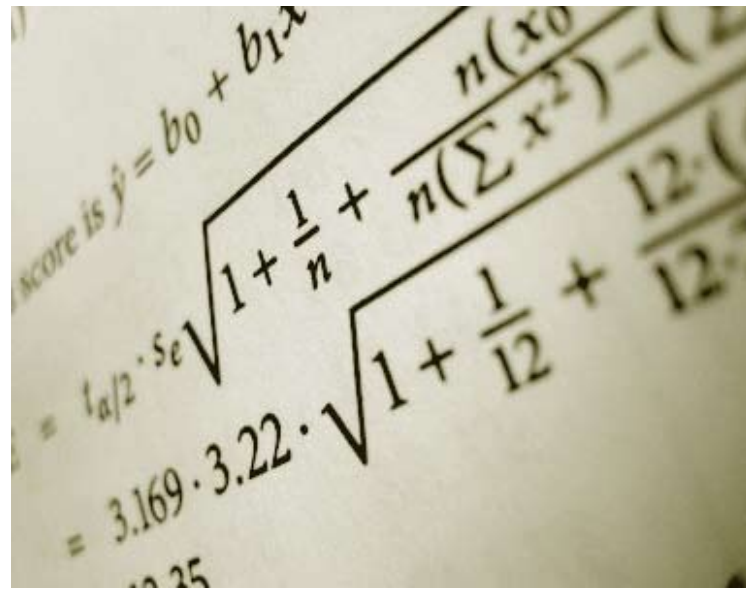




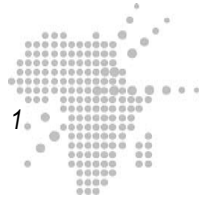
# Probability And Statistics



Prepared by Paul CHEGE


$$\text{score is } \hat{y} = b_0 + b_1x$$
$$= t_{\alpha/2} \cdot se \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}}$$
$$= 3.169 \cdot 3.22 \cdot \sqrt{1 + \frac{1}{12} + \frac{12 \cdot (\dots)^2}{\dots}}$$





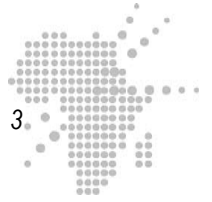
## NOTICE

This document is published under the conditions of the Creative Commons  
[http://en.wikipedia.org/wiki/Creative\\_Commons](http://en.wikipedia.org/wiki/Creative_Commons)  
Attribution  
<http://creativecommons.org/licenses/by/2.5/>  
License (abbreviated “cc-by”), Version 2.5.



## TABLE OF CONTENTS

I. Probability and Statistics	3
II. Prerequisite Course or Knowledge	3
III. Time	3
IV. Materials	3
V. Module Rationale	3
VI. Content	4
6.1 Overview	4
6.2 Outline	5
6.3 Graphic Organizer	6
VII. General Objective(s)	7
VIII. Specific Learning Activities	7
IX. Teaching and Learning Activities	9
X. Compiled List of all Key Concepts (Glossary)	12
XI. Compiled List of Compulsory Readings	18
XII. Compiled List of Resources	19
XIII. Compiled List of Useful Links	20
XIV. Learning Activities	21
XV. Synthesis of the Module	112
XVI. Summative Evaluation	113
XVII. References	121
XVIII. Student records	122
XIX. Main Author of the Module	123



## **I. Probability and Statistics**

by Paul Chege

## **II. Prerequisite courses or knowledge**

Secondary school statistics and probability.

## **III. Time**

The total time for this module is 120 study hours.

## **IV. Material**

Students should have access to the core readings specified later. Also, they will need a computer to gain full access to the core readings. Additionally, students should be able to install the computer software wxMaxima and use it to practice algebraic concepts.

## **V. Module Rationale**

Probability and Statistics, besides being a key area in the secondary schools' teaching syllabuses, it forms an important background to advanced mathematics at tertiary level. Statistics is a fundamental area of Mathematics that is applied across many academic subjects and is useful in analysis in industrial production. The study of statistics produces statisticians that analyse raw data collected from the field to provide useful insights about a population. The statisticians provide governments and organizations with concrete backgrounds of a situation that helps managers in decision making. For example, rate of spread of diseases, rumours, bush fires, rainfall patterns, and population changes.

