^2 p^3 \\ P(X = 4) &= {}^5C_4 p^4 q^1 = 5qp^4 \\ P(X = 5) &= {}^5C_5 p^5 q^0 = p^5 \end{aligned}$$

But  $q^5, 5q^4p, \dots, p^5$  are terms of the binomial expansion of  $(q + p)^5$ , in which

$$(q + p)^5 = q^5 + 5q^4p + 10q^3p^2 + 10q^2p^3 + 5qp^4 + p^5$$

or

$$1 = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$

Then in general we note that the values  $P(X = x)$ , for  $x = 0, 1, 2, \dots, n$  can be obtained by considering the terms in the binomial expansion of  $(q + p)^n$  noting that  $(q + p) = 1$  Thus

$$\begin{aligned} (q + p)^n &= {}^nC_0 q^n p^0 + {}^nC_1 q^{n-1} p^1 + {}^nC_2 q^{n-2} p^2 + \dots + {}^nC_r q^{n-r} p^r + \dots + {}^nC_n q^0 p^n \\ \Rightarrow 1 &= P(X = 0) + P(X = 1) + P(X = 2) + \dots + P(X = r) + \dots + P(X = n) \end{aligned}$$

And if  $X$  is distributed in this way then it is said to be binomially distributed, with parameters  $n$  and  $p$ .

**Example 6.1.2** The probability that a kind of component survives a given shock test is  $\frac{3}{4}$ . Find the probability that exactly two of the next four components tested survived.

Observe that this is a binomial experiment. We are given

- Probability of success  $p = 3/4$
- 4 repeated trials of tests
- $x = 2$  number of success in 4 trials.

Then

$$P(X = 2) = b\left(2, 4, \frac{3}{4}\right) = {}^4C_2 \left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right)^2 = \frac{4!}{2!2!} \left[\frac{3^2}{4^4}\right] = \frac{27}{128}$$

**Example 6.1.3** Five coins are tossed. If the outcomes are assumed independent find the probability that exactly 3 heads will appear.

This is a binomial experiment in which

- 5 repeated trials
- probability of a success  $p = 1/2$  [a head to appear in one experiment, one throw]

Then

$$P(X = 3) = {}^5C_3 \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^2 = (10) \cdot \left(\frac{1}{2^5}\right) = \frac{10}{32} = \frac{5}{16}$$

**Example 6.1.4** The probability that a patient recovers from a rare blood disease is 0.4. If 15 people are known to have contracted the disease, what is the probability that

1.) exactly 5 survive?

Let  $X$  be the number of people that survive. Then  $p = 0.4, n = 15, x = 5$  and

$$P(x = 5) = b(5, 15, 0.4) = {}^{15}C_5 \cdot (0.4)^5 \cdot (0.6)^{10} = 0.1859$$

2.) at least 10 survive

$$\begin{aligned} P(x \geq 10) &= P(X = 10) + P(X = 11) + P(X = 12) + P(X = 13) + P(X = 14) + P(X = 15) \\ &= {}^{15}C_{10} \cdot (0.4)^{10} \cdot (0.6)^5 + {}^{15}C_{11} \cdot (0.4)^{11} \cdot (0.6)^4 + {}^{15}C_{12} \cdot (0.4)^{12} \cdot (0.6)^3 \\ &\quad + {}^{15}C_{13} \cdot (0.4)^{13} \cdot (0.6)^2 + {}^{15}C_{14} \cdot (0.4)^{14} \cdot (0.6)^1 + {}^{15}C_{15} \cdot (0.4)^{15} \cdot (0.6)^0 \\ &= 0.0338 \end{aligned}$$

**Example 6.1.5** In a certain town, the need for money to buy drugs is given as the reason for 75% of all thefts. What is the probability that exactly 2 of the next 4 theft cases reported resulted from the need for money?

Let  $X$  be the number of thefts successful, then

$$\begin{aligned} n &= 4 \text{ (number of thefts)} \\ x &= 2 \text{ (number of successes)} \\ p &= 75\% = \frac{3}{4} \\ P(X = 2) &= b\left(2, 4, \frac{3}{4}\right) = {}^4C_2 \cdot \left(\frac{3}{4}\right)^2 \cdot \left(\frac{1}{4}\right)^2 = 0.2109 \end{aligned}$$

### 6.1.3 Reading Binomial Tables

Frequently, we are interested in problems where it is necessary to find ( $P(X < a)$ ,  $P(a \leq X \leq b)$  or  $P(X > b)$ .) Fortunately, binomial sums are available in statistical tables. To obtain the desired probabilities from tables we use the following definitions:

1.)  $P(X \leq a) = \sum_{x=0}^a b(x, n, p)$

2.)

$$\begin{aligned} P(X \geq a) &= 1 - P(X < a) \\ &= 1 - \sum_{x=0}^{a-1} b(x, n, p) \end{aligned}$$

3.)

$$\begin{aligned} P(X = a) &= b(a, n, p) \\ &= P(X \leq a) - P(X \leq a - 1) \\ &= \sum_{x=0}^a b(x, n, p) - \sum_{x=0}^{a-1} b(x, n, p) \end{aligned}$$

4.)

$$\begin{aligned} P(a \leq x \leq b) &= \sum_{x=a}^b b(x, n, p) \\ &= P(x \leq b) - P(x < a) \\ &= \sum_{x=0}^b b(x, n, p) - \sum_{x=0}^{a-1} b(x, n, p) \end{aligned}$$

5.)

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a)$$

6.)

$$P(a \leq X < b) = P(X < b) - P(X < a)$$

7.)

$$P(a < X < b) = P(X < b) - P(X \leq a)$$

8.)

$$P(X < a) = P(X \leq a - 1) = \sum_{x=0}^{a-1} b(x, n, p)$$

9.)

$$P(X > a) = 1 - P(X \leq a) = 1 - \sum_{x=0}^a b(x, n, p)$$



**Example 6.1.6** A new car dealer knows from past experience that on the average, she will make sale to about 20 of her 100 customers. What is the probability that in five randomly selected presentations, she makes a sale to

- 1.) exactly three customers    2.) at most one customer    3.) at least 2 customers

It is a Binomial distribution with  $p = 20\% = 0.2, n = 5$

- 1.) exactly three

$$\begin{aligned}P(X = 3) &= b(3, 5, 0.2) \\&= \sum_{x=0}^3 b(x, 5, 0.2) - \sum_{x=0}^2 b(x, 5, 0.2) \\&= 0.9933 - 0.9421 \\&= 0.0512\end{aligned}$$

- 2.) at most one

$$P(X \leq 1) = \sum_{x=0}^1 b(x, 5, 0.2) = 0.7373$$

- 3.) at least two

$$\begin{aligned}P(X \geq 2) &= 1 - P(X < 2) \\&= 1 - P(X \leq 1) \\&= 1 - \sum_{x=0}^1 b(x, 5, 0.2) \\&= 1 - 0.7373 \\&= 0.2627\end{aligned}$$

**Exercise 6.1** A die is thrown four times. Find the probability that a 5 is obtained each throw.

**Exercise 6.2** Repeat Example 6.1.4 using the Binomial Table

- 1.) exactly 5 survive?

$$\begin{aligned}P(x = 5) &= P(X \leq 5) - P(X \leq 4) \\&= \sum_{x=0}^5 b(x, 15, 0.4) - \sum_{x=0}^4 b(x, 15, 0.4) \\&= 0.4032 - 0.2173 = 0.1859\end{aligned}$$

- 2.) at least 10 survive

$$\begin{aligned}P(x \geq 10) &= 1 - P(X \leq 9) \\&= 1 - \sum_{x=0}^9 b(x, 15, 0.4) \\&= 1 - 0.9662 = 0.0338\end{aligned}$$

**Example 6.1.7** The probability that a patient recovers from a rare blood disease is 0.4. If 15 people are known to have contracted the disease, what is the probability that

- |                                       |                        |
|---------------------------------------|------------------------|
| 1.) at least 10 survive               | 4.) exactly 4 survive? |
| 2.) between 3 and 8 inclusive survive |                        |
| 3.) exactly 5 survive?                | 5.) exactly 10 die?    |

$n = 15, p = 0.4$ , therefore,

1.)

$$\begin{aligned}P(X \geq 10) &= 1 - P(X < 10) \\&= 1 - P(X \leq 9) \\&= 1 - \sum_{x=0}^9 b(x, 15, 0.4) \\&= 1 - 0.9662 = 0.0338\end{aligned}$$

2.)

$$\begin{aligned}P(3 \leq X \leq 8) &= \sum_{x=3}^8 b(x, 15, 0.4) \\&= P(X \leq 8) - P(X < 3) \\&= P(X \leq 8) - P(X \leq 2) \\&= \sum_{x=0}^8 b(x, 15, 0.4) - \sum_{x=0}^2 b(x, 15, 0.4) \\&= 0.9050 - 0.0271 \\&= 0.8779\end{aligned}$$

3.)

$$\begin{aligned}P(X = 5) &= b(5, 15, 0.4) \\&= \sum_{x=0}^5 b(x, 15, 0.4) - \sum_{x=0}^4 b(x, 15, 0.4) \\&= 0.4032 - 0.2173 \\&= 0.1859\end{aligned}$$

**Exercise 6.3** Answer parts (4) and (5) for Example 6.1.7 using both the probability mass function  $f(x)$  and the Binomial tables. Compare the two solutions for each part.

**Exercise 6.4** Repeat Example 6.1.7 using the Binomial definition Equation 6.1 not the table. Compare the two methods in terms of accuracy and computational difficulty.

