General, Biological and Biomedical Statistics

General, Biological and Biomedical Statistics

By

Waleed Al-Murrani

Edited by Richard Handy

Cambridge
Scholars
Publishing

General, Biological and Biomedical Statistics

By Waleed Al-Murrani

Edited by Richard Handy

This book published 2024 (self published by the author 2021)

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

The material contained in this book is set in good faith for general use and application and no liability can be accepted for any loss or expense incurred as a result of relying in particular circumstances on topics included in this book.

ISBN: 978-1-0364-1114-5
ISBN (Ebook): 978-1-0364-1115-2

In memory of my aunt, Najiya Murrani, and my mother.

TABLE OF CONTENTS

LIST OF FIGURES

PREFACE

MY STORY:
WHY I WROTE THIS BOOK

Professor Waleed K. Al-Murrani
BVM&S (University of Baghdad), Iraq
Res. Cert. (University of Bradford), UK
Dip. & PhD (University of Edinburgh, UK)
Professor of Genetics and Biostatistics (University of Baghdad, Iraq)
Professor and Honorary Professor (University of Plymouth, UK)

I graduated from the College of Veterinary Medicine and Surgery, University of Baghdad, in 1961. In 1970, I was awarded a scholarship from the British Council and University of Baghdad to do a PhD in Animal Genetics at the Institute of Animal Genetics, Faculty of Science, University of Edinburgh, UK. Upon enrolment, I was asked to take a Diploma in Genetics as a first step towards the PhD.

I was hit by many modules majoring in statistics for ‘population’ studies, as well as quantitative genetics and animal breeding. This was my first encounter with statistics. I found myself able to master all other eight modules while facing difficulty with statistics. Luckily, the first semester of statistics, given by Professor R.C. Roberts, was easygoing. The associated Quantitative Genetics module was given by Professor D.S. Falconer through his book *Introduction to Quantitative Genetics*. Both presented the material in an easy comprehensible manner relying very little on mathematics and calculus.

The second statistics module, in the second semester, was given by Professor W.G. Hill in association with a module in Animal Breeding delivered by Professor Alan Robertson; both modules were like a journey through 'Hell' because of the super-complicated formulae, which leaned heavily on mathematics, particularly algebra and calculus. At that time, I must confess, I lived the drama of my early life and decided to quit and change speciality to animal hygiene. To achieve that, I made an appointment with the professor of Animal Hygiene at the University of Edinburgh and went to see him. He was very nice and sympathised, but he advised me against wasting six months of my scholarship and pumped me up to face the challenge. I went back to the diploma and to the statistics module with a decision to fight and, honourably die, rather than quit.

Under this pressure, I arranged with an undergraduate mathematics student to lecture me in algebra and some calculus. The chap was very smart and enabled me to comprehend the material; I paid him a hamburger and a cup of tea for each two-hour lecture. With the extra knowledge and information, I managed to do very well in the final exam to the extent that Professor Falconer and Professor Roberts offered me to continue with a PhD under joint supervision. I started my research and analysed my data using an Olivetti desk computer and the University Frame Computer with punched cards.

After completing the PhD in 1973, I decided to do two things upon my return to Iraq: first, to introduce, for the first time, teaching biomedical statistics module for MSc students at the College of Veterinary Medicine and the College of Medicine; second, to make statistics very easy to comprehend and enjoy. This decision was based on a fact I discovered after surviving 'Hell'; that statistics does not need much of the mathematics tools, and it is a life philosophy rather than mechanics.

I introduced and taught statistics at the College of Veterinary Medicine, the College of Science, the College of Medicine, the College of Education for Girls, and the Institute of Genetic Engineering and Biotechnology for Higher Studies in Iraq. I enjoyed teaching statistics and helping students in data analysis. I am delighted that students I taught over 40 years still remember my statistics lectures and the relevant jokes.

In Iraq, I, along with a group of three crop scientists, authored a textbook *Principles of Biological Statistics* in Arabic (1986, Pub: Baghdad University) for students and researchers in agricultural and veterinary sciences, and for general use. We divided responsibility for the chapters, but later authorised one of us to edit the whole volume and put it into a unified style. This co-author did a great job, but the book was edited to serve mainly students of agriculture. Though I feel proud to be one of the authors of a book still in use in Iraq and the region, I have always thought that the final edited version masked the personality of each author.

