Basic Concepts of Statistics.

Prasanta Mahapatra[1]

Statistics:

The word statistics is derived from the Italian word *stato*, which means "state" and *statista* which means a state official. Statistics originally meant facts useful to the state and collected by the *statista.* Today, we view statistics as a science of information. It deals with collection, analysis and interpretation of data. Three key areas of statistical methods are; descriptive statistics, exploratory data analysis, and inference. All three are connected. Statistical inference requires descriptive statistics at least on some samples. Most formal statistical inference methods are preceded by exploratory data analysis, testing conjectures, searching for possible patterns in data, and working towards formulation of testable hypotheses. Statistical inference is based on the concept of probability. Plausibility of hypotheses are tested on the basis of information contained in a sample.

Study Population or Universe:

At the very basic level statistics allows for concise description of an entity, or a phenomenon. The entity that we seek to describe may be a single event, an object, a person, or a collection, respectively, of events, objects and persons. A set of statistics may describe many aspects and / or give us fuller description of an aspect of the object of study. Objects can be described with help of a language such as English, Hindi or Telugu. For example, when we write an essay on a subject, we essentially, are describing that subject. Statistics on the subject allows for clear and succinct description. Both clarity and efficiency are important here. A statistic enables us to describe an aspect of the object of study efficiently[2]. We do use statistics to describe properties of a single event, an object or a person. Most commonly, however, statistics is about a collection or a class of events, objects and persons. The collection or class of objects[3] that we seek to describe and understand with help of statistics is referred to as the *study population* or simply *population*. Some refer to it as the *study universe*, *universe of study* or simply the *universe*. Where a population is large or its boundaries are not easily defined, statistics on a sample may be feasible. The sample statistics describe the sample and usually allow us to draw some inferences about the population. Here the *population* described by the statistics is the sample. Depending on the nature of the sample and the type of statistics, we are able to draw some inferences about the study population. The dual use of the word population in such situations has the potential for confusion. Moreover, statistics, these days rarely limits itself to the descriptive role alone. Hence statisticians usually use population to mean the larger population or the universe about which inferences are made with help of sample statistics. Expressions like the *descriptive statistics of the sample* are used to connote the descriptive role of a set of sample statistics. A population described by a set of statistics has to be real. We may, however, be able to draw inferences about a conceptual population based on statistics from a sample conceivably from

---

[1] President, The Institute of Health Systems, HACA Bhavan, Hyderabad, AP 5004, India.

[2] Note, however, that statistics usually do not describe the object in its entirety. Many attributes and aspects of an entity or phenomenon are not easily captured by statistics. We have a range of language based descriptive tools such as the essay, the story, the poem, and the performing arts. However, these are not always adequate for the job at hand, leaving scope for creative appearance of new forms of description and communication!

[3] For sake of brevity, the word object(s) will be used from now on to include events, objects, persons as well as phenomena that constitutes the object of a study.

the said population. Thus from an inferential perspective, the population or the universe may be real or conceptual. Suppose, for example, we observe the effect of a treatment on a sample of patients and find it to work. We would conclude that the treatment would work on similar patients in future. Here the population about which we seek to draw inference is conceived to constitute currently encountered patients as well as similar patients we are likely to encounter in future. We visualize the currently encountered patients as a sample of the universe consisting of all similar patients likely to exist ever. In other words, the universe of all similar patients likely to exist ever is the conceptual population of which the currently encountered patients are a sample.

We use the word *subject*, a *case*, an *object* or an *observation* to denote individual elements of a population. A subject is the smallest or the basic unit of study, a single item, an event or a person. Subjects have properties. Some properties may be same for all subjects in the population. Suppose, for example, we are interested in the survival of all infants born in a given area. All infants born and likely to be born in the said area would constitute the study population. Every member of this population is surely a human being. Thus the species of all subjects in this population is a *constant*. But the infants borne in the area would vary with respect to many other property such as the mothers literacy, gestation period, parents economic status, place of delivery etc. The property or characteristic with respect to which subjects differ in some measurable way is called a *variable*. Where the study population is conceived as the state of the subjects over time, characteristics of the same subject that differ from time to time will also be a variable. Note that, *variable* refers to the property, characteristics or attribute of subjects. A variable is a condition or quality that can vary from one case to another. It is the characteristic being measured on a set of people, objects or events. Each member of this set may take on different values. The property or characteristic that does not vary between subjects constituting belonging to a study population is called a *constant* with respect to that population. The term *variable* is most commonly used in general statistics. Synonyms exist and may be used by specific branches of science more commonly. For example, evolutionary and systematic biologist use the term *character* instead of the term *variable*. Another synonym of the word *variable* is *attribute*, commonly used in psychology mostly to denote properties that are either there or not there. Note that, to qualify as a variable, the attribute or property under consideration must consist of at least two values, for example, male and female (Pedhazur and Schmelkin, 1991, p174). Studying males or females only converts the variable sex into a constant. Further, at any given time it must be possible to assign to each element in the population one and only one value on the variable under consideration. For example, at any given time, a person can be assigned one and only one value on height, weight, age, etc. The following three examples illustrate the concept of population, and variables.

1. An nutritional epidemiologist wants to assess the extent of malnutrition among the adult population of an area. (S)he collects height and weight measurements of all adults in the area to compute body mass index (BMI) which is an indicator of malnutrition among adults. Here, the universe to be described consist of all adults living in the area. They are the study population. Here the study population consist of persons. Each person contributing one unit to the study population. Variables are the height, weight and BMI of each adult in the area.
2. A District Health Authority wants to describe the type of cases doctors are likely to encounter during the course of a year in the out patient department health care institutions of the area. Such a description would be an useful training for newly recruited medical officers about to be posted in out patient departments. Here the

universe to be described is the set of all clinical encounters in the outpatient department. The population consists of events, namely the out patient visits. A person making three outpatient visits for the various reasons contributes three units to the study population. Another person making a single outpatient visit contributes one unit to the study population. Each out patient clinical encounter would have many aspects. For example, the age, sex of the patient, presenting symptom, season or month of the year in which presenting, Billing Category (Free, Insurance, Employer Reimbursement, Out-of-Pocket, etc.), Emergency or not, etc. These are the variables of the out patient visit event.

3. A hospital administrator wants to describe the state of all water taps in the hospital. Each water tap is assigned an unique identification number. The administrator wants to describe various characteristics of the water tap population in the hospital. For example, dry (does not yield water) or not (yields water), leaking-or-not, tap positioning is appropriate or not, tap design is appropriate or not, etc. Here all water taps in the hospital constitute the population or study universe. Each water tap is a study unit. Variables are; dry-or-not, leaking-or-not, etc.

Type of Variables:

Variables are primarily classified from the measurement perspective[4]. From the measurement perspective, we first distinguish between (a) qualitative or categorical variables, and (d) quantitative variables.

Variables that are expressed qualitatively by classification of subjects to a set of categories without any magnitude relationship between them, are called categorical variables, nominal variables or attributes. Classification of subjects into different categories is the foundation of categorical variables. There is no size relationship between different categories. Objects assigned to different categories differ in kind but not in degree. Hence categorical variables are also referred to as qualitative variables. Since the labels assigned to each category are essentially names, the categorical variables are also referred to as nominal variables. Note however, that a qualitative variable arises only if subjects can be classified into mutually exclusive and collectively exhaustive categories. Mutual exclusivity means that if a subject is assigned to one category, then it can not be assigned to another category. For example, the variable sex usually has two categories male and female. A person is either a male or a female. Collective exhaustiveness means that every subject can be assigned to some category. Examples of categorical variables are sex, blood group, disease entity, treatment options, causes of death, etc. Different system of blood groups are a good example of qualitative variables in medical sciences. According to the ABO system, a person's blood group may be A, B, AB or O. According to the Rh blood grouping system, a persons blood group may be Rh+ or Rh-. Antecedent cause of death is a another example of a categorical variable in health sciences. The antecedent cause of death varies from one case to another. No

[4] Pedhazur and Schmelkin (1991, p174-179) discuss about classification of variables from (a) the measurement perspective and (b) the research question or inferential perspective. From the inferential perspective they recognise dependent and independent variables. They do however, recognise that variables are not inherently dependent or independent. The same variable may be conceived of as independent in one study, or even in one phase of the same study, and as dependent in another study, or in another phase of the same study. According to Tiryakian (1968) a typology should explicitly identify the dimension(s) along which items are assigned to different types. Typology of variables from a measurement perspective meets this criteria. In contrast, classification of variables as independent or dependent helps describe parts of a specific cause and effect model but does not help differentiate the variables per se along any dimension. Hence we prefer to present the primary typology of variables form the measurement perspective.

cause of death is any way more or less in quantity than another cause of death. Here the study universe consists of all deaths experienced by a given a population over a certain period of time. The International Classification of Diseases (ICD) issued by the World Health Organization from time to time contains an exhaustive classification of causes of death. Although many factors may contribute to death, the current official convention is to assign a single category from out of the ICD. Thus mutual exclusivity is ensured by convention. Description of the type of cases doctors are likely to encounter in a health centre, mentioned earlier can use the ICD chapters on classification of clinical encounters. A special type of categorical variables is the logical variable or attributes. These variables refer to the existence or lack of a property. For example, a case has a disease or does not have it. A case received the treatment or did not receive the treatment. The Dry-or-Not, Leaking-or-Not status of water taps in the water tap survey described earlier are examples of logical variables. Another class of categorical variables are dichotomous variables having a maximum of two categories. Examples are; (a) Sex with Male and Female as the only two categories, (b) Outpatient, Inpatient; etc. Polytomous qualitative variables have more than two categories. List-1 General Mortality - Condensed list of the ICD-10 has 103 cause of death categories. Polytomous categorical variables can be dichotomized for purposes of measurement and analysis. For example, Malaria and Other causes of death. Such a classification would be justified in case of a study on the mortality attributable to Malaria. Similarly, ownership of health care institutions may be conceived as a dichotomous variable with public and private as the two categories or a polytomous categorical variable with three categories such as public, nonprofit and forprofit.

Type	Sub types	Examples
Qualitative, Categorical or Nominal		
	Logical, Indicator, or Dummy Variables	Diseased or Not.
		Water tap is Dry or Not.
	Categorical Variables	Presenting Symptom of Outpatients
		Billing Category
		Classification of Cause of Death
Quantitative		
	Rank Order	Order of Birth
	Discrete or Meristic	Children Ever Born.
	Continuous	Height, Weight

Quantitative variables refer to characteristics having a magnitude. The magnitude may be ordinal, interval or continuous. The most basic notion of magnitude is rank order. Here subjects can be ordered according to their rank. But the difference between two subjects with successive ranks may not be the same as the difference between another pair of subjects with successive ranks. For example, take the rank order of students on the basis of marks obtained in an examination. The difference between the first and second rank holder is not necessarily the same as the difference between say the 11th and 12th rank holder. Alan Agresti (1990) recognises that “the position of ordinal variables on the quantitative / qualitative classification is fuzzy. They are often treated as qualitative, being analyzed using methods for nominal variables.” But in many respects, Agresti (1990) opines, ordinal variables are closer to discrete or continuous variables than nominal variables. They possess important

quantitative information. Each level has a greater or smaller magnitude of the characteristic than another level.

Some variables assume only discrete values. For example number of children ever born to a mother can only be a whole number. Continuous variables may take on any value in an interval of numbers. For example, height, weight, blood pressure, quantity of water consumed in a health care organisation, etc. A discrete variable has a countable number of values. A continuous variables can vary in quantity by infinitesimally small degrees (Argyrous, 2000 p13). Some authors classify quantitative variables as interval variables and ratio variables. But this is primarily a measurement issue and hence discussed later. Quantitative variables are either, ordinal, discrete or continuous.

Measurement:

Recall that the definition of a variable refers to differences between subjects in some measurable way. Measurability is a key requirement for us to recognise a characteristic as a variable. Without measurability, we do not even know if the characteristic remains constant or varies between subject in the population. Unless we are able to measure a characteristics in some way, we can not distinguish it from others nor can we draw any inferences about the population with an ill defined characteristic. Here measurement is to be interpreted in its broadest sense and includes recognition of the existence of an attribute, classification of a quality etc. Any measurement systems would have at least three distinguishable components, namely; (a) a scale or an instrument, (b) an observer, and (c) a measurement protocol. We will use a simple example to illustrate these three components. Suppose we are asked to measure the body weight of a group of people. To do so you will need a weighing balance. Choice of the weighing balance is an important issue linked to desired level of accuracy and margin of error. Analogue bathroom weighing scales are used in most clinical and epidemiological survey settings. These scales may have least counts of upto 0.5 Kg and error margins upto one Kg. Depending on the research question at hand and availability of funds, you may choose more or less accurate weighing machines. You and your coworkers have to learn how to use the balance to measure body weight. You are the observer. For example, rotating dial type bath room weighing scales usually require that the weight be read vertically straight off the dial. Reading from a side would give erroneous readings. Thirdly, you need to follow a measurement protocol. Apparel and footwear add to the measured weight and may be a source of errors. So you may want to follow a protocol to record weight after shoes are taken-off. Thus many factors affect the result of measurement on a variable, such as designed accuracy of the instrument, observer training, measurement protocol. These are maters of fact to be ascertained and factored into statistical analysis of data. A key aspect of the measurement process is the type or level of scale used to measure a variable. The level of measurement scale used ultimately determines the type of data available for further analysis. Hence we will review this aspect of the measurement process in some detail. There are four different type of scales or levels of measurement such as; (a) the nominal, (b) the ordinal, (c) the interval, and (d) the ratio scale. We will examine each of them in some detail.

The essential measurement act for a nominal scale is classification of objects into different categories. Nominal scaling requires appropriately developed classification techniques. Classification involves the ordering of cases in terms of their similarity. The terms Typology (mostly used by social scientists) and Taxonomy (mostly used by biologists) are synonymous with classification. Tiryakian (1968) provides a brief but comprehensive treatment of typology as an analytical tool in social sciences. Bailey (1994) gives a more

detailed account of various classification techniques. Very briefly, any valid classification system must be exhaustive and mutually exclusive. As mentioned earlier mutual exclusivity means that a subject can not be assigned to more than one category. Exhaustiveness means that every subject must be assigned to some category. Where the interest is on a few of many possible categories, a residual category such as “miscellaneous” or “others” is created. Whatever the classification rules, objects classified in different categories are treated as different in kind, not in degree. In other words, classes of a nominal scale are not ordered (Pedhazur and Schmelkin, 1991, p19). Thus, measurement with a nominal scale really amounts to classifying the objects and giving them the name (hence “nominal” scale) of the category to which they belong (Robert Pagano, 1994, p23). A fundamental property of nominal scales is that of equivalence. This means that all members of a given class are the same from the standpoint of the classification variable. Some times categories in a classification system may be given numerical codes. Assignment of any numerical code to a set of categories is purely to facilitate data handling and should not confused with any sense of magnitude.

Figure-1: Five point ordered categories used in a patient satisfaction survey questionnaire.

Negatively framed question:

You are usually kept waiting for a long time when you need doctor's attention / consultation.