**Example 6.1.8** A fair coin is tossed 14 times. What is the probability that the head will appear

- |                      |  |
|----------------------|--|
| 1.) at least 8 times | 4.) less than 12 times                   |
| 2.) at most 6 times  |  |
| 3.) exactly 5 times  | 5.) less than 9 and greater than 4 times |

Let  $X$  = number of heads. With  $n = 14$ ,  $p = 0.5$

1.)

$$\begin{aligned}P(X \geq 8) &= 1 - P(X < 8) = 1 - P(X \leq 7) \\&= 1 - \sum_{x=0}^7 b(x, 14, 0.5) \\&= 1 - 0.6047 \\&= 0.3953\end{aligned}$$

2.)

$$P(X \leq 6) = \sum_{x=0}^6 b(x, 14, 0.5) = 0.3953$$

3.)

$$\begin{aligned}P(x = 5) &= P(X \leq 5) - P(X \leq 4) \\&= \sum_{x=0}^5 b(x, 14, 0.5) - \sum_{x=0}^4 b(x, 14, 0.5) \\&= 0.2121 - 0.0898 \\&= 0.1222\end{aligned}$$

4.)

$$P(X < 12) = P(X \leq 11) = \sum_{x=0}^{11} b(x, 14, 0.5) = 0.9935$$

5.)

$$\begin{aligned}P(4 < X < 9) &= P(X < 9) - P(X \leq 4) \\&= P(X \leq 8) - P(X \leq 4) \\&= \sum_{x=0}^8 b(x, 14, 0.5) - \sum_{x=0}^4 b(x, 14, 0.5) \\&= 0.7880 - 0.0898 \\&= 0.6982\end{aligned}$$

**Note 6.1.2** Do not get confused with binomial tables. In these tables, the number of trials,  $n$  is given in the far left column, the number of successes  $x$  (or  $r$  in some tables); in the next column to the right and the probability of success along the top most row. But in all can also go discrete, ie, adding each, but its too tedious.

**Example 6.1.9** A cadet fires shots at a target at distances ranging from 25 m to 90 m. The probability of hitting the target with a single shot is  $p$ . When firing from a distance  $d$  m,  $p = \frac{3}{200}(90 - d)$ . Each shot is fired independently.

The cadet fires 10 shots from a distance of 40 m.

- 1.) (a) Find the probability that exactly 6 shot hit the target.

**Solution :** *A finite number of shots (10). They are only two possible outcomes (hit or miss a target). Each shot is fired independently, and probability remain constant. This is a typical Binomial distribution.*

*The probability of hitting the target (Cadet always standing at 40 m) is given by*

$$p = \frac{3}{200}(90 - d) = \frac{3}{200}(90 - 40) = \frac{3}{4}$$

*Therefore  $X \sim B\left(10, \frac{3}{4}\right)$  So the probability*

$$P(X = 6) = \binom{10}{6} \left(\frac{3}{4}\right)^6 \left(\frac{1}{4}\right)^4 = 0.1459$$

■

- (b) Find the probability that at least 8 shots hit the target.

**Solution :**

$$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10) = 0.5255$$

■

- 2.) The cadet fires 20 shots from a distance of  $x$  m. Find, to nearest integer, the value of  $x$  if the cadet has an 80% chance of hitting the target at least once.

**Solution :** *Let  $X \sim B(20, p)$*

$$\begin{aligned} P(X \geq 1) &= 0.8 \\ 1 - P(X < 1) &= 0.8 \\ 1 - P(X = 0) &= 0.8 \end{aligned}$$

*Therefore  $P(X = 0) = 0.2$ . But*

$$\begin{aligned} P(X = 0) &= 0.2 \\ \binom{10}{0} p^0 q^{20} &= 0.2 \\ q^{20} &= 0.2 \\ (1 - p)^{20} &= 0.2 \\ (1 - p) &= 0.2^{1/20} \\ 1 - p &= 0.9226 \\ p &= 0.0773 \end{aligned}$$

When  $d = x$

$$p = \frac{3}{200}(90 - d)$$

$$p = \frac{3}{200}(90 - x)$$

$$0.0773 = \frac{3}{200}(90 - x)$$

to have

$$x = 90 - \frac{200}{3}(0.0773) = 84.84 = 85$$

to nearest integer.

■

**6.1.4 The Mean (Expectation) of a Binomial Distribution**

The mean  $\mu$  or expectation of a binomially distributed random variable  $X$  is

$$E(X) = \mu = np \quad (6.3)$$

where  $n$  is the number of repeated trials and  $p$ , where  $0 \leq p \leq 1$  is the probability of success.

Method 1: Using the Bernoulli trials:

Let the binomial random variable  $X$  be a sum of  $n$  independent Bernoulli trials  $X_1, X_2, \dots, X_n$ . Then  $X = X_1 + X_2 + \dots + X_n$ . If each of these Bernoulli trials is assigned values 1 for success and 0 for failure then

$$E(X_i) = 1 \cdot p + 0 \cdot q = p$$

**Proof :**

$$\begin{aligned} E(X) &= E(X_1 + X_2 + \dots + X_n) \\ &= E(X_1) + E(X_2) + \dots + E(X_n) \\ &= p + p + \dots + p \\ &= np \text{ since } n \text{ times} \end{aligned}$$

■

Method 2: **Proof :**

$$\begin{aligned} E(X) &= \sum_{x=0}^n x f(x) \\ &= \sum_{x=0}^n x \cdot \frac{n!}{x!(n-x)!} p^x q^{n-x} \\ &= \sum_{x=0}^n x \cdot \frac{n(n-1)!}{x(x-1)!(n-x)!} p \cdot p^{x-1} q^{n-x} \\ &= np \sum_{x=0}^n \frac{(n-1)!}{(x-1)!(n-x)!} \cdot p^{x-1} q^{n-x} \\ &= np \sum_{x=0}^n f(x) \\ &= np \end{aligned}$$

■

**6.1.5 The Variance of Binomial Distribution**

The variance of a binomially distributed random variable  $X$  is given by

$$\text{Var}(X) = \sigma^2 = npq \quad (6.4)$$

Thus the *standard deviation* of a binomial distribution is  $\sigma = \sqrt{npq}$ .

Method 1: For a Bernoulli trial  $X_i, \mu = p$

$$\begin{aligned} \text{Var}(X_i) &= E[(X_i - \mu)^2] = E[(X_i - p)^2] = E(X_i^2 - 2pX_i + p^2) = E(X_i^2) - p^2 \\ &= 1^2 \cdot p + 0^2 \cdot q - p^2 = p - p^2 = p(1 - p) = pq \end{aligned}$$

Then if a binomially distributed random variable  $X$  is made of  $n$  independent Bernoulli trials  $X_1, X_2, \dots, X_n$  then

**Proof :**

$$\begin{aligned} X &= X_1 + X_2 + X_3 + \dots + X_n \\ \text{Var}(X) &= \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n) \\ &= pq + pq + \dots + pq \quad (\text{for } n \text{ times}) \\ &= npq \end{aligned}$$

■

Method 2: **Proof :**

$$\begin{aligned} \text{Var}(X) &= E(X^2) - (E(X))^2 \\ &= E[X(X-1)] + E(X) - (E(X))^2 \\ &= \left[ \sum_{x=0}^n x(x-1) \cdot \frac{n!}{x!(n-x)!} \cdot p^x q^{n-x} \right] + np - (np)^2 \\ &= \left[ \sum_{x=0}^n x(x-1) \cdot \frac{n(n-1)(n-2)!}{x(x-1)(x-2)!(n-x)!} p^2 \cdot p^{x-2} q^{n-x} \right] + np - (np)^2 \\ &= \left[ n(n-1)p^2 \sum_{x=0}^n \frac{(n-2)!}{(x-2)!(n-x)!} \cdot p^{x-2} q^{n-x} \right] + np - (np)^2 \\ &= \left[ n(n-1)p^2 \sum_{x=0}^n f(x) \right] + np - (np)^2 \\ &= n(n-1)p^2 + np - (np)^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np - np^2 \\ &= np(1 - p) \\ &= npq \end{aligned}$$

■

**Example 6.1.10** In a family the probability of having a boy is 0.8. If there are seven children in the family determine

- 1.) the expected number of boys
- 2.) the expected number of girls
- 3.) the probability that at least four are girls
- 4.) the probability that they are all boys

Define  $X$  = number of boys

- 1.) then with  $n = 7$  and  $p = 0.8$

$$E(X) = np = (7)(0.8) = 5.6$$

- 2.)  $X$  = number of girls then with  $n = 7$  and  $p = 0.2$

$$E(X) = np = (7)(0.2) = 1.4$$

- 3.) probability that there are at least 4 girls

$$\begin{aligned}P(X \geq 4) &= 1 - P(X < 4) \\&= 1 - P(X \leq 3) \\&= 1 - \sum_{x=0}^3 b(x, 7, 0.2) \\&= 1 - 0.9667 \\&= 0.0333\end{aligned}$$

- 4.)

$$\begin{aligned}P(\text{all are boys}) &= P(X = 7) \\&= P(X \leq 7) - P(X \leq 6) \\&= \sum_{x=0}^7 b(x, 7, 0.8) - \sum_{x=0}^6 b(x, 7, 0.8) \\&= 1.0000 - 0.7903 \\&= 0.2097\end{aligned}$$

**Example 6.1.11** A multiple-choice quiz has ten questions each with four alternative answers of which one is correct. Determine the expected number of correct answers and variance, if someone does the quiz by sheer guesswork.

$$n = 10, p = \frac{1}{4} = 0.25$$

Let  $X$  = be correct answers

$$\begin{aligned}E(X) &= np = (10) \left(\frac{1}{4}\right) = 2.5 \\ \text{Var}(X) &= npq = (10) \left(\frac{1}{4}\right) \left(\frac{3}{4}\right) = 1.875\end{aligned}$$



**Example 6.1.12** A multiple-choice exam consists of twenty questions each with five alternative answers of which only one is correct. Five marks are awarded for each correct answer and one mark is subtracted for each incorrect or un attempted question. If the exam is answered by sheer guess work, find

- 1.) the expected number of correct and incorrect answers

$$\begin{aligned}E(\text{correct answers}) &= np = (20) \left(\frac{1}{5}\right) = 4 \\E(\text{incorrect answers}) &= np = (20) \left(\frac{4}{5}\right) = 16\end{aligned}$$

- 2.) the expected overall mark

Expected mark for correct answers =  $4 \times 5 = 20$ .

Expected mark for incorrect answers =  $16 \times 1 = 16$

Therefore expected overall mark =  $20 - 16 = 4$  marks.

