Sanni M. Ali, Sileshi Lulseged, Girmay Medhin. *Ethiop Med J, 2018,Vol. 56, No. 3*

TEACHING ARTICLE

EMJ SERIES ON STATISTICS AND METHODS: VARIABLES, POPULATIONS AND SAMPLES

Sanni M. Ali, DVM, MSc, PhD[1,2], Sileshi Lulseged, MD, MMed,[3*] Girmay Medhin, MSc, PhD.[4]

INTRODUCTION

This the first article in the EMJ *Series on Statistics and Methods* describes "variables" which represent different demographic and clinical characteristics of several individuals in a data set. Various attributes of quantitative and qualitative variables, including type, scale of measurement and values they may take are presented. Some descriptive statistics and graphic summaries are introduced. Pertinent examples drawn from journal articles are provided. The article also defines "population" as used in statistics and epidemiology and highlights population in size and scope in the context of the research question to be answered. It emphasizes the need for explicitly defining sampling based on the research question in a particular study. In addition, the articles introduces concepts and notations related to parameter and statistic. It highlights the need for using appropriate sampling method in selecting a representative sample of the study population and to be able to make valid inferences and generalizations. It is indicated in this articles that specific issues requiring further details will be addressed in the *Series* articles in subsequent Issues of EMJ.

VARIABLES

Epidemiological studies involve collection and analysis of new or existing datasets, which consist of different demographic and clinical characteristics of several individuals. Examples are age, sex, presence of a disease, blood glucose level and blood pressure measurements of individuals. These characteristics are known as "Variables" and they represent quantities or attributes that may vary from individual to individual (1,2). Here we distinguish between the variable (e.g. SEX) which is often denoted by a capital letter from the values of the variable (e.g., "male" and "female" or 0 and 1) usually denoted by a lowercase letter. On the other hand, "SEX" will be considered as a fixed attribute if the population under investigation is only males or only females because the values will not vary from individual to individual within that study.

If we consider "sex" as an example of a variable, its values can vary between male and female depending on the study participant, which we can label numerically as "male" = 0 and "female" = 1. Such variables that take on values that are intrinsically non-numerical or categorical are called qualitative variables. These variables describe the phenomena of our interest qualitatively. The values of qualitative variables can be counted, categorized in to different groups or levels, and summarized using proportions or percentages. The summary of these variables provide answers to important clinical or public health questions such as the proportion of females among hospital admissions, or incidence of HIV-AIDS or diabetes mellitus in a certain place and over a specific period of time.

The scale of measurement of a qualitative variable is termed *nominal* if the number assigned to it is purely arbitrary and chosen by the researcher without regard to any order of ranking. "Marital status" with possible values of 1= "Never married", 2= "Currently Married" and 3 = "Other" is a good example of a variable measured on a *nominal* scale. On the other hand, presence and severity of a disease may be assigned 0 if "absent", 1 if "mild", 2 if "moderate" or 3 if "severe". Such assignment has some ranking associated with it although there is no consistent level of magnitude of difference between ranks; the scale is called *ordinal* (Table 1). Bar charts and pie charts can be used to graphically present the summary measures of these variables. For example, EMJ articles on intimate partners testing (3) and pediatric limb amputation (4) used a bar chart to summarize results on personal habits during pregnancy and an example on use of pie charts can be found in the EMJ (5) summarizing results on use of imaging modalities.

[1,2]Faculty of Epidemiology and Population Health, Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK and [2]Center for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK.
[3]Department of Pediatric and Child Health, Faculty of Medicine, School of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia [4]Aklilu Lemma Institute of Pathobiology, Addis Ababa University Addis Ababa, Ethiopia
*Corresponding author. sileshilulseged@gmail.com