For this book, I started to collect material and formulate concepts and philosophies as early as the 1970s. The intention is to offer students and researchers at all levels, mainly in biological and biomedical sciences, but also in other disciplines, a simple approach to understand the concept and philosophy first. My aim is to enable students and researchers to select the appropriate methods to design and analyse data, and to reach fruitful conclusions. I firmly believe that excessive use of statistics is as dangerous as not using statistics at all.

From the onset of studying and understanding statistics in the early 1970s, I relied heavily on great statistics textbooks and references such as: Snedecor and Cochran, Steel and Tory, Bradford Hill, the lecture notes of Professor Roberts and Professor Hill during the Diploma (1970); in addition to the richness of published papers by eminent animal breeders and hundreds of other published works. I always supplemented my knowledge by reading widely, including articles on the internet which, if you know what you need, can be very helpful. My thorough understanding and application of many statistical methodologies, concepts and philosophies drawn from such textbooks and references over 40 years of teaching and research helped me to construct this book and select its content. Supporting hundreds of undergraduate and postgraduate research students, staff members and researchers in their experimental design, data analyses, presentation, and publishing in a wide range of disciplines in biological, biomedical, medical, veterinary, and agricultural sciences, pure sciences and even engineering helped in formulating my own concepts, philosophy, and approach to ‘statistics’.

This book is not a traditional one. It does not reference every statement and methodology because such are the properties of those who laid the foundation for statistics and its application. However, some key general references are listed at the end of this book. Also, at the end of each chapter, there are selected references targeted at the content therein. My contribution, when new or different from what is already published, is made clear in the text and for which I bear full responsibility.

When examples were needed, I used either my own data or hypothetical data, tailored to show the case. The examples were carefully selected to be used on a personal level by the reader. These examples were often worked by hand to show the reader the steps in the calculations. However, most students and researchers currently use computer software statistical packages to analyse data. However, when computers and software programmes are not in use or available, one can organise the data in worksheets and analyse by hand. I have chosen to show illustrations drawn from Minitab® Statistical Software (Version 17 onwards) which is provided by the University of Plymouth to all students and staff. In this respect, it should be noted that all statistical packages do nearly the same job. I have endeavoured to present the material in a standard format and in chapters, especially for basic statistics; but in a problem-oriented way that enables debate and consideration of the philosophical arguments. In this respect, I hope this book is useful for beginners as well as researchers at a more advanced level.

Waleed K. Al-Murrani, 2024, Plymouth, UK

ACKNOWLEDGEMENTS

I wish to thank the thousands of undergraduate and postgraduate students I taught in my long teaching and research career and the many researchers and staff members I helped develop experimental designs, conduct data analyses and publish. Their views and suggestions regarding the selection of topics and the need for a text that is easy to read and understand, yet comprehensive enough for their studies, have been extremely valuable.

In finalizing this book and preparing it for publication, I must thank the colleagues who contributed, in one way or another, towards improving the scientific merit and presentation. Statistician Associate Professor Dr Julian Stander from the School of Engineering and Mathematics, University of Plymouth, read an early draft of this book. His comments and advice led to clear and sound improvements in both the science and the presentation. Dr Stander noted: *“My impression is that it is an extremely interesting and thorough piece of work. It covers many topics, including some quite advanced ones.*”

Associate Professor Dr John Moody from the School of Biological and Marine Sciences, University of Plymouth, read the whole draft and gave an impression as a scientist who frequently uses statistics to analyse his own research and that of his students. I was incredibly pleased to receive this comment from him: "*I think the coverage of the book is fine, particularly for biomedical sciences. I do think it is written clearly. There seems to be a good range of examples – personally when using statistics books in the past I have always liked it when I can find an example that is similar to the sort of thing I had been doing.*”

Mr. Eric Cooke from the University of Southampton read the first five chapters and suggested improvements. He smartly drew the sketch of the relation between samples and population in Chapter 2. Associate Professor Dr Andrew Foey from the School of Biomedical Sciences, University of Plymouth, thoroughly read Chapter 2 on Sampling and made many suggestions, most of which were considered and led to improvements. Associate Professor Dr Lynn McCallum from the School of Biomedical Sciences, University of Plymouth, looked at Sampling for Cellular Studies (Chapter 2) and commented: *“I’ve read from section 2.4 onward, and this is very good, easy to read and follow the principles and your guidance, this is what we do mostly within tissue culture currently.”*