- 3.) the variance of the mark in (ii)

$$Var(X) = npq = (4) \left(\frac{1}{4}\right) \left(\frac{3}{4}\right) = 0.75$$

**Exercise 6.5** In a certain city, rain falls on average one day out of three. If three dates are selected at random, what is the probability that rain will fall on at least one of the three dates?

**Exercise 6.6** A coin is biased in such a way that the head is twice as likely to occur as the tail. If it is tossed 5 times in succession, what is the probability that at least one head appears?

**Exercise 6.7** Five dice are tossed. What is the probability that

- 1.)  $x$  6s will turn up
- 2.) at least  $y$  6s will turn up
- 3.) at most  $y$  6s will turn up

**Exercise 6.8** Based on the data available in Mukono, the probability that a newly born baby will be a girl is about 0.487. In the next three births,

- 1.) what is the probability that exactly one is a girl?
- 2.) what is the probability that at most one is a girl?
- 3.) what is the probability that at least one is a girl?
- 4.) what is the probability that exactly two boys?
- 5.) Determine the probability distribution of the random variable  $X$ , “the number of girls” born in the next three births.

**Exercise 6.9** Seedlings are planted in 10 rows of six each. The probability of a seedling dying before it flowers is  $\frac{1}{8}$ . Calculate the expected number of rows and variance, in which all seedlings flower.

**Exercise 6.10** Studies show that 60% of Buganda families use physical aggression to resolve conflict. If 20 families are selected at random, find the probability that the number that use physical aggression to resolve conflict is

- 1.) exactly 10
- 2.) between 10 and 15, inclusive
- 3.) over 75% of those surveyed

**Exercise 6.11** According to Uganda Bankers Association, only one in 10 people are dissatisfied with their local bank. If 15 people are selected at random what is the probability that the number dissatisfied with their local bank is

- 1.) exactly two?
- 2.) at most two?
- 3.) at least two?

**Exercise 6.12** A multiple choice quiz has 15 questions each with 4 possible answers in which only 1 is correct. What is the probability that sheer guess work yields from 5 to 10 correct answers, inclusive?

**Exercise 6.13** It is known that 75% of mice inoculated with a serum are protected from a certain disease. If three mice are inoculated what is the probability that at most two contract the disease?

**Exercise 6.14** A multiple-choice quiz has 15 questions each with four possible answers of which only one is correct. A student answers the quiz by guesswork and gets one mark for the correct answer and zero for the wrong answer.

- 1.) the probability that he gets at most five incorrect answers
- 2.) the probability that he gets between five and ten correct answers inclusive
- 3.) his expected overall mark

**Exercise 6.15** A farmer learns that 40 out of every 50 seeds germinate when sown. If 12 seeds are sown what is the probability that

- 1.) exactly two thirds of these seeds germinate
- 2.) at most 9 seeds germinate

**Exercise 6.16** The probability that a pupil arrives late on any given day is 0.1.

- 1.) What is the probability of being punctual for a week (that is, 5 school days)?
- 2.) Calculate the number of days he will be late in school term of 14 weeks
- 3.) Also calculate the expected number of completely punctual weeks in the term

**Exercise 6.17** A new vaccine was tested on 15 people to determine its effectiveness. The drug company claims that a random person who is given the vaccine will develop immunity with probability of 0.8. What is the mean of this distributions?

**Exercise 6.18** A doctor has found out that 14% of the children treated with a particular drug survive. Ten children are known to have contracted the disease. Find the probability that

- 1.) at least seven will survive
- 2.) between three and five children, inclusive, survive
- 3.) determine the variance of the distribution

**Example 6.1.13** Seventeen percent of victims of financial fraud know the perpetrator of the fraud personally.

- 1.) Use the formula to construct the probability distribution for the number  $X$  of people in a random sample of five victims of financial fraud who knew the perpetrator personally.

**Solution :** *The random variable  $X$  is binomial with parameters  $n = 5$  and  $p = 0.17$ ;  $q = 1 - p = 0.83$ . The possible values of  $X$  are 0, 1, 2, 3, 4, and 5.*

$$\begin{aligned}P(0) &= \frac{5!}{0!5!}(0.17)^0(0.83)^5 \\&= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1)(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5)} 1 \cdot (0.3939040643) \\&= 0.3939040643 \approx 0.3939\end{aligned}$$

$$\begin{aligned}P(1) &= \frac{5!}{1!4!}(0.17)^1(0.83)^4 \\&= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1)(1 \cdot 2 \cdot 3 \cdot 4)} (0.17) \cdot (0.47458321) \\&= 5 \cdot (0.17) \cdot (0.47458321) \\&= 0.4033957285 \approx 0.4034\end{aligned}$$

$$\begin{aligned}P(2) &= \frac{5!}{2!3!}(0.17)^2(0.83)^3 \\&= \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1 \cdot 2)(1 \cdot 2 \cdot 3)} (0.0289) \cdot (0.571787) \\&= 10 \cdot (0.0289) \cdot (0.571787) \\&= 0.165246443 \approx 0.1652\end{aligned}$$

*The remaining three probabilities are computed similarly, to give the probability distribution*

$x$	0	1	2	3	4	5
$P(x)$	0.3939	0.4034	0.1652	0.0338	0.0035	0.0001

*The probabilities do not add up to exactly 1 because of rounding.* ■

- 2.) A investigator examines five cases of financial fraud every day. Find the most frequent number of cases each day in which the victim knew the perpetrator.

**Solution :** This probability distribution is represented by the histogram in Figure 6.1, which graphically illustrates just how improbable the events  $X = 4$  and  $X = 5$  are. The corresponding bar in the histogram above the number 4 is barely visible, if visible at all, and the bar above 5 is far too short to be visible.

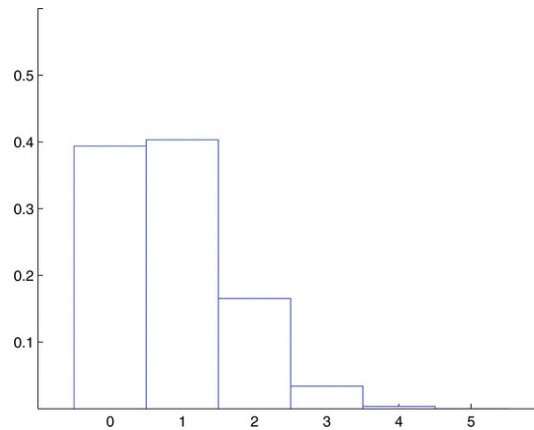


Figure 6.1: Probability Distribution of the Binomial Random Variable in Example 6.1.13

The value of  $X$  that is most likely is  $X = 1$ , so the most frequent number of cases seen each day in which the victim knew the perpetrator is one. ■

- 3.) A investigator examines five cases of financial fraud every day. Find the average number of cases per day in which the victim knew the perpetrator.

**Solution :** The average number of cases per day in which the victim knew the perpetrator is the mean of  $X$ , which is

$$\begin{aligned}\mu &= \sum xf(x) \\ &= 0 \cdot 0.3939 + 1 \cdot 0.4034 + 2 \cdot 0.1652 + 3 \cdot 0.0338 + 4 \cdot 0.0035 + 5 \cdot 0.0001 \\ &= 0.8497\end{aligned}$$

■

**Example 6.1.14** Find the mean and standard deviation of the random variable  $X$  of Example 6.1.13.

**Solution :** The random variable  $X$  is binomial with parameters  $n = 5$  and  $p = 0.17$ , and  $q = 1 - p = 0.83$ . Thus its mean and standard deviation are

$$\mu = np = (5) \cdot (0.17) = 0.85 \text{ (exactly)}$$

and

$$\sigma = \sqrt{npq} = \sqrt{(5) \cdot (0.17) \cdot (0.83)} = \sqrt{0.7055} \approx 0.8399$$

■

**Exercise 6.19** In an experiment, two unbiased dice are tossed. When they are tossed, the greater score (or either score if they are the same) is recorded.

- 1.) Obtain the probability distribution table for this experiment
- 2.) If the experiment is repeated 10 times, determine
  - (a) the probability that a score of 5 appears at least two times but at most 5 times
  - (b) the expected number of times a score of 2 will appear

**Example 6.1.15** A student takes a ten-question true/false exam.

- 1.) Find the probability that the student gets exactly six of the questions right simply by guessing the answer on every question.

**Solution :** Let  $X$  denote the number of questions that the student guesses correctly. Then  $X$  is a binomial random variable with parameters  $n = 10$  and  $p = 0.50$ .

The probability sought is  $P(6)$ . The formula gives

$$P(6) = 10!(6!)(4!)(.5)^6 = 0.205078125$$

Using the table,

$$P(6) = P(X \leq 6) - P(X \leq 5) = 0.8281 - 0.6230 = 0.2051$$

■

- 2.) Find the probability that the student will obtain a passing grade of 60% or greater simply by guessing.

**Solution :** The student must guess correctly on at least 60% of the questions, which is  $(0.60) \cdot (10) = 6$  questions. The probability sought is not  $P(6)$  (an easy mistake to make), but

$$P(X \geq 6) = P(6) + P(7) + P(8) + P(9) + P(10)$$

Instead of computing each of these five numbers using the formula and adding them we can use the table to obtain

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.6230 = 0.3770$$

which is much less work and of sufficient accuracy for the situation at hand. ■

**Example 6.1.16** An appliance repairman services five washing machines on site each day. One-third of the service calls require installation of a particular part.

- 1.) The repairman has only one such part on his truck today. Find the probability that the one part will be enough today, that is, that at most one washing machine he services will require installation of this particular part.

**Solution :** Let  $X$  denote the number of service calls today on which the part is required. Then  $X$  is a binomial random variable with parameters  $n = 5$  and  $p = 1/3 = 0.\bar{3}$

Note that the probability in question is not  $P(1)$ , but rather  $P(X \leq 1)$ . Using the cumulative distribution table,

$$P(X \leq 1) = 0.4609$$

■

- 2.) Find the minimum number of such parts he should take with him each day in order that the probability that he have enough for the day's service calls is at least 95%.

