Biostatistics II
Data, variables and measurement scales

Tariq Zaman, Abbas Raza*

Department of Dermatology, FMH College of Medicine & Dentistry and Fatima Memorial Hospital, Lahore
* Department of Medicine Allama Iqbal Medical College and Jinnah Hospital, Lahore.

Abstract Biostatistics is part and parcel of biomedical research. The data from biomedical studies require statistical treatment for its effective presentation and correct interpretation. The choice of appropriate statistical methods to be applied depends upon the kind of data and variables. Different types of variables (characteristics) generate different types of data. Variables can be classified in many ways; qualitative and quantitative, nominal and ordinal or discrete and continuous. Different measurement scales are used to measure different types of variables; nominal & ordinal scales for qualitative and interval & ratio scale for quantitative variables. Different statistical methods/tests are employed for presentation and analysis of different types of data. For qualitative data; percentage, proportions, rate, ratios, standard error of proportions, chi-square test etc. are applied. While mean, range, standard deviation, coefficient of variation, correlation coefficient, etc. are employed in case of quantitative data. This article, second of the series, describes various types of data, variables and measurement scales.

Learning objective At the completion of this article, the reader should have a clear understanding of different types of data, variables and their measurement scales, and be able to utilize these concepts in organizing the research of his own and for interpreting the results of other studies.

Key word Data, variables, measurement scales.

Introduction

The relationship of bricks and nails to a building is the same as of data, variables and measurement scales to a research project. Building blocks of inappropriate size or arrangement lead to defective structure which cannot be covered by final polishing. Clear concepts of different types of data, variable and measurement scale is mandatory to choose the appropriate statistical methods to be employed for effective presentation and correct interpretation of the results. Mistakes in these basics can not be rectified later at the stage of data analysis and conclusion making.

An understanding of the variables and measurement scales is even more important as computerized statistical packages are being increasingly used for statistical analysis. Computers can do complex statistical calculations at an astonishing speed but the output they produce is only as sensible as the input we feed. If we apply inappropriate statistical tests, not meant for the type of variables and measurement scales under consideration, the results will

Address for correspondence
Dr. Tariq Zaman
210 G.T. Road, Baghban Pura, Lahore, Pakistan.

be illegitimate or confusing. Its output simply becomes a thoroughly processed nonsense. For this, computer programmers often use the term "GIGO", an acronym standing for "Garbage In, Garbage Out".

It is a common observation that the scientific research is started without a thoughtful consideration of types of data and variables involved and the scales of their measurement. This article, second of the series, therefore, is dedicated to building blocks of statistics, i.e. data, variables, their types and measurement scales.

Data

Generally the term data is applied to the information collected about any thing (persons, objects, places, etc.). Statistical data is a set of information systematically collected by observation/counting/measurement about characteristic(s) of more than one subjects/cases. Data is pleural of *datum* which represents information about a characteristic of one subject only.

The subjects or cases upon which the data are collected are called ***statistical units***. The characteristics of statistical units on which the information is collected are known as ***variables***. The information recorded is called ***observation***. Single observation about a variable of a subject is the ***datum*** and set of all the observations about one or more variables of all the subjects is known as ***data***.

For example, in **Table 1** observations are collected from ten students of a particular institute by a questionnaire survey about their various characteristics. Ten students (I.D. No. 1 to 10) are the subjects, cases or statistical units. Their nine characteristics, age, sex, weight, height, smoking, exercise, residence, pulse rate and blood group are the variables. All the ten values/figures of one or more variables of all the subjects are the data.

Medical statistics or data are collected from three main sources:

1. From existing ***records***, e.g. census data or hospital medical record.
2. By ***surveys***, e.g. by observing, measuring or administering a questionnaire.
3. By conducting ***experiments*** and recording the results.

Types of statistical data

The statistical data can be classified into two broad categories: qualitative data and quantitative data.

Qualitative data

The qualitative data have no magnitude or size of the characteristics. The subjects are classified by counting, having some characteristic or not, and not by measurement. The observations collected from characteristics such as sex, young, adult, vaccinated, diseased, cured, died, cause of death, taking drug or placebo, etc. are examples of qualitative data. In **Table 1** the data collected about the characteristics, sex, smoking, exercise, residence and blood group constitute the qualitative data.