My son, Dr Samer Al-Murrani, CEO of Babylon BioConsulting, read Chapter 2 on Sampling and made many suggestions on sampling for cellular studies, which I took on board. My sincere thanks and appreciation to: Dr Anas Al-Hiti for drawing the normal distribution with its confidence intervals and limits (Figure 1-12) and Dr Salam Al-Sabah for adjusting graph pixels and tables; my daughter Associate Professor Dr Sana Murrani for designing the previous book cover when the latter was self-published; Dr Rabab Alkutbe and Associate Professor Dr Gail Rees for providing the data and drawings of the percentile (1-10 b, c), and my daughter Dr Sally Murrani for reading various drafts and indexing.

My sincere thanks and appreciation to Professor Richard Handy in his capacity as an independent editor. He is a Professor at the School of Biological Sciences, University of Plymouth. He has made an in-depth edit of book content and the English language, as well as the presentation, to bring clarity and additional understanding to the scientific concepts.

Writing this book has been challenging and without the support of my family over many years, this book would not be available to students and researchers. My wife, Selma Al-Sam, my son Samer and my daughters, Sally, Sana, and Sura offered me much support and encouragement; they all contributed in different ways, which I wish to acknowledge with sincere love and affection.

I used Minitab Software to conduct all the statistical analysis needed to solve the examples and to print all the associated figures. I wish to acknowledge the use of Minitab software and in keeping with Minitab policy to state that: *"Portions of information contained in this book are printed with permission of Minitab, LLC. All material remains the exclusive property and copyright of Minitab, LLC. All rights reserved."*

CHAPTER 1

BASIC STATISTICS AND NOTATION

1.1 Why do we need statistics?

We can simply answer this question by saying: We need statistics because there is always variability and uncertainty in nearly all aspects of nature and life. In nature, and in all kinds of scientific research, ‘difference’ or ‘variability’ is the rule; ‘resemblance’ and ‘similarity’ are the exceptions. This is true of biology. Statistics is the science of uncertainty. It is a ‘Philosophy’ rather than ‘Mechanics,’ so please do not simply associate statistics with computer packages, because these are only tools to help you perform complex mathematics. Luckily, you do not need a great amount of mathematics to understand and perform statistical manipulations and procedures.

Statistics revolves around the idea of probability; the likelihood of an outcome or event occurring in the data. Statistics helps us pass judgment on our data by allowing us to understand the variability in the data and therefore the potential sources of error or uncertainty. Statistics is an integral part of the scientific process, and while research might start with an idea or observation, statistics can inform on how to proceed from the very beginning. Research generally follows this sequence of events and thinking:

I. A ‘scientific question’ or ‘hypothesis’ and ‘statistical inference’.

All research work, whether it is intended to be experimental in the laboratory or observation-driven in a survey, is always initiated to address a certain scientific question or hypothesis. This is true for all fields of knowledge. The scientific question or hypothesis needs to be framed in your experimental design in a way which leads to data or observations which, when statistically analysed, will give the answers that will help you reach ‘meaningful conclusions’. Most of research or studies are based on samples taken at random from the target population with the aim of making decisions about that population, this process is termed *‘inference statistics’*. In doing so, we must avoid bias in statistical analysis. To achieve this one should assume a ‘null hypothesis’ which is taken from a Latin word '*nullus*' which means 'none' and is often given the mathematical notation of 'N0'. The statistical analysis will either approve or disprove the null hypothesis and will show the confidence you have in rejecting your null hypothesis; thus, proving your idea or hypothesis is correct. The major *inferences* that statisticians can draw from the data are significance, estimation of parameters, generalisation, and conclusion.

II. An ‘experimental design’ or ‘approach’.

Your data will have variability or differences, so you must design your experiment so that any sources of variability unrelated to your hypothesis are controlled, or at least known, and that the key measurements that relate to your hypothesis are collected, so that you can properly test the null hypothesis (i.e., determine if it is true or not), without errors that might lead to incorrect or false conclusions. The experimental design considerations include, deciding the sample type or primary unit of replication in the study design, sample size (how many replicates), how the samples are grouped or associated with each other to make treatments, controls needed for comparison with the treatments, what to measure and when, and so forth.