**Solution :** The answer is the smallest number  $x$  such that the table entry  $P(X \leq x)$  is at least 0.9500. Since  $P(X \leq 2) = 0.7901$  is less than 0.95, two parts are not enough. Since  $P(X \leq 3) = 0.9547$  is as large as 0.95, three parts will suffice at least 95% of the time. Thus the minimum needed is three. ■

## 6.2 The Poisson Distribution

### Introduction

The Poisson distribution, named after S.D Poisson (1781 - 1840), a French Mathematician, is an important discrete process. Its experiments yield numerical values of a random variable  $X$  whose outcomes occur during given time intervals or specific regions. For instance, an experiment that records the number of telephone calls per hour by an office or the number of days a school is closed.

This lecture will introduce you to the characteristics of a Poisson experiment and the corresponding distribution.

#### 6.2.1 The Poisson Experiment and Distribution

**Definition 6.2.1** A Poisson experiment is one that possesses the following properties.

1. The number of successes occurring in one time interval or specific region are independent of those occurring in any other disjoint time interval or region of space.
2. The probability of a single success occurring at very short time interval or in a small region is proportional to the length of the time interval or the size of the region and does not depend on the number of success occurring in such a short time interval or falling in such small regions is negligible.

**Definition 6.2.2** The number  $X$  of successes occurring in a Poisson experiment is called a Poisson random variable. The probability distribution of the Poisson random variable  $X$  is called the Poisson distribution and is denoted by  $P(x, \lambda)$  since its values depend on the average number of successes occurring in the given time interval or specified region.

#### 6.2.2 Poisson Probability Distribution

The probability distribution of the Poisson random variable  $X$ , representing the number of outcomes occurring in a given time interval or specified region, is

$$P(X = x) = p(x, \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} ; x = 0, 1, 2, \dots, \lambda > 0 \quad (6.5)$$

where  $\lambda$  is the average number of outcomes in the given time interval or specified region and  $e = 2.718$ . Thus,  $p(x, \lambda)$  is a Poisson probability distribution with parameter  $\lambda$ .

**Note 6.2.1** A Poisson is applied when talking of a specified region, time, area. E.g **per** page, **per** month. Where  $\lambda$  is its number (on average) of its occurrence in that specified area.

**Note 6.2.2** For this study, lower case  $p$  means Poisson, say  $p(a, \lambda)$  as was with  $b$  to mean Binomial in previous section.

### 6.2.3 Examples of Poisson Distribution

- 1.) Car accidents on a particular stretch of road in a day.
- 2.) Flaws **in a given** length of material
- 3.) Accidents in a factory in a ( per) week.
- 4.) Telephone calls made to a switchboard **per** (in a given) minute.
- 5.) Insurance claims made to a company **in a given** time.
- 6.) Particles emitted by a radioactive source **in a given** time.
- 7.) Misprints **per** page in a book of a given number of pages.

**Definition 6.2.3** The probability that a Poisson random variable  $X$  takes on a value, say  $a$ , in a given time interval, is given by

$$P(X = a) = p(a, \lambda) = \frac{e^{-\lambda} \lambda^a}{a!} \quad (6.6)$$

**Example 6.2.1** The average number of days a school is closed due to bad weather **in a term** is 4. What is the probability that the school will close for 6 days **in a term**?

$X = 6$  and  $\lambda = 4$

Thus

$$\begin{aligned} P(X = 6) &= p(6, 4) \\ &= \frac{e^{-4} \cdot 4^6}{6!} \\ &= 0.1042 \end{aligned}$$

Alternatively, using Poisson table,

$$\begin{aligned} P(X = 6) &= P(X \leq 6) - P(X \leq 5) \\ &= \sum_{x=0}^6 p(x, 4) - \sum_{x=0}^5 p(x, 4) \\ &= 0.8893 - 0.7851 \\ &= 0.1042 \end{aligned}$$

**Example 6.2.2** The average number of field mice **per acre** of a field is estimated to be 10. Find the probability that **a given acre** will contain more than 15 mice.

Here  $x > 15$ ,  $\lambda = 10$

Then

$$\begin{aligned} P(X > 15) &= 1 - P(X \leq 15) \\ &= 1 - \sum_{x=0}^{15} p(x, 10) \\ &= 1 - 0.9513 \\ &= 0.0487 \end{aligned}$$

**Example 6.2.3** Telephone calls enter a school switchboard on the average of two **every 3 minutes**. What is the probability of

1.) Exactly 5 calls arriving in a 9-minute period?

$\lambda$	Region
2	3
??	9

$$\Rightarrow \lambda = \frac{(2)(9)}{3} = 6$$
$$P(X = 5) = \frac{e^{-6}(6)^5}{5!}$$

2.) Exactly 5 calls arriving in a 12-minute period?

$\lambda$	Region
2	3
??	12

$$\Rightarrow \lambda = \frac{(2)(12)}{3} = 8$$
$$P(X = 5) = \frac{e^{-8}(8)^5}{5!}$$

3.) At least 5 calls arriving in a 9-minute period?

$$\begin{aligned}P(X \geq 5) &= 1 - P(X \leq 4) \\&= 1 - \sum_{x=0}^4 p(x, 6) \\&= 1 - 0.2851 \\&= 0.7149\end{aligned}$$

**Exercise 6.20** The average number of loaves of bread put on a shelf in a bakery in a half-hour period is 12. Of interest is the number of loaves of bread put on the shelf in five minutes. The time interval of interest is five minutes. What is the probability that the number of loaves, selected randomly, put on the shelf in five minutes is three?

**Solution :** *Let  $X$  = the number of loaves of bread put on the shelf in five minutes. If the average number of loaves put on the shelf in 30 minutes (half-hour) is 12, then the average number of loaves put on the shelf in five minutes is*

$$\left(\frac{5}{30}\right)(12) = 2$$

*loaves of bread.*

*The probability question asks you to find  $P(x = 3)$ .*

■



**Example 6.2.4** Leah's answering machine receives about six telephone calls between 8 a.m. and 10 a.m. What is the probability that Leah receives more than one call in the next 15 minutes?

**Solution :** Let  $X$  = the number of calls Leah receives in 15 minutes. (The *interval of interest* is 15 minutes.)

If Leah receives, on the average, six telephone calls in two hours, and there are eight 15 minute intervals in two hours, then Leah receives

$$\frac{1}{8}(6) = 0.75$$

calls in 15 minutes, on average. So,  $E(X) = \lambda = 0.75$  for this problem. Such that

$$P(x > 1) = 0.1734$$

■

**Example 6.2.5** According to Baydin, an email management company, an email user gets, on average, 147 emails per day. Let  $X$  = the number of emails an email user receives per day. The discrete random variable  $X$  takes on the values  $x = 0, 1, 2, \dots$ . The random variable  $X$  has a Poisson distribution:  $X \sim P(147)$ . The mean is 147 emails.

1.) What is the probability that an email user receives exactly 160 emails per day?

**Solution :**

$$P(X = 160) = \text{poissonpdf}(147, 160) \approx 0.0180$$

■

2.) What is the probability that an email user receives at most 160 emails per day?

**Solution :**

$$P(X \leq 160) = \text{poissoncdf}(147, 160) \approx 0.8666$$

■

3.) What is the standard deviation?

**Solution :**

$$\sigma = \sqrt{\lambda} = \sqrt{147} \approx 12.1244$$

■

**Example 6.2.6** According to a survey a university professor gets, on average, 7 emails per day. Let  $X$  = the number of emails a professor receives per day. The discrete random variable  $X$  takes on the values  $x = 0, 1, 2, \dots$ . The random variable  $X$  has a Poisson distribution:  $X \sim P(7)$ . The mean is 7 emails.

1.) What is the probability that an email user receives exactly 2 emails per day?

**Solution :**

$$P(X = 2) = \frac{\lambda^{x_e - \lambda}}{x!} = \frac{7^2 e^{-7}}{2!} = 0.022$$

■

2.) What is the probability that an email user receives at most 2 emails per day?

**Solution :**

$$P(X \leq 2) = \frac{7^0 e^{-7}}{0!} + \frac{7^1 e^{-7}}{1!} + \frac{7^2 e^{-7}}{2!} = 0.029$$

■

3.) What is the standard deviation?

**Solution :**

$$\sigma = \sqrt{\lambda} = \sqrt{7} \approx 2.65$$

■

**Example 6.2.7** Text message users receive or send an average of 41.5 text messages per day.

1.) How many text messages does a text message user receive or send per hour?

**Solution :** Let  $X$  = the number of texts that a user sends or receives in one hour. The average number of texts received per hour is  $\frac{41.5}{24} \approx 1.7292$ .

■

2.) What is the probability that a text message user receives or sends two messages per hour?

**Solution :**

$$P(x = 2) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1.7292^2 e^{-1.7292}}{2!} = 0.265$$

■

3.) What is the probability that a text message user receives or sends more than two messages per hour?

**Solution :**

$$P(X > 2) = 1 - P(X \leq 2) = 1 - \left[ \frac{1.7292^0 e^{-1.7292}}{0!} + \frac{1.7292^1 e^{-1.7292}}{1!} + \frac{1.7292^2 e^{-1.7292}}{2!} \right] = 0.250$$

■

### 6.2.4 Expectation of Poisson Distribution

**Theorem 6.2.1** For a Poisson random variable  $X$  with parameter  $\lambda$  and probability mass function defined by

$$P(x, \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots; \quad \lambda > 0$$

we define the mean or expectation of  $X$ ,  $E(X)$  as

$$E(X) = \lambda \tag{6.7}$$

**Proof :**

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x \rho(x) \\ &= \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &\quad \left[ \text{since } \frac{x}{x!} = \frac{x}{x(x-1)!} = \frac{1}{(x-1)!} \right] \\ &= \lambda \sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda \sum_{x=0}^{\infty} f(x) \\ &= \lambda[1] \\ &= \lambda \end{aligned}$$

■

### 6.2.5 Variance of Poisson Distribution

**Theorem 6.2.2** For a Poisson random variable  $X$ , with parameter  $\lambda$ , the variance of  $X$ ,  $Var(X)$ , is given by

$$Var(X) = \lambda \quad (6.8)$$

**Proof :**

$$Var(X) = E[X(X-1)] + E[X] - [E(X)]^2$$

*We use the definition of  $E(X)$  and evaluate  $E[X(X-1)]$*

*But*

$$\begin{aligned} E[X(X-1)] &= \sum x(x-1)\rho(x) \\ &= \sum x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} \\ &= \lambda^2 \sum e^{-\lambda} \frac{\lambda^{x-2}}{(x-2)!} \quad \text{since } \frac{x(x-1)}{x!} = \frac{1}{(x-2)!} \\ &= \lambda^2 \sum f(x) \\ &= \lambda^2 \end{aligned}$$

*The variance*

$$\begin{aligned} Var(X) &= E[X(X-1)] + E[X] - [E(X)]^2 \\ &= \lambda^2 + \lambda - \lambda^2 \\ &= \lambda \end{aligned}$$

■

**Exercise 6.21** Makerere University is hit on average by a strike 2 times *a year* . Assuming the frequency of such strikes follows a Poisson distribution. Find the probability that Makerere University will be hit by

- 1.) exactly 6 strikes *in a year*,
- 2.) between 5 and 7 strikes *in 4 years*.