In qualitative data, also called ***categorical data***, the subjects are counted and classified into various categories depending upon

Table 1 Students questionnaire data

I.D. No.	*Age (yrs)*	*Sex*	*Weight (Kg)*	*Height (cm)*	*Smoking*	*Exercise*	*Residence*	*Pulse Rate*	*Blood Group*
1	18.0	M	55.8	172.0	Yes	2	2	74	O
2	21.0	M	70.5	168.0	No	3	1	77	B
3	19.5	F	46.0	155.7	Yes	1	1	82	O
4	17.0	M	52.0	160.5	No	2	2	68	AB
5	18.0	M	64.3	180.0	Yes	1	1	70	A
6	20.0	F	54.8	158.8	No	2	2	79	A
7	19.5	F	48.9	161.3	No	3	1	80	B
8	18.0	M	74.0	178.0	No	2	2	76	O
9	19.0	M	67.5	167.0	Yes	1	2	80	A
10	20.5	F	50.2	145.4	No	1	1	75	O

Exercise: no =1, light = 2, heavy = 3; residence: urban = 1, rural = 2

Table 2 Classification of Students According To Gender and Exercise Status

Exercise Status	*No. of Students*		
	Male	*Female*	*Total*
Not	2	2	4
Light	3	1	4
Heavy	1	1	2
Total	6	4	10

alternative values of a specific characteristic (two or more than two). For example in **Table 1** for the characteristic *sex,* 6 are males and 4 are females (two categories) and for the characteristics *exercise,* 4 do not exercise, 4 perform light and 2 heavy exercise (three categories). The subjects may also be classified on more than one characteristic simultaneously. For example, the students in **Table 1** may be classified on gender and exercise status simultaneously in six categories: no exercise male, no exercise female, light exercise male, light exercise female, heavy exercise male and heavy exercise female, as shown in **Table 2**.

Statistically the qualitative data are expressed in ratio, proportion, percentage or rate. For its analysis, the commonly employed statistical methods are standard error of proportions and chi-square tests.

Quantitative data

The quantitative data have a magnitude, which can be measured. In qualitative data the subjects are classified or categorized into groups of a characteristic, while in qualitative data the characteristic of the subjects is measured or quantified. The characteristics or the variables, such as age, pulse rate, temperature, blood pressure, hemoglobin level, etc. can be measured and their values generate quantitative data. The observation on each subject is represented by a number or quantity. In **Table 1** the data recorded about the characteristics age, weight, height and pulse rate represent the quantitative data.

The statistical methods applied in the analysis of quantitative data are mean, range, standard deviation, coefficient of variation or correlation coefficient, etc.

Variables

A variable is a characteristic that varies in different subjects or cases. It has different values for different subjects. For example age is a variable as it varies from individual to individual. Similarly, heart rate, blood

pressure, leukocytes count, blood glucose, or a question (scabies is a contagious disease; True/False/Don't know), etc. are other examples of variables. The nine characteristics in student's questionnaire data in **Table 1** are all variables. A variable may also have different values in the same subject at different times, e.g. heart rate of a subject before and after exercise, a disease parameter before and after a specific therapy in a subject, etc.

The quantity that does not vary in different subjects or at different times in the same subject is called a ***constant.*** It has a single fixed value, e.g. π is a constant having a fixed value of 3.141. For a particular population the values of mean, proportion, standard deviation, standard error and correlation coefficient are considered as constants.

Types of variables

Variables are categorized into different types by many ways, some of which are as follow:

Qualitative vs. quantitative variables

The qualitative and quantitative data are described earlier. These data are generated by qualitative and quantitative variables, respectively.

The qualitative variables are also called categorical variables as observations in these variables fall into distinct and mutually exclusive categories. In **Table 1**, the characteristics of sex, smoking, exercise, residence and blood group are qualitative or categorical variables.

While observations in qualitative variables are counted, the observations in quantitative variables are measured and have certain magnitude. The quantitative variables are also called as measured variables. In **Table 1** the characteristics age, weight, height and pulse rate are examples of quantitative or measured variables.

Nominal vs. ordinal variables

Both nominal and ordinal variables are types of categorical variables. In nominal variables the categories are unordered while in ordinal variables they are arranged in a definite order.

In nominal (unordered) variables there is no natural order between the categories, as in variable sex it does not make any difference whether male comes first or female in data presentation. Similarly, any arrangement or order in blood groups A, B, AB or O has no significance. The variables sex, smoking, residence and blood group in **Table 1** are examples of nominal categorical variables.

In ordinal (ordered) variables the categories of the variables can be arranged in a definite order, and the order does matter, e.g. light/heavy, short/tall, mild/moderate/severe, etc. In **Table 1** the exercise status; no exercise, light exercise and heavy exercise, is an example of ordinal categorical variable.

Discrete vs. continuous variables

The variables which can have only integral values (whole numbers) over a range and not fractions of integers are called discrete or integral variables. Population of different cities, number of red blood cells in a blood sample, number of vaccinated individuals in

a group, number of births and deaths, etc. are examples of discrete or integral variable. The discrete variables may be qualitative or quantitative. In **Table 1** the variables of sex, smoking, exercise, residence, blood groups are examples of discrete qualitative variables and the pulse rate is a discrete quantitative variable.

While discrete variable can have only intermittent values, the continuous variable can have any value over a continuous spectrum or range such as, weight, hemoglobin level, blood protein level, temperature, etc. The continuous variables are always quantitative in nature. In **Table 1** the values of characteristics age, weight and height are the examples of continuous quantitative variables.