III. Samples and sample size.

You will nearly always be unable to get your data for the whole population in which you are interested. Therefore, you must choose a sample of the population and you must do so to avoid bias in the data. You need some statistical insight to choose a sample and a sample size that will be big enough to let you test your null hypothesis. Sometimes in surveys, we can work on the whole population as in a population census.

IV. Methodology and experimental manipulations.

This is what you will do to execute your experimental design and to make the measurements needed for the intended study. The experiment could be conducted in the laboratory or in the field, or the study might be at your desk. For example, where data is mined from scientific literature or databases.

V. Data collection and recording.

The data collection might take many forms, such as a printout of numerical values from an instrument, numbers that you carefully record in your notebook during observations, values that you noted from a display on an instrument, the results of a biochemical assay, etc., careful, and correct observation is important. The raw data should then be compiled into an electronic file (e.g., a series of organised spreadsheets) or hand recorded in lists and then carefully checked again to ensure the values are still correct. Only when you are sure the raw data is compiled correctly can the statistical analysis begin.

VI. Statistical analyses of the data.

This may involve initially describing and exploring the data by summarising it or presenting it in simple tables, graphs, or charts, etc. Then, statistical analysis tests may be done to compare the data and determine any statistical significance, either by a computer package or by hand.

VII. Drawing conclusions and writing up.

Statistics will help you draw conclusions from your study. Was your null hypothesis correct and if not, what else might explain the data? The process of writing up your work begins, but it is an iterative process and can involve more statistics. As you think about your data, you may have new questions to ask of it, and thus more statistics to refine the details of your thinking and data interpretation.

We need statistics to do all that!

1.2 Statistical notations and definitions

The statistical notations and expressions used in this book are commonly used in nearly all published books, though you may encounter a few different ones. If you clearly indicate what your symbols represent, others will have no problem in understanding your statistical jargon.

X_i^n or Y_i^n and so forth: '*i*' indicates a particular data point (an observation, measurement, or trait) in a sequence of '*n*' data values you obtained; where: $i = 1\ldots n$ which is from the first to the last measurement in your sample of '*n*' measurements or observations. This notation is illustrated in the example below.

Example 1-1: A study on chickens collected data on the proportions of different types of blood cells. Measurements include the percentages of the blood Lymphocyte (L) and another blood cell called Heterophils (H), in a random sample of $n = 11$ individual chickens from a flock of more than 157 chickens (Al-Murrani, et. al. 1997). Thus the % Lymphocyte data (X) and the % Heterophil data (Y) might be written as:

X_i^n : 70 69 58 60 60 50 35 49 54 70 63
Y_i^n : 26 36 36 21 31 41 52 38 35 20 27

Where: i = 1…11 and n = 11 (‘n’ represents the number of observations in this subset sample of the observations). In this example, one refers to the value of the 3rd X observation by writing, X_3; so X_3 equals 58. Similarly, X_{11}, is the value of the 11th X observation, which is 63.

Population size (*N*): This is usually written as ‘*N*’ (capital letter). It is the total number of observations or measurements of a trait for the whole population. The whole population may be for humans in one country such as the British population living in the United Kingdom, the population of humans living in Iraq. Or populations of species of animals or plants, the Merino Sheep population. Or in the case of our worked example above, the whole population is a flock of 157 chickens, so *N* = 157 (Raw data, Appendix II). So, the population in statistical sense is very 'fluid' and range from living organisms to measurements.

Sample Size (*n*): This is usually written as ‘*n*’ (small and in italics). It is the number of observations in a sample which is part of the population; in the above worked example on chickens, *n* = 11.

Σ (Sigma): Sum of all observations.

$\boldsymbol{\sum_i^n X}$**:** In our example $\Sigma_1^{11}X$ is the summation from the 1st to the 11th L% count. So:

$\Sigma_1^{11}X$ = 70 + 60 + 58 +.... + 63 = 638

If we have $\boldsymbol{\sum_i^n X}$ with i ≠ 3**:** It means, the sum of all observations except the 3rd observation. So, in our example we have 638 – 58 = 580

$\boldsymbol{\bar{X}}$ **(X bar) and** $\boldsymbol{\bar{Y}}$ **(Y bar):** This is the ‘**mean**’ or ‘**average**’ of the **sample** or ‘sample mean’ of X and Y data, respectively. It equals the sum of all observations in the sample divided by the **sample size (*n*).** It is sometimes called the ‘arithmetic mean’ of the sample. In our example, sample mean $\bar{X}$ = 58%).