**Exercise 6.22** Let  $X$  have a Poisson distribution with mean 4. Find

- 1.)  $P(2 \leq X \leq 5)$
- 2.)  $P(X \geq 3)$
- 3.)  $P(X \leq 3)$
- 4.)  $P(X = 4)$

**Exercise 6.23** Customers arrive at a bank at an average of 11 per hour. Assume that the number of arrivals per hour has a Poisson distribution, find the probability that more than 10 customers arrive in any given hour.

**Exercise 6.24** If a random variable  $X$  has a Poisson distribution. So that

$$3P(X = 1) = P(X = 2)$$

Find

$$P(X = 4)$$

**Exercise 6.25** A hotel prepares a toast salad containing on the average 5 vegetables on a given day. Find the probability that the salad contains more than 3 vegetables

- 1.) on a given day
- 2.) on 3 of the next 4 days

**Exercise 6.26** If a Poisson random variable has an expected value of 3.0 what is its variance?

**Exercise 6.27** Calls to a particular center occur at an average rate of 30 per hour. Find the probability that there are at least two calls in a given minute. 0.0902

# Chapter 7

## Probability Tables

## 7.1 Cumulative Binomial Probabilities $P(X \leq c)$ Table

$n$	$x$	$p$	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1	0		0.9900	0.9800	0.9700	0.9600	0.9500	0.9400	0.9300	0.9200	0.9100
	1		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0		0.9801	0.9604	0.9409	0.9216	0.9025	0.8836	0.8649	0.8464	0.8281
	1		0.9999	0.9996	0.9991	0.9984	0.9975	0.9964	0.9951	0.9936	0.9919
	2		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0		0.9703	0.9412	0.9127	0.8847	0.8574	0.8306	0.8044	0.7787	0.7536
	1		0.9997	0.9988	0.9974	0.9953	0.9928	0.9896	0.9860	0.9818	0.9772
	2		1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997	0.9995	0.9993
	3		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	0		0.9606	0.9224	0.8853	0.8493	0.8145	0.7807	0.7481	0.7164	0.6857
	1		0.9994	0.9977	0.9948	0.9909	0.9860	0.9801	0.9733	0.9656	0.9570
	2		1.0000	1.0000	0.9999	0.9998	0.9995	0.9992	0.9987	0.9981	0.9973
	3		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	4		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0		0.9510	0.9039	0.8587	0.8154	0.7738	0.7339	0.6957	0.6591	0.6240
	1		0.9990	0.9962	0.9915	0.9852	0.9774	0.9681	0.9575	0.9456	0.9326
	2		1.0000	0.9999	0.9997	0.9994	0.9988	0.9980	0.9969	0.9955	0.9937
	3		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997
	4		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n$	$x$	$p$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
1	0		0.9000	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000
	1		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0		0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1		0.9900	0.9775	0.9600	0.9375	0.9100	0.8775	0.8400	0.7975	0.7500
	2		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0		0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1		0.9720	0.9393	0.8960	0.8438	0.7840	0.7183	0.6480	0.5748	0.5000
	2		0.9990	0.9966	0.9920	0.9844	0.9730	0.9571	0.9360	0.9089	0.8750
	3		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	0		0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1		0.9477	0.8905	0.8192	0.7383	0.6517	0.5630	0.4752	0.3910	0.3125
	2		0.9963	0.9880	0.9728	0.9492	0.9163	0.8735	0.8208	0.7585	0.6875
	3		0.9999	0.9995	0.9984	0.9961	0.9919	0.9850	0.9744	0.9590	0.9375
	4		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0		0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0312
	1		0.9185	0.8352	0.7373	0.6328	0.5282	0.4284	0.3370	0.2562	0.1875
	2		0.9914	0.9734	0.9421	0.8965	0.8369	0.7648	0.6826	0.5931	0.5000
	3		0.9995	0.9978	0.9933	0.9844	0.9692	0.9460	0.9130	0.8688	0.8125
	4		1.0000	0.9999	0.9997	0.9990	0.9976	0.9947	0.9898	0.9815	0.9688
6	0		0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1		0.8857	0.7765	0.6554	0.5339	0.4202	0.3191	0.2333	0.1636	0.1094
	2		0.9842	0.9527	0.9011	0.8306	0.7443	0.6471	0.5443	0.4415	0.3437
	3		0.9987	0.9941	0.9830	0.9624	0.9295	0.8826	0.8208	0.7447	0.6563
	4		0.9999	0.9996	0.9984	0.9954	0.9891	0.9777	0.9590	0.9308	0.8906
	5		1.0000	1.0000	0.9999	0.9998	0.9993	0.9982	0.9959	0.9917	0.9844
7	0		0.9321	0.8681	0.8080	0.7514	0.6983	0.6485	0.6017	0.5578	0.5168
	1		0.9980	0.9921	0.9829	0.9706	0.9556	0.9382	0.9187	0.8974	0.8745
$n$	$x$	$p$	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
6	0		0.9415	0.8858	0.8330	0.7828	0.7351	0.6899	0.6470	0.6064	0.5679
	1		0.9985	0.9943	0.9875	0.9784	0.9672	0.9541	0.9392	0.9227	0.9048
	2		1.0000	0.9998	0.9995	0.9988	0.9978	0.9962	0.9942	0.9915	0.9882
	3		1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9995	0.9992
	4		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	0		0.9321	0.8681	0.8080	0.7514	0.6983	0.6485	0.6017	0.5578	0.5168
	1		0.9980	0.9921	0.9829	0.9706	0.9556	0.9382	0.9187	0.8974	0.8745

# 7.1. CUMULATIVE BINOMIAL PROBABILITIES $P(X \leq c)$ TABLE

8	2		1.0000	0.9997	0.9991	0.9980	0.9962	0.9937	0.9903	0.9860	0.9807
	3		1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9993	0.9988	0.9982
	4		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
	5		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0		0.9227	0.8508	0.7837	0.7214	0.6634	0.6096	0.5596	0.5132	0.4703
9	1		0.9973	0.9897	0.9777	0.9619	0.9428	0.9208	0.8965	0.8702	0.8423
	2		0.9999	0.9996	0.9987	0.9969	0.9942	0.9904	0.9853	0.9789	0.9711
	3		1.0000	1.0000	0.9999	0.9998	0.9996	0.9993	0.9987	0.9978	0.9966
	4		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997
	5		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0		0.9135	0.8337	0.7602	0.6925	0.6302	0.5730	0.5204	0.4722	0.4279
	1		0.9966	0.9869	0.9718	0.9522	0.9288	0.9022	0.8729	0.8417	0.8088
	2		0.9999	0.9994	0.9980	0.9955	0.9916	0.9862	0.9791	0.9702	0.9595
	3		1.0000	1.0000	0.9999	0.9997	0.9994	0.9987	0.9977	0.9963	0.9943
	4		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9995
	5		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n$	$x$	$p$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
7	0		0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1		0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625
	2		0.9743	0.9262	0.8520	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266
	3		0.9973	0.9879	0.9667	0.9294	0.8740	0.8002	0.7102	0.6083	0.5000
	4		0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734
8	5		1.0000	0.9999	0.9996	0.9987	0.9962	0.9910	0.9812	0.9643	0.9375
	6		1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9984	0.9963	0.9922
	7		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0		0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1		0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.0632	0.0352
9	2		0.9619	0.8948	0.7969	0.6785	0.5518	0.4278	0.3154	0.2201	0.1445
	3		0.9950	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.4770	0.3633
	4		0.9996	0.9971	0.9896	0.9727	0.9420	0.8939	0.8263	0.7396	0.6367
	5		1.0000	0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555
	6		1.0000	1.0000	0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648
	7		1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9983	0.9961
	8		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0		0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1		0.7748	0.5995	0.4362	0.3003	0.1960	0.1211	0.0705	0.0385	0.0195
	2		0.947	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.08984
	3		0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539
	4		0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5000
	5		0.9999	0.9994	0.9969	0.9900	0.9747	0.9464	0.9006	0.8342	0.7461
	6		1.0000	1.0000	0.9997	0.9987	0.9957	0.9888	0.9750	0.9502	0.9102
	7		1.0000	1.0000	1.0000	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805
	8		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980
	9		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n$	$x$	$p$	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
10	0		0.9044	0.8171	0.7374	0.6648	0.5987	0.5386	0.4840	0.4344	0.3894
	1		0.9957	0.9838	0.9655	0.9418	0.9139	0.8824	0.8483	0.8121	0.7746
	2		0.9999	0.9991	0.9972	0.9938	0.9885	0.9812	0.9717	0.9599	0.9460
	3		1.0000	1.0000	0.9999	0.9996	0.9990	0.9980	0.9964	0.9942	0.9912
	4		1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9994	0.9990
11	5		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	6		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0		0.8953	0.8007	0.7153	0.6382	0.5688	0.5063	0.4501	0.3996	0.3544
	1		0.9948	0.9805	0.9587	0.9308	0.8981	0.8618	0.8228	0.7819	0.7399
	2		0.9998	0.9988	0.9963	0.9917	0.9848	0.9752	0.9630	0.9481	0.9305
	3		1.0000	1.0000	0.9998	0.9993	0.9984	0.9970	0.9947	0.9915	0.9871
	4		1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9995	0.9990	0.9983
	5		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	6		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000