Choice ->	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Score ->	**1**	**2**	**3**	**4**	**5**

Positively framed question:

You have easy access to the medical specialists in the hospital (MPSQ25)

Choice ->	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Score ->	**5**	**4**	**3**	**2**	**1**

[1] Source: Mahapatra Prasanta; Srinivas Kallam. APVVP Patient Satisfaction Survey, December 2001. Hyderabad: Institute of Health Systems, RP21/2002.

An ordinal scale represents the next higher level of measurement. It possesses a relatively low level of the property of magnitude. With an ordinal scale we rank order the objects being measured according to whether they posses more, less or the same amount of the variable being measured (Robert Pagano, 1994, p24). An ordinal level of measurement, in addition to the function of classification, allows cases to be ordered by degree according to measurements of the variable (Argyrous, 2000, p11). Since both nominal and ordinal scales categorize cases, they are sometimes called categorical scales. Ordinal scales allow for ranking of subjects. Variables referring to human attitude towards particular issues can be viewed to fall into ordered categories. Attitudes are measured usually with the help of three, five or seven ordered categories. The following figures shows a five point ordered categories to ascertain patient experience about accessibility of doctor. These two questions are taken from a patient satisfaction survey instrument (Mahapatra, Srilatha and Sridhar, 2001). Note that category labels remain the same but ordinal rankings are reversed depending on the positive or negative frame of the question.

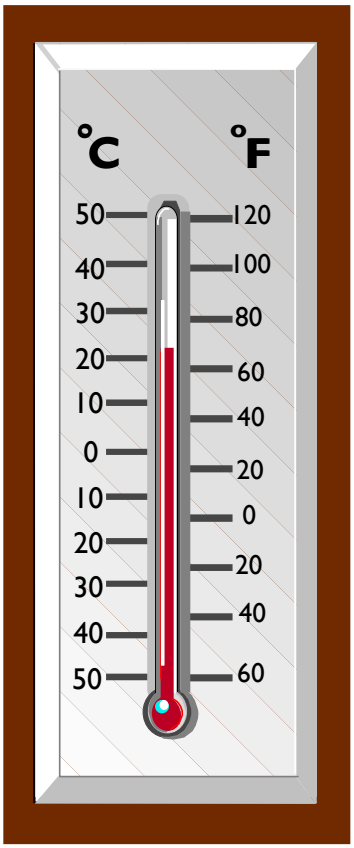
Health State Valuation by Card Sort

Community valuation of the relative severity of various health states usually starts with a card sorting exercise. Here the valuer works with a pack of health state cards. Each health state card describes a condition along six dimensions. The valuer is asked to order the cards from best health state in the pack to the worst health state in the pack. The data derived from this is a rank order within the respective set of health states. Following rank order was assigned to 11 health states including the valuers own health state, by a valuer. The valuer was randomly chosen from the AP Health State Valuation Study Data Set (Mahapatra and others, 1999).

Rank	Health State / Condition
1	Own Health Today
2	Mild Diabetes, no Symptoms
3	Watery Diarrhea 5 times a day
4	Mild Tuberculosis with Treatment
5	Below the Knee Amputation(one leg)
6	Peptic Ulcer
7	Below the Knee Amputation (two legs)
8	Two Broken Arms in Cast
9	Unipolar Major Depression
10	Severe Continuous Migraine
11	Quadriplegia

The interval scale represents a much higher level of measurement than the ordinal scale. It possesses the properties of magnitude and equal interval between adjacent units but does not have an absolute zero point. Thus, the interval scale possesses the properties of ordinal scale and, in addition, has equal intervals between adjacent units. Equal intervals between adjacent units means that there are equal amounts of the variable being measured between adjacent units on the scale (Robert Pagano, 1994, p24). Temperature recorded by the clinical thermometer is a good example of an interval scale. Clinical thermometers are either calibrated with a Fahrenheit scale or a Celsius scale. Both of these are interval scales. The intervals in the scale are equally placed. Although the scales have an interval labeled zero, they do not have a true zero point. A reading of 20^{o}C is not twice as hot as 10^{o}C. Interval scales allow arithmetic operations such as addition, subtraction, but do not allow for multiplication, or division. However, differences between two interval scale measurements can be treated as ratio data. For example the change in temperature measured with the Celsius or Fahrenheit scale can be divided by the change in temperature in another instance measured with the same scale.

Figure-2:
Interval-Scale Thermometer



The next and highest, level of measurement is called a ratio scale. It has all the properties of an interval scale and, in addition, has an absolute zero point. Without an absolute zero point, it is not legitimate to do ratios with the scale readings. Since the ratio scale has an absolute zero point, ratios are permissible (hence the name ratio scale). A good example to illustrate the difference between interval and ratio scales is to compare the Celsius scale of temperature with the Kelvin scale. Zero on the Kelvin scale is an absolute zero i.e. complete absence of heat. Zero on the Celsius scale is the temperature at which water freezes. It is an arbitrary zero point that actually occurs at 273° Kelvin. The Celsius scale is an interval scale and the Kelvin scale is a ratio scale. The difference in heat between 8° and 9° is the same as between 99° and 100° whether the scale is Celsius or Kelvin. However, we can not do ratios with the Celsius scale. A reading of 20° Celsius is really 273° + 20° = 293° Kelvin and 10° Celsius is 273° + 10° = 283° Kelvin. Clearly 293° Kelvin would not be twice as hot 283°. A reading of 20° Kelvin would be twice as hot as 10° Kelvin. Examples of ratio scale measurements are, time, length, weight, age, frequency counts, etc. Ratio scales allow for all arithmetic operations such as addition, subtraction, multiplication, and division.

Measurement instruments bearing an interval or ratio scales have a least count. The least count of an instrument is the smallest unit on the scale. Thus all measurements made on a continuous variable are approximate. For example some of the analogue bathroom weighing scales have a least count of 0.5 Kg. Suppose the exact weight of a person has a fraction of less than 0.5 Kg. The weighing machine can not distinguish this quantity. The observer will record the weight to the nearest 0.5 Kg. Hence the real limits of a continuous variable would be (the recorded measurement $\pm$ 0.5 $\times$ least count). Suppose a study records

10 people's body weight as 64.5 Kg. The real limit of these weights would be 64.25 to 64.75 Kg.

A Clinical Scale with Least Count of lbs or 0.5 Kg.

[1] These weighing scales are known under a variety of names such as; Fitness scales, Personal scales, Bathroom scales, Professional Dial Scales, etc.

Data:

Table-1: Different measurements on literacy giving different types of data.

Nominal	Ordinal	Interval
High School Pass or Not	Just Literate	Years of Schooling
	Primary	
Literate or Not	Secondary School	
	High School	
	College	
	Post Graduate	

Measurement on a variable gives data. The data can be nominal, ordinal, interval or ratio. A variable that is intrinsically nominal, such as Sex, can only be measured by a nominal scale i.e. a classification system. An ordinal variable when measured with help of a nominal

scale yields nominal data. It can yield ordinal data, if the measurement scale is ordinal. Similarly an interval type quantitative variable yields nominal data when measured with nominal scale, can yield both nominal and ordinal data when measured with an ordinal scale and similarly can give nominal, ordinal or interval data when measured with an interval scale. Operationally speaking, data is the starting point of statistical analyses. Naturally many statistical texts introduce the concept of qualitative, quantitative data and distinguish between nominal, ordinal, interval or ratio data and then proceed to illustrate various statistical operations.

Parameters:

A parameter quantifies a characteristic of the population. It is the true, but usually unknown, state of nature (i.e. the universe of study) about which we want to make an inference. Recall that population or the study universe can be real or conceptual. In fact, the study population or universe we deal with in inferential statistics are mostly conceptual rather than real. Appropriate conceptualization of the population is the fundamental research design act required before data can be analysed to draw any conclusion. Suppose we are interested to know about a population, we have that population physically available to us, and do not have any time or resource constraint to measure any characteristic of interest. Then we would not need to make any inference. We would simply measure the characteristic of interest and compute the parameters. But ideal measurement systems do not exist. Any measurement system will have some margin of error. So, at the least, we have to make some inference about the size of this error, to arrive at the exact parameter value. Since we can not know with certainty the size of the error, we can not compute with certainty the value of the parameter. The only opportunity available to us is to estimate the parameter value. More commonly, however, we conceive of study universes, all elements of which may not physically exist for measurement to happen. Suppose we are interested to assess the mortality risk to which infants born and brought up in a given village are exposed. Here the mortality risk is a state of nature. It is the variable to which people living in the village are exposed. We know that infant mortality risk is non negative non zero in any area. Theoretically, we can not have a situation where infant mortality risk can be said to be zero. Even if all infants actually born in the area survive until their first birth day, we can not say that infant mortality risk is zero. The risk of death is always there, howsoever small it may be! That is clearly our current understanding of the state of nature as far as biological life is concerned.

Table-2: Probability of observing no infant death during the course of a year in communities with different population size living under different levels of infant mortality risk (IMR).

Population Size	IMR-> Births	10	30	50	70
		Probability of at least one death in a year			
500	10	0.90	0.74	0.60	0.48
1000	20	0.82	0.54	0.36	0.23
2000	40	0.67	0.30	0.13	0.05
3000	60	0.55	0.16	0.05	0.01
4000	80	0.45	0.09	0.02	0.00
5000	100	0.37	0.05	0.01	0.00
10000	200	0.13	0.00	0.00	0.00
20000	400	0.02	0.00	0.00	0.00
30000	600	0.00	0.00	0.00	0.00
40000	800	0.00	0.00	0.00	0.00

[1] IMR is shown as number of deaths / 1000 child birth.
[2] Crude birth rate is assumed to remain constant at 20 births / 1000 population.
[3] The probabilities have been calculated using Lotus 123 @Binomial(Births,1,IMR,2)

The above table shows results from a thought experiment. It computes the probability of observing at least one infant death during the course of a year in a hypothetical community of different population size and experiencing different infant mortality risks. The infant mortality risk (IMR) is the population parameter. Four population parameter values are chosen. Remember that a population parameter is in truth a fixed value. It does not change since the population parameter is essentially a descriptive measure of the study population. The four population parameters imply four different populations. The table shows that for a village with 1000 population and truly experiencing infant mortality risk of 30 infant deaths / 1000 live births, the probability of observing a single infant death during the course of a year is less than half. Suppose we happen to study such a village, observe 20 child births here during the course of a year, and treat those child births as the population of our study. For any given year the chance that we will not encounter any infant death is a little more than half. Considering that chance is usually lumpy (Abelson, 1995 p17-38) one may not come across a single infant death for two to three years in a row. By the same logic, one may come across upto three infant deaths in a year given the same infant morality risk of 30 infant deaths / 1000 live births acting on the same village with 1000 persons. Of course, the probability of such an event would be very low, some where around 0.0183 implying that such an event may occur, on average, once in 55 years. In the first scenario, if we assume our study universe to consist of all infants born in this hypothetical village of 1000 people, we would conclude that the infants in this village are not exposed to any mortality risk. In the second scenario of 3 observed infant deaths, we might conclude that the infant mortality risk in this village is 150 infant deaths / 1000 live births! Note that we started with an assumption that the true IMR for this village to be 30 infant deaths / 1000 live births. Instead, visualizing a larger hypothetical population of which the 20 infants born in the village during a year is a sample, allows us to view the sample statistic to be an estimate the true infant mortality risk. Even then, the second scenario with three observed deaths will give us a point estimate of 150 infant deaths / 1000 live births. But this time, since we have assumed the 20 births to be a sample from a large number of potential births, we will calculate statistical confidence limits and give out a range of values that straddles the point estimate of 150. Also note that the population of the village is different from the study population. Here the study population consists of infants born or likely to be born in the village. The only link of the village

population to the study population is that the village population contribute to the sample by giving birth to infants. We are using the same word population to mean two different concepts. Firstly the village population means actual human beings living in a village. The study population here is conceived to consist of the infants actually born as well as the infants that could have born in the village. A less confusing term would be the study universe, which is synonymous with study population.

Parameters are usually represented by Greek alphabets. This is a convention. For example the Greek letter μ is traditionally used to denote arithmetic mean of a variable. The Greek letter σ is used to represent population variance of the variable. Europe's ancient philosophy developed in the Graeco-Roman world. Ancient Roman philosophy sprang from the Greek tradition. Most of the philosophical writings were in the Greek language. Of all the branches of human knowledge, philosophy involves a lot of reflection and thinking about the world. The tradition of using Greek alphabets to denote population parameters appears to reinforce the idea that they refer to state of nature, and the truth about a variable, that we may not ever know with complete certainty.

Sample:

A sample is a subset of the study universe.

Statistic:

A statistic is any function of the data. Data invariable come from samples. Thus statistic is a number calculated on sample data that quantifies a characteristic of the sample. A sample statistic gives information about a corresponding population parameter. For example, the sample mean for a sample of data would enable us to estimate the population mean.

References:

Abelson Robert P. Statistics as Principled Argument. Hillsdale, NJ, USA//Hove UK: Lawrence Erlbaum Associates, Publishers; 1995.

Agresti Alan. Categorical data analysis. New York: John Wiley and Sons; 1990.

Argyrous George. Statistics for Social and Health Research. With a Guide to SPSS. New Delhi, London, Thousand Oaks: SAGE Publications; 2000.

Bailey Kenneth D. Typologies and taxonomies. An introduction to classification techniques. Newburry Park//London//New Delhi: Sage; 1994.

Mahapatra Prasanta, Srilatha S., Sridhar P. A patient satisfaction survey in public hospitals. Journal. Academy of Hospital Administration 2001 Jul-2001 Dec 31;13(2):11-5.

Pagano Robert R. Understanding statistics in the behavioral sciences. Fourth edition. Minneapolis-St Paul: West publishing co.; 1994.

Pedhazur Elazar J.; Liora Pedhazur Schmelkin. Measurement, Design, and Analysis: An Integrated Approach. Hillsdale, New Jersey//Hove//London: Lawrence Erlbaum Associates; 1991.

Tiryakian E.A. Typologies. in: Sills D.L., Editor. International encyclopedia of the social sciences. Vol 16.1968. pp. 177-86.

Exercises

1. Identify the study population and sample in the following situations:
   i. Patient Satisfaction Survey: The AP Vaidya Vidhana Parishad (APVVP) manages a large network of public hospitals in Andhra Pradesh. A Patient Satisfaction Surveys (PSS) was conducted by the Institute in the year 1999. All of the 19 District Hospitals, and all of the six Area Hospitals were covered by the survey. The study team wanted to gather a sample of about 50 inpatients from each hospital. The survey team visited the hospital with short notice to the hospital authorities. All patients who had completed three days of stay in the hospital were listed. If the number exceeded 50 a simple random sample was taken to select 50 patients. If the list was equal to 50, all in the list were interviewed. If less than 50, the team recruited additional patients as they completed three day stay in the hospital. A patient satisfaction questionnaire was administered to the patient or an attendant.
   ii. Study on the Structure and Dynamics of Private Health Sector in Andhra Pradesh (SDPHAP): This study sought to understand the structure and dynamics of the private health sector in Andhra Pradesh, in order to provide insights for meaningful policy intervention. Three areas in Andhra Pradesh were selected for the study. The area around the state capital namely Hyderabad - Ranga Reddy districts was chosen purposively. In addition, Visakhapatnam district was chosen from the economically developed districts and Warangal was selected from the economically backward districts. Thus the sample areas are; (a) Hyderabad - Ranga Reddy, (b) Visakhapatnam, and (c) Warangal districts. Altogether 256 health care institutions were surveyed consisting of 150 HCIs in the private and non profit sector and 106 HCIs in the public sector. Three types of HCIs were studied, namely (a) large hospitals, (b) small hospitals and (c) clinics or primary health centres.
2. Following is an extract from the IHS Director's report to the eighth annual general meeting (2000-2001), on 06 December, 2001. Identify the descriptive and / or inferential role of the statistics cited by the Institute Director.