μ or u: **Mean of the whole population or the Population mean:** Using all 157 observations of Lymphocyte % from the population of chickens in our example above: μ_x = 58.23 %

$\boldsymbol{x_i}$; $\boldsymbol{y_i}$**:** is the 'deviation' of any and each observation in the sample from the mean, $\bar{X}$ or $\bar{Y}$, or from the population mean, μ. The deviation can be positive or negative (i.e., a number with a '+' or a '-' sign). For example, $x_{5,}$ is the deviation of the 5th observation in the sample from the mean $\bar{X}$ of 58%:

x_5= 60 – 58 = +2

The sum of deviations $\boldsymbol{\sum_i^n x}$ must be zero. In our example on chickens:

$\sum_1^{11} x$ = (70 – 58) + (60 – 58) +…+ (63 – 58) = 0

Remember: the sum of deviations of observations from the overall mean of the population, μ, or from the mean of the sample, $\bar{X}$, must always equal zero; if not, you must re-check your arithmetic.

$\boldsymbol{\sum X^2}$**: 'sum of squares'.** Sometimes written **'*SS*'.** It is the sum of squared '*n*' observations, starting from the 1st:
In the example on chickens above:

$\sum X^2$ = SS = $70^2 + 60^2$ + ...+ 63^2 = 38,244
$\sum Y^2$ = SS = $26^2 + 36^2$ + …+ 27^2 = 12,853

$\boldsymbol{\sum_i^n x^2}$ ***or ss*****: 'Sum of Squared Deviations'.** This is the sum of all squared deviations from the overall mean of the population, $\boldsymbol{\mu}$, or the sample mean, $\bar{X}$, it is also called the '**Corrected Sum of Squares**'.

$\boldsymbol{ss = \sum_i^n x^2 = \sum_i^n (X - \mu)^2}$ (for the population) and

$\boldsymbol{ss = \sum_i^n x^2 = \sum_i^n (X - \bar{X})^2}$ (for the sample) and this can be shown as:

$\boldsymbol{ss = \sum X^2 - \frac{(\Sigma X)^2}{n}}$

$\frac{(\Sigma_{i=1}^{n} X)^2}{n}$: is the **Correction Factor**, which equals the squared sum of all observations in the sample divided by sample size (*n*).

$\boldsymbol{\sum_{i=1}^n XiYi}$**: the 'Sum of Cross Products'** of the two variables, *X* and *Y*.

$\boldsymbol{\sum_{i=1}^{n} xy}$**: the 'Corrected Sum of Cross Products'** or 'sum of cross products of deviations' of each observation from its appropriate mean, $\bar{X}$ and $\bar{Y}$. The equation is as follows:

$$\sum_{i=1}^{n} xy = \sum_{i=1}^{n} XiYi - \frac{\{\sum_{i=1}^{n} X \cdot \sum_{i=1}^{n} Y\}}{n}$$

Negative sign indicates an inverse relation between *X* and *Y* as one increases the other decreases. If the sign is positive, it indicates a positive relation where *X* and *Y* vary in the same direction.

$\mathbf{s^2}$**: 'V': Sample variance.** The sample variance.

$\boldsymbol{\sigma^2}$ **: 'Sigma square'; Population Variance**. This represents the variance of the population.

σ: **Population Standard Deviation.** This is the standard deviation of the population of interest. Sometimes it is used for the standard deviation of a sample.

s (or *SD* or *sd*): **Sample Standard Deviation**. It is closely related to the sample variance, as will be noted in the coming sections.

$\boldsymbol{SE_{\bar{X}}}$: **Standard Error of the sample mean**. This is a character of the sample; the population mean is the true mean with no error.

CV%: **Coefficient of Variation expressed as percent**: The coefficient of variation % shows the size of variability of the observations in your sample in term of standard deviation as a percentage of the mean. Where:

CV% = (SD/$\bar{X}$) x 100

$\boldsymbol{Cov_{XY}}$: **Covariance** which equals $\frac{\sum_{i=1}^{n} xy}{n-1}$ and is called the covariance of variables *X* and *Y* (or any two sets of paired variables). It represents the associated variation of the two variables for *n* paired observations.