On the other hand, the study of probability helps decision making in government agents and organizations based on the theory of chance. For example:- predicting the male and female children born within a given period and projecting the amount of rainfall that regions expect to receive based on some historical data on rainfall patterns. Probability has also been extensively used in the determination of high, middle and low quality products in industrial production e.g the number of good and defective parts expected in an industrial manufacturing process.



## VI. Content

### 6.1 Overview

This module consists of three units:

#### **Unit 1: Descriptive Statistics and Probability Distributions**

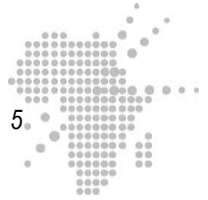
Descriptive statistics in unit one is developed either as an extension of secondary mathematics or as an introduction to first time learners of statistics. It introduces the measures of dispersion in statistics. The unit also introduces the concept of probability and the theoretical treatment of probability.

#### **Unit 2: Random variables and Test Distributions**

This unit requires Unit 1 as a prerequisite. It develops from the moment and moment generating functions, Markov and Chebychev inequalities, special univariate distributions, bivariate probability distributions and analyses conditional probabilities. The unit gives insights into the analysis of correlation coefficients and distribution functions of random variables such as the Chi-square, t and F.

#### **Unit 3: Probability Theory**

This unit builds up from unit 2. It analyses probability using indicator functions. It introduces Bonferoni inequality random vectors,, generating functions, characteristic functions and statistical independence random samples. It develops further the concepts of functions of several random variables and independence of  $X$  and  $S^2$  in normal samples order statistics. The unit summarises with the treatment of convergence and limit theorems.



## 6.2 Outline: Syllabus

### Unit 1 ( 40 hours): Descriptive Statistics and Probability Distributions

#### Level 1. Priority A. No prerequisite.

Frequency distributions relative and cumulative distributions, various frequency curves, mean, Mode Median. Quartiles and Percentiles, Standard deviation, symmetrical and skewed distributions. Probability; sample space and events; definition of probability, properties of probability; random variables; probability distributions, expected values of random variables; particular distributions; Bernoulli, binomial, Poisson, geometric, hypergeometric, uniform, exponential and normal. Bivariate frequency distributions. Joint probability tables and marginal probabilities.

### Unit 2 ( 40 hours): Random Variables and Test Distributions

#### Level 2. Priority B. Statistics 1 is prerequisite.

Moment and moment generating function. Markov and Chebychev inequalities, special Univariate distributions. Bivariate probability distribution; Joint Marginal and conditional distributions; Independence; Bivariate expectation Regression and Correlation; Calculation of regression and correlation coefficient for bivariate data. Distribution function of random variables, Bivariate normal distribution. Derived distributions such as Chi-Square. t. and F.