7.1. CUMULATIVE BINOMIAL PROBABILITIES  $P(X \leq c)$  TABLE

$n$	$x$	$p$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
10	0		0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1		0.7361	0.5443	0.3758	0.2440	0.1493	0.0860	0.0464	0.0233	0.0107
	2		0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996	0.0547
	3		0.9872	0.9500	0.8791	0.7759	0.6496	0.5138	0.3823	0.2660	0.1719
	4		0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.3770
	5		0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.6230
	6		1.0000	0.9999	0.9991	0.9965	0.9894	0.9740	0.9452	0.8980	0.8281
	7		1.0000	1.0000	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453
	8		1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983	0.9955	0.9893
	9		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990
	10		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	0		0.3138	0.1673	0.0859	0.0422	0.0198	0.0088	0.0036	0.0014	0.0004
	1		0.6974	0.4922	0.3221	0.1971	0.1130	0.0606	0.0302	0.0139	0.0059
	2		0.9104	0.7788	0.6174	0.4552	0.3127	0.2001	0.1189	0.0652	0.0327
	3		0.9815	0.9306	0.8389	0.7133	0.5696	0.4256	0.2963	0.1911	0.1133
	4		0.9972	0.9841	0.9496	0.8854	0.7897	0.6683	0.5328	0.3971	0.2744
	5		0.9997	0.9973	0.9883	0.9657	0.9218	0.8513	0.7535	0.6331	0.5000
	6		1.0000	0.9997	0.9980	0.9924	0.9784	0.9499	0.9006	0.8262	0.7256
	7		1.0000	1.0000	0.9998	0.9988	0.9957	0.9878	0.9707	0.9390	0.8867
	8		1.0000	1.0000	1.0000	0.9999	0.9994	0.9980	0.9941	0.9852	0.9673
	9		1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9993	0.9978	0.9941
	10		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9995
12	11		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0		0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002
	1		0.6590	0.4435	0.2749	0.1584	0.0850	0.0424	0.0196	0.0083	0.0032
	2		0.8891	0.7358	0.5583	0.3907	0.2528	0.1513	0.0834	0.0421	0.0193
	3		0.9744	0.9078	0.7946	0.6488	0.4925	0.3467	0.2253	0.1345	0.0730
	4		0.9957	0.9761	0.9274	0.8424	0.7237	0.5833	0.4382	0.3044	0.1938
	5		0.9995	0.9954	0.9806	0.9456	0.8822	0.7873	0.6652	0.5269	0.3872
	6		0.9999	0.9993	0.9961	0.9857	0.9614	0.9154	0.8418	0.7393	0.6128
	7		1.0000	0.9999	0.9994	0.9972	0.9905	0.9745	0.9427	0.8883	0.8062
	8		1.0000	1.0000	0.9999	0.9996	0.9983	0.9944	0.9847	0.9644	0.9270
	9		1.0000	1.0000	1.0000	1.0000	0.9998	0.9992	0.9972	0.9921	0.9807
	10		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9968
	11		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	12		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n$	$x$	$p$	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
12	0		0.8864	0.7847	0.6938	0.6127	0.5404	0.4759	0.4186	0.3677	0.3225
	1		0.9938	0.9769	0.9514	0.9191	0.8816	0.8405	0.7967	0.7513	0.7052
	2		0.9998	0.9985	0.9952	0.9893	0.9804	0.9684	0.9532	0.9348	0.9134
	3		1.0000	0.9999	0.9997	0.9990	0.9978	0.9957	0.9925	0.9880	0.9820
	4		1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9991	0.9984	0.9973
	5		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997
	6		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
13	0		0.8775	0.7690	0.6730	0.5882	0.5133	0.4474	0.3893	0.3383	0.2935
	1		0.9928	0.9730	0.9436	0.9068	0.8646	0.8186	0.7702	0.7206	0.6707
	2		0.9997	0.9980	0.9938	0.9865	0.9755	0.9608	0.9422	0.9201	0.8946
	3		1.0000	0.9999	0.9995	0.9986	0.9969	0.9940	0.9897	0.9837	0.9758
	4		1.0000	1.0000	1.0000	0.9999	0.9997	0.9993	0.9987	0.9976	0.9959
	5		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9995
	6		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
14	7		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0		0.8687	0.7536	0.6528	0.5647	0.4877	0.4205	0.3620	0.3112	0.2670
	1		0.9916	0.9690	0.9355	0.8941	0.8470	0.7963	0.7436	0.6900	0.6368
	2		0.9997	0.9975	0.9923	0.9833	0.9699	0.9522	0.9302	0.9042	0.8745
	3		1.0000	0.9999	0.9994	0.9981	0.9958	0.9920	0.9864	0.9786	0.9685
	4		1.0000	1.0000	1.0000	0.9998	0.9996	0.9990	0.9980	0.9965	0.9941
	5		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9992

7.1. CUMULATIVE BINOMIAL PROBABILITIES  $P(X \leq c)$  TABLE

	6		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	7		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n$	$x$	$p$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
13	0		0.2542	0.1209	0.0550	0.0238	0.0097	0.0037	0.0013	0.0004	0.0001
	1		0.6213	0.3983	0.2336	0.1267	0.0637	0.0296	0.0126	0.0050	0.0017
	2		0.8661	0.6920	0.5017	0.3326	0.2025	0.1132	0.0579	0.0269	0.0112
	3		0.9658	0.8820	0.7473	0.5843	0.4206	0.2783	0.1686	0.0929	0.0461
	4		0.9935	0.9658	0.9009	0.7940	0.6543	0.5005	0.3530	0.2279	0.1334
	5		0.9991	0.9925	0.9700	0.9198	0.8346	0.7159	0.5744	0.4268	0.2905
	6		0.9999	0.9987	0.9930	0.9757	0.9376	0.8705	0.7712	0.6437	0.5000
	7		1.0000	0.9998	0.9988	0.9944	0.9818	0.9538	0.9023	0.8212	0.7095
	8		1.0000	1.0000	0.9998	0.9990	0.9960	0.9874	0.9679	0.9302	0.8666
	9		1.0000	1.0000	1.0000	0.9999	0.9993	0.9975	0.9922	0.9797	0.9539
	10		1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9987	0.9959	0.9888
	11		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983
	12		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	13		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
14	0		0.2288	0.1028	0.0440	0.0178	0.0068	0.0024	0.0008	0.0002	0.0000
	1		0.5846	0.3567	0.1979	0.1010	0.0475	0.0205	0.0081	0.0029	0.0009
	2		0.8416	0.6479	0.4481	0.2811	0.1608	0.0839	0.0398	0.0170	0.0065
	3		0.9559	0.8535	0.6982	0.5213	0.3552	0.2205	0.1243	0.0632	0.0287
	4		0.9908	0.9533	0.8702	0.7415	0.5842	0.4227	0.2793	0.1672	0.0898
	5		0.9985	0.9885	0.9561	0.8883	0.7805	0.6405	0.4859	0.3373	0.2120
	6		0.9998	0.9978	0.9884	0.9617	0.9067	0.8164	0.6925	0.5461	0.3953
	7		1.0000	0.9997	0.9976	0.9897	0.9685	0.9247	0.8499	0.7414	0.6047
	8		1.0000	1.0000	0.9996	0.9978	0.9917	0.9757	0.9417	0.8811	0.7880
	9		1.0000	1.0000	1.0000	0.9997	0.9983	0.9940	0.9825	0.9574	0.9102
	10		1.0000	1.0000	1.0000	1.0000	0.9998	0.9989	0.9961	0.9886	0.9713
	11		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9978	0.9935
	12		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9991
	13		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	14		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n$	$x$	$p$	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
15	0		0.8601	0.7386	0.6333	0.5421	0.4633	0.3953	0.3367	0.2863	0.2430
	1		0.9904	0.9647	0.9270	0.8809	0.8290	0.7738	0.7168	0.6597	0.6035
	2		0.9996	0.9970	0.9906	0.9797	0.9638	0.9429	0.9171	0.8870	0.8531
	3		1.0000	0.9998	0.9992	0.9976	0.9945	0.9896	0.9825	0.9727	0.9601
	4		1.0000	1.0000	0.9999	0.9998	0.9994	0.9986	0.9972	0.9950	0.9918
	5		1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9993	0.9987
	6		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	7		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0		0.8515	0.7238	0.6143	0.5204	0.4401	0.3716	0.3131	0.2634	0.2211
	1		0.9891	0.9601	0.9182	0.8673	0.8108	0.7511	0.6902	0.6299	0.5711
	2		0.9995	0.9963	0.9887	0.9758	0.9571	0.9327	0.9031	0.8689	0.8306
	3		1.0000	0.9998	0.9989	0.9968	0.9930	0.9868	0.9779	0.9658	0.9504
	4		1.0000	1.0000	0.9999	0.9997	0.9991	0.9981	0.9962	0.9932	0.9889
	5		1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9995	0.9990	0.9981
	6		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997
	7		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n$	$x$	$p$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
15	0		0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000
	1		0.5490	0.3186	0.1671	0.0802	0.0353	0.0142	0.0052	0.0017	0.0005
	2		0.8159	0.6042	0.3980	0.2361	0.1268	0.0617	0.0271	0.0107	0.0037
	3		0.9444	0.8227	0.6482	0.4613	0.2969	0.1727	0.0905	0.0424	0.0176
	4		0.9873	0.9383	0.8358	0.6865	0.5155	0.3519	0.2173	0.1204	0.0592
	5		0.9978	0.9832	0.9389	0.8516	0.7216	0.5643	0.4032	0.2608	0.1509
	6		0.9997	0.9964	0.9819	0.9434	0.8689	0.7548	0.6098	0.4522	0.3036
	7		1.0000	0.9994	0.9958	0.9827	0.9500	0.8868	0.7869	0.6535	0.5000
	8		1.0000	0.9999	0.9992	0.9958	0.9848	0.9578	0.9050	0.8182	0.6964