$\boldsymbol{b_{xy}}$ $\boldsymbol{_{or}}$ $\boldsymbol{b_{yx}}$ **: The** R**egression Coefficient** of *Y* on *X* or *X* on *Y,* respectively. $\boldsymbol{b_{yx}}$ represents a quantitative measure for the dependence or association of the dependent variable *Y* on the independent variable *X*; and $\boldsymbol{b_{xy}}$ represents the dependence or association of the *X* dependent variable on *Y* independent variable, respectively. It can take any positive or negative value depending on the size and direction of association (the covariance) and have quantitative units. In the subscript of $'\boldsymbol{b}'$ the dependent variable comes first.

r: **Pearson Correlation Coefficient**. A qualitative measure of association between two variables. It is a descriptive measure and carries no units.

$\boldsymbol{r_s}$: **Spearman's Rank correlation coefficient**. A qualitative measure of association between the 'ranks' of two variables. It is a descriptive measure of relation and has no units.

$\boldsymbol{R^2}$: **Degree of Determination**. It explains how much of the variation was taken care of by your experimental data. It also reflects on the precision of the research work. Unity (1) or 100% is the maximum accuracy possible. The scientific community generally accepts an $\boldsymbol{R^2}$ of 80% and higher.

Intra-class Correlation Coefficient: It is a measure of reliability when tests are measuring quantitative parameters. It describes the degree of agreement between measurements or ratings of two groups.

***DF (df)*: Degrees of Freedom**. This is the number (*N*) or (*n*) of measurements or observations that respectively describes the population or sample completely.

P: **Probability** means how likely an event is to occur.

Event: Any particular 'observed value' of a variable or measurement or a trait which is usually denoted by *X* or *Y*, etc., Examples of events include: getting a head or a tail when tossing a coin, getting any particular number from 1 to 6 when throwing a dice, getting a systolic blood pressure measurement of 145 mm Hg, a body weight of 55 kg, the occurrence of a certain disease, and so forth.

Data: Raw information, quantitative (continuous or discrete) or qualitative categorical and proportions, gathered in research of an experimental nature or from surveys and records. When data are arranged and analysed, they can convey useful information. Examples include records of blood pressure in human subjects, birth rate, gender, incidence of diseases, gene frequencies, gene expression, rainfall, bacterial growth rate, red blood cell counts, death rate, body weight at birth or at any age, height in centimetres, body mass index (*BMI*) measurements, hospital records, houses in a specific area etc.

***$P_?$*: Percentiles.** It is the value which divides a 'ranked' data set into 100 equal parts; commonly used to construct standard growth charts for children from birth onwards. For example, an observed value with P_{75} indicates that 75% of the population of values is less than that observation or percentage.

***$Q_?$*: Quartiles.** The values which divide the 'ranked data' set into four equal parts. The 1st quartile is termed $Q1$ which corresponds to all the values up to the percentile P_{25}; the 2nd quartile $Q2$ corresponds to values between P_{25} and P_{50} and the 3rd quartile $Q3$ corresponds to values between P_{50} and P_{75}. The 4th quartile, therefore, contains the highest 25% of the values in the ranked data. The ranked data may represent the ranking of students' grades in certain subject, for example.

Standard Score or Z-score: this expresses the deviation or difference of any value from the mean in terms of standard deviations, so:

$$\mathrm{z} = \frac{(X - \mu)}{\sigma}$$

Inference statistics: using data generated from working on samples to generalize about the population from which the sample is taken. All statistical analysis of research data is 'Inference Statistics'. This will be discussed later in the chapter.

Attribute: a specific value for a variable or trait.
Accuracy: how close the measurement is to its true value.
Precision: how close the repeated measures are to each other.

1.3 Basic statistical concepts you should know

1.3.1 Probability: concept, scale, laws, and some applications

There is a huge body of literature dealing with probability theory and applications. I will limit myself to what is relevant to the structure of this book. **Probability** which is commonly symbolised by '***P***' and is defined as 'the likelihood of an event or observation occurring in several trials'. It can also be defined as the 'proportion of repetitions of an event or observation in a series of possible outcomes'. Some statisticians define probability as 'how likely is an event to occur?'. Probability distribution is a mathematical function that describes the probability of different possible values of a variable.

If we denote the number of ways or outcomes in which an 'event' or 'happening' can occur by '*f* ', and the total number of trials by '*N*', then the probability of an event *f* occurring is: