

STATISTICS IN THE INFORMATION AGE: A RE-EXAMINATION OF BASIC CONCEPTS IN STATISTICS WITH EMPHASIS ON MEASUREMENT, INFORMATION, AND VARIABLE

V.R.N. Mutabihirwa*

Summary

Statistics started as a practice of collecting and keeping facts on the various aspects of social-economic activities of society. This art of collecting and keeping facts can be traced some number of years B.C. in the ancient states of Europe, Asia, and Egypt. In our times, from this practice of compiling facts, to the tabulation and public publication of data arising from the 1662 London bill of mortality and other works, Statistics has developed into a body of knowledge as we know it in its complexity today. However, the basic concepts of the science of Statistics tend to be either overlooked or taken for granted even by the Statisticians themselves, a phenomenon related to what the author refers to as the ‘numbers paradigm’ centred on the difference between facts, figures and numbers. It is the opinion of the author that in the information age the phenomenon could lead to Statisticians being nabbed as experts who would assure that “if one puts one’s head in the freezer and one’s legs in the oven, on the average one would be O.K.’. In this paper, the author re-examines the concepts of measurement, information, and variable in Statistics and other related concepts, leading to the definition of Statistics as an information-based discipline, with an aim of doing away with the numbers paradigm in the applications of Statistics as a science.

1 INTRODUCTION

1.1 Statistics as a Science and the Numbers-Paradigm

Statistics started as a practice of collecting and keeping facts on the various aspects of social-economic activities of society, an art that can be traced some number of years B.C. in the ancient states of Europe, Asia, and Egypt. For example, an organised collection of facts on land in Egypt can be traced as far back as 1400 B.C. (Medhi (1992)). Also, in India an organised framework of the registration of births and deaths can be traced as far back as before 300 B.C. (Gupta & Kapoor (1973)). However, as time went on and the economies of those states grew, inevitably more and more aspects arose on which the facts had to be collected and kept, and the mass of facts and records collected became immerse. This lead to three things: first, the users of statistics (as facts) could not utilise them as they are, and as such the statistics had to be summarised; second, better ways of collecting, organising, summarising, and presenting the facts and records had to be developed; third, methods of analysing the facts in order to have the required organisation and summaries had to be developed. For example, the start of tabulation and public publication of data arising from the 1662 London bill of mortality (Medhi(ibid)) and the work of Adolphe Quetelet on collection and tabulation of data in the Royal Belgium Observatory and his concept of averages (Stigler

* V.R.N. Mutabihirwa, B.Sc.(Hons), M.A. (Stat) is a Lecturer in Statistics at the National University of Lesotho, Kingdom of Lesotho.

(1986)). Subsequently, these developments of concepts and methods in Statistics, together with their fusion with mathematical tools of analysis, has translated into Statistics in the form we know it today. And, in our times Statistics inevitably – both indirectly as a science and directly as a reference to facts – encompass every aspect of our day-to-day activities.

Lastrucci (1963:3) defines science as ‘an objective, logical, and systematic method of analysis of phenomena, devised to permit the accumulation of reliable knowledge’. In concurrence with Lastrucci, we can define science as a ‘continuous epistemological accumulation of knowledge through systematic and logical studies of identified characteristics of phenomena in human related activities’. We can say then from the definition that any body of knowledge will be referred to as a science, only if it is capable of systematically and logically contributing to the epistemological accumulation of knowledge. Further, we observe from the definition that the essence of the epistemological accumulation of knowledge is availability of facts on the characteristics of phenomena. Consequently, Statistics – both historically and in its application today as a discipline through which facts can be made available and analysed – comes out as a science that intersects with almost all the other sciences in our modern times.

The study of celestial bodies by Isaac Newton and that of human behaviour by John Locke entailed observing, coming out with facts, and using the facts to arrive at the conclusions through a logical framework of analysis. However, Newton, by virtue of being one of the supposed to be mathematical-logic-inclined natural scientists, analysed the facts in terms of mathematical-logic tools of analysis - numbers and formulae. On the other hand Locke, by virtue of being one of the supposed to be descriptive-logic-inclined social scientists, analysed the facts in terms of the descriptive-logical analysis. The essence of the difference in approaches in historical terms is generally attributed to the belief and conviction that unlike facts in natural science, facts in social science could not or were difficult to present in terms of the mathematical-logic tools of analysis. We shall refer to this paradigm as the ‘numbers-for-natural-science paradigm’ or simply the ‘numbers paradigm’.

In our times the numbers paradigm is manifested in two main ways as far as the science of Statistics is concerned. First, we have - especially in social sciences - ill-defined classifications of the ways and approaches in which the accumulation of knowledge is attained through research. The classifications themselves include for example the classifications of research into qualitative research and quantitative research, social science research, educational research, biological sciences research, et cetera, which raises the question of what research means by definition. Further, such classifications are subjective, as for example while Bryman (1988:11)’s classification of social science research into qualitative and quantitative research is centred on the methods and instruments used in compilation of facts, the same classification of social science research by Bernard (1995:19) and Phillips (1976:4) emphasises the methods used to analyse the facts. This manifestation, together with transcendence of Statistics into other disciplines, has lead Statistics at basic level to develop incoherently as a science, and thereby rendering its application in the accumulation of knowledge to depend subjectively on the classifications and those who create them.

The second manifestation of the numbers paradigm in our times is the lack of clarity of the basic concepts of Statistics like data, population, sample, and variable, which tend to depend on the author’s background and understanding of Statistics, with even Statisticians tending to either ignore them or leave them half-defined. For example the following: the treatment of

the term data in Weiss (1989:chapter 2); “measurements recorded in a data set are basic pieces of information nature conveys to the investigator”, say Bhattacharyya and Johnson (1977:11); Daniel and Terrel (1995:4) defines a variable as “a characteristic”; Harper (1988:chapter 10) seems not to differentiate between the measures of central tendency and averages, and so do Moser and Kalton (1971:7) when they say “a useful summary measure to describe the population in terms of a variable is an average of some form, usually the arithmetic mean”; “any set of quantifiable data can be referred to as a population if that set consists all values of interest” say Harnett and Soni(1991:2); Spiegel (1992:1) correctly considers a population as ‘a collection of units’ but defines a variable as “a symbol”; Grant and Leavenworth(1996:7) brings in “variables and attributes data” as “variables data are measured on a continuous scale with individual measurements rounded to some desired number of decimal points. Attributes data involve counts of articles or counts of events”.

The two manifestations of the numbers paradigm would in the long run render Statistics fail to stand as a science. This is because for it to be regarded as a science, Statistics ought to be build up through in the first instance well established primitive terms, which would logically lead to definitions, which would in turn build into axioms and theorems, which would ultimately build into the various techniques in the science. And, the primitive terms, definitions, axioms, theorems and techniques should not depend on individuals’ perception of Statistics but consistently hold universally at all levels of knowledge.

The numbers-paradigm manifestations themselves arise from authors and those who apply Statistics to either take for granted or not consider the difference between facts, figures, and numbers; a consideration which ought to be emphasised in the development of Statistics as a science. That is to say, for example using the same number 20 we can constitute different figures like ‘R20’, ‘US\$20’, ‘shillingi 20’, while we can use different numbers 1.5 and 1500000 to constitute the same figure as ‘EUR1.5 million’ and ‘EUR1500000’. As such numbers and figures cannot be the same. On the other hand, there is definitely a difference between facts and the figures/descriptions used to express the facts. For example there is a difference between the figure “US\$1500000” and the fact “the company made a profit of US\$1500000 in 2003”, and the description “male” is different from the fact “the respondent was a male”. Apparently, an emphasis on the differences between facts, figures, and numbers would lead to a shift in paradigm. And, as we move deeper into the twentieth century in the information age, for Statistics to be sustained as a science, the need for the shift from the numbers paradigm in the applications of Statistics becomes more and more inevitable. The change in paradigm itself can be achieved by a re-examination of the concepts of measurement, information, and variable, which would form a basis for definitions of other concepts and uphold Statistics as an information-based science.

1.2 Considerations in Development and Applying Statistics as a Science

The essence of the historical development of Statistics is the variation of the facts on characteristics from one entity of reference to another. That is for example if the consumption patterns of all households was the same, so that we have the same facts on the characteristics of consumption patterns, then we do not need Statistics techniques. As such, the development and application of Statistics, three considerations need to be taken into account. The first of such consideration is that the intended analysis of the facts involves a combination of two or more facts in question. For example, in finding out how tall a certain Mr. Sikwane is we don't require Statistics, since the intended purpose involves only one fact. On the other hand if we want to find out the typical height of the five sons of Mr. Sikwane, we may need Statistics

because we may need to combine the five facts on the heights of the sons.

The second consideration is that the intended purpose of analysis should involve facts of the same kind in the sense that they are on the same characteristic. For example, in finding how much profit was realised at a certain company in the year 2003, whereby we would combine facts on cost and expenses together with the facts on revenue to arrive at the fact on profit for the year, Statistics techniques are not the best to use, if ever we can use them. On the other hand, if we had to find out what the general level of the monthly profit was in a year, we may need Statistics techniques, since we would be dealing with twelve facts of the same kind on profit.

The third consideration is that the combination of the facts in question should lead to a single result, representative of the individual facts, in such a way that the result can be used in place of the facts to realise the intended purpose in the analysis. In other words, the facts would be considered collectively at the same time, and the individual significance of the facts becomes relatively not significant as far as the realisation of intended purpose is concerned. For example, in establishing the general level of productivity of an employee in the public service, we would combine the facts of the same kind on productivity of employees in the public service to arrive at a single result - the general level of productivity, whereby the application of Statistics becomes inevitable. We note that the productivity levels of the individual employees would not be significant as far as the realisation of the intended purpose of combining the facts is concerned.

2 INFORMATION AND DATA

We start the re-examination of the basic concepts as mentioned in subsection 1.1 with two basic considerations, that is to say: first, numbers, figures, and facts are distinctively different; and second, phenomena arising from activities of human interaction with nature are identified through characteristics, and the knowledge on the phenomena is accumulated through availability of facts on the characteristics. Then, the terms '**phenomenon**', '**characteristic**', and '**fact**', whose detail discussion is not entered into here, are introduced as 'primitive terms'. Further, we observe that facts on the characteristics of phenomena exist independent of our presence or intended purposes in the epistemological accumulation of knowledge. That is to say, for example consider a researcher who identifies the phenomenon 'HIV/AIDS Awareness in Primary Schools in Botswana' as worth studying for accumulation of knowledge on HIV/AIDS. The phenomenon itself is identified through various characteristics, and irrespective of the presence or absence and intentions of the researcher, facts on the various characteristics of 'HIV/AIDS Awareness in Primary Schools in Botswana' arise as the activities of human interaction with nature in the primary schools in Botswana take place. Also, there would be numerous facts on the characteristics of phenomena arising from the day-to-day activities in a supermarket, which are independent of the decision making process in the supermarket. And, an experiment on characteristics of the phenomenon 'reproduction behaviour of HIV virus', would lead to various facts on the characteristics, and these facts are independent of the experimenter although in this case the experimenter may control the activities involved.

However, the researcher would have to identify, capture and record facts on those characteristics of the phenomenon he/she considers as worth researching on in relation to the specified purpose of the research; as for example facts on the three characteristics 'the

geographical position of the primary school in Botswana', 'the level of awareness among the pupils', and 'the level of awareness among teachers'. Also, it is important for the supermarket to identify, record, and keep those facts that are required for specified purposes in the day-to-day decision making process in the supermarket. And, the experimenter would have to identify, record, and keep those facts that are required for the specified purpose of the experiment. In general, the activities of human interaction with nature lead to numerous facts on the characteristics of phenomena arising from the activities. In any environment of such interaction there would be those facts that are required for a specified purpose within the environment. We refer to such facts as original facts. In other words, we define **original facts** as **the facts on specified characteristics of phenomena arising from the activities of human interaction with nature in an environment, which are required for a specified purpose in the environment**

The ultimate aim of the researcher or experimenter mentioned above is the dissemination of required knowledge on the phenomenon being studied at the right time, while the ultimate aim in the supermarket case is to make the right decisions at the right time. We refer to this as '**effective and time-efficient realisation of the specified purpose**', and we observe that this is achieved through analysis of facts. The original facts recorded and kept in an environment would normally not be in a form that when analysed would lead to effective and time-efficient realisation of the specified purpose in the environment. In other words, we would normally need to put the original facts in the form useful for analysis in relation to the mentioned realisation, and this useful form of original facts is what we refer to as information. That is to say, we define **information** as **the form of original facts in an environment, which when analysed lead to effective and time-efficient realisation of the specified purpose**. We note that the generation of information items from original facts is a logical transformation. That is for example, the name of a person in an original fact item for payroll would be the same name in an information item on the payslip for the person.

We refer to either original facts or information or both as data. That is to say, **data** is defined as **the facts on specified characteristics of phenomena arising from the activities of human interaction with nature in an environment, which are required for a specified purpose, and intended to be used for the effective and time-efficient realisation of the specified purpose in the environment**.

We observe from the definition that data items are purposeful compiled facts, and neither are they figures nor description nor numbers. And, unlike as portrayed in the vast literature on the theory of measurement, the individual or individuals compiling data do not create the data but as we shall see in the next section compile the facts in a specified form.

3 MEASUREMENT

3.1 Introduction

There are as many definitions of the term measurement as there are authors on the concept. For examples the following: Francis (1967:187) defines the concept as "the assignment of numbers according to some rule"; Bohrnstedt (1983:70) defines measurement as "the assignment of numbers to observed phenomenon according to certain rules", and goes further to say that "one might assign the number 1 to all males and 0 to all females"; according to Bernard (1995:24) "variables are measured by their indicators, and indicators are defined by

their values” and “measurement is deciding which value to record”, presumably on the variables, while the variable is defined as “something that can take more than one value, and values can be words or numbers”; Philips (1976:137) comes out with a better definition of measurement as “a **process** by which the individual obtains information”. However, almost all such definitions and explanations of measurement lack consistence in terms of what is measured, the basis of the process of measurement, and the results from the process. The inconsistency can be traced to two main considerations, namely: first, the difference between the data items (original facts/information) and the figures/descriptions through which the data items can be expressed; and second, for data items expressed through figures, the difference between the figures and the numbers with which the figures are created. This perplexity on measurement, as noted earlier in section 1 of the paper, originates basically from the numbers paradigm.

The process of generating data items for epistemological accumulation of knowledge involves four main things: first, the identification of the phenomenon on which some given knowledge is to be accumulated; second, identification of the characteristics for which data items are to be produced; third, identification of the form in which the data items are to be expressed; and fourth, making the data items available. For example, let us consider the researcher on the phenomenon ‘HIV/AIDS Awareness in Primary Schools in Botswana’ mentioned in section 2, who then depending on the specified purpose identified the mentioned three characteristics. Then, third, the researcher would decide on the form in which the data items on the respective characteristics would be made available, for example for the data items on ‘the level of awareness among teachers’ in terms of the awareness being very adequate, adequate, and not adequate. Fourth, the researcher would generate the data items on the three characteristics in the required form. This **process** of generating data items on identified characteristics is what we refer to as measurement, and we say we are measuring for example the three characteristics of phenomenon ‘HIV/AIDS awareness in Primary Schools in Botswana’.

However, as noted in subsection 1.2, the measurement in Statistics will apparently involve generation of two or more data items on the same characteristic. Secondly, from third consideration, the generation of the data items is such that from one entity of reference to another we have one and only one data item on the characteristic being measured. For example in measuring the characteristics ‘the level of awareness among teachers’ mentioned above, we will have one and only one data item on the characteristic from one ‘primary school’ to another. Then, we define **measurement as a process of generating data items, basing on an entity of reference and in a specified form, for an identified characteristic of a given phenomenon in such that on each of such entity of reference one and only one data item is generated on the characteristic.**

We refer to the entity of reference in measurement as **the object of measurement**. As from the definition of data, we note that the definition of measurement incorporates the specified purpose for generation of the data, and for any given characteristic being measured data items would be generated on the objects of measurement pertaining to the specified purpose. These objects of measurement then would collectively constitute what is called a **population**. Then, any part of the population is referred to as the **sample**. This remains in line with the way the two concepts population and sample are currently generally defined. For example, considering the research on the phenomenon ‘HIV/AIDS Awareness in Primary Schools in Botswana’ mentioned in section 2 above we would have ‘a primary school’ as the object of measurement, and for any one primary school the researcher will then have one and only one

data item generated on each of the three characteristics in the process of measurement. Further, ‘the collection of all primary schools in Botswana’ will constitute the population. It is important to emphasise three things on the population and sample concepts. First, both population and sample are collection of objects of measurement, and neither of them is constituted by values. Secondly, both concepts are first and foremost to do with the process of measurement, and as such are to do with data items (facts) rather than the numbers used in the analysis as portrayed in most literature. Thirdly, unlike as portrayed in most literature again, the characteristics being measured identify the phenomenon and not the population.

We observe that we can carry out measurement under uncertainty. However, the results of measurement in this case will first be looked at in terms of possible data items and the measurement process itself is on characteristics of phenomena involving statistical random experiments. We define a **statistical random experiment** as an **operation with the following properties: it results in known distinct outcomes; one and only one outcome can result when the operation is carried out; and the outcome to result is not known a priori the operation is carried out**. We refer to the collection of the outcomes of a statistical experiment as a **sample space**. That is to say, for example consider a statistical random experiment of ‘tossing a dice and observing which side turns upright’ and a game whereby ‘a player has to toss a dice 5 times, the side which turns upright on each of the tosses is noted, and a player has a success if the side numbered either 1 or 3 or 5 turns upright when the statistical random experiment is carried out. Considering ‘a player’ as the object of measurement, we can measure the characteristic ‘successes in the five tosses of the dice’, with the possible data items expressed in terms of ‘no success’, ‘1 success’,, ‘5 successes’. And, we have the **possible data items** given as {‘the player gets no success in the 5 tosses’, ‘the player gets one success in the 5 tosses’, ‘the player gets 2 successes out of the 5 tosses’,, ‘the player gets 5 successes in the 5 tosses’}. However, in case of more than one player being involved in the game, and assuming each player plays the game once, we note that we would have one and one actual data item on each of the players.

3.2 Scales of Measurement

In order to satisfy the definition of measurement given in subsection 3.1 above, the specified form in which a data item is to be generated has to have three properties. First, the data items are expressed in such a way that objects of measurement can be distinctly related in regard to the characteristic being measured and in such a way that the data items can be put into well-specified categories. We shall refer to the categories as the **scale panes** for the data items. The second consideration is that the scale panes are mutually exclusive, in the sense that every data item can be expressed through one and only one of the scale panes. Third, the scale panes are exhaustive in the sense that they cover all possible categories through which the data items can be expressed in relation to the specified purpose in the measurement. We refer to such a **specified form in which a data item is to be expressed in the process of measurement** as a **scale of measurement**.

The relationships for any scale of measurement, which can be established between the objects of measurement, depend on the chosen scale panes. Then, we have two groups in which the scales of measurement can be classified. The first group are the scales of measurements that lead to data items being expressed in terms of descriptions of qualities or attributes. The mutually exclusive and exhaustive scale panes in this group are countable and normally relatively few. We refer to such scales of measurement as **qualitative scales of measurement**, and we refer to the data items generated under this group as **qualitative data**

items.

The second group of scales of measurement are those that lead to data items being expressed in terms of figures. The mutually exclusive and exhaustive scale panes in this case are themselves figures. We note that the scale panes then are relatively many in number, and in most cases either infinitely countable or on a continuous range of figures. We refer to such scales of measurement as **quantitative scales of measurement**, and we refer to the data items generated under this group as **quantitative data items**.

We can now adopt the classification first introduced by Stevens(1951) to further classify the scales of measurement. That is to say, for a qualitative scale of measurement the scale panes can be chosen in such a way that they have the same value of importance in relation to the characteristic being measured and the specified purpose in the measurement. Consequently, the only relationship which can be formed between the objects of measurement is to group them according to the given scale panes for the data items. For example, in the measurement of ‘the geographical position of the primary school’ in the research on HIV/AIDS awareness in subsection 3.1 above, let us say the researcher chooses the scale panes as ‘east of Botswana’, ‘west of Botswana’, ‘north of Botswana’, and ‘south of Botswana’. Let us assume the scale panes have the same value of importance in relation to the ‘geographical position of the primary school’ in the research. Then, the only relationship we can have is to group the objects of measurement (primary schools) according to the scale panes. We cannot for example form any order relationship between the objects of measurement based on the scale panes, like ranking them in a certain order. As Stevens, we refer to such a scale of measurement as **nominal scale of measurement**.

On the other hand, under a qualitative scale of measurement, the scale panes could be chosen such that they have different value of importance in relation to the characteristic being measured. As such, apart from the classification, an order relationship between the objects of measurement can be established basing on value of importance of the scale panes. For example, in the measurement of ‘level of awareness among teachers’ in the research on HIV/AIDS awareness in subsection 3.1 above, we have the scale panes as ‘very adequate’, ‘adequate’, and ‘not adequate’. Then, apart from grouping of the objects of measurement in relation to the scale panes, we can order them for example according to the ‘level of awareness among teachers’ being as ‘very adequate’, or ‘adequate’ or ‘not adequate’. We refer to such a scale of measurement as **ordinal scale of measurement**.

In case of quantitative scales of measurement, we observe two things, namely: first, as noted before, the figures in the data items constitute the scale panes; and second, since the figures are constructed through numbers, the relationships between the objects of measurement are in terms of the magnitudes reflected by the figures in the respective data items. As such, in addition to the relationships under ordinal scale of measurement, relationships in terms of differences in the magnitudes and comparisons between the differences can be established between the objects of measurement.

Consequently, there are two levels of relationships under quantitative scales of measurement. First, is whereby the establishment of the scale panes is based on an arbitrary starting point, for example in the measurement of temperature the starting point is the ‘freezing point of water’. In such cases, in addition to the relationships at ordinal scale of measurement, relationships between the objects can be established in terms of the differences between the magnitudes. However, we observe that under such scale of measurement we have scale panes

below and above the starting point, and as such the differences between the magnitudes themselves cannot be compared. That is to say, for example in the analysis of data items on temperature we can establish the difference between the temperatures of two objects of measurement, but we cannot establish whether or not one object is twice as warm as the other. We refer such a scale of measurement as the **interval scale of measurement**.

The second and higher level of relationships under quantitative scales of measurement is whereby the starting or reference point for the scales panes is fixed and well defined. For example in the measurement of age of individuals the starting point is 'the moment one is born', which would in this case lead to scales panes (figures) on a continuous range. Also, in the measurement of 'level of awareness among pupils' in the research on HIV/AIDS awareness in subsection 3.1 above, let us say the researcher chooses to express data items in terms of 'number of pupils aware of HIV/AIDS'. Then, the starting point is 'no pupil', and we have countable scales panes (figures) in the data items as 'no pupil', '1 pupil', '2 pupils' et cetera. We note that in such a scale of measurement, because of the well-defined starting point, we cannot have scale panes below the starting point. As such, the scales panes are such that the differences in magnitudes themselves can in this case be compared, normally through ratios and proportions. As Stevens named it, we refer to such scale of measurement as the **ratio scale of measurement**.

3.3 A Note on the Current Literature on the Theory of Measurement

Apparently, there is a vast literature on the Theory of Measurement, ranging from theory in basic texts on Research Methods and Statistics, to for example that by Anderson et alas (1983), to the measurement theory in education, to that in Narens (1985)'s Abstract Measurement Theory. There are three main observations to be made on the measurement theory in the literature, which can be dated as far back as to the Greek's theory of magnitudes by Eudoxus and Thaeatetus. The first observation is that most of the literature - if not all - generally portray measurement as 'an assignment of values/numbers), which in turn lead to two issues. That is to say: first, the definition raises the questions 'assignment to what?', 'based on what?', 'to produce what?'; and second, is the overall purpose of the process, which is supposed to be the epistemological accumulation of knowledge. As pointed out earlier the accumulation of knowledge is based on facts, and as such the immediate product of measurement as a process should not be numbers – and not descriptions and figures but facts. This anomaly is streamlined through the re-examination of the concept of the variable as in section 4 below.

The second observation is that most literature with all the embodied philosophical and mathematical complexities are to do with mainly the following: the procedures to be used in constructing scale panes – specifically for qualitative scales of measurement; the instruments used in the measurement, and the implication on the data items produced, of such construction and use of the instruments in terms measurement error, reliability, validity, and precision. However, these concepts ought not to be directly tied to the measurement, which by definition remains the same as the process of generating facts as discussed above, but they should be tied to the way one chooses to carry out measurement. As such, as much as most of the concepts on measurement advanced in the current theory remain useful, they need to be revisited with emphasis that what is produced from measurement are data items and not figures/descriptions/numbers.

The third observation is related to the fact that the techniques of analysis in Statistics are

basically in terms of the mathematical-logic tools of analysis - numbers and formulae, while measurement leads to facts. As such, the central issue in the current theory of measurement is how the facts could be analysed through the techniques like those in Statistics, which borrow from the already established number theory. And, depending on the background of the author, the connection between the concepts in measurement and Statistics techniques used in the analysis tend to be either not well presented or created to suite the understanding of the author. As mentioned earlier, this is a result of the manifestation of the numbers paradigm, and it can be taken care of through the concept of the variable as introduced below.

4 VARIABLES

4.1 Introduction

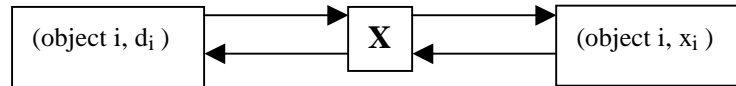
The variable has as many definitions as authors who have written anything about the concept. For example, the following: “a **symbol** that **can be replaced** by one of the elements of some specified set” from Hays (1963:36); “a **symbol**, such as X,Y, H, x, or B, that **can assume** any of a prescribed set of values” from Spiegel (ibid); “any **factor** in educational research which influences the observation or management of an educational phenomenon” from Page et alas(1977:357); “a **characteristic distinguished** by measurements” from Gupta and Kapoor(1973:8); “**characteristic that assumes** different values for different entities” by Daniel and Terrell(ibid); “a **characteristic having intrinsic value**, a zero-point, and equal intervals” by Zeldith (1966:8); “**characteristic** that is being **measured**” from Harper(1988:5). These inconsistent multi definitions of the same concept from authors in Statistics, which have roots in the perplexity of measurement pointed out in subsection 3.1, not only in essence leave the concept of the variable undefined, but also necessitate a re-examination of the concept.

The re-examination of the concept of the variable needs three things to be taken into account. First, as emphasised earlier, data items are facts and not just figures or descriptions, and for data items expressed through figures, we ought to differentiate between the figures and the numbers through which the figures are constituted. Second, Statistics techniques bear on the combination of the data items as discussed in subsection 1.2, and we note that the single result referred to is itself a data item and not a figure, a description or a number. That is for example, the single result would be the data item “the monthly average profit in 2003 was US\$1.5 million”, and not the figure “US\$1.5 million” or the number 1500000. Third, on the other hand, Statistics techniques themselves involve essentially mathematical manipulation of numbers, and we need not emphasise the fact that we basically cannot carry out mathematical operations on figures and descriptions through which the data items are expressed.

Apparently, the discussion above steers the necessity of having a relationship between data items and numbers, and this relationship is what we refer to as a variable. And, we define a **variable as a relation between data items and numbers which is such that for every object of measurement there exists one and only one number representing the data item on the object of measurement.**

Given data items obtained on a population or sample and a variable defined for the data items, we shall refer to the number related to a data item through the variable as an **observation** of the variable. We shall refer to all the observations of a variable collectively as **statistical data**, and when referring to statistical data for a variable we shall refer to the data

items (facts) represented by the observations as **factual data**. And, let d_i be the factual data item on the i^{th} object in the population/sample, let X be the defined variable, and let x_i be the respective observation of the variable representing the factual data item. Then, the relationship between the factual data item and the observation of X can be presented as below:



We observe that there are three considerations in constructing a variable. First, the object of measurement has to be reflected, and it is then referred to as the **object** of the variable. Second, the characteristic for which the data items were obtained is to be reflected, and we shall refer to the indication of this as the **characteristic indicator** of the variable. The characteristic indicator is erroneously referred to as the variable, as in some of the literature sighted above. The third consideration is the indication of how different numbers are representing the different scale panes for the data items. We shall refer to this part of a variable as the **relation indicator**. The relation indicator of a variable for qualitative data items are to be explicitly constructed – as noted in subsection 4.2 below, while in case of variables for quantitative data items the relation indicator is directly related and depends on the figures in the data items.

4.2 A Note on the Types of Variables

The main classification of variables is based on the data items for which the respective variables are defined. That is to say, we have **qualitative variables** as **variables defined for qualitative data items** and **quantitative variables** as **variables defined for quantitative data items**. Further classifications include the following: discrete and continuous variables, based on the possible observations of the variables; random variables, which are variables defined for data items based on the sample space of a statistical random experiment; derived variables, which are variables defined as transformations based on observations of other variables.

As mentioned earlier, the scale panes of qualitative data items are count ably finite and the relation indicators of qualitative variables are to be explicitly constructed. As such the numbers constituting possible observations in the relation indicators for such variables are to be explicitly chosen, while taking into consideration the following: first, the numbers chosen have to be in the logical or natural sequence of the scale panes for data items; second, to avoid problems in the interpretation of results, the numbers have to be chosen in a systematic manner, with successive numbers differing by 1; third, to avoid unnecessary computational problems, the numbers should be nonnegative integers.

For example, consider the research on HIV/AIDS awareness referred to in subsection 3.1 above. We note that the object of the variables that would be defined for the data items on all the three characteristics is ‘a primary school’. Then, for the data items on characteristic “geographical position of the primary school” we could have a qualitative variable as below:

“the relationship which associates the geographical position of a primary school with a number as follows: 1 with ‘east of Botswana’, 2 with ‘west of Botswana’, 3 with ‘north of Botswana’, and 4 with ‘south of Botswana’”

We note that since the scale panes have the same value of relative importance the logical order of the numbers in the relation indicator does not matter. On the other hand, the logical

order of the numbers will have significance in the relation indicator of the qualitative variable below, for the data items on the characteristic “the HIV/AIDS awareness level of teachers”:

“the relationship which associates the level of HIV/AIDS awareness of teachers in a primary school with a number as follows: 1 with ‘very adequate’, 2 with ‘adequate’, 3 with ‘inadequate’”

This is because the data items would be at ordinal scale of measurement.

The construction of quantitative variables is less involving than that of qualitative variables. This is because as mentioned above the relation indicators of such variables are directly related and depends on the scale panes (figures) of the data items. For example, for the data items on the characteristic ‘level of awareness among pupils’ in the same research on HIV/AIDS awareness we could have a quantitative variable for the data items as **“the number of pupils in a primary school aware of HIV/AIDS”**. We note that because the scale panes are built through nonnegative integers we have the relation indicator of the variable as “number of”. Other examples of quantitative variables are as follows: **“the production cost in a month in thousands of US\$”**, whose object is ‘a month-period’, characteristic indicator is ‘production cost’, and relation indicator is ‘in thousands of US\$’; and **“population of a country in millions of people”**, whose object is ‘a country’, characteristic indicator is ‘population’, and relation indicator is ‘in millions of people’.

Consider the possible data items on the characteristic ‘successes in the five tosses of the dice’ mentioned in subsection 3.1 above. We could define a random variable for the possible data items as **“the number of successes a player gets out of the five tosses”**, in which we have the relation indicator as ‘number of’, characteristic indicator as ‘successes out of the five tosses’ and the object of the variable as ‘a player’. We observe that in this case we would have possible observations as {0,1,2,3,4,and 5} and not actual data items until the game is played.

5 CONCLUSIONS

The re-examination and defining of the concept of measurement presented in the paper brings to bear an historical eminence that Statistics as a science is based on data items (facts compiled for a specified purpose), and not mere figures, descriptions or numbers. And, we can then define **Statistics as science involving methods and techniques for compilation, presentation, and analysis of data items through which we can arrive at a realisation of the specified purpose of measurement involving combinations of two or more data items of the same kind to arrive at respective single results, in such a way that the data items are considered collectively at the same time, and their individual significance becomes relatively not important as far as the realisation of the specified purpose is concerned.**

The process of measurement itself as defined in the paper is not abstractly independent of the individual carrying out measurement, for the individual determines the specified purpose in the measurement and identifies the form in which the facts are to be available – the scale of measurement. This is where the current theory of measurement and the theory of constructs of sociologists could come in, to assist the individual to have the right procedure and instruments to carry out correctly and consistently the process of measurement as defined, and specifically in the construction of scale panes for the chosen scale of measurement. However the theories, which tend to mix up data items (facts) with the numbers used either to express or in the analysis of the data items, need to be re-visited to streamline the essential

difference between measurement as a process on one hand and the way it is carried out on the other.

We further observe from the paper that although the starting point in the application of Statistics is the availability of appropriate data items generated through measurement, Statistics techniques themselves involve mathematical-logic tools of analysis – numbers and formulae. As such the concept of the variable is created as a relation that appropriately bridges the data items on one hand and numbers on the other, and the specific techniques to be used – which are applied on the statistical data (the numbers) – would be selected to suite the scale of measurement for data items. For example, we can now correctly classify Quantitative and Qualitative Analysis Techniques in terms of techniques applied to data items at quantitative and qualitative scales of measurement respectively. We note that both categories of techniques will involve numbers and formulae.

The consistent distinction between factual data items (facts) and statistical data (numbers), with the two connected through the variable as a relation comes to streamline the entire spectrum of Statistics as a science. That is to say, for example the following:

- (a) Statistical Data Compilation, would involve two stages namely: generation of data items through measurement together with the associated procedures like sample survey frameworks and instruments like questionnaires; and production of statistical data through appropriate variables.
- (b) Presentation of Statistical Data like Tables and Pictorial Presentations will no longer be taken just as tabulations and pictures, but communicative technical constructions involving statistical data with titles and other parts to reflect the objects, characteristic indicator(s), and relation indicator(s) of the variable(s) involved.
- (c) Presentations of Statistical Data like Frequency Distributions will no longer be just lists of figures and how often they occur, but communicative technical constructions involving statistical data as well. Further, the concepts in the constructions would be well defined, as for example frequency as “as the number of objects on which an observation of a variable was obtained”, class not as range of figures but “range of numbers in which observations of a variable fall” and class frequency as “the number objects on which observations falling within a given class were obtained”.
- (d) The day-to-day concepts of average and average fluctuation from averages – which involve factual data (facts), will no longer be mixed up with the measures of central tendency and dispersion – which are results of mathematical manipulation of the statistical data (numbers). And, we would be able to talk of averages in case of qualitative data items without offending the mathematicians.
- (e) Probability Distributions would be able to be studied independent of their applications, even at basic level. That is to say, for example the normal probability distribution is just a statistical structure independent of the characteristic(s) being measured. Then, from the structure of the observations of the variable in question, we can purposefully associate the normal probability distribution with the variable in the analysis.
- (f) The more advanced techniques like categorical data analysis would no longer be tied to the factual data (facts) and their structures, and they can be developed in terms of only number theory properties. This is because the techniques would involve only tuples of numbers and their associated frequencies, with the tuples of numbers themselves related to the factual data through the respective variables, and not to do directly with the factual data items.

In general, we can say that the redefining of the basic concepts of measurement, data (information), and variable as we have done in this paper, and then re-building the other concepts and techniques based on the new concepts, will definitely strengthen Statistics as a Science as we move deeper into the information age. And, more important Statisticians will inevitably be a necessity and cannot be replaced by computers and associated software.

6 REFERENCES

1. Anderson A.B., Basilevsky A. and Hum D.P.J., *Measurement: Theory and Techniques*, in *Handbook of Survey Research* edited by Peter H. Rossi et al., Academic Press, Inc., 1983
2. Bernard H.R., *Research Methods in Anthropology*, Alta Mira Press, 1995
3. Bhattacharyya G.K. and Johnson R.A., *Statistical Concepts and Methods*, John Wiley and Sons, 1977
4. Bohrnsted G.W., *Measurement*, in *Handbook of Survey Research* edited by Peter H. Rossi et al., Academic Press, Inc., 1983
5. Bryman A., *Quantity and Quality Social Research*, Unwin Hyman Ltd., 1988
6. Daniel W.D. and Terrell J.C., *Business Statistics for Management and Economics*, Houghton Mifflin Company, 1995
7. Francis R.G., *Scaling Techniques*, in *An Introduction to Social Science Research* edited by John T. Doby, Meredith Publishing Company, 1967
8. Grant E.L. and Leavenworth R.S., *Statistical Quality Control*, McGraw-Hill Companies Inc., 1996
9. Gupta S.C. and Kapoor V.K., *Fundamental of Mathematical Statistics*, Sultan Chand & Sons Publishers, 1973
10. Harnett D.L. and Soni A.K., *Statistical Methods for Business and Economics*, Addison-Wesley Publishing Company, 1991
11. Harper W.M., *Statistics*, Longman Group UK Ltd, 1988
12. Hays W.L., *Statistics*, Rinehart and Winston Inc., 1963
13. Lastrucci C.L., *The Scientific Approach: Basic Principles of the Scientific Method*, Schenkman Publishing Company Inc., 1963
14. Medhi J., *Statistical Methods: An Introductory Text*, John Wiley and Sons, 1992
15. Moser, C.A. and Kalton G., *Survey Methods in Social Investigation*, Heinemann Educational Books, 1971
16. Narens L., *Abstract Measurement Theory*, The MIT Press, 1985
17. Page G.T., Thomas J.B.' and Marshall A.R., *International Dictionary of Education*
18. Phillips, B.S., *Social Research: Strategy and Tactics*, Macmillan Publishing Co., 1976
19. Spiegel M.R., *Schaum's Outline of Theory and Problems of Statistics*, McGraw-Hill Book Company, 1992
20. Stevens S.S., *Mathematics, Measurement, and Psychophysics* in *Handbook of Experimental Psychology* edited by S.S. Stevens, Wiley Publishing Company, 1951
21. Stigler S.M., *The History of Statistics*, Harvard University Press, 1986
22. Zeldith M. (jr.), *Sociological Statistics*, Rinehart and Winston Inc., 1966
23. Weiss N.A., *Elementary Statistics*, Addison-Wesley Publishing Company, 1989