Table 1. Summary of Different Scales of Measurement

Scale	Characteristics
Nominal	Numbers selected are purely arbitrary. Data cannot be ranked in a specific order. Examples: Sex as male = 0 or female =1, Diabetes as no = 0 yes = 1, Illness type as Renal failure =1, heart failure = 2, gastric ulcer =3
Ordinal	Numbers assigned could be arbitrary but indicates direction. Data can be ranked in a specific order - from low to high or from high to low. There is no consistent level of magnitude of difference between ranks or no precise differences between two measurements (e.g. difference between severe and moderate is not the same as the difference between moderate and mild). Calculation of statistic on the values such as averages and standard deviations is not appropriate. Examples: stages of cancer (stage I, II, III, IV), pain level (measured on 1-10 scale), satisfaction level (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied), social status (upper, middle, lower), BMI (body mass index)-based nutritional status (sever thin, thin, normal, overweight, and obese).
Interval	Has predetermined order on the scale and clearly defined measurement units. Difference between two measurements has meaning, but their ratio does not, i.e. there is consistent level of magnitude of difference between observed units. There is no absolute zero: the zero is arbitrary and does not indicate total absence of the characteristics measured (e.g. 0 degree Celsius does not correspond to total absence of heat). Examples: Body temperature in degree Fahrenheit or Celsius,
Ratio	An interval with true ratio and absolute zero. True ratio: A heartbeat of 80-beats per minute is truly “twice” as fast as 40 beats per second but a temperature of 30degree Celsius does not mean twice as much heat as 15 degrees Celsius (but it does in degree kelvin) Absolute zero: the zero point indicates the absence of the quantity being measured. Examples: Temperature in degree Kelvin, weight, height, pulse rate, respiratory rate, blood pressure, blood glucose level, cholesterol level.

Other group of variables called *quantitative* variables. This group of variables take on numerical values that are either integers (whole numbers) or real numbers (integers, fractions or decimals). *Quantitative* variables can be discrete if they take on integer values (for example, duration of hospital stay in days) or continuous (2) if the values they take on are real numbers (for example, blood glucose level measured in mmol/L). It is common to see the values of continuous variables rounded to the nearest integers. For example, age of a patient may be recorded as 40 years instead of 39.5 years. The scales of measurement for *quantitative* variables could be *interval* if there is a consistent level of magnitude of difference between values. *Interval* scale has both a clearly defined order to the values and a unit of measure; however, the “zero” point on the scale is arbitrary. For example, body temperature measured in Fahrenheit scale - 0 degree Fahrenheit does not correspond a total absence of heat.

On the other hand, when the scale of measurement has an absolute zero, it is called a *ratio* scale, e.g. heart rate measured in beats/minute. Several numerical and graphical methods can be used for description and visualization of such variables. For example, large number of blood glucose measurements from patients in a diabetic care clinic can be summarized using two quantities. The mean indicating where most of the blood glucose measurements are located on the scale (hence called measure of central tendency or averages) and standard deviation indicating the presence of some variability in blood glucose measurements around this single value (hence called measure of dispersion or variability) (6,7) among the patients in the clinic. One can also use median and interquartile range to summarize quantitative variables whose values are not symmetrical. (Summery statistics and graphical representations will be presented in detail in the next Issue of EMJ).

POPULATION AND SAMPLES

In statistics and epidemiology, the term "population" has a wider meaning. It refers to a collection of not only people but also animals, objects, events, procedures, observations, or measurements that possess some common characteristics (8). The size and scope of a "population" vary depending on the research question. For example, a researcher may be interested in the average blood glucose level of adults in Addis Ababa, Ethiopia or in selected characteristics of patients visiting a specific diabetic care centre. The population in the first example refers to entire adults, healthy or sick, residing in the city and in the second example it refers to patients who would actually visit the diabetic care centre if they were ill during a specific period.

The researcher may not be able to enlist the entire population he or she is studying. However, he/she should explicitly define it based on the research question that is being addressed in that particular study. For example, in the blood glucose research question described above, it may not be possible to enumerate every adult in the whole city and measure their blood glucose level even if there might be a census of residents of the city from the Central Statistical Agency. In this case, appropriate sampling method need to be employed to select a representative sample of the Addis Ababa adult population using the Census Registry as a sampling frame. If the intent of the study is blood glucose level at health facility, the list of patients who visited the diabetes care centre with their blood glucose measurements may be available from the clinic's registry.

Both populations in the previous examples can further be specified using additional attributes or variables such as gender, age (e.g.40-65 years), health conditions (e.g., kidney failure patients), socio-economic characteristics (e.g., middle-income individuals living in big cities). This is particularly important because, often, one cannot study the entire population for several reasons such as cost, time, ethics, or logistics. However, researchers can collect information on a subset of the population of interest, called a sample, and can make generalizations or inferences, about the numerical characteristics (e.g., the average blood glucose levels) of the population being considered (7,8) (Figure 1). In other words, study sample is the group of individuals who actually participate in a study. The study participants who provide required information is called study sample and the total number of sampling units invited to participate in the study is called sample size (2,9).

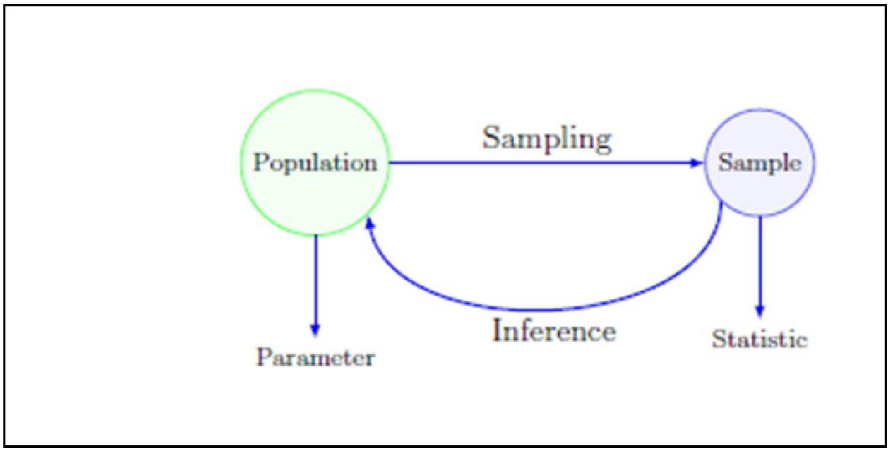

Figure 1. Relationship between population and sample

Any summary of a measurable characteristic such as the mean blood glucose level of the population is called a *parameter* and a summary of measurable characteristic of study participants or a sample is called a *statistic*(2,3). Parameter and statistic also differ in their commonly used notation and the formulae used to compute their numerical values (Table 2). Parameter is usually denoted by Greek letters (for example, population mean is denoted by μ or mu and the standard deviation is denoted by σ, lower case sigma) whereas corresponding statistic is often denoted by Latin letters (for example, sample mean is denoted by $\overline{X}$ and the standard deviation is denoted by s or SD). The formula used to calculate standard deviation of a given population $\left(\sum_{i=1}^{N}(x_i-\mu)^2/N\right)$ is different from the formula used to calculate the standard deviation of a sample $\left(\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)\right)$ where x_i. is measurement on subject i, N is the size of the population, n is sample size, μ is population mean.

Table 2. Comparison of a Population and a Sample Selected from the Population

	Population	Sample
Definition	All individuals with common characteristics defined by the research question.	Part of the population selected to represent the entire population defined by the research question.
Data availability	Often not feasible to get from every individual in the population.	For all study participants
Characteristic	Parameter	Statistic or estimate
Notation	Greek letters	Latin letters
Data collection	Census or complete enumeration.	Sample survey or sampling
Aim	Identify characteristics of all individuals that constitute the population defined by the research question.	Make generalizations about the characteristics of the population based on the information generated from the sample.

A sample selected from a population has fewer observations and contain less information than the full population. Hence, estimates (statistic) obtained from the sample always involve some uncertainty when they are used to represent population parameters- the smaller the sample size, the lesser information that exist in the sample about the population resulting in higher degree of uncertainty. However, sample can also contain the same number of individuals as the population or even more individuals than the population depending on the sampling method (for example, bootstrap sampling with replacement, which will be discussed in the coming series).