### Unit 3 ( 40 hours): Probability Theory

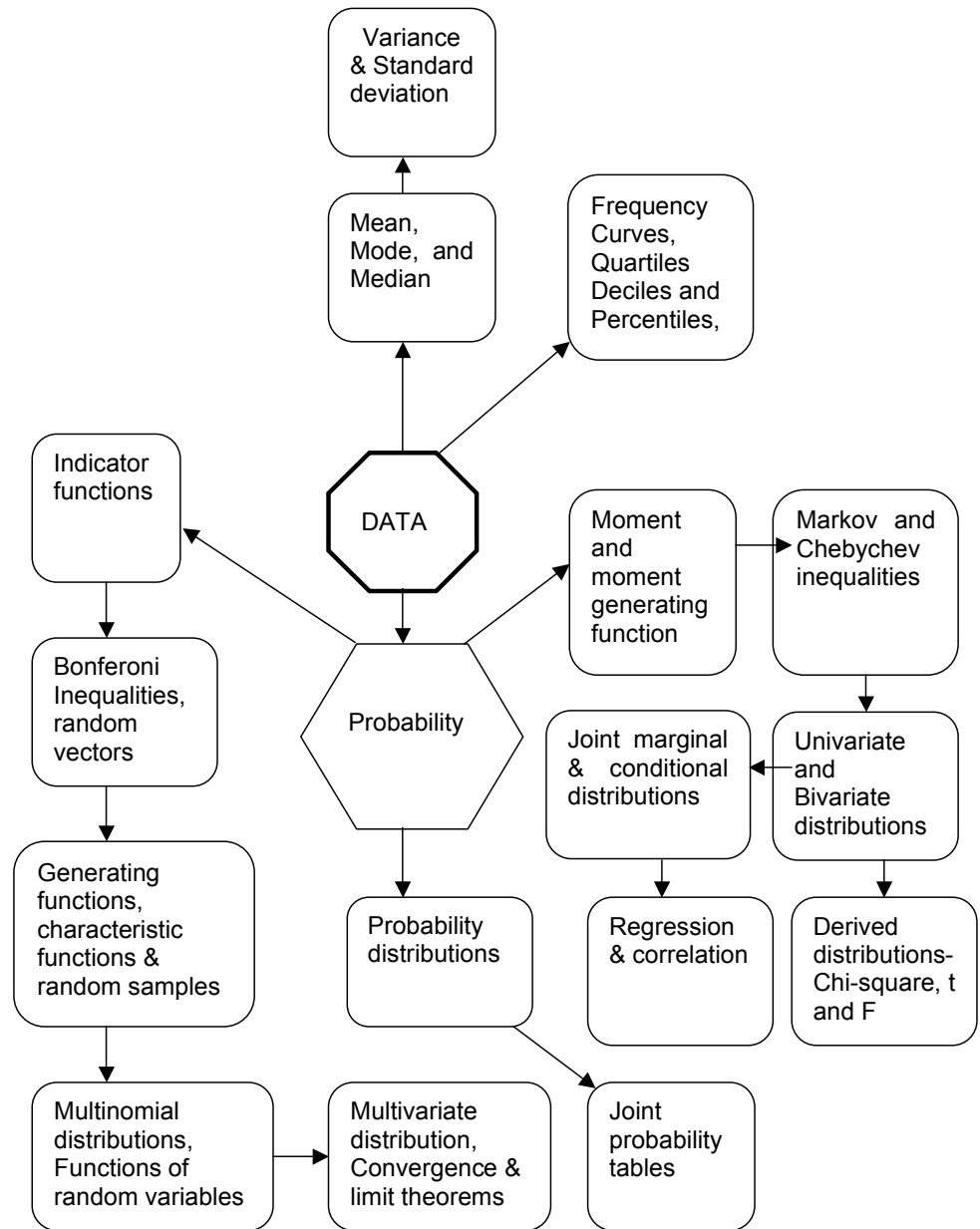
#### Level 3. Priority C. Statistics 2 is prerequisite.

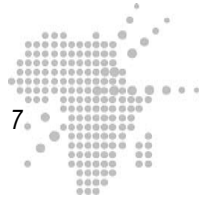
Probability: Use of indicator functions. Bonferoni inequality Random vectors. Generating functions. Characteristics functions. Statistical independence Random samples. Multinomial distribution. Functions of several random variables.

The independence of  $X$  and  $S^2$  in normal samples Order statistics Multivariate normal distribution. Convergence and limit theorems. Practical exercises.



### 6.3 Graphic Organiser





## VII. General Objective(s)

By the end of this module, the trainee should be able to compute the various measures of dispersions in statistics and work out probabilities based on laws of probability and carry out tests on data using the theories of probability

## VIII. Specific Learning Objectives (Instructional Objectives)

### Unit 1: Descriptive Statistics and Probability Distributions ( 40 Hours)

By the end of unit 1, the trainee should be able to:

- Draw various frequency curves
- Work out the mean, mode, median, quartiles, percentiles and standard deviations of discrete and grouped data
- Define and state the properties of probability
- Illustrate random variables, probability distributions, and expected values of random variables.
- Illustrate Bernoulli, Binomial, Poisson, Geometric, Hypergeometric, Uniform, Exponential and Normal distributions
- Investigate Bivariate frequency distributions
- Construct joint probability tables and marginal probabilities.

### Unit 2: Random Variables and Test Distributions ( 40 Hours)

By the end of unit 2, the trainee should be able to:

- Illustrate moment and moment generating functions
- Analyse Markov and Chebychev inequalities
- Examine special Univariate distributions, bivariate probability distributions, Joint marginal and conditional distributions.
- Show Independence, Bivariate expectation, regression and correlation
- Calculate regression and correlation coefficient for bivariate data
- Show distribution function of random variables.
- Examine Bivariate normal distribution
- Illustrate derived distributions such as Chi-Square, t, and F.