# 7.1. CUMULATIVE BINOMIAL PROBABILITIES $P(X \leq c)$ TABLE

16	9		1.0000	1.0000	0.9999	0.9992	0.9963	0.9876	0.9662	0.9231	0.8491
	10		1.0000	1.0000	1.0000	0.9999	0.9993	0.9972	0.9907	0.9745	0.9408
	11		1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9981	0.9937	0.9824
	12		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9963
	13		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995
	14		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0		0.1853	0.0743	0.0281	0.0100	0.0033	0.0010	0.0003	0.0000	0.0000
	1		0.5147	0.2839	0.1407	0.0635	0.0261	0.0098	0.0033	0.0010	0.0003
	2		0.7892	0.5614	0.3518	0.1971	0.0994	0.0451	0.0183	0.0066	0.0021
	3		0.9316	0.7899	0.5981	0.4050	0.2459	0.1339	0.0651	0.0281	0.0106
	4		0.9830	0.9209	0.7982	0.6302	0.4499	0.2892	0.1666	0.0853	0.0384
	5		0.9967	0.9765	0.9183	0.8103	0.6598	0.4900	0.3288	0.1976	0.1051
	6		0.9995	0.9944	0.9733	0.9204	0.8247	0.6881	0.5272	0.3660	0.2272
	7		0.9999	0.9989	0.9930	0.9729	0.9256	0.8406	0.7161	0.5629	0.4018
	8		1.0000	0.9998	0.9985	0.9925	0.9743	0.9329	0.8577	0.7441	0.5982
	9		1.0000	1.0000	0.9998	0.9984	0.9929	0.9771	0.9417	0.8759	0.7728
10		1.0000	1.0000	1.0000	0.9997	0.9984	0.9938	0.9809	0.9514	0.8949	
11		1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9851	0.9616	
12		1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9991	0.9965	0.9894	
13		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9979	
14		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	
15		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	
$n$	$x$	$p$	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
18	17	0	0.8429	0.7093	0.5958	0.4996	0.4181	0.3493	0.2912	0.2423	0.2012
	1		0.9877	0.9554	0.9091	0.8535	0.7922	0.7283	0.6638	0.6005	0.5396
	2		0.9994	0.9956	0.9866	0.9714	0.9497	0.9218	0.8882	0.8497	0.8073
	3		1.0000	0.9997	0.9986	0.9960	0.9912	0.9836	0.9727	0.9581	0.9397
	4		1.0000	1.0000	0.9999	0.9996	0.9988	0.9974	0.9949	0.9911	0.9855
	5		1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9993	0.9985	0.9973
	6		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996
	7		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0		0.8345	0.6951	0.5780	0.4796	0.3972	0.3283	0.2708	0.2229	0.1831
	1		0.9862	0.9505	0.8997	0.8393	0.7735	0.7055	0.6378	0.5719	0.5091
	2		0.9993	0.9948	0.9843	0.9667	0.9419	0.9102	0.8725	0.8298	0.7832
	3		1.0000	0.9996	0.9982	0.9950	0.9891	0.9799	0.9667	0.9494	0.9277
	4		1.0000	1.0000	0.9998	0.9994	0.9985	0.9966	0.9933	0.9884	0.9814
	5		1.0000	1.0000	1.0000	0.9999	0.9998	0.9995	0.9990	0.9979	0.9962
	6		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9994
	7		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
8		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	
$n$	$x$	$p$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
18	17	0	0.1668	0.0631	0.0230	0.0075	0.0023	0.0007	0.0002	0.0000	0.0000
	1		0.4818	0.2525	0.1182	0.0501	0.0193	0.0067	0.0021	0.0006	0.0001
	2		0.7618	0.5198	0.3096	0.1637	0.0774	0.0327	0.0123	0.0041	0.0012
	3		0.9174	0.7556	0.5489	0.3530	0.2019	0.1028	0.0464	0.0184	0.0064
	4		0.9779	0.9013	0.7582	0.5739	0.3887	0.2348	0.126	0.05958	0.0245
	5		0.9953	0.9681	0.8943	0.7653	0.5968	0.4197	0.2639	0.1471	0.0717
	6		0.9992	0.9917	0.9623	0.8929	0.7752	0.6188	0.4478	0.2902	0.1662
	7		0.9999	0.9983	0.9891	0.9598	0.8954	0.7872	0.6405	0.4743	0.3145
	8		1.0000	0.9997	0.9974	0.9876	0.9597	0.9006	0.8011	0.6626	0.5000
	9		1.0000	1.0000	0.9995	0.9969	0.9873	0.9617	0.9081	0.8166	0.6855
	10		1.0000	1.0000	0.9999	0.9994	0.9968	0.988	0.9652	0.9174	0.8338
	11		1.0000	1.0000	1.0000	0.9999	0.9993	0.997	0.9894	0.9699	0.9283
	12		1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9975	0.9914	0.9755
	13		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9981	0.9936
	14		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9988
	15		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
16		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	
18	0		0.1501	0.0536	0.0180	0.0056	0.0016	0.0004	0.0001	0.0000	0.0000

7.1. CUMULATIVE BINOMIAL PROBABILITIES  $P(X \leq c)$  TABLE

	1		0.4503	0.2241	0.0991	0.0395	0.0142	0.0046	0.0013	0.0003	0.0000
	2		0.7338	0.4797	0.2713	0.1353	0.0600	0.0236	0.0082	0.0025	0.0007
	3		0.9018	0.7202	0.5010	0.3057	0.1646	0.0783	0.0328	0.0120	0.0038
	4		0.9718	0.8794	0.7164	0.5187	0.3327	0.1886	0.0942	0.0411	0.0154
	5		0.9936	0.9581	0.8671	0.7175	0.5344	0.355	0.2088	0.1077	0.04813
	6		0.9988	0.9882	0.9487	0.8610	0.7217	0.5491	0.3743	0.2258	0.1189
	7		0.9998	0.9973	0.9837	0.9431	0.8593	0.7283	0.5634	0.3915	0.2403
	8		1.0000	0.9995	0.9957	0.9807	0.9404	0.8609	0.7368	0.5778	0.4073
	9		1.0000	0.9999	0.9991	0.9946	0.9790	0.9403	0.8653	0.7473	0.5927
	10		1.0000	1.0000	0.9998	0.9988	0.9939	0.9788	0.9424	0.872	0.7597
	11		1.0000	1.0000	1.0000	0.9998	0.9986	0.9938	0.9797	0.9463	0.8811
	12		1.0000	1.0000	1.0000	1.0000	0.9997	0.9986	0.9942	0.9817	0.9519
	13		1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9846
	14		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9990	0.9962
	15		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993
	16		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	17		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n$	$x$	$p$	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
19	0		0.8262	0.6812	0.5606	0.4604	0.3774	0.3086	0.2519	0.2051	0.1666
	1		0.9847	0.9454	0.8900	0.8249	0.7547	0.6829	0.6121	0.5440	0.4798
	2		0.9991	0.9939	0.9817	0.9616	0.9335	0.8979	0.8561	0.8092	0.7585
	3		1.0000	0.9995	0.9978	0.9939	0.9868	0.9757	0.9602	0.9398	0.9147
	4		1.0000	1.0000	0.9998	0.9993	0.9980	0.9956	0.9915	0.9853	0.9765
	5		1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9986	0.9971	0.9949
	6		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9991
	7		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
20	0		0.8179	0.6676	0.5438	0.4420	0.3585	0.2901	0.2342	0.1887	0.1516
	1		0.9831	0.9401	0.8802	0.8103	0.7358	0.6605	0.5869	0.5169	0.4516
	2		0.9990	0.9929	0.9790	0.9561	0.9245	0.8850	0.8390	0.7879	0.7334
	3		1.0000	0.9994	0.9973	0.9926	0.9841	0.9710	0.9529	0.9294	0.9007
	4		1.0000	1.0000	0.9997	0.9990	0.9974	0.9944	0.9893	0.9817	0.9710
	5		1.0000	1.0000	1.0000	0.9999	0.9997	0.9991	0.9981	0.9962	0.9932
	6		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9994	0.9987
	7		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	8		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n$	$x$	$p$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
19	0		0.1351	0.0456	0.0144	0.0042	0.0011	0.0003	0.0000	0.0000	0.0000
	1		0.4203	0.1985	0.0829	0.0310	0.0104	0.0031	0.0008	0.0002	0.0000
	2		0.7054	0.4413	0.2369	0.1113	0.0462	0.0170	0.0055	0.0015	0.0004
	3		0.8850	0.6841	0.4551	0.2631	0.1332	0.0591	0.0230	0.0077	0.0022
	4		0.9648	0.8556	0.6733	0.4654	0.2822	0.1500	0.0696	0.0280	0.0096
	5		0.9914	0.9463	0.8369	0.6678	0.4739	0.2968	0.1629	0.0777	0.0318
	6		0.9983	0.9837	0.9324	0.8251	0.6655	0.4812	0.3081	0.1727	0.0835
	7		0.9997	0.9959	0.9767	0.9225	0.8180	0.6656	0.4878	0.3169	0.1796
	8		1.0000	0.9992	0.9933	0.9713	0.9161	0.8145	0.6675	0.4940	0.3238
	9		1.0000	0.9999	0.9984	0.9911	0.9674	0.9125	0.8139	0.6710	0.5000
	10		1.0000	1.0000	0.9997	0.9977	0.9895	0.9653	0.9115	0.8159	0.6762
	11		1.0000	1.0000	1.0000	0.9995	0.9972	0.9886	0.9648	0.9129	0.8204
	12		1.0000	1.0000	1.0000	0.9999	0.9994	0.9969	0.9884	0.9658	0.9165
	13		1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9969	0.9891	0.9682
	14		1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9972	0.9904
	15		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9978
	16		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996
	17		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	0		0.1216	0.0388	0.0115	0.0032	0.0008	0.0001	0.0000	0.0000	0.0000
	1		0.3917	0.1756	0.0692	0.0243	0.0076	0.0021	0.0005	0.0001	0.0000
	2		0.6769	0.4049	0.2061	0.0913	0.0355	0.0121	0.0036	0.0009	0.0002
	3		0.8670	0.6477	0.4114	0.2252	0.1071	0.0444	0.0160	0.0049	0.0013