Therefore, qualitative variables are always discrete in nature, but quantitative variables may be continuous or discrete.

Measurement scales

Scaling is a process of measuring and scale of measurement is a particular way of assigning numbers or symbols to observations, e.g. kilogram for weight, meters for height, numbers for counts, etc.

The aim of measurement is to provide objective, accurate, and communicable descriptions of a variable. There are four basic scales of measurement as (a) nominal, (b) ordinal, (c) interval or (d) ratio.

Nominal scale

The nominal scale is the weakest level of measurement. Observations on this scale are classified into mutually exclusive qualitative categories only. For example the variables sex, vaccination status and color of eyes may be classified on nominal scale as male / female, vaccinated/non-vaccinated, green/blue/brown/black, respectively. Categories in nominal scale can be assigned numbers (1, 2, 3…) or symbols (A, B, C…) for identification purposes which however, carries no particular order or numerical significance. In **Table 1** the variables sex, smoking, residence and blood group are measured on nominal scale.

Ordinal scale

The ordinal scale is the next higher level of measurement and has the characteristics of nominal scale with the additional property of rank ordering the categories according to some criterion. For example, the variables socioeconomic status and disease severity may be categorized on ordinal scale as poor/lower middle/upper middle/rich, and mild/moderate/severe, respectively. Numbers or symbols may be assigned to these categories according to their natural order. The symbols A, B, C & D and numbers 1, 2 & 3 may be used for successive categories of socioeconomic status and disease severity, respectively. In table I the variable exercise is measured on ordinal scale into no, light and heavy exercise and assigned numbers 1, 2 & 3. The ordinal scale represents difference between categories in some direction but it does not describe the magnitude of difference between two categories.

Interval scale

The level of measurement of interval scale is higher than that of ordinal scale as the distance between any two intervals on the scale is defined and constant. Interval scale

not only elaborates the rank order between the observations that are measured, but also compares and quantifies the sizes of differences between them. However, in this scale the point designated as 'zero' is arbitrary and not absolute which is a characteristic of ratio scale. Therefore, its level of measurement is in between ordinal and ratio scale. The classical example of interval scale is temperature measured in degree Fahrenheit or Celsius. Here it can be said that a temperature of 40 degrees Fahrenheit or Celsius is higher than a temperature of 30 degrees, and that an increase from 20 to 40 degrees is twice as much as an increase from 30 to 40 degrees. But we cannot say that temperature of 40 degrees Fahrenheit or Celsius is twice as warm as that of 20 degrees, because the temperature of 0 degree Fahrenheit or Celsius does not represent absolute absence of warmth. However, in Kelvin scale, which is a ratio scale, 0 degree represent absolute absence of warmth and temperature below 0 degree is not possible. This property of absolute zero is not present in interval scale.

Ratio scales

Ratio scale is the highest level of measurement. In this scale, all the properties of interval scale are present along with an absolute zero point. In this way one can compare not only the difference between two observations (as in interval scale), but also compare the size of one observation with the other, i.e. the ratio. As Kelvin scale is a ratio scale, it can be said that temperature of 80 degrees Kelvin is twice as warm as temperature of 40 degrees, but this statement will not be legitimate for an interval scale like Fahrenheit or Celsius. In addition to Kelvin scale of temperature; kilograms for weight, meters for distance, years for age, etc. are all examples of ratio scales. In these examples zero means absence of the property being measured, as zero kilograms, meters or years means absolutely no weight, length or age, respectively. In **Table 1** age, weight, height and pulse rate are the variables measured on ratio scale.

In summary, nominal and ordinal scales are used to measure qualitative data while interval and ratio scales are used to measure qualitative data. However, interval scale is utilized infrequently in routine except for measuring temperature (in Fahrenheit and Celsius). The commonly employed scale for quantitative data is the ratio scale.

Further reading

1. Bluman AG, editor. *Elementary Statistics – step by step approach,* 5th edn. New York: McGraw-Hill; 2004.
2. Krishnamurty GB, Kasovia-Schmitt P, Ostroff DJ, editors. *Statistics – An interactive text for health and life sciences*. London: Jones & Bartlett Publishers International; 1995.
3. Mann PS, editor. *Introductory Statistics, 5th edn*. Singapore: John Willy & Sons, Inc; 2004.
4. Snedecor GW, Cochran WG, editors. *Statistical Methods, 7th edn*. Ames: Iowa State Univ. Press; 1998.
5. Armitage P, Berry G, editors. *Statistical Methods in Medical Research, 3rd edn*. Cambridge: Blackwell Scientific; 1994.
6. Swinscow TDV, Campbell MJ, editors. *Statistics at square one, 10th edn*. London: BMJ Books; 2003.
7. Appleton DR. What statistics should we teach medical undergraduates and graduates? *Stat Med* 1990; **9**:1013-21.
8. Zaman T, Raza A. Biostatistics-I: Introduction, role and applications in medicine. *J Pak Assoc Dermatol* 2004; **14**: 147-51.