### **Unit 3: Probability Theory ( 40 Hours)**

By the end of unit 3, the trainee should be able to:

- Use indicator functions in probability
- Show Bonferoni inequality random vectors
- Illustrate generating and characteristic functions
- Examine statistical independence random samples and multinomial distribution
- Evaluate functions of several random variables
- Illustrate the independence of  $X$  and  $S^2$  in normal samples order statistics
- Show multivariate normal distribution
- Illustrate convergence and limit theorems.
- Work out practical exercises.



## IX. Teaching and Learning Activities

### 9.1 Pre-assessment

Basic mathematics is a pre-requisite for Probability and Statistics.

#### Questions

- 1) When a die is rolled, the probability of getting a number greater than 4 is
  - A.  $\frac{1}{6}$
  - B.  $\frac{1}{3}$
  - C.  $\frac{1}{2}$
  - D. 1
- 2) A single card is drawn at random from a standard deck of cards. Find the probability that is a queen.
  - A.  $\frac{1}{13}$
  - B.  $\frac{1}{52}$
  - C.  $\frac{4}{13}$
  - D.  $\frac{1}{2}$
- 3) Out of 100 numbers, 20 were 4's, 40 were 5's, 30 were 6's and the remainder were 7's. Find the arithmetic mean of the numbers.
  - A. 0.22
  - B. 0.53
  - C. 2.20
  - D. 5.30



4) Calculate the mean of the following data.

Height (cm)	Class mark (x)
60 - 62	61
63 - 65	64
66 - 68	67
69 - 71	70
72 - 74	73

- A. 57.40  
B. 62.00  
C. 67.45  
D. 72.25
- 5) Find the mode of the following data: 5, 3, 6, 5, 4, 5, 2, 8, 6, 5, 4, 8, 3, 4, 5, 4, 8, 2, 5, and 4.
- A. 4  
B. 5  
C. 6  
D. 8
- 6) The range of the values a probability can assume is
- A. From 0 to 1  
B. From -1 to +1  
C. From 1 to 100  
D. From 0 to  $\frac{1}{2}$
- 7) Find the median of the following data: 8, 7, 11, 5, 6, 4, 3, 12, 10, 8, 2, 5, 1, 6, 4.
- A. 12  
B. 5  
C. 8  
D. 6
- 8) Find the range of the set of numbers: 7, 4, 10, 9, 15, 12, 7, 9.
- A. 9  
B. 11  
C. 7  
D. 8.88



9) When two coins are tossed, the sample space is

- A. H, T and HT
- B. HH, HT, TH, TT
- C. HH, HT, TT
- D. H, T

10) If a letter is selected at random from the word “Mississippi”, find the probability that it is an “i”

- A.  $\frac{1}{8}$
- B.  $\frac{1}{2}$
- C.  $\frac{3}{11}$
- D.  $\frac{4}{11}$

### Answer Key

1. B      2. A      3. D      4. C      5. B  
6. A      7. D      8. B      9. B      10. D

### Pedagogical Comment For Learners

This pre-assessment is meant to give the learners an insight into what they can remember regarding Probability and Statistics. A score of less than 50% in the pre-assessment indicates the learner needs to revise Probability and Statistics covered in secondary mathematics. The pre-assessment covers basic concepts that trainees need to be familiar with before progressing with this module. Please revise Probability and Statistics covered in secondary mathematics to master the basics if you have problems with this pre-assessment.



## X. Key Concepts ( Glossary)

**Mutually Exclusive:** Two events are mutually exclusive if they cannot occur at the same time.

**Variance** of a set of data is defined as the square of the standard deviation i.e variance =  $s^2$ .

**A trial:** This refers to an activity of carrying out an experiment like picking a card from a deck of cards or rolling a die or dices

**Sample space:** This refers to all possible outcomes of a probability experiment. e.g. in tossing a coin, the outcomes are either Head(H) or tail(T)

**A random variable:** is a function that assigns a real number to every possible result of a random experiment.

**Random sample** is one chosen by a method involving an unpredictable component.

**Bernoulli distribution:** is a discrete probability distribution, which takes value 1 with success probability  $p$  and value 0 with failure probability  $q = 1 - p$ .

**Binomial distribution** is the discrete probability distribution of the number of successes in a sequence of  $n$  independent yes/no experiments, each of which yields success with probability  $p$

**Hypergeometric distribution:** is a discrete probability distribution that describes the number of successes in a sequence of  $n$  draws from a finite population without replacement.

**Poisson distribution:** is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate, and are independent of the time since the last event

**Correlation:** is a measure of association between two variables.

**Regression:** is a measure used to examine the relationship between one dependent and one independent variable.

**Chi-square test** is any statistical hypothesis test in which the test statistic has a chi-square distribution when the null hypothesis is true, or any in which the probability distribution of the test statistic (assuming the null hypothesis is true) can be made to approximate a chi-square distribution as closely as desired by making the sample size large enough.

**Multivariate normal distribution** is a specific probability distribution, which can be thought of as a generalization to higher dimensions of the one-dimensional normal distribution.

**$t$ -test** is any statistical hypothesis test for two groups in which the test statistic has a Student's  $t$  distribution if the null hypothesis is true



## Statistical Terms

1. **Raw data:** Data that has not been organised numerically.
2. **Arrays:** An arrangement of raw data numerical data in ascending order of magnitude.
3. **Range:** the difference between the largest and the smallest numbers in a data.
4. **Class intervals:** In a range of grouped data e.g 21-30, 31-40 etc, then 21-30 is called the class interval.
5. **Class limits:** In a class interval of 21-30, then 21 and 30 are called class limits.
6. **Lower class limits (l.c.l) :** In the class interval 21-30, the lower class limit is 21
7. **Upper class limit (u.c.l):** in the class interval 21-30, the upper class limit is 30
8. **Lower and upper class boundaries:** In the class interval 21-30, the lower class boundary is 20.5 and the upper class boundary is 30.5. These boundaries assume that theoretically measurements for a class interval 21-30 includes all the numbers from 20.5 to 30.5
9. **Class Interval:** In a class 21-30, then the class interval is the difference between the upper class limit and the lower class limit i.e.  $30.5 - 20.5 = 10$ . The class interval is also known as class width or class size.
10. **Class Mark or Mid-point:** In a class interval 21-30, the class mark is the average

$$\text{of 21 and 30 i.e. } \frac{21 + 30}{2} = 25.5$$

11. **Frequency Distributions:** large masses of raw data maybe arranged in classes in tabular form with their corresponding frequencies. e.g.

Mass (kg)	10-19	20-29	30-39	40-49
Number of pupils (f)	5	7	10	6

This tabular arrangement is called a frequency distribution or frequency table.

12. **Cumulative Frequency:** For the following frequency distribution, the cumulative frequencies are calculated as additions of individual frequencies

Mass ( X)	20-24	25-29	30-34	35-39	40-44
Frequency (f)	4	10	16	8	2
Cumulative Frequency( C.F)	4	4+10=14	14+16=30	30+8=38	38+2=40



Hence the cumulative frequency of a value is its frequency plus frequencies of all smaller values.

The above table is called a **Cumulative Frequency** table.

### 13. Relative – Frequency Distributions: In a frequency distribution

Mass ( X)	20-24	25-29	30-34	35-39	40-44
Frequency (f)	4	10	16	8	2

$$\sum f = 40$$

The **relative frequency** of a class 25-29 is the frequency of the class divided by the total frequency of all classes (cumulative frequency) and generally expressed as a percentage.

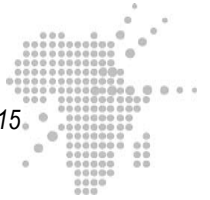
#### Example:

$$\text{The relative frequency of the class 25-29} = \frac{f}{\sum f} \times 100\% = \frac{10}{40} \times 100 = 25\%$$

Note: the sum of relative frequencies is 100% or 1.

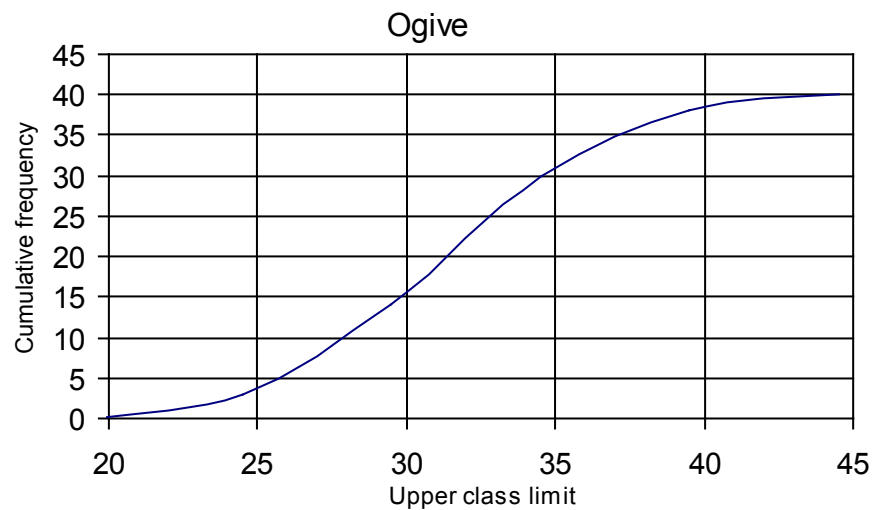
### 14. Cumulative Frequency Curve ( Ogive)

Mass ( X)	20-24	25-29	30-34	35-39	40-44
Frequency (f)	4	10	16	8	2
Cumulative Frequency( C.F)	4	4+10=14	14+16=30	30+8=38	38+2=40



From the above cumulative frequency table, we can draw a graph of cumulative frequency verses the upper class boundaries.

Upper class boundaries	24.5	29.5	34.5	39.5	44.5
Cumulative frequencies	3	14	30	38	40



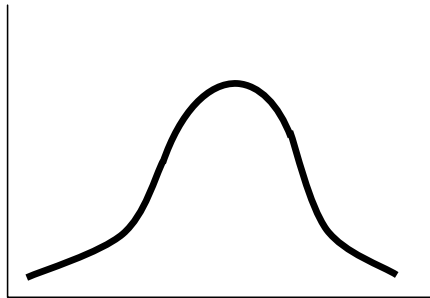
**Note:** From the cumulative frequency data, the first plotting point is ( 24.5, 3). If we started our graph at this point, it would remain hanging on the y-axis. We create another point (19.5, 0) as a starting point. 19.5 is the projected upper class boundary of the preceding class.





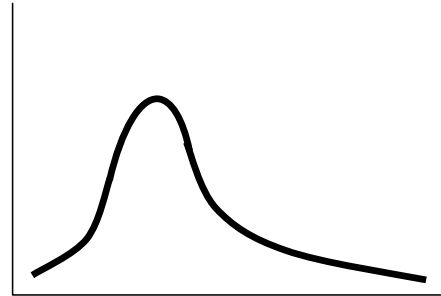
## Shapes of Frequency Curves

Symmetrical or bell-shaped.



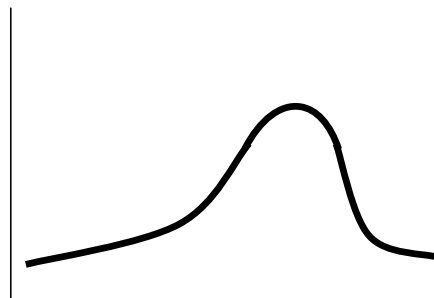
Has equal frequency to the left and right of the central maximum e.g. normal curve

Skewed to the right ( positive skewness)



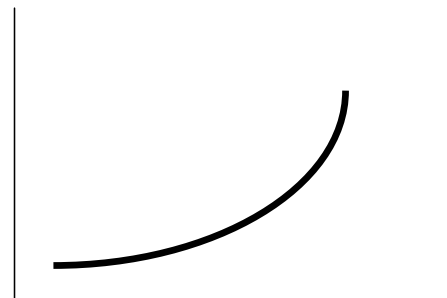
Has the maximum towards the left and the longer tail to the right

Skewed to the left ( Negative skewness)



Has the maximum towards the right of the and the longer tail to the left

J -Shaped



Has the maximum occurring at the right end

## Thank You for previewing this eBook

You can read the full version of this eBook in different formats:

- HTML (Free /Available to everyone)
- PDF / TXT (Available to V.I.P. members. Free Standard members can access up to 5 PDF/TXT eBooks per month each month)
- Epub & Mobipocket (Exclusive to V.I.P. members)

To download this full book, simply select the format you desire below