A well-chosen sample will contain most of the information about a particular population and is best obtained by random sampling (9). A sample is called representative of the population when it reflects the distribution of the population's characteristics such as age distribution or gender composition from which the sample is drawn. Sample statistic (e.g. mean age or proportion of females) calculated from representative sample will be close to corresponding population values (parameters) which can be obtained by summarizing the population data. Such statistic is said to be valid estimate of population parameter. The validity of the inference from the sample to the population from which the sample was taken or even to other populations is called external validity or generalizability (10). Whether a sample is representative of the population should be judged based on the knowledge of subject matter and it is not a statistical question in itself. The word "representative" refers to the characteristic of a sample, whereas the word "random" refers to the process by which the sample is selected from the population - and does not describe the sample as such (8).

Randomly drawn samples must have two characteristics: 1) every subject in the population has a known non-zero chance to be included in the sample; 2) selection of one individual from the population is independent of selection of another individual from the same population (2,4). The ideal random sampling method is called simple random sampling that gives equal chance of selection for every individual. However, it is often impossible or impractical to use simple random sample technique unless a list of all individuals in the population, also called a sampling frame, is available as well as manageable in terms of size.

When the sample is selected using random sampling, the characteristics of the individuals or different variables of interest are called random variables (2). The values of random variable are determined by chance. In order to make a connection between population, sample and random variable, let us consider a study conducted using a random sample selected from a population defined the research question. The raw data in such a study will consist of observations made on individuals included in the random sample of study participants. Any aspect of an individual in the sample that is measured or recorded, such as blood glucose level or gender is called a random variable. It is random in the sense that when we randomly select an individual from a defined population, inherently, we are also randomly selecting the measurements or records of this individual (e.g. his or her blood glucose measurement or gender) from a pool of measurements of the population.

For example, let us assume that a physician will randomly select a patient for a study by tossing a coin from the population of patients visiting her or his clinic for medical reasons. It is well known that for a fair coin, the probability or the chance that the coin tossed will result in a "head" and "tail" is equal to 0.5. The physician tosses a coin and if it is resulted in a "head", the next patient entering his/her office will be selected for the study and if it is resulted in a "tail", the patient will not be selected. Now, the probability that the next patient to be selected for the study will be a "female" is 0.5 as is the probability for any other patient visiting the physician to be included in the study during the recruitment period. By randomly selecting an individual for a sample, we are also randomly selecting attributes associated with the individual in the sample, which is why we call these attributes random variables. To draw valid inferences about the population from which the sample is derived, the reader needs precise information on the population, the sample, and the sampling method. The key point here is that the researchers' and readers' main interest is not the sample itself but the population or the information that the sample can provide about the population which the investigator cannot enumerate and obtain full information from (8). The population which the researcher wants to generalize the findings obtained from the sample is called the target population.

From a given population of size N, one can draw several random samples of the same size (say n) and each sample may consist of different individuals. Assume that the population size (i.e. N) is bigger compared to the size of the sample ($N > n$). Each sample will have its own mean which is slightly different from other sample means. This discrepancy arises due to chance or random variation, a phenomenon called sampling variation. Because of sampling variation, each sample mean is unlikely to be exactly equal to the population mean but it will be close to the corresponding value from the population. The extent to which the means of the different samples are close to each other, irrespective of the true population mean, is called precision. When a statistic is both valid and precise, it is termed as accurate (Figure 2). Note that the actual population parameter, for example μ, is a fixed and often an unknown quantity. However, this is what we want to estimate from the sample using sample statistic. The difference between a sample statistic used to estimate a population parameter and the actual but unknown value of the parameter is called sampling error.