	4	0.9568	0.8298	0.6296	0.4148	0.2375	0.1182	0.0510	0.0189	0.0059
	5	0.9887	0.9327	0.8042	0.6172	0.4164	0.2454	0.1256	0.0553	0.0207
	6	0.9976	0.9781	0.9133	0.7858	0.6080	0.4166	0.2500	0.1299	0.0577
	7	0.9996	0.9941	0.9679	0.8982	0.7723	0.6010	0.4159	0.2520	0.1316
	8	0.9999	0.9987	0.9900	0.9591	0.8867	0.7624	0.5956	0.4143	0.2517
	9	1.0000	0.9998	0.9974	0.9861	0.9520	0.8782	0.7553	0.5914	0.4119
10	1.0000	1.0000	0.9994	0.9961	0.9829	0.9468	0.8725	0.7507	0.5881	
11	1.0000	1.0000	0.9999	0.9991	0.9949	0.9804	0.9435	0.8692	0.7483	
12	1.0000	1.0000	1.0000	0.9998	0.9987	0.9940	0.9790	0.9420	0.8684	
13	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9935	0.9786	0.9423	
14	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9936	0.9793	
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9941	
16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	
17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	
18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	

## 7.2 Poisson Distribution Table

The table for  $P(X \leq x) = \sum_{r=0}^x e^{-\lambda} \frac{\lambda^r}{r!}$  gives the probability of that a Poisson random variable  $X$  with mean  $= \lambda$

$\lambda =$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.2	1.4	1.6	1.8
$x =$	0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679	0.3012	0.2466	0.2019
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725	0.7358	0.6626	0.5918	0.5249	0.4628
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371	0.9197	0.8795	0.8335	0.7834	0.7306
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865	0.9810	0.9662	0.9463	0.9212	0.8913
4	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977	0.9963	0.9923	0.9857	0.9763	0.9636
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9994	0.9985	0.9968	0.9940	0.9896
6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9994	0.9987	0.9974
7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9994
8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\lambda =$	2.0	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0	4.5	5.0	5.5
$x =$	0	0.1353	0.1108	0.0907	0.0743	0.0608	0.0498	0.0408	0.0334	0.0273	0.0224	0.0183	0.0111	0.0067
1	0.4060	0.3546	0.3084	0.2674	0.2311	0.1991	0.1712	0.1468	0.1257	0.1074	0.0916	0.0611	0.0404	0.0266
2	0.6767	0.6227	0.5697	0.5184	0.4695	0.4232	0.3799	0.3397	0.3027	0.2689	0.2381	0.1736	0.1247	0.0884
3	0.8571	0.8194	0.7787	0.7360	0.6919	0.6472	0.6025	0.5584	0.5152	0.4735	0.4335	0.3423	0.2650	0.2017
4	0.9473	0.9275	0.9041	0.8774	0.8477	0.8153	0.7806	0.7442	0.7064	0.6678	0.6288	0.5321	0.4405	0.3575
5	0.9834	0.9751	0.9643	0.9510	0.9349	0.9161	0.8946	0.8705	0.8441	0.8156	0.7851	0.7029	0.6160	0.5289
6	0.9955	0.9925	0.9884	0.9828	0.9756	0.9665	0.9554	0.9421	0.9267	0.9091	0.8893	0.8311	0.7622	0.6860
7	0.9989	0.9980	0.9967	0.9947	0.9919	0.9881	0.9832	0.9769	0.9692	0.9599	0.9489	0.9134	0.8666	0.8095
8	0.9998	0.9995	0.9991	0.9985	0.9976	0.9962	0.9943	0.9917	0.9883	0.9840	0.9786	0.9597	0.9319	0.8944
9	1.0000	0.9999	0.9998	0.9996	0.9993	0.9989	0.9982	0.9973	0.9960	0.9942	0.9919	0.9829	0.9682	0.9462
10	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9995	0.9992	0.9987	0.9981	0.9972	0.9933	0.9863	0.9747
11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9996	0.9994	0.9991	0.9976	0.9945	0.9890
12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997	0.9992	0.9980	0.9955
13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9993	0.9983
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\lambda =$	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	11.0	12.0	13.0	14.0	15.0
$x =$	0	0.0025	0.0015	0.0009	0.0006	0.0003	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0174	0.0113	0.0073	0.0047	0.0030	0.0019	0.0012	0.0008	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000
2	0.0620	0.0430	0.0296	0.0203	0.0138	0.0093	0.0062	0.0042	0.0028	0.0012	0.0005	0.0002	0.0001	0.0000

3	0.1512	0.1118	0.0818	0.0591	0.0424	0.0301	0.0212	0.0149	0.0103	0.0049	0.0023	0.0011	0.0005	0.0002
4	0.2851	0.2237	0.1730	0.1321	0.0996	0.0744	0.0550	0.0403	0.0293	0.0151	0.0076	0.0037	0.0018	0.0009
5	0.4457	0.3690	0.3007	0.2414	0.1912	0.1496	0.1157	0.0885	0.0671	0.0375	0.0203	0.0107	0.0055	0.0028
6	0.6063	0.5265	0.4497	0.3782	0.3134	0.2562	0.2068	0.1649	0.1301	0.0786	0.0458	0.0259	0.0142	0.0076
7	0.7440	0.6728	0.5987	0.5246	0.4530	0.3856	0.3239	0.2687	0.2202	0.1432	0.0895	0.0540	0.0316	0.0180
8	0.8472	0.7916	0.7291	0.6620	0.5925	0.5231	0.4557	0.3918	0.3328	0.2320	0.1550	0.0998	0.0621	0.0374
9	0.9161	0.8774	0.8305	0.7764	0.7166	0.6530	0.5874	0.5218	0.4579	0.3405	0.2424	0.1658	0.1094	0.0699
10	0.9574	0.9332	0.9015	0.8622	0.8159	0.7634	0.7060	0.6453	0.5830	0.4599	0.3472	0.2517	0.1757	0.1185
11	0.9799	0.9661	0.9467	0.9208	0.8881	0.8487	0.8030	0.7520	0.6968	0.5793	0.4616	0.3532	0.2600	0.1848
12	0.9912	0.9840	0.9730	0.9573	0.9362	0.9091	0.8758	0.8364	0.7916	0.6887	0.5760	0.4631	0.3585	0.2676
13	0.9964	0.9929	0.9872	0.9784	0.9658	0.9486	0.9261	0.8981	0.8645	0.7813	0.6815	0.5730	0.4644	0.3632
14	0.9986	0.9970	0.9943	0.9897	0.9827	0.9726	0.9585	0.9400	0.9165	0.8540	0.7720	0.6751	0.5704	0.4657
15	0.9995	0.9988	0.9976	0.9954	0.9918	0.9862	0.9780	0.9665	0.9513	0.9074	0.8444	0.7636	0.6694	0.5681
16	0.9998	0.9996	0.9990	0.9980	0.9963	0.9934	0.9889	0.9823	0.9730	0.9441	0.8987	0.8355	0.7559	0.6641
17	0.9999	0.9998	0.9996	0.9992	0.9984	0.9970	0.9947	0.9911	0.9857	0.9678	0.9370	0.8905	0.8272	0.7489
18	1.0000	0.9999	0.9999	0.9997	0.9993	0.9987	0.9976	0.9957	0.9928	0.9823	0.9626	0.9302	0.8826	0.8195
19	1.0000	1.0000	1.0000	0.9999	0.9997	0.9995	0.9989	0.9980	0.9965	0.9907	0.9787	0.9573	0.9235	0.8752
20	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9991	0.9984	0.9953	0.9884	0.9750	0.9521	0.9170
21	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9993	0.9977	0.9939	0.9859	0.9712	0.9469
22	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9990	0.9970	0.9924	0.9833	0.9673
23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9995	0.9985	0.9960	0.9907	0.9805
24	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9993	0.9980	0.9950	0.9888
25	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990	0.9974	0.9938
26	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9987	0.9967
27	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9983
28	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9991
29	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996
30	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
31	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
32	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 7.2: The Poisson cumulative probabilities

## 7.3 Normal Distribution Table

### 7.3.1 Negative Z-values Table

Cumulative probabilities for **NEGATIVE** Z-values are shown in the following table:

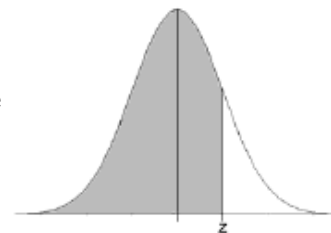


Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641



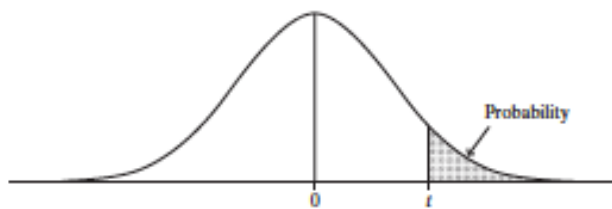
7.3.2 Positive  $Z$ -values Table

Cumulative probabilities for **POSITIVE**  $Z$ -values are shown in the following table:



$Z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
<b>0.0</b>	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
<b>0.1</b>	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
<b>0.2</b>	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
<b>0.3</b>	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
<b>0.4</b>	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
<b>0.5</b>	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
<b>0.6</b>	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
<b>0.7</b>	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
<b>0.8</b>	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
<b>0.9</b>	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
<b>1.0</b>	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
<b>1.1</b>	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
<b>1.2</b>	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
<b>1.3</b>	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
<b>1.4</b>	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
<b>1.5</b>	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
<b>1.6</b>	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
<b>1.7</b>	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
<b>1.8</b>	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
<b>1.9</b>	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
<b>2.0</b>	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
<b>2.1</b>	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
<b>2.2</b>	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
<b>2.3</b>	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
<b>2.4</b>	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
<b>2.5</b>	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
<b>2.6</b>	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
<b>2.7</b>	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
<b>2.8</b>	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
<b>2.9</b>	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
<b>3.0</b>	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
<b>3.1</b>	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
<b>3.2</b>	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
<b>3.3</b>	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
<b>3.4</b>	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

## 7.4 Student's -t Distribution Table



<i>t</i> -Distribution Critical Values												
df, $\nu$	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
$z^*$	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level $C$											

## Chapter 8

# Probability and Statistics By Python