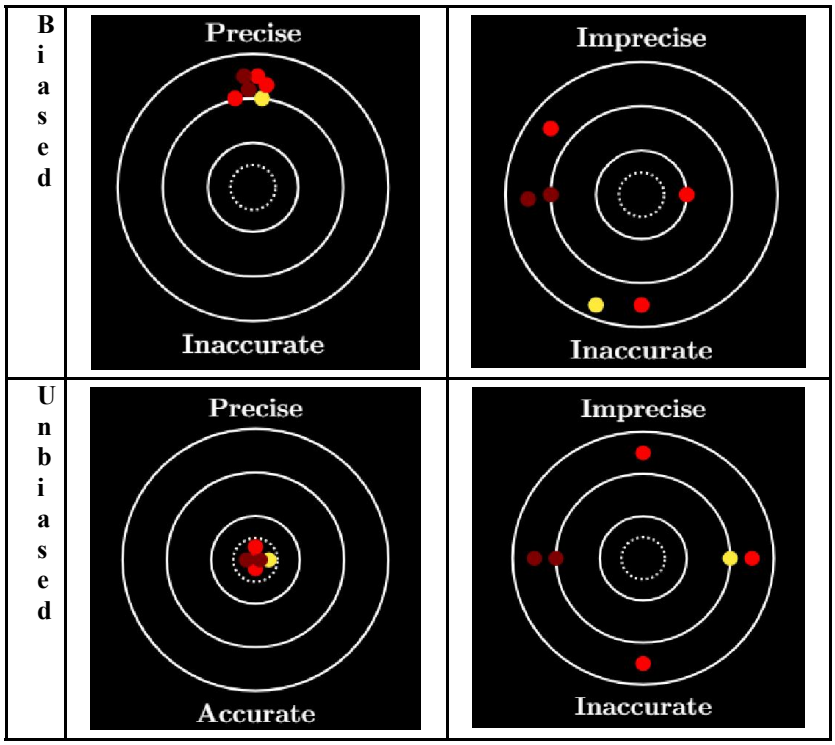

Figure 2: Differences between unbiased (valid), precise, and accurate estimates. The coloured small circles indicate estimates and the centre of the dotted circle is the true population parameter.

Now, if we collect all the means computed from several random samples and define a variable to take these sample means as its possible values, that variable can be considered as random variable (2) and will have a distribution of its own often called sampling distribution of the sample means. As the number of samples gets larger (at least more than 30 different samples of size n) the sampling distribution of the sample means will be indistinguishable from a Normal or Gaussian distribution. It is possible to compute the mean of this random variable (i.e. mean of the different sample means) and its value will be equal to the population mean as long as the number of samples of size n is very large. This assertion is based on the central limit theorem (2,6) which will be discussed in the next teaching article entitled "Distributions and central limit theorem". It is also possible to quantify the variability of the different sample means relative to the population mean, called the standard error (SE) of the mean. While standard deviation (sd) measures the variability of individual response in a given sample relative to the sample mean, the standard error of the mean (SE) measures the variability of means of several samples of size n relative to the population mean.

In order to understand how to make inference about the population parameter based on the information obtained from the sample and summarized in a statistic, we have to know the sampling distribution of the statistic of interest such as the mean, the proportion, and the standard deviation. The next article in the EMJ series will focus on probability distributions in general and sampling distributions of selected statistic and relevance of sampling distribution in statistical inference in particular.

REFERENCES

1. Altman D, J Bland. Statistics notes. Variables and parameters. In: BMJ 1999, 318: 7199.
2. Van Belle G, et al. Biostatistics: a methodology for the health sciences. Vol. 519. John Wiley & Sons, 2004.
3. Dendir E, Deyessa N. Intimate maternal partner violence and low birth weight in Addis Ababa public hospitals. Ethiop Med J 2018;56(2):5.
4. Yasin S, Ayana B, Bezabih B, Wamisho B. causes of pediatric limb amputations at Tikur Anbessa Specialized Hospital and the role of traditional bone-setters ("wogeshas"). Ethiop Med J, 2018, 56 (2):162-63.

5. Kebede T, Khalili K, Habtamu A, Henok F. The safety of image-guided biopsy and its impact on patient care: the first Ethiopian experience Ethiop Med J, 2018;56(1):6.
6. Betty R Kirkwood and Jonathan AC Sterne. Essential medical statistics. John Wiley & Sons, 2010.
7. Bland M.An introduction to medical statistics. Oxford University Press (UK), 2015
8. Michael J Campbell and Thomas Douglas Victor Swinscow. Statistics at square one. John Wiley & Sons, 2011.
9. Whitley E, Ball J. Statistics review 2: Samples and populations. Critical Care2002,6: 143.
10. Rothman K, Greenland S, Lash T. Modern epidemiology. 2008.