"People are gradually recognising the bibliographic niche being cultivated by the IHS library. Although our library is small, it has some collections in the area of health economics, health system research etc. not easily available elsewhere in Hyderabad. As of January 2000 we had 25 associate members. Currently we have 37 associate members including one life associate member. Most of these memberships are taken to access the Institutes library services. Currently the library services about 323 retrievals per month."

3. Indicate which of the following represent a variable and which a constant:
   i. The number of letters in the Devanagari script.
   ii. The number of letters in various languages of the world.
   iii. The number of hours in a day.
   iv. The time at which you eat dinner.
   v. The number of students admitted every year to the same academic program in which you are enrolled.
   vi. The amount of sleep you get each night.
   vii. Body weight of people in your class.
   viii. The number of playing cards in a pack.

4. Following is an extract of selected questions from the NFHS-2 Women's questionnaire[5]. NFHS-2 instructions about codes, if any, to be assigned to the responses are also given. Identify the type of variable (qualitative, quantitative, etc.) implied by respective questions.

No.	Questions and Filters	Coding Categories	
107	What is your marital status?	Currently Married	1
		Married but Gauna not performed	2
		Separated	3
		Deserted	4
		Divorced	5
		Widowed	6
		Never Married	7
119	Can you read and write?	Yes .....1 No 2	
201	Have you ever given birth?	Yes	1
		No	2
230	Are you pregnant now?	Yes	1
		No	2
		Unsure	8
231	How many months pregnant are you?	Months	
902	Respondent's Haemoglobin level G/DL		

5. Survey of water taps: A survey of water taps in a public building asks the following question for each tap. Identify the variable types and measurement scales.
   i. TapId: Alphanumeric identification given by you to uniquely identify each tap.
   ii. Location: Area where located. For example; Corridor, Patio, Gents Toilet -1, etc.
   iii. Date: Date of the survey.
   iv. Time: Time of the survey.
   v. Yield: Open the tap and observe for water flow and report; Dry, Trickle, or Flow.
   vi. Leaking: Applicable only if Yield # Dry. Open the tap and let water flow for a few seconds. Then close the tap. Observe if flow stops completely (Leaking = False) or there is some leak (Leaking = True).
   vii. Positioning: Consider the purpose for which the tap has been provided? Give your assessment about positioning of the tap from the perspective of the intended use. For example, a tap meant for hand wash should be is the tap positioning appropriate for its intended use? If yes, then Positioning = Y else, Positioning = N.

[5] The first National Family Health Survey (NFHS-1) in India was conducted during 1992-93. The second survey (NFHS-2) was conducted during 1998-99. The principal objective of NFHS is to provide state and national level estimates fertility, the practice of family planning, infant and child mortality, maternal and child health, and the utilization of health services provided to mothers and children. The NFHS in India is in many respects similar in scope to the Demographic and Health Surveys (DHS) elsewhere in the world. NFHS-2 India report has been published by International Institute for Population Sciences (IIPS); ORC Macro; Roy TK, et al. National Family Health Survey 1998-99 (NFHS-2). India. Mumbai (Bombay): International Institute for Population Sciences (IIPS); 2000 Oct.

viii. Design: Consider the purpose for which the tap has been provided? Give your assessment about design of the tap from the perspective of the intended use. For example, a tap in operation theatre area for hand wash and scrubbing should have an elbow operated lever to open and close the tap. A tap serving a sink should have a long enough neck, so that a person does not have to bend a lot while washing things in the sink. Thus ask; is the tap design appropriate for its intended use? If yes, then Design = Y else, Design = N.

ix. Tapwork: Assess if the tap needs any work. Choose one of the following: Repair, Replace, Reposition, Redesign, Reposition And Redesign, Connect (with water source), or None.

Basic Concepts of Statistics.

Prasanta Mahapatra[1]

Statistics:

The word statistics is derived from the Italian word *stato*, which means "state" and *statista* which means a state official. Statistics originally meant facts useful to the state and collected by the *statista.* Today, we view statistics as a science of information. It deals with collection, analysis and interpretation of data. Three key areas of statistical methods are; descriptive statistics, exploratory data analysis, and inference. All three are connected. Statistical inference requires descriptive statistics at least on some samples. Most formal statistical inference methods are preceded by exploratory data analysis, testing conjectures, searching for possible patterns in data, and working towards formulation of testable hypotheses. Statistical inference is based on the concept of probability. Plausibility of hypotheses are tested on the basis of information contained in a sample.

Study Population or Universe:

At the very basic level statistics allows for concise description of an entity, or a phenomenon. The entity that we seek to describe may be a single event, an object, a person, or a collection, respectively, of events, objects and persons. A set of statistics may describe many aspects and / or give us fuller description of an aspect of the object of study. Objects can be described with help of a language such as English, Hindi or Telugu. For example, when we write an essay on a subject, we essentially, are describing that subject. Statistics on the subject allows for clear and succinct description. Both clarity and efficiency are important here. A statistic enables us to describe an aspect of the object of study efficiently[2]. We do use statistics to describe properties of a single event, an object or a person. Most commonly, however, statistics is about a collection or a class of events, objects and persons. The collection or class of objects[3] that we seek to describe and understand with help of statistics is referred to as the *study population* or simply *population*. Some refer to it as the *study universe*, *universe of study* or simply the *universe*. Where a population is large or its boundaries are not easily defined, statistics on a sample may be feasible. The sample statistics describe the sample and usually allow us to draw some inferences about the population. Here the *population* described by the statistics is the sample. Depending on the nature of the sample and the type of statistics, we are able to draw some inferences about the study population. The dual use of the word population in such situations has the potential for confusion. Moreover, statistics, these days rarely limits itself to the descriptive role alone. Hence statisticians usually use population to mean the larger population or the universe about which inferences are made with help of sample statistics. Expressions like the *descriptive statistics of the sample* are used to connote the descriptive role of a set of sample statistics. A population described by a set of statistics has to be real. We may, however, be able to draw inferences about a conceptual population based on statistics from a sample conceivably from

---

[1] President, The Institute of Health Systems, HACA Bhavan, Hyderabad, AP 5004, India.

[2] Note, however, that statistics usually do not describe the object in its entirety. Many attributes and aspects of an entity or phenomenon are not easily captured by statistics. We have a range of language based descriptive tools such as the essay, the story, the poem, and the performing arts. However, these are not always adequate for the job at hand, leaving scope for creative appearance of new forms of description and communication!

[3] For sake of brevity, the word object(s) will be used from now on to include events, objects, persons as well as phenomena that constitutes the object of a study.

the said population. Thus from an inferential perspective, the population or the universe may be real or conceptual. Suppose, for example, we observe the effect of a treatment on a sample of patients and find it to work. We would conclude that the treatment would work on similar patients in future. Here the population about which we seek to draw inference is conceived to constitute currently encountered patients as well as similar patients we are likely to encounter in future. We visualize the currently encountered patients as a sample of the universe consisting of all similar patients likely to exist ever. In other words, the universe of all similar patients likely to exist ever is the conceptual population of which the currently encountered patients are a sample.

We use the word *subject*, a *case*, an *object* or an *observation* to denote individual elements of a population. A subject is the smallest or the basic unit of study, a single item, an event or a person. Subjects have properties. Some properties may be same for all subjects in the population. Suppose, for example, we are interested in the survival of all infants born in a given area. All infants born and likely to be born in the said area would constitute the study population. Every member of this population is surely a human being. Thus the species of all subjects in this population is a *constant*. But the infants borne in the area would vary with respect to many other property such as the mothers literacy, gestation period, parents economic status, place of delivery etc. The property or characteristic with respect to which subjects differ in some measurable way is called a *variable*. Where the study population is conceived as the state of the subjects over time, characteristics of the same subject that differ from time to time will also be a variable. Note that, *variable* refers to the property, characteristics or attribute of subjects. A variable is a condition or quality that can vary from one case to another. It is the characteristic being measured on a set of people, objects or events. Each member of this set may take on different values. The property or characteristic that does not vary between subjects constituting belonging to a study population is called a *constant* with respect to that population. The term *variable* is most commonly used in general statistics. Synonyms exist and may be used by specific branches of science more commonly. For example, evolutionary and systematic biologist use the term *character* instead of the term *variable*. Another synonym of the word *variable* is *attribute*, commonly used in psychology mostly to denote properties that are either there or not there. Note that, to qualify as a variable, the attribute or property under consideration must consist of at least two values, for example, male and female (Pedhazur and Schmelkin, 1991, p174). Studying males or females only converts the variable sex into a constant. Further, at any given time it must be possible to assign to each element in the population one and only one value on the variable under consideration. For example, at any given time, a person can be assigned one and only one value on height, weight, age, etc. The following three examples illustrate the concept of population, and variables.

1. An nutritional epidemiologist wants to assess the extent of malnutrition among the adult population of an area. (S)he collects height and weight measurements of all adults in the area to compute body mass index (BMI) which is an indicator of malnutrition among adults. Here, the universe to be described consist of all adults living in the area. They are the study population. Here the study population consist of persons. Each person contributing one unit to the study population. Variables are the height, weight and BMI of each adult in the area.
2. A District Health Authority wants to describe the type of cases doctors are likely to encounter during the course of a year in the out patient department health care institutions of the area. Such a description would be an useful training for newly recruited medical officers about to be posted in out patient departments. Here the

universe to be described is the set of all clinical encounters in the outpatient department. The population consists of events, namely the out patient visits. A person making three outpatient visits for the various reasons contributes three units to the study population. Another person making a single outpatient visit contributes one unit to the study population. Each out patient clinical encounter would have many aspects. For example, the age, sex of the patient, presenting symptom, season or month of the year in which presenting, Billing Category (Free, Insurance, Employer Reimbursement, Out-of-Pocket, etc.), Emergency or not, etc. These are the variables of the out patient visit event.

3. A hospital administrator wants to describe the state of all water taps in the hospital. Each water tap is assigned an unique identification number. The administrator wants to describe various characteristics of the water tap population in the hospital. For example, dry (does not yield water) or not (yields water), leaking-or-not, tap positioning is appropriate or not, tap design is appropriate or not, etc. Here all water taps in the hospital constitute the population or study universe. Each water tap is a study unit. Variables are; dry-or-not, leaking-or-not, etc.

Type of Variables:

Variables are primarily classified from the measurement perspective[4]. From the measurement perspective, we first distinguish between (a) qualitative or categorical variables, and (d) quantitative variables.

Variables that are expressed qualitatively by classification of subjects to a set of categories without any magnitude relationship between them, are called categorical variables, nominal variables or attributes. Classification of subjects into different categories is the foundation of categorical variables. There is no size relationship between different categories. Objects assigned to different categories differ in kind but not in degree. Hence categorical variables are also referred to as qualitative variables. Since the labels assigned to each category are essentially names, the categorical variables are also referred to as nominal variables. Note however, that a qualitative variable arises only if subjects can be classified into mutually exclusive and collectively exhaustive categories. Mutual exclusivity means that if a subject is assigned to one category, then it can not be assigned to another category. For example, the variable sex usually has two categories male and female. A person is either a male or a female. Collective exhaustiveness means that every subject can be assigned to some category. Examples of categorical variables are sex, blood group, disease entity, treatment options, causes of death, etc. Different system of blood groups are a good example of qualitative variables in medical sciences. According to the ABO system, a person's blood group may be A, B, AB or O. According to the Rh blood grouping system, a persons blood group may be Rh+ or Rh-. Antecedent cause of death is a another example of a categorical variable in health sciences. The antecedent cause of death varies from one case to another. No

[4] Pedhazur and Schmelkin (1991, p174-179) discuss about classification of variables from (a) the measurement perspective and (b) the research question or inferential perspective. From the inferential perspective they recognise dependent and independent variables. They do however, recognise that variables are not inherently dependent or independent. The same variable may be conceived of as independent in one study, or even in one phase of the same study, and as dependent in another study, or in another phase of the same study. According to Tiryakian (1968) a typology should explicitly identify the dimension(s) along which items are assigned to different types. Typology of variables from a measurement perspective meets this criteria. In contrast, classification of variables as independent or dependent helps describe parts of a specific cause and effect model but does not help differentiate the variables per se along any dimension. Hence we prefer to present the primary typology of variables form the measurement perspective.

cause of death is any way more or less in quantity than another cause of death. Here the study universe consists of all deaths experienced by a given a population over a certain period of time. The International Classification of Diseases (ICD) issued by the World Health Organization from time to time contains an exhaustive classification of causes of death. Although many factors may contribute to death, the current official convention is to assign a single category from out of the ICD. Thus mutual exclusivity is ensured by convention. Description of the type of cases doctors are likely to encounter in a health centre, mentioned earlier can use the ICD chapters on classification of clinical encounters. A special type of categorical variables is the logical variable or attributes. These variables refer to the existence or lack of a property. For example, a case has a disease or does not have it. A case received the treatment or did not receive the treatment. The Dry-or-Not, Leaking-or-Not status of water taps in the water tap survey described earlier are examples of logical variables. Another class of categorical variables are dichotomous variables having a maximum of two categories. Examples are; (a) Sex with Male and Female as the only two categories, (b) Outpatient, Inpatient; etc. Polytomous qualitative variables have more than two categories. List-1 General Mortality - Condensed list of the ICD-10 has 103 cause of death categories. Polytomous categorical variables can be dichotomized for purposes of measurement and analysis. For example, Malaria and Other causes of death. Such a classification would be justified in case of a study on the mortality attributable to Malaria. Similarly, ownership of health care institutions may be conceived as a dichotomous variable with public and private as the two categories or a polytomous categorical variable with three categories such as public, nonprofit and forprofit.

Type	Sub types	Examples
Qualitative, Categorical or Nominal		
	Logical, Indicator, or Dummy Variables	Diseased or Not.
		Water tap is Dry or Not.
	Categorical Variables	Presenting Symptom of Outpatients
		Billing Category
		Classification of Cause of Death
Quantitative		
	Rank Order	Order of Birth
	Discrete or Meristic	Children Ever Born.
	Continuous	Height, Weight

Quantitative variables refer to characteristics having a magnitude. The magnitude may be ordinal, interval or continuous. The most basic notion of magnitude is rank order. Here subjects can be ordered according to their rank. But the difference between two subjects with successive ranks may not be the same as the difference between another pair of subjects with successive ranks. For example, take the rank order of students on the basis of marks obtained in an examination. The difference between the first and second rank holder is not necessarily the same as the difference between say the 11th and 12th rank holder. Alan Agresti (1990) recognises that “the position of ordinal variables on the quantitative / qualitative classification is fuzzy. They are often treated as qualitative, being analyzed using methods for nominal variables.” But in many respects, Agresti (1990) opines, ordinal variables are closer to discrete or continuous variables than nominal variables. They possess important

quantitative information. Each level has a greater or smaller magnitude of the characteristic than another level.

Some variables assume only discrete values. For example number of children ever born to a mother can only be a whole number. Continuous variables may take on any value in an interval of numbers. For example, height, weight, blood pressure, quantity of water consumed in a health care organisation, etc. A discrete variable has a countable number of values. A continuous variables can vary in quantity by infinitesimally small degrees (Argyrous, 2000 p13). Some authors classify quantitative variables as interval variables and ratio variables. But this is primarily a measurement issue and hence discussed later. Quantitative variables are either, ordinal, discrete or continuous.

Measurement:

Recall that the definition of a variable refers to differences between subjects in some measurable way. Measurability is a key requirement for us to recognise a characteristic as a variable. Without measurability, we do not even know if the characteristic remains constant or varies between subject in the population. Unless we are able to measure a characteristics in some way, we can not distinguish it from others nor can we draw any inferences about the population with an ill defined characteristic. Here measurement is to be interpreted in its broadest sense and includes recognition of the existence of an attribute, classification of a quality etc. Any measurement systems would have at least three distinguishable components, namely; (a) a scale or an instrument, (b) an observer, and (c) a measurement protocol. We will use a simple example to illustrate these three components. Suppose we are asked to measure the body weight of a group of people. To do so you will need a weighing balance. Choice of the weighing balance is an important issue linked to desired level of accuracy and margin of error. Analogue bathroom weighing scales are used in most clinical and epidemiological survey settings. These scales may have least counts of upto 0.5 Kg and error margins upto one Kg. Depending on the research question at hand and availability of funds, you may choose more or less accurate weighing machines. You and your coworkers have to learn how to use the balance to measure body weight. You are the observer. For example, rotating dial type bath room weighing scales usually require that the weight be read vertically straight off the dial. Reading from a side would give erroneous readings. Thirdly, you need to follow a measurement protocol. Apparel and footwear add to the measured weight and may be a source of errors. So you may want to follow a protocol to record weight after shoes are taken-off. Thus many factors affect the result of measurement on a variable, such as designed accuracy of the instrument, observer training, measurement protocol. These are maters of fact to be ascertained and factored into statistical analysis of data. A key aspect of the measurement process is the type or level of scale used to measure a variable. The level of measurement scale used ultimately determines the type of data available for further analysis. Hence we will review this aspect of the measurement process in some detail. There are four different type of scales or levels of measurement such as; (a) the nominal, (b) the ordinal, (c) the interval, and (d) the ratio scale. We will examine each of them in some detail.

The essential measurement act for a nominal scale is classification of objects into different categories. Nominal scaling requires appropriately developed classification techniques. Classification involves the ordering of cases in terms of their similarity. The terms Typology (mostly used by social scientists) and Taxonomy (mostly used by biologists) are synonymous with classification. Tiryakian (1968) provides a brief but comprehensive treatment of typology as an analytical tool in social sciences. Bailey (1994) gives a more

detailed account of various classification techniques. Very briefly, any valid classification system must be exhaustive and mutually exclusive. As mentioned earlier mutual exclusivity means that a subject can not be assigned to more than one category. Exhaustiveness means that every subject must be assigned to some category. Where the interest is on a few of many possible categories, a residual category such as “miscellaneous” or “others” is created. Whatever the classification rules, objects classified in different categories are treated as different in kind, not in degree. In other words, classes of a nominal scale are not ordered (Pedhazur and Schmelkin, 1991, p19). Thus, measurement with a nominal scale really amounts to classifying the objects and giving them the name (hence “nominal” scale) of the category to which they belong (Robert Pagano, 1994, p23). A fundamental property of nominal scales is that of equivalence. This means that all members of a given class are the same from the standpoint of the classification variable. Some times categories in a classification system may be given numerical codes. Assignment of any numerical code to a set of categories is purely to facilitate data handling and should not confused with any sense of magnitude.

Figure-1: Five point ordered categories used in a patient satisfaction survey questionnaire.

Negatively framed question:

You are usually kept waiting for a long time when you need doctor's attention / consultation.

Choice ->	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Score ->	**1**	**2**	**3**	**4**	**5**

Positively framed question:

You have easy access to the medical specialists in the hospital (MPSQ25)

Choice ->	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Score ->	**5**	**4**	**3**	**2**	**1**

[1] Source: Mahapatra Prasanta; Srinivas Kallam. APVVP Patient Satisfaction Survey, December 2001. Hyderabad: Institute of Health Systems, RP21/2002.

An ordinal scale represents the next higher level of measurement. It possesses a relatively low level of the property of magnitude. With an ordinal scale we rank order the objects being measured according to whether they posses more, less or the same amount of the variable being measured (Robert Pagano, 1994, p24). An ordinal level of measurement, in addition to the function of classification, allows cases to be ordered by degree according to measurements of the variable (Argyrous, 2000, p11). Since both nominal and ordinal scales categorize cases, they are sometimes called categorical scales. Ordinal scales allow for ranking of subjects. Variables referring to human attitude towards particular issues can be viewed to fall into ordered categories. Attitudes are measured usually with the help of three, five or seven ordered categories. The following figures shows a five point ordered categories to ascertain patient experience about accessibility of doctor. These two questions are taken from a patient satisfaction survey instrument (Mahapatra, Srilatha and Sridhar, 2001). Note that category labels remain the same but ordinal rankings are reversed depending on the positive or negative frame of the question.

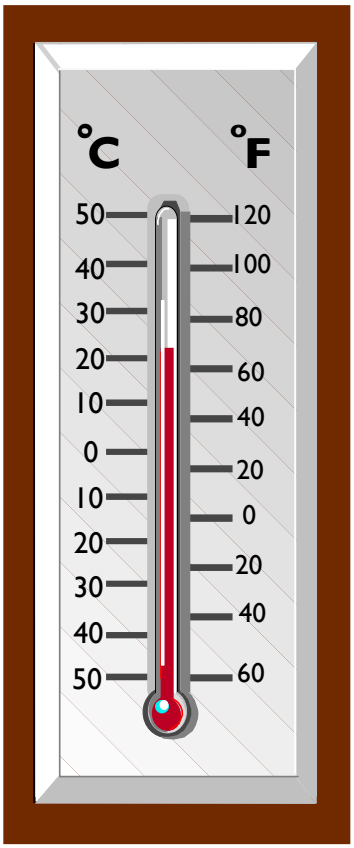
Health State Valuation by Card Sort

Community valuation of the relative severity of various health states usually starts with a card sorting exercise. Here the valuer works with a pack of health state cards. Each health state card describes a condition along six dimensions. The valuer is asked to order the cards from best health state in the pack to the worst health state in the pack. The data derived from this is a rank order within the respective set of health states. Following rank order was assigned to 11 health states including the valuers own health state, by a valuer. The valuer was randomly chosen from the AP Health State Valuation Study Data Set (Mahapatra and others, 1999).

Rank	Health State / Condition
1	Own Health Today
2	Mild Diabetes, no Symptoms
3	Watery Diarrhea 5 times a day
4	Mild Tuberculosis with Treatment
5	Below the Knee Amputation(one leg)
6	Peptic Ulcer
7	Below the Knee Amputation (two legs)
8	Two Broken Arms in Cast
9	Unipolar Major Depression
10	Severe Continuous Migraine
11	Quadriplegia

The interval scale represents a much higher level of measurement than the ordinal scale. It possesses the properties of magnitude and equal interval between adjacent units but does not have an absolute zero point. Thus, the interval scale possesses the properties of ordinal scale and, in addition, has equal intervals between adjacent units. Equal intervals between adjacent units means that there are equal amounts of the variable being measured between adjacent units on the scale (Robert Pagano, 1994, p24). Temperature recorded by the clinical thermometer is a good example of an interval scale. Clinical thermometers are either calibrated with a Fahrenheit scale or a Celsius scale. Both of these are interval scales. The intervals in the scale are equally placed. Although the scales have an interval labeled zero, they do not have a true zero point. A reading of 20^{o}C is not twice as hot as 10^{o}C. Interval scales allow arithmetic operations such as addition, subtraction, but do not allow for multiplication, or division. However, differences between two interval scale measurements can be treated as ratio data. For example the change in temperature measured with the Celsius or Fahrenheit scale can be divided by the change in temperature in another instance measured with the same scale.

Figure-2:
Interval-Scale Thermometer



The next and highest, level of measurement is called a ratio scale. It has all the properties of an interval scale and, in addition, has an absolute zero point. Without an absolute zero point, it is not legitimate to do ratios with the scale readings. Since the ratio scale has an absolute zero point, ratios are permissible (hence the name ratio scale). A good example to illustrate the difference between interval and ratio scales is to compare the Celsius scale of temperature with the Kelvin scale. Zero on the Kelvin scale is an absolute zero i.e. complete absence of heat. Zero on the Celsius scale is the temperature at which water freezes. It is an arbitrary zero point that actually occurs at 273° Kelvin. The Celsius scale is an interval scale and the Kelvin scale is a ratio scale. The difference in heat between 8° and 9° is the same as between 99° and 100° whether the scale is Celsius or Kelvin. However, we can not do ratios with the Celsius scale. A reading of 20° Celsius is really 273° + 20° = 293° Kelvin and 10° Celsius is 273° + 10° = 283° Kelvin. Clearly 293° Kelvin would not be twice as hot 283°. A reading of 20° Kelvin would be twice as hot as 10° Kelvin. Examples of ratio scale measurements are, time, length, weight, age, frequency counts, etc. Ratio scales allow for all arithmetic operations such as addition, subtraction, multiplication, and division.

Measurement instruments bearing an interval or ratio scales have a least count. The least count of an instrument is the smallest unit on the scale. Thus all measurements made on a continuous variable are approximate. For example some of the analogue bathroom weighing scales have a least count of 0.5 Kg. Suppose the exact weight of a person has a fraction of less than 0.5 Kg. The weighing machine can not distinguish this quantity. The observer will record the weight to the nearest 0.5 Kg. Hence the real limits of a continuous variable would be (the recorded measurement $\pm$ 0.5 $\times$ least count). Suppose a study records

10 people's body weight as 64.5 Kg. The real limit of these weights would be 64.25 to 64.75 Kg.

A Clinical Scale with Least Count of lbs or 0.5 Kg.

[1] These weighing scales are known under a variety of names such as; Fitness scales, Personal scales, Bathroom scales, Professional Dial Scales, etc.

Data:

Table-1: Different measurements on literacy giving different types of data.

Nominal	Ordinal	Interval
High School Pass or Not	Just Literate	Years of Schooling
	Primary	
Literate or Not	Secondary School	
	High School	
	College	
	Post Graduate	

Measurement on a variable gives data. The data can be nominal, ordinal, interval or ratio. A variable that is intrinsically nominal, such as Sex, can only be measured by a nominal scale i.e. a classification system. An ordinal variable when measured with help of a nominal

scale yields nominal data. It can yield ordinal data, if the measurement scale is ordinal. Similarly an interval type quantitative variable yields nominal data when measured with nominal scale, can yield both nominal and ordinal data when measured with an ordinal scale and similarly can give nominal, ordinal or interval data when measured with an interval scale. Operationally speaking, data is the starting point of statistical analyses. Naturally many statistical texts introduce the concept of qualitative, quantitative data and distinguish between nominal, ordinal, interval or ratio data and then proceed to illustrate various statistical operations.

Parameters:

A parameter quantifies a characteristic of the population. It is the true, but usually unknown, state of nature (i.e. the universe of study) about which we want to make an inference. Recall that population or the study universe can be real or conceptual. In fact, the study population or universe we deal with in inferential statistics are mostly conceptual rather than real. Appropriate conceptualization of the population is the fundamental research design act required before data can be analysed to draw any conclusion. Suppose we are interested to know about a population, we have that population physically available to us, and do not have any time or resource constraint to measure any characteristic of interest. Then we would not need to make any inference. We would simply measure the characteristic of interest and compute the parameters. But ideal measurement systems do not exist. Any measurement system will have some margin of error. So, at the least, we have to make some inference about the size of this error, to arrive at the exact parameter value. Since we can not know with certainty the size of the error, we can not compute with certainty the value of the parameter. The only opportunity available to us is to estimate the parameter value. More commonly, however, we conceive of study universes, all elements of which may not physically exist for measurement to happen. Suppose we are interested to assess the mortality risk to which infants born and brought up in a given village are exposed. Here the mortality risk is a state of nature. It is the variable to which people living in the village are exposed. We know that infant mortality risk is non negative non zero in any area. Theoretically, we can not have a situation where infant mortality risk can be said to be zero. Even if all infants actually born in the area survive until their first birth day, we can not say that infant mortality risk is zero. The risk of death is always there, howsoever small it may be! That is clearly our current understanding of the state of nature as far as biological life is concerned.

Table-2: Probability of observing no infant death during the course of a year in communities with different population size living under different levels of infant mortality risk (IMR).

Population Size	IMR-> Births	10	30	50	70
		Probability of at least one death in a year			
500	10	0.90	0.74	0.60	0.48
1000	20	0.82	0.54	0.36	0.23
2000	40	0.67	0.30	0.13	0.05
3000	60	0.55	0.16	0.05	0.01
4000	80	0.45	0.09	0.02	0.00
5000	100	0.37	0.05	0.01	0.00
10000	200	0.13	0.00	0.00	0.00
20000	400	0.02	0.00	0.00	0.00
30000	600	0.00	0.00	0.00	0.00
40000	800	0.00	0.00	0.00	0.00

[1] IMR is shown as number of deaths / 1000 child birth.
[2] Crude birth rate is assumed to remain constant at 20 births / 1000 population.
[3] The probabilities have been calculated using Lotus 123 @Binomial(Births,1,IMR,2)

The above table shows results from a thought experiment. It computes the probability of observing at least one infant death during the course of a year in a hypothetical community of different population size and experiencing different infant mortality risks. The infant mortality risk (IMR) is the population parameter. Four population parameter values are chosen. Remember that a population parameter is in truth a fixed value. It does not change since the population parameter is essentially a descriptive measure of the study population. The four population parameters imply four different populations. The table shows that for a village with 1000 population and truly experiencing infant mortality risk of 30 infant deaths / 1000 live births, the probability of observing a single infant death during the course of a year is less than half. Suppose we happen to study such a village, observe 20 child births here during the course of a year, and treat those child births as the population of our study. For any given year the chance that we will not encounter any infant death is a little more than half. Considering that chance is usually lumpy (Abelson, 1995 p17-38) one may not come across a single infant death for two to three years in a row. By the same logic, one may come across upto three infant deaths in a year given the same infant morality risk of 30 infant deaths / 1000 live births acting on the same village with 1000 persons. Of course, the probability of such an event would be very low, some where around 0.0183 implying that such an event may occur, on average, once in 55 years. In the first scenario, if we assume our study universe to consist of all infants born in this hypothetical village of 1000 people, we would conclude that the infants in this village are not exposed to any mortality risk. In the second scenario of 3 observed infant deaths, we might conclude that the infant mortality risk in this village is 150 infant deaths / 1000 live births! Note that we started with an assumption that the true IMR for this village to be 30 infant deaths / 1000 live births. Instead, visualizing a larger hypothetical population of which the 20 infants born in the village during a year is a sample, allows us to view the sample statistic to be an estimate the true infant mortality risk. Even then, the second scenario with three observed deaths will give us a point estimate of 150 infant deaths / 1000 live births. But this time, since we have assumed the 20 births to be a sample from a large number of potential births, we will calculate statistical confidence limits and give out a range of values that straddles the point estimate of 150. Also note that the population of the village is different from the study population. Here the study population consists of infants born or likely to be born in the village. The only link of the village

population to the study population is that the village population contribute to the sample by giving birth to infants. We are using the same word population to mean two different concepts. Firstly the village population means actual human beings living in a village. The study population here is conceived to consist of the infants actually born as well as the infants that could have born in the village. A less confusing term would be the study universe, which is synonymous with study population.

Parameters are usually represented by Greek alphabets. This is a convention. For example the Greek letter μ is traditionally used to denote arithmetic mean of a variable. The Greek letter σ is used to represent population variance of the variable. Europe's ancient philosophy developed in the Graeco-Roman world. Ancient Roman philosophy sprang from the Greek tradition. Most of the philosophical writings were in the Greek language. Of all the branches of human knowledge, philosophy involves a lot of reflection and thinking about the world. The tradition of using Greek alphabets to denote population parameters appears to reinforce the idea that they refer to state of nature, and the truth about a variable, that we may not ever know with complete certainty.

Sample:

A sample is a subset of the study universe.

Statistic:

A statistic is any function of the data. Data invariable come from samples. Thus statistic is a number calculated on sample data that quantifies a characteristic of the sample. A sample statistic gives information about a corresponding population parameter. For example, the sample mean for a sample of data would enable us to estimate the population mean.

References:

Abelson Robert P. Statistics as Principled Argument. Hillsdale, NJ, USA//Hove UK: Lawrence Erlbaum Associates, Publishers; 1995.

Agresti Alan. Categorical data analysis. New York: John Wiley and Sons; 1990.

Argyrous George. Statistics for Social and Health Research. With a Guide to SPSS. New Delhi, London, Thousand Oaks: SAGE Publications; 2000.

Bailey Kenneth D. Typologies and taxonomies. An introduction to classification techniques. Newburry Park//London//New Delhi: Sage; 1994.

Mahapatra Prasanta, Srilatha S., Sridhar P. A patient satisfaction survey in public hospitals. Journal. Academy of Hospital Administration 2001 Jul-2001 Dec 31;13(2):11-5.

Pagano Robert R. Understanding statistics in the behavioral sciences. Fourth edition. Minneapolis-St Paul: West publishing co.; 1994.

Pedhazur Elazar J.; Liora Pedhazur Schmelkin. Measurement, Design, and Analysis: An Integrated Approach. Hillsdale, New Jersey//Hove//London: Lawrence Erlbaum Associates; 1991.

Tiryakian E.A. Typologies. in: Sills D.L., Editor. International encyclopedia of the social sciences. Vol 16.1968. pp. 177-86.

Exercises

1. Identify the study population and sample in the following situations:
   i. Patient Satisfaction Survey: The AP Vaidya Vidhana Parishad (APVVP) manages a large network of public hospitals in Andhra Pradesh. A Patient Satisfaction Surveys (PSS) was conducted by the Institute in the year 1999. All of the 19 District Hospitals, and all of the six Area Hospitals were covered by the survey. The study team wanted to gather a sample of about 50 inpatients from each hospital. The survey team visited the hospital with short notice to the hospital authorities. All patients who had completed three days of stay in the hospital were listed. If the number exceeded 50 a simple random sample was taken to select 50 patients. If the list was equal to 50, all in the list were interviewed. If less than 50, the team recruited additional patients as they completed three day stay in the hospital. A patient satisfaction questionnaire was administered to the patient or an attendant.
   ii. Study on the Structure and Dynamics of Private Health Sector in Andhra Pradesh (SDPHAP): This study sought to understand the structure and dynamics of the private health sector in Andhra Pradesh, in order to provide insights for meaningful policy intervention. Three areas in Andhra Pradesh were selected for the study. The area around the state capital namely Hyderabad - Ranga Reddy districts was chosen purposively. In addition, Visakhapatnam district was chosen from the economically developed districts and Warangal was selected from the economically backward districts. Thus the sample areas are; (a) Hyderabad - Ranga Reddy, (b) Visakhapatnam, and (c) Warangal districts. Altogether 256 health care institutions were surveyed consisting of 150 HCIs in the private and non profit sector and 106 HCIs in the public sector. Three types of HCIs were studied, namely (a) large hospitals, (b) small hospitals and (c) clinics or primary health centres.
2. Following is an extract from the IHS Director's report to the eighth annual general meeting (2000-2001), on 06 December, 2001. Identify the descriptive and / or inferential role of the statistics cited by the Institute Director.

"People are gradually recognising the bibliographic niche being cultivated by the IHS library. Although our library is small, it has some collections in the area of health economics, health system research etc. not easily available elsewhere in Hyderabad. As of January 2000 we had 25 associate members. Currently we have 37 associate members including one life associate member. Most of these memberships are taken to access the Institutes library services. Currently the library services about 323 retrievals per month."

3. Indicate which of the following represent a variable and which a constant:
   i. The number of letters in the Devanagari script.
   ii. The number of letters in various languages of the world.
   iii. The number of hours in a day.
   iv. The time at which you eat dinner.
   v. The number of students admitted every year to the same academic program in which you are enrolled.
   vi. The amount of sleep you get each night.
   vii. Body weight of people in your class.
   viii. The number of playing cards in a pack.

4. Following is an extract of selected questions from the NFHS-2 Women's questionnaire[5]. NFHS-2 instructions about codes, if any, to be assigned to the responses are also given. Identify the type of variable (qualitative, quantitative, etc.) implied by respective questions.

No.	Questions and Filters	Coding Categories	
107	What is your marital status?	Currently Married	1
		Married but Gauna not performed	2
		Separated	3
		Deserted	4
		Divorced	5
		Widowed	6
		Never Married	7
119	Can you read and write?	Yes .....1 No 2	
201	Have you ever given birth?	Yes	1
		No	2
230	Are you pregnant now?	Yes	1
		No	2
		Unsure	8
231	How many months pregnant are you?	Months	
902	Respondent's Haemoglobin level G/DL		

5. Survey of water taps: A survey of water taps in a public building asks the following question for each tap. Identify the variable types and measurement scales.
   i. TapId: Alphanumeric identification given by you to uniquely identify each tap.
   ii. Location: Area where located. For example; Corridor, Patio, Gents Toilet -1, etc.
   iii. Date: Date of the survey.
   iv. Time: Time of the survey.
   v. Yield: Open the tap and observe for water flow and report; Dry, Trickle, or Flow.
   vi. Leaking: Applicable only if Yield # Dry. Open the tap and let water flow for a few seconds. Then close the tap. Observe if flow stops completely (Leaking = False) or there is some leak (Leaking = True).
   vii. Positioning: Consider the purpose for which the tap has been provided? Give your assessment about positioning of the tap from the perspective of the intended use. For example, a tap meant for hand wash should be is the tap positioning appropriate for its intended use? If yes, then Positioning = Y else, Positioning = N.

[5] The first National Family Health Survey (NFHS-1) in India was conducted during 1992-93. The second survey (NFHS-2) was conducted during 1998-99. The principal objective of NFHS is to provide state and national level estimates fertility, the practice of family planning, infant and child mortality, maternal and child health, and the utilization of health services provided to mothers and children. The NFHS in India is in many respects similar in scope to the Demographic and Health Surveys (DHS) elsewhere in the world. NFHS-2 India report has been published by International Institute for Population Sciences (IIPS); ORC Macro; Roy TK, et al. National Family Health Survey 1998-99 (NFHS-2). India. Mumbai (Bombay): International Institute for Population Sciences (IIPS); 2000 Oct.

viii. Design: Consider the purpose for which the tap has been provided? Give your assessment about design of the tap from the perspective of the intended use. For example, a tap in operation theatre area for hand wash and scrubbing should have an elbow operated lever to open and close the tap. A tap serving a sink should have a long enough neck, so that a person does not have to bend a lot while washing things in the sink. Thus ask; is the tap design appropriate for its intended use? If yes, then Design = Y else, Design = N.

ix. Tapwork: Assess if the tap needs any work. Choose one of the following: Repair, Replace, Reposition, Redesign, Reposition And Redesign, Connect (with water source), or None.

Basic Concepts of Statistics.

Prasanta Mahapatra[1]

Statistics:

The word statistics is derived from the Italian word *stato*, which means "state" and *statista* which means a state official. Statistics originally meant facts useful to the state and collected by the *statista.* Today, we view statistics as a science of information. It deals with collection, analysis and interpretation of data. Three key areas of statistical methods are; descriptive statistics, exploratory data analysis, and inference. All three are connected. Statistical inference requires descriptive statistics at least on some samples. Most formal statistical inference methods are preceded by exploratory data analysis, testing conjectures, searching for possible patterns in data, and working towards formulation of testable hypotheses. Statistical inference is based on the concept of probability. Plausibility of hypotheses are tested on the basis of information contained in a sample.

Study Population or Universe:

At the very basic level statistics allows for concise description of an entity, or a phenomenon. The entity that we seek to describe may be a single event, an object, a person, or a collection, respectively, of events, objects and persons. A set of statistics may describe many aspects and / or give us fuller description of an aspect of the object of study. Objects can be described with help of a language such as English, Hindi or Telugu. For example, when we write an essay on a subject, we essentially, are describing that subject. Statistics on the subject allows for clear and succinct description. Both clarity and efficiency are important here. A statistic enables us to describe an aspect of the object of study efficiently[2]. We do use statistics to describe properties of a single event, an object or a person. Most commonly, however, statistics is about a collection or a class of events, objects and persons. The collection or class of objects[3] that we seek to describe and understand with help of statistics is referred to as the *study population* or simply *population*. Some refer to it as the *study universe*, *universe of study* or simply the *universe*. Where a population is large or its boundaries are not easily defined, statistics on a sample may be feasible. The sample statistics describe the sample and usually allow us to draw some inferences about the population. Here the *population* described by the statistics is the sample. Depending on the nature of the sample and the type of statistics, we are able to draw some inferences about the study population. The dual use of the word population in such situations has the potential for confusion. Moreover, statistics, these days rarely limits itself to the descriptive role alone. Hence statisticians usually use population to mean the larger population or the universe about which inferences are made with help of sample statistics. Expressions like the *descriptive statistics of the sample* are used to connote the descriptive role of a set of sample statistics. A population described by a set of statistics has to be real. We may, however, be able to draw inferences about a conceptual population based on statistics from a sample conceivably from

---

[1] President, The Institute of Health Systems, HACA Bhavan, Hyderabad, AP 5004, India.

[2] Note, however, that statistics usually do not describe the object in its entirety. Many attributes and aspects of an entity or phenomenon are not easily captured by statistics. We have a range of language based descriptive tools such as the essay, the story, the poem, and the performing arts. However, these are not always adequate for the job at hand, leaving scope for creative appearance of new forms of description and communication!

[3] For sake of brevity, the word object(s) will be used from now on to include events, objects, persons as well as phenomena that constitutes the object of a study.

the said population. Thus from an inferential perspective, the population or the universe may be real or conceptual. Suppose, for example, we observe the effect of a treatment on a sample of patients and find it to work. We would conclude that the treatment would work on similar patients in future. Here the population about which we seek to draw inference is conceived to constitute currently encountered patients as well as similar patients we are likely to encounter in future. We visualize the currently encountered patients as a sample of the universe consisting of all similar patients likely to exist ever. In other words, the universe of all similar patients likely to exist ever is the conceptual population of which the currently encountered patients are a sample.

We use the word *subject*, a *case*, an *object* or an *observation* to denote individual elements of a population. A subject is the smallest or the basic unit of study, a single item, an event or a person. Subjects have properties. Some properties may be same for all subjects in the population. Suppose, for example, we are interested in the survival of all infants born in a given area. All infants born and likely to be born in the said area would constitute the study population. Every member of this population is surely a human being. Thus the species of all subjects in this population is a *constant*. But the infants borne in the area would vary with respect to many other property such as the mothers literacy, gestation period, parents economic status, place of delivery etc. The property or characteristic with respect to which subjects differ in some measurable way is called a *variable*. Where the study population is conceived as the state of the subjects over time, characteristics of the same subject that differ from time to time will also be a variable. Note that, *variable* refers to the property, characteristics or attribute of subjects. A variable is a condition or quality that can vary from one case to another. It is the characteristic being measured on a set of people, objects or events. Each member of this set may take on different values. The property or characteristic that does not vary between subjects constituting belonging to a study population is called a *constant* with respect to that population. The term *variable* is most commonly used in general statistics. Synonyms exist and may be used by specific branches of science more commonly. For example, evolutionary and systematic biologist use the term *character* instead of the term *variable*. Another synonym of the word *variable* is *attribute*, commonly used in psychology mostly to denote properties that are either there or not there. Note that, to qualify as a variable, the attribute or property under consideration must consist of at least two values, for example, male and female (Pedhazur and Schmelkin, 1991, p174). Studying males or females only converts the variable sex into a constant. Further, at any given time it must be possible to assign to each element in the population one and only one value on the variable under consideration. For example, at any given time, a person can be assigned one and only one value on height, weight, age, etc. The following three examples illustrate the concept of population, and variables.

1. An nutritional epidemiologist wants to assess the extent of malnutrition among the adult population of an area. (S)he collects height and weight measurements of all adults in the area to compute body mass index (BMI) which is an indicator of malnutrition among adults. Here, the universe to be described consist of all adults living in the area. They are the study population. Here the study population consist of persons. Each person contributing one unit to the study population. Variables are the height, weight and BMI of each adult in the area.
2. A District Health Authority wants to describe the type of cases doctors are likely to encounter during the course of a year in the out patient department health care institutions of the area. Such a description would be an useful training for newly recruited medical officers about to be posted in out patient departments. Here the

universe to be described is the set of all clinical encounters in the outpatient department. The population consists of events, namely the out patient visits. A person making three outpatient visits for the various reasons contributes three units to the study population. Another person making a single outpatient visit contributes one unit to the study population. Each out patient clinical encounter would have many aspects. For example, the age, sex of the patient, presenting symptom, season or month of the year in which presenting, Billing Category (Free, Insurance, Employer Reimbursement, Out-of-Pocket, etc.), Emergency or not, etc. These are the variables of the out patient visit event.

3. A hospital administrator wants to describe the state of all water taps in the hospital. Each water tap is assigned an unique identification number. The administrator wants to describe various characteristics of the water tap population in the hospital. For example, dry (does not yield water) or not (yields water), leaking-or-not, tap positioning is appropriate or not, tap design is appropriate or not, etc. Here all water taps in the hospital constitute the population or study universe. Each water tap is a study unit. Variables are; dry-or-not, leaking-or-not, etc.

Type of Variables:

Variables are primarily classified from the measurement perspective[4]. From the measurement perspective, we first distinguish between (a) qualitative or categorical variables, and (d) quantitative variables.

Variables that are expressed qualitatively by classification of subjects to a set of categories without any magnitude relationship between them, are called categorical variables, nominal variables or attributes. Classification of subjects into different categories is the foundation of categorical variables. There is no size relationship between different categories. Objects assigned to different categories differ in kind but not in degree. Hence categorical variables are also referred to as qualitative variables. Since the labels assigned to each category are essentially names, the categorical variables are also referred to as nominal variables. Note however, that a qualitative variable arises only if subjects can be classified into mutually exclusive and collectively exhaustive categories. Mutual exclusivity means that if a subject is assigned to one category, then it can not be assigned to another category. For example, the variable sex usually has two categories male and female. A person is either a male or a female. Collective exhaustiveness means that every subject can be assigned to some category. Examples of categorical variables are sex, blood group, disease entity, treatment options, causes of death, etc. Different system of blood groups are a good example of qualitative variables in medical sciences. According to the ABO system, a person's blood group may be A, B, AB or O. According to the Rh blood grouping system, a persons blood group may be Rh+ or Rh-. Antecedent cause of death is a another example of a categorical variable in health sciences. The antecedent cause of death varies from one case to another. No

[4] Pedhazur and Schmelkin (1991, p174-179) discuss about classification of variables from (a) the measurement perspective and (b) the research question or inferential perspective. From the inferential perspective they recognise dependent and independent variables. They do however, recognise that variables are not inherently dependent or independent. The same variable may be conceived of as independent in one study, or even in one phase of the same study, and as dependent in another study, or in another phase of the same study. According to Tiryakian (1968) a typology should explicitly identify the dimension(s) along which items are assigned to different types. Typology of variables from a measurement perspective meets this criteria. In contrast, classification of variables as independent or dependent helps describe parts of a specific cause and effect model but does not help differentiate the variables per se along any dimension. Hence we prefer to present the primary typology of variables form the measurement perspective.

cause of death is any way more or less in quantity than another cause of death. Here the study universe consists of all deaths experienced by a given a population over a certain period of time. The International Classification of Diseases (ICD) issued by the World Health Organization from time to time contains an exhaustive classification of causes of death. Although many factors may contribute to death, the current official convention is to assign a single category from out of the ICD. Thus mutual exclusivity is ensured by convention. Description of the type of cases doctors are likely to encounter in a health centre, mentioned earlier can use the ICD chapters on classification of clinical encounters. A special type of categorical variables is the logical variable or attributes. These variables refer to the existence or lack of a property. For example, a case has a disease or does not have it. A case received the treatment or did not receive the treatment. The Dry-or-Not, Leaking-or-Not status of water taps in the water tap survey described earlier are examples of logical variables. Another class of categorical variables are dichotomous variables having a maximum of two categories. Examples are; (a) Sex with Male and Female as the only two categories, (b) Outpatient, Inpatient; etc. Polytomous qualitative variables have more than two categories. List-1 General Mortality - Condensed list of the ICD-10 has 103 cause of death categories. Polytomous categorical variables can be dichotomized for purposes of measurement and analysis. For example, Malaria and Other causes of death. Such a classification would be justified in case of a study on the mortality attributable to Malaria. Similarly, ownership of health care institutions may be conceived as a dichotomous variable with public and private as the two categories or a polytomous categorical variable with three categories such as public, nonprofit and forprofit.

Type	Sub types	Examples
Qualitative, Categorical or Nominal		
	Logical, Indicator, or Dummy Variables	Diseased or Not.
		Water tap is Dry or Not.
	Categorical Variables	Presenting Symptom of Outpatients
		Billing Category
		Classification of Cause of Death
Quantitative		
	Rank Order	Order of Birth
	Discrete or Meristic	Children Ever Born.
	Continuous	Height, Weight

Quantitative variables refer to characteristics having a magnitude. The magnitude may be ordinal, interval or continuous. The most basic notion of magnitude is rank order. Here subjects can be ordered according to their rank. But the difference between two subjects with successive ranks may not be the same as the difference between another pair of subjects with successive ranks. For example, take the rank order of students on the basis of marks obtained in an examination. The difference between the first and second rank holder is not necessarily the same as the difference between say the 11th and 12th rank holder. Alan Agresti (1990) recognises that “the position of ordinal variables on the quantitative / qualitative classification is fuzzy. They are often treated as qualitative, being analyzed using methods for nominal variables.” But in many respects, Agresti (1990) opines, ordinal variables are closer to discrete or continuous variables than nominal variables. They possess important

quantitative information. Each level has a greater or smaller magnitude of the characteristic than another level.

Some variables assume only discrete values. For example number of children ever born to a mother can only be a whole number. Continuous variables may take on any value in an interval of numbers. For example, height, weight, blood pressure, quantity of water consumed in a health care organisation, etc. A discrete variable has a countable number of values. A continuous variables can vary in quantity by infinitesimally small degrees (Argyrous, 2000 p13). Some authors classify quantitative variables as interval variables and ratio variables. But this is primarily a measurement issue and hence discussed later. Quantitative variables are either, ordinal, discrete or continuous.

Measurement:

Recall that the definition of a variable refers to differences between subjects in some measurable way. Measurability is a key requirement for us to recognise a characteristic as a variable. Without measurability, we do not even know if the characteristic remains constant or varies between subject in the population. Unless we are able to measure a characteristics in some way, we can not distinguish it from others nor can we draw any inferences about the population with an ill defined characteristic. Here measurement is to be interpreted in its broadest sense and includes recognition of the existence of an attribute, classification of a quality etc. Any measurement systems would have at least three distinguishable components, namely; (a) a scale or an instrument, (b) an observer, and (c) a measurement protocol. We will use a simple example to illustrate these three components. Suppose we are asked to measure the body weight of a group of people. To do so you will need a weighing balance. Choice of the weighing balance is an important issue linked to desired level of accuracy and margin of error. Analogue bathroom weighing scales are used in most clinical and epidemiological survey settings. These scales may have least counts of upto 0.5 Kg and error margins upto one Kg. Depending on the research question at hand and availability of funds, you may choose more or less accurate weighing machines. You and your coworkers have to learn how to use the balance to measure body weight. You are the observer. For example, rotating dial type bath room weighing scales usually require that the weight be read vertically straight off the dial. Reading from a side would give erroneous readings. Thirdly, you need to follow a measurement protocol. Apparel and footwear add to the measured weight and may be a source of errors. So you may want to follow a protocol to record weight after shoes are taken-off. Thus many factors affect the result of measurement on a variable, such as designed accuracy of the instrument, observer training, measurement protocol. These are maters of fact to be ascertained and factored into statistical analysis of data. A key aspect of the measurement process is the type or level of scale used to measure a variable. The level of measurement scale used ultimately determines the type of data available for further analysis. Hence we will review this aspect of the measurement process in some detail. There are four different type of scales or levels of measurement such as; (a) the nominal, (b) the ordinal, (c) the interval, and (d) the ratio scale. We will examine each of them in some detail.

The essential measurement act for a nominal scale is classification of objects into different categories. Nominal scaling requires appropriately developed classification techniques. Classification involves the ordering of cases in terms of their similarity. The terms Typology (mostly used by social scientists) and Taxonomy (mostly used by biologists) are synonymous with classification. Tiryakian (1968) provides a brief but comprehensive treatment of typology as an analytical tool in social sciences. Bailey (1994) gives a more

detailed account of various classification techniques. Very briefly, any valid classification system must be exhaustive and mutually exclusive. As mentioned earlier mutual exclusivity means that a subject can not be assigned to more than one category. Exhaustiveness means that every subject must be assigned to some category. Where the interest is on a few of many possible categories, a residual category such as “miscellaneous” or “others” is created. Whatever the classification rules, objects classified in different categories are treated as different in kind, not in degree. In other words, classes of a nominal scale are not ordered (Pedhazur and Schmelkin, 1991, p19). Thus, measurement with a nominal scale really amounts to classifying the objects and giving them the name (hence “nominal” scale) of the category to which they belong (Robert Pagano, 1994, p23). A fundamental property of nominal scales is that of equivalence. This means that all members of a given class are the same from the standpoint of the classification variable. Some times categories in a classification system may be given numerical codes. Assignment of any numerical code to a set of categories is purely to facilitate data handling and should not confused with any sense of magnitude.

Figure-1: Five point ordered categories used in a patient satisfaction survey questionnaire.

Negatively framed question:

You are usually kept waiting for a long time when you need doctor's attention / consultation.

Choice ->	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Score ->	**1**	**2**	**3**	**4**	**5**

Positively framed question:

You have easy access to the medical specialists in the hospital (MPSQ25)

Choice ->	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Score ->	**5**	**4**	**3**	**2**	**1**

[1] Source: Mahapatra Prasanta; Srinivas Kallam. APVVP Patient Satisfaction Survey, December 2001. Hyderabad: Institute of Health Systems, RP21/2002.

An ordinal scale represents the next higher level of measurement. It possesses a relatively low level of the property of magnitude. With an ordinal scale we rank order the objects being measured according to whether they posses more, less or the same amount of the variable being measured (Robert Pagano, 1994, p24). An ordinal level of measurement, in addition to the function of classification, allows cases to be ordered by degree according to measurements of the variable (Argyrous, 2000, p11). Since both nominal and ordinal scales categorize cases, they are sometimes called categorical scales. Ordinal scales allow for ranking of subjects. Variables referring to human attitude towards particular issues can be viewed to fall into ordered categories. Attitudes are measured usually with the help of three, five or seven ordered categories. The following figures shows a five point ordered categories to ascertain patient experience about accessibility of doctor. These two questions are taken from a patient satisfaction survey instrument (Mahapatra, Srilatha and Sridhar, 2001). Note that category labels remain the same but ordinal rankings are reversed depending on the positive or negative frame of the question.

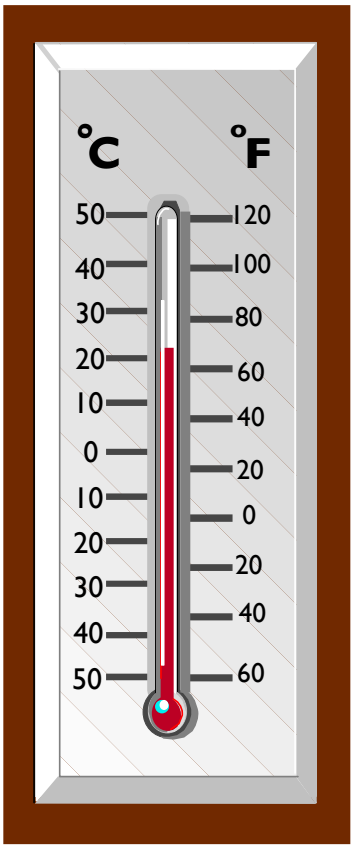
Health State Valuation by Card Sort

Community valuation of the relative severity of various health states usually starts with a card sorting exercise. Here the valuer works with a pack of health state cards. Each health state card describes a condition along six dimensions. The valuer is asked to order the cards from best health state in the pack to the worst health state in the pack. The data derived from this is a rank order within the respective set of health states. Following rank order was assigned to 11 health states including the valuers own health state, by a valuer. The valuer was randomly chosen from the AP Health State Valuation Study Data Set (Mahapatra and others, 1999).

Rank	Health State / Condition
1	Own Health Today
2	Mild Diabetes, no Symptoms
3	Watery Diarrhea 5 times a day
4	Mild Tuberculosis with Treatment
5	Below the Knee Amputation(one leg)
6	Peptic Ulcer
7	Below the Knee Amputation (two legs)
8	Two Broken Arms in Cast
9	Unipolar Major Depression
10	Severe Continuous Migraine
11	Quadriplegia

The interval scale represents a much higher level of measurement than the ordinal scale. It possesses the properties of magnitude and equal interval between adjacent units but does not have an absolute zero point. Thus, the interval scale possesses the properties of ordinal scale and, in addition, has equal intervals between adjacent units. Equal intervals between adjacent units means that there are equal amounts of the variable being measured between adjacent units on the scale (Robert Pagano, 1994, p24). Temperature recorded by the clinical thermometer is a good example of an interval scale. Clinical thermometers are either calibrated with a Fahrenheit scale or a Celsius scale. Both of these are interval scales. The intervals in the scale are equally placed. Although the scales have an interval labeled zero, they do not have a true zero point. A reading of 20^{o}C is not twice as hot as 10^{o}C. Interval scales allow arithmetic operations such as addition, subtraction, but do not allow for multiplication, or division. However, differences between two interval scale measurements can be treated as ratio data. For example the change in temperature measured with the Celsius or Fahrenheit scale can be divided by the change in temperature in another instance measured with the same scale.

Figure-2:
Interval-Scale Thermometer



The next and highest, level of measurement is called a ratio scale. It has all the properties of an interval scale and, in addition, has an absolute zero point. Without an absolute zero point, it is not legitimate to do ratios with the scale readings. Since the ratio scale has an absolute zero point, ratios are permissible (hence the name ratio scale). A good example to illustrate the difference between interval and ratio scales is to compare the Celsius scale of temperature with the Kelvin scale. Zero on the Kelvin scale is an absolute zero i.e. complete absence of heat. Zero on the Celsius scale is the temperature at which water freezes. It is an arbitrary zero point that actually occurs at 273° Kelvin. The Celsius scale is an interval scale and the Kelvin scale is a ratio scale. The difference in heat between 8° and 9° is the same as between 99° and 100° whether the scale is Celsius or Kelvin. However, we can not do ratios with the Celsius scale. A reading of 20° Celsius is really 273° + 20° = 293° Kelvin and 10° Celsius is 273° + 10° = 283° Kelvin. Clearly 293° Kelvin would not be twice as hot 283°. A reading of 20° Kelvin would be twice as hot as 10° Kelvin. Examples of ratio scale measurements are, time, length, weight, age, frequency counts, etc. Ratio scales allow for all arithmetic operations such as addition, subtraction, multiplication, and division.

Measurement instruments bearing an interval or ratio scales have a least count. The least count of an instrument is the smallest unit on the scale. Thus all measurements made on a continuous variable are approximate. For example some of the analogue bathroom weighing scales have a least count of 0.5 Kg. Suppose the exact weight of a person has a fraction of less than 0.5 Kg. The weighing machine can not distinguish this quantity. The observer will record the weight to the nearest 0.5 Kg. Hence the real limits of a continuous variable would be (the recorded measurement $\pm$ 0.5 $\times$ least count). Suppose a study records

10 people's body weight as 64.5 Kg. The real limit of these weights would be 64.25 to 64.75 Kg.

A Clinical Scale with Least Count of lbs or 0.5 Kg.

[1] These weighing scales are known under a variety of names such as; Fitness scales, Personal scales, Bathroom scales, Professional Dial Scales, etc.

Data:

Table-1: Different measurements on literacy giving different types of data.

Nominal	Ordinal	Interval
High School Pass or Not	Just Literate	Years of Schooling
	Primary	
Literate or Not	Secondary School	
	High School	
	College	
	Post Graduate	

Measurement on a variable gives data. The data can be nominal, ordinal, interval or ratio. A variable that is intrinsically nominal, such as Sex, can only be measured by a nominal scale i.e. a classification system. An ordinal variable when measured with help of a nominal

scale yields nominal data. It can yield ordinal data, if the measurement scale is ordinal. Similarly an interval type quantitative variable yields nominal data when measured with nominal scale, can yield both nominal and ordinal data when measured with an ordinal scale and similarly can give nominal, ordinal or interval data when measured with an interval scale. Operationally speaking, data is the starting point of statistical analyses. Naturally many statistical texts introduce the concept of qualitative, quantitative data and distinguish between nominal, ordinal, interval or ratio data and then proceed to illustrate various statistical operations.

Parameters:

A parameter quantifies a characteristic of the population. It is the true, but usually unknown, state of nature (i.e. the universe of study) about which we want to make an inference. Recall that population or the study universe can be real or conceptual. In fact, the study population or universe we deal with in inferential statistics are mostly conceptual rather than real. Appropriate conceptualization of the population is the fundamental research design act required before data can be analysed to draw any conclusion. Suppose we are interested to know about a population, we have that population physically available to us, and do not have any time or resource constraint to measure any characteristic of interest. Then we would not need to make any inference. We would simply measure the characteristic of interest and compute the parameters. But ideal measurement systems do not exist. Any measurement system will have some margin of error. So, at the least, we have to make some inference about the size of this error, to arrive at the exact parameter value. Since we can not know with certainty the size of the error, we can not compute with certainty the value of the parameter. The only opportunity available to us is to estimate the parameter value. More commonly, however, we conceive of study universes, all elements of which may not physically exist for measurement to happen. Suppose we are interested to assess the mortality risk to which infants born and brought up in a given village are exposed. Here the mortality risk is a state of nature. It is the variable to which people living in the village are exposed. We know that infant mortality risk is non negative non zero in any area. Theoretically, we can not have a situation where infant mortality risk can be said to be zero. Even if all infants actually born in the area survive until their first birth day, we can not say that infant mortality risk is zero. The risk of death is always there, howsoever small it may be! That is clearly our current understanding of the state of nature as far as biological life is concerned.

Table-2: Probability of observing no infant death during the course of a year in communities with different population size living under different levels of infant mortality risk (IMR).

Population Size	IMR-> Births	10	30	50	70
		Probability of at least one death in a year			
500	10	0.90	0.74	0.60	0.48
1000	20	0.82	0.54	0.36	0.23
2000	40	0.67	0.30	0.13	0.05
3000	60	0.55	0.16	0.05	0.01
4000	80	0.45	0.09	0.02	0.00
5000	100	0.37	0.05	0.01	0.00
10000	200	0.13	0.00	0.00	0.00
20000	400	0.02	0.00	0.00	0.00
30000	600	0.00	0.00	0.00	0.00
40000	800	0.00	0.00	0.00	0.00

[1] IMR is shown as number of deaths / 1000 child birth.
[2] Crude birth rate is assumed to remain constant at 20 births / 1000 population.
[3] The probabilities have been calculated using Lotus 123 @Binomial(Births,1,IMR,2)

The above table shows results from a thought experiment. It computes the probability of observing at least one infant death during the course of a year in a hypothetical community of different population size and experiencing different infant mortality risks. The infant mortality risk (IMR) is the population parameter. Four population parameter values are chosen. Remember that a population parameter is in truth a fixed value. It does not change since the population parameter is essentially a descriptive measure of the study population. The four population parameters imply four different populations. The table shows that for a village with 1000 population and truly experiencing infant mortality risk of 30 infant deaths / 1000 live births, the probability of observing a single infant death during the course of a year is less than half. Suppose we happen to study such a village, observe 20 child births here during the course of a year, and treat those child births as the population of our study. For any given year the chance that we will not encounter any infant death is a little more than half. Considering that chance is usually lumpy (Abelson, 1995 p17-38) one may not come across a single infant death for two to three years in a row. By the same logic, one may come across upto three infant deaths in a year given the same infant morality risk of 30 infant deaths / 1000 live births acting on the same village with 1000 persons. Of course, the probability of such an event would be very low, some where around 0.0183 implying that such an event may occur, on average, once in 55 years. In the first scenario, if we assume our study universe to consist of all infants born in this hypothetical village of 1000 people, we would conclude that the infants in this village are not exposed to any mortality risk. In the second scenario of 3 observed infant deaths, we might conclude that the infant mortality risk in this village is 150 infant deaths / 1000 live births! Note that we started with an assumption that the true IMR for this village to be 30 infant deaths / 1000 live births. Instead, visualizing a larger hypothetical population of which the 20 infants born in the village during a year is a sample, allows us to view the sample statistic to be an estimate the true infant mortality risk. Even then, the second scenario with three observed deaths will give us a point estimate of 150 infant deaths / 1000 live births. But this time, since we have assumed the 20 births to be a sample from a large number of potential births, we will calculate statistical confidence limits and give out a range of values that straddles the point estimate of 150. Also note that the population of the village is different from the study population. Here the study population consists of infants born or likely to be born in the village. The only link of the village

population to the study population is that the village population contribute to the sample by giving birth to infants. We are using the same word population to mean two different concepts. Firstly the village population means actual human beings living in a village. The study population here is conceived to consist of the infants actually born as well as the infants that could have born in the village. A less confusing term would be the study universe, which is synonymous with study population.

Parameters are usually represented by Greek alphabets. This is a convention. For example the Greek letter μ is traditionally used to denote arithmetic mean of a variable. The Greek letter σ is used to represent population variance of the variable. Europe's ancient philosophy developed in the Graeco-Roman world. Ancient Roman philosophy sprang from the Greek tradition. Most of the philosophical writings were in the Greek language. Of all the branches of human knowledge, philosophy involves a lot of reflection and thinking about the world. The tradition of using Greek alphabets to denote population parameters appears to reinforce the idea that they refer to state of nature, and the truth about a variable, that we may not ever know with complete certainty.

Sample:

A sample is a subset of the study universe.

Statistic:

A statistic is any function of the data. Data invariable come from samples. Thus statistic is a number calculated on sample data that quantifies a characteristic of the sample. A sample statistic gives information about a corresponding population parameter. For example, the sample mean for a sample of data would enable us to estimate the population mean.

References:

Abelson Robert P. Statistics as Principled Argument. Hillsdale, NJ, USA//Hove UK: Lawrence Erlbaum Associates, Publishers; 1995.

Agresti Alan. Categorical data analysis. New York: John Wiley and Sons; 1990.

Argyrous George. Statistics for Social and Health Research. With a Guide to SPSS. New Delhi, London, Thousand Oaks: SAGE Publications; 2000.

Bailey Kenneth D. Typologies and taxonomies. An introduction to classification techniques. Newburry Park//London//New Delhi: Sage; 1994.

Mahapatra Prasanta, Srilatha S., Sridhar P. A patient satisfaction survey in public hospitals. Journal. Academy of Hospital Administration 2001 Jul-2001 Dec 31;13(2):11-5.

Pagano Robert R. Understanding statistics in the behavioral sciences. Fourth edition. Minneapolis-St Paul: West publishing co.; 1994.

Pedhazur Elazar J.; Liora Pedhazur Schmelkin. Measurement, Design, and Analysis: An Integrated Approach. Hillsdale, New Jersey//Hove//London: Lawrence Erlbaum Associates; 1991.

Tiryakian E.A. Typologies. in: Sills D.L., Editor. International encyclopedia of the social sciences. Vol 16.1968. pp. 177-86.

Exercises

1. Identify the study population and sample in the following situations:
   i. Patient Satisfaction Survey: The AP Vaidya Vidhana Parishad (APVVP) manages a large network of public hospitals in Andhra Pradesh. A Patient Satisfaction Surveys (PSS) was conducted by the Institute in the year 1999. All of the 19 District Hospitals, and all of the six Area Hospitals were covered by the survey. The study team wanted to gather a sample of about 50 inpatients from each hospital. The survey team visited the hospital with short notice to the hospital authorities. All patients who had completed three days of stay in the hospital were listed. If the number exceeded 50 a simple random sample was taken to select 50 patients. If the list was equal to 50, all in the list were interviewed. If less than 50, the team recruited additional patients as they completed three day stay in the hospital. A patient satisfaction questionnaire was administered to the patient or an attendant.
   ii. Study on the Structure and Dynamics of Private Health Sector in Andhra Pradesh (SDPHAP): This study sought to understand the structure and dynamics of the private health sector in Andhra Pradesh, in order to provide insights for meaningful policy intervention. Three areas in Andhra Pradesh were selected for the study. The area around the state capital namely Hyderabad - Ranga Reddy districts was chosen purposively. In addition, Visakhapatnam district was chosen from the economically developed districts and Warangal was selected from the economically backward districts. Thus the sample areas are; (a) Hyderabad - Ranga Reddy, (b) Visakhapatnam, and (c) Warangal districts. Altogether 256 health care institutions were surveyed consisting of 150 HCIs in the private and non profit sector and 106 HCIs in the public sector. Three types of HCIs were studied, namely (a) large hospitals, (b) small hospitals and (c) clinics or primary health centres.
2. Following is an extract from the IHS Director's report to the eighth annual general meeting (2000-2001), on 06 December, 2001. Identify the descriptive and / or inferential role of the statistics cited by the Institute Director.

"People are gradually recognising the bibliographic niche being cultivated by the IHS library. Although our library is small, it has some collections in the area of health economics, health system research etc. not easily available elsewhere in Hyderabad. As of January 2000 we had 25 associate members. Currently we have 37 associate members including one life associate member. Most of these memberships are taken to access the Institutes library services. Currently the library services about 323 retrievals per month."

3. Indicate which of the following represent a variable and which a constant:
   i. The number of letters in the Devanagari script.
   ii. The number of letters in various languages of the world.
   iii. The number of hours in a day.
   iv. The time at which you eat dinner.
   v. The number of students admitted every year to the same academic program in which you are enrolled.
   vi. The amount of sleep you get each night.
   vii. Body weight of people in your class.
   viii. The number of playing cards in a pack.

4. Following is an extract of selected questions from the NFHS-2 Women's questionnaire[5]. NFHS-2 instructions about codes, if any, to be assigned to the responses are also given. Identify the type of variable (qualitative, quantitative, etc.) implied by respective questions.

No.	Questions and Filters	Coding Categories	
107	What is your marital status?	Currently Married	1
		Married but Gauna not performed	2
		Separated	3
		Deserted	4
		Divorced	5
		Widowed	6
		Never Married	7
119	Can you read and write?	Yes .....1 No 2	
201	Have you ever given birth?	Yes	1
		No	2
230	Are you pregnant now?	Yes	1
		No	2
		Unsure	8
231	How many months pregnant are you?	Months	
902	Respondent's Haemoglobin level G/DL		

5. Survey of water taps: A survey of water taps in a public building asks the following question for each tap. Identify the variable types and measurement scales.
   i. TapId: Alphanumeric identification given by you to uniquely identify each tap.
   ii. Location: Area where located. For example; Corridor, Patio, Gents Toilet -1, etc.
   iii. Date: Date of the survey.
   iv. Time: Time of the survey.
   v. Yield: Open the tap and observe for water flow and report; Dry, Trickle, or Flow.
   vi. Leaking: Applicable only if Yield # Dry. Open the tap and let water flow for a few seconds. Then close the tap. Observe if flow stops completely (Leaking = False) or there is some leak (Leaking = True).
   vii. Positioning: Consider the purpose for which the tap has been provided? Give your assessment about positioning of the tap from the perspective of the intended use. For example, a tap meant for hand wash should be is the tap positioning appropriate for its intended use? If yes, then Positioning = Y else, Positioning = N.

[5] The first National Family Health Survey (NFHS-1) in India was conducted during 1992-93. The second survey (NFHS-2) was conducted during 1998-99. The principal objective of NFHS is to provide state and national level estimates fertility, the practice of family planning, infant and child mortality, maternal and child health, and the utilization of health services provided to mothers and children. The NFHS in India is in many respects similar in scope to the Demographic and Health Surveys (DHS) elsewhere in the world. NFHS-2 India report has been published by International Institute for Population Sciences (IIPS); ORC Macro; Roy TK, et al. National Family Health Survey 1998-99 (NFHS-2). India. Mumbai (Bombay): International Institute for Population Sciences (IIPS); 2000 Oct.

viii. Design: Consider the purpose for which the tap has been provided? Give your assessment about design of the tap from the perspective of the intended use. For example, a tap in operation theatre area for hand wash and scrubbing should have an elbow operated lever to open and close the tap. A tap serving a sink should have a long enough neck, so that a person does not have to bend a lot while washing things in the sink. Thus ask; is the tap design appropriate for its intended use? If yes, then Design = Y else, Design = N.

ix. Tapwork: Assess if the tap needs any work. Choose one of the following: Repair, Replace, Reposition, Redesign, Reposition And Redesign, Connect (with water source), or None.

Basic Concepts of Statistics.

Prasanta Mahapatra[1]

Statistics:

The word statistics is derived from the Italian word *stato*, which means "state" and *statista* which means a state official. Statistics originally meant facts useful to the state and collected by the *statista.* Today, we view statistics as a science of information. It deals with collection, analysis and interpretation of data. Three key areas of statistical methods are; descriptive statistics, exploratory data analysis, and inference. All three are connected. Statistical inference requires descriptive statistics at least on some samples. Most formal statistical inference methods are preceded by exploratory data analysis, testing conjectures, searching for possible patterns in data, and working towards formulation of testable hypotheses. Statistical inference is based on the concept of probability. Plausibility of hypotheses are tested on the basis of information contained in a sample.

Study Population or Universe:

At the very basic level statistics allows for concise description of an entity, or a phenomenon. The entity that we seek to describe may be a single event, an object, a person, or a collection, respectively, of events, objects and persons. A set of statistics may describe many aspects and / or give us fuller description of an aspect of the object of study. Objects can be described with help of a language such as English, Hindi or Telugu. For example, when we write an essay on a subject, we essentially, are describing that subject. Statistics on the subject allows for clear and succinct description. Both clarity and efficiency are important here. A statistic enables us to describe an aspect of the object of study efficiently[2]. We do use statistics to describe properties of a single event, an object or a person. Most commonly, however, statistics is about a collection or a class of events, objects and persons. The collection or class of objects[3] that we seek to describe and understand with help of statistics is referred to as the *study population* or simply *population*. Some refer to it as the *study universe*, *universe of study* or simply the *universe*. Where a population is large or its boundaries are not easily defined, statistics on a sample may be feasible. The sample statistics describe the sample and usually allow us to draw some inferences about the population. Here the *population* described by the statistics is the sample. Depending on the nature of the sample and the type of statistics, we are able to draw some inferences about the study population. The dual use of the word population in such situations has the potential for confusion. Moreover, statistics, these days rarely limits itself to the descriptive role alone. Hence statisticians usually use population to mean the larger population or the universe about which inferences are made with help of sample statistics. Expressions like the *descriptive statistics of the sample* are used to connote the descriptive role of a set of sample statistics. A population described by a set of statistics has to be real. We may, however, be able to draw inferences about a conceptual population based on statistics from a sample conceivably from

---

[1] President, The Institute of Health Systems, HACA Bhavan, Hyderabad, AP 5004, India.

[2] Note, however, that statistics usually do not describe the object in its entirety. Many attributes and aspects of an entity or phenomenon are not easily captured by statistics. We have a range of language based descriptive tools such as the essay, the story, the poem, and the performing arts. However, these are not always adequate for the job at hand, leaving scope for creative appearance of new forms of description and communication!

[3] For sake of brevity, the word object(s) will be used from now on to include events, objects, persons as well as phenomena that constitutes the object of a study.

the said population. Thus from an inferential perspective, the population or the universe may be real or conceptual. Suppose, for example, we observe the effect of a treatment on a sample of patients and find it to work. We would conclude that the treatment would work on similar patients in future. Here the population about which we seek to draw inference is conceived to constitute currently encountered patients as well as similar patients we are likely to encounter in future. We visualize the currently encountered patients as a sample of the universe consisting of all similar patients likely to exist ever. In other words, the universe of all similar patients likely to exist ever is the conceptual population of which the currently encountered patients are a sample.

We use the word *subject*, a *case*, an *object* or an *observation* to denote individual elements of a population. A subject is the smallest or the basic unit of study, a single item, an event or a person. Subjects have properties. Some properties may be same for all subjects in the population. Suppose, for example, we are interested in the survival of all infants born in a given area. All infants born and likely to be born in the said area would constitute the study population. Every member of this population is surely a human being. Thus the species of all subjects in this population is a *constant*. But the infants borne in the area would vary with respect to many other property such as the mothers literacy, gestation period, parents economic status, place of delivery etc. The property or characteristic with respect to which subjects differ in some measurable way is called a *variable*. Where the study population is conceived as the state of the subjects over time, characteristics of the same subject that differ from time to time will also be a variable. Note that, *variable* refers to the property, characteristics or attribute of subjects. A variable is a condition or quality that can vary from one case to another. It is the characteristic being measured on a set of people, objects or events. Each member of this set may take on different values. The property or characteristic that does not vary between subjects constituting belonging to a study population is called a *constant* with respect to that population. The term *variable* is most commonly used in general statistics. Synonyms exist and may be used by specific branches of science more commonly. For example, evolutionary and systematic biologist use the term *character* instead of the term *variable*. Another synonym of the word *variable* is *attribute*, commonly used in psychology mostly to denote properties that are either there or not there. Note that, to qualify as a variable, the attribute or property under consideration must consist of at least two values, for example, male and female (Pedhazur and Schmelkin, 1991, p174). Studying males or females only converts the variable sex into a constant. Further, at any given time it must be possible to assign to each element in the population one and only one value on the variable under consideration. For example, at any given time, a person can be assigned one and only one value on height, weight, age, etc. The following three examples illustrate the concept of population, and variables.

1. An nutritional epidemiologist wants to assess the extent of malnutrition among the adult population of an area. (S)he collects height and weight measurements of all adults in the area to compute body mass index (BMI) which is an indicator of malnutrition among adults. Here, the universe to be described consist of all adults living in the area. They are the study population. Here the study population consist of persons. Each person contributing one unit to the study population. Variables are the height, weight and BMI of each adult in the area.
2. A District Health Authority wants to describe the type of cases doctors are likely to encounter during the course of a year in the out patient department health care institutions of the area. Such a description would be an useful training for newly recruited medical officers about to be posted in out patient departments. Here the

universe to be described is the set of all clinical encounters in the outpatient department. The population consists of events, namely the out patient visits. A person making three outpatient visits for the various reasons contributes three units to the study population. Another person making a single outpatient visit contributes one unit to the study population. Each out patient clinical encounter would have many aspects. For example, the age, sex of the patient, presenting symptom, season or month of the year in which presenting, Billing Category (Free, Insurance, Employer Reimbursement, Out-of-Pocket, etc.), Emergency or not, etc. These are the variables of the out patient visit event.

3. A hospital administrator wants to describe the state of all water taps in the hospital. Each water tap is assigned an unique identification number. The administrator wants to describe various characteristics of the water tap population in the hospital. For example, dry (does not yield water) or not (yields water), leaking-or-not, tap positioning is appropriate or not, tap design is appropriate or not, etc. Here all water taps in the hospital constitute the population or study universe. Each water tap is a study unit. Variables are; dry-or-not, leaking-or-not, etc.

Type of Variables:

Variables are primarily classified from the measurement perspective[4]. From the measurement perspective, we first distinguish between (a) qualitative or categorical variables, and (d) quantitative variables.

Variables that are expressed qualitatively by classification of subjects to a set of categories without any magnitude relationship between them, are called categorical variables, nominal variables or attributes. Classification of subjects into different categories is the foundation of categorical variables. There is no size relationship between different categories. Objects assigned to different categories differ in kind but not in degree. Hence categorical variables are also referred to as qualitative variables. Since the labels assigned to each category are essentially names, the categorical variables are also referred to as nominal variables. Note however, that a qualitative variable arises only if subjects can be classified into mutually exclusive and collectively exhaustive categories. Mutual exclusivity means that if a subject is assigned to one category, then it can not be assigned to another category. For example, the variable sex usually has two categories male and female. A person is either a male or a female. Collective exhaustiveness means that every subject can be assigned to some category. Examples of categorical variables are sex, blood group, disease entity, treatment options, causes of death, etc. Different system of blood groups are a good example of qualitative variables in medical sciences. According to the ABO system, a person's blood group may be A, B, AB or O. According to the Rh blood grouping system, a persons blood group may be Rh+ or Rh-. Antecedent cause of death is a another example of a categorical variable in health sciences. The antecedent cause of death varies from one case to another. No

[4] Pedhazur and Schmelkin (1991, p174-179) discuss about classification of variables from (a) the measurement perspective and (b) the research question or inferential perspective. From the inferential perspective they recognise dependent and independent variables. They do however, recognise that variables are not inherently dependent or independent. The same variable may be conceived of as independent in one study, or even in one phase of the same study, and as dependent in another study, or in another phase of the same study. According to Tiryakian (1968) a typology should explicitly identify the dimension(s) along which items are assigned to different types. Typology of variables from a measurement perspective meets this criteria. In contrast, classification of variables as independent or dependent helps describe parts of a specific cause and effect model but does not help differentiate the variables per se along any dimension. Hence we prefer to present the primary typology of variables form the measurement perspective.

cause of death is any way more or less in quantity than another cause of death. Here the study universe consists of all deaths experienced by a given a population over a certain period of time. The International Classification of Diseases (ICD) issued by the World Health Organization from time to time contains an exhaustive classification of causes of death. Although many factors may contribute to death, the current official convention is to assign a single category from out of the ICD. Thus mutual exclusivity is ensured by convention. Description of the type of cases doctors are likely to encounter in a health centre, mentioned earlier can use the ICD chapters on classification of clinical encounters. A special type of categorical variables is the logical variable or attributes. These variables refer to the existence or lack of a property. For example, a case has a disease or does not have it. A case received the treatment or did not receive the treatment. The Dry-or-Not, Leaking-or-Not status of water taps in the water tap survey described earlier are examples of logical variables. Another class of categorical variables are dichotomous variables having a maximum of two categories. Examples are; (a) Sex with Male and Female as the only two categories, (b) Outpatient, Inpatient; etc. Polytomous qualitative variables have more than two categories. List-1 General Mortality - Condensed list of the ICD-10 has 103 cause of death categories. Polytomous categorical variables can be dichotomized for purposes of measurement and analysis. For example, Malaria and Other causes of death. Such a classification would be justified in case of a study on the mortality attributable to Malaria. Similarly, ownership of health care institutions may be conceived as a dichotomous variable with public and private as the two categories or a polytomous categorical variable with three categories such as public, nonprofit and forprofit.

Type	Sub types	Examples
Qualitative, Categorical or Nominal		
	Logical, Indicator, or Dummy Variables	Diseased or Not.
		Water tap is Dry or Not.
	Categorical Variables	Presenting Symptom of Outpatients
		Billing Category
		Classification of Cause of Death
Quantitative		
	Rank Order	Order of Birth
	Discrete or Meristic	Children Ever Born.
	Continuous	Height, Weight

Quantitative variables refer to characteristics having a magnitude. The magnitude may be ordinal, interval or continuous. The most basic notion of magnitude is rank order. Here subjects can be ordered according to their rank. But the difference between two subjects with successive ranks may not be the same as the difference between another pair of subjects with successive ranks. For example, take the rank order of students on the basis of marks obtained in an examination. The difference between the first and second rank holder is not necessarily the same as the difference between say the 11th and 12th rank holder. Alan Agresti (1990) recognises that “the position of ordinal variables on the quantitative / qualitative classification is fuzzy. They are often treated as qualitative, being analyzed using methods for nominal variables.” But in many respects, Agresti (1990) opines, ordinal variables are closer to discrete or continuous variables than nominal variables. They possess important

quantitative information. Each level has a greater or smaller magnitude of the characteristic than another level.

Some variables assume only discrete values. For example number of children ever born to a mother can only be a whole number. Continuous variables may take on any value in an interval of numbers. For example, height, weight, blood pressure, quantity of water consumed in a health care organisation, etc. A discrete variable has a countable number of values. A continuous variables can vary in quantity by infinitesimally small degrees (Argyrous, 2000 p13). Some authors classify quantitative variables as interval variables and ratio variables. But this is primarily a measurement issue and hence discussed later. Quantitative variables are either, ordinal, discrete or continuous.

Measurement:

Recall that the definition of a variable refers to differences between subjects in some measurable way. Measurability is a key requirement for us to recognise a characteristic as a variable. Without measurability, we do not even know if the characteristic remains constant or varies between subject in the population. Unless we are able to measure a characteristics in some way, we can not distinguish it from others nor can we draw any inferences about the population with an ill defined characteristic. Here measurement is to be interpreted in its broadest sense and includes recognition of the existence of an attribute, classification of a quality etc. Any measurement systems would have at least three distinguishable components, namely; (a) a scale or an instrument, (b) an observer, and (c) a measurement protocol. We will use a simple example to illustrate these three components. Suppose we are asked to measure the body weight of a group of people. To do so you will need a weighing balance. Choice of the weighing balance is an important issue linked to desired level of accuracy and margin of error. Analogue bathroom weighing scales are used in most clinical and epidemiological survey settings. These scales may have least counts of upto 0.5 Kg and error margins upto one Kg. Depending on the research question at hand and availability of funds, you may choose more or less accurate weighing machines. You and your coworkers have to learn how to use the balance to measure body weight. You are the observer. For example, rotating dial type bath room weighing scales usually require that the weight be read vertically straight off the dial. Reading from a side would give erroneous readings. Thirdly, you need to follow a measurement protocol. Apparel and footwear add to the measured weight and may be a source of errors. So you may want to follow a protocol to record weight after shoes are taken-off. Thus many factors affect the result of measurement on a variable, such as designed accuracy of the instrument, observer training, measurement protocol. These are maters of fact to be ascertained and factored into statistical analysis of data. A key aspect of the measurement process is the type or level of scale used to measure a variable. The level of measurement scale used ultimately determines the type of data available for further analysis. Hence we will review this aspect of the measurement process in some detail. There are four different type of scales or levels of measurement such as; (a) the nominal, (b) the ordinal, (c) the interval, and (d) the ratio scale. We will examine each of them in some detail.

The essential measurement act for a nominal scale is classification of objects into different categories. Nominal scaling requires appropriately developed classification techniques. Classification involves the ordering of cases in terms of their similarity. The terms Typology (mostly used by social scientists) and Taxonomy (mostly used by biologists) are synonymous with classification. Tiryakian (1968) provides a brief but comprehensive treatment of typology as an analytical tool in social sciences. Bailey (1994) gives a more

detailed account of various classification techniques. Very briefly, any valid classification system must be exhaustive and mutually exclusive. As mentioned earlier mutual exclusivity means that a subject can not be assigned to more than one category. Exhaustiveness means that every subject must be assigned to some category. Where the interest is on a few of many possible categories, a residual category such as “miscellaneous” or “others” is created. Whatever the classification rules, objects classified in different categories are treated as different in kind, not in degree. In other words, classes of a nominal scale are not ordered (Pedhazur and Schmelkin, 1991, p19). Thus, measurement with a nominal scale really amounts to classifying the objects and giving them the name (hence “nominal” scale) of the category to which they belong (Robert Pagano, 1994, p23). A fundamental property of nominal scales is that of equivalence. This means that all members of a given class are the same from the standpoint of the classification variable. Some times categories in a classification system may be given numerical codes. Assignment of any numerical code to a set of categories is purely to facilitate data handling and should not confused with any sense of magnitude.

Figure-1: Five point ordered categories used in a patient satisfaction survey questionnaire.

Negatively framed question:

You are usually kept waiting for a long time when you need doctor's attention / consultation.

Choice ->	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Score ->	**1**	**2**	**3**	**4**	**5**

Positively framed question:

You have easy access to the medical specialists in the hospital (MPSQ25)

Choice ->	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Score ->	**5**	**4**	**3**	**2**	**1**

[1] Source: Mahapatra Prasanta; Srinivas Kallam. APVVP Patient Satisfaction Survey, December 2001. Hyderabad: Institute of Health Systems, RP21/2002.

An ordinal scale represents the next higher level of measurement. It possesses a relatively low level of the property of magnitude. With an ordinal scale we rank order the objects being measured according to whether they posses more, less or the same amount of the variable being measured (Robert Pagano, 1994, p24). An ordinal level of measurement, in addition to the function of classification, allows cases to be ordered by degree according to measurements of the variable (Argyrous, 2000, p11). Since both nominal and ordinal scales categorize cases, they are sometimes called categorical scales. Ordinal scales allow for ranking of subjects. Variables referring to human attitude towards particular issues can be viewed to fall into ordered categories. Attitudes are measured usually with the help of three, five or seven ordered categories. The following figures shows a five point ordered categories to ascertain patient experience about accessibility of doctor. These two questions are taken from a patient satisfaction survey instrument (Mahapatra, Srilatha and Sridhar, 2001). Note that category labels remain the same but ordinal rankings are reversed depending on the positive or negative frame of the question.

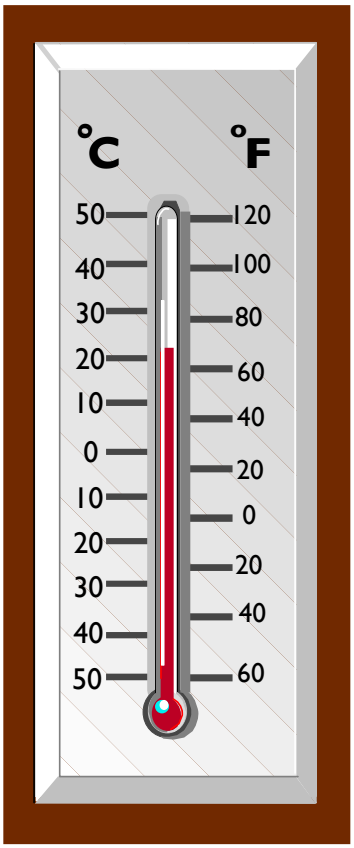
Health State Valuation by Card Sort

Community valuation of the relative severity of various health states usually starts with a card sorting exercise. Here the valuer works with a pack of health state cards. Each health state card describes a condition along six dimensions. The valuer is asked to order the cards from best health state in the pack to the worst health state in the pack. The data derived from this is a rank order within the respective set of health states. Following rank order was assigned to 11 health states including the valuers own health state, by a valuer. The valuer was randomly chosen from the AP Health State Valuation Study Data Set (Mahapatra and others, 1999).

Rank	Health State / Condition
1	Own Health Today
2	Mild Diabetes, no Symptoms
3	Watery Diarrhea 5 times a day
4	Mild Tuberculosis with Treatment
5	Below the Knee Amputation(one leg)
6	Peptic Ulcer
7	Below the Knee Amputation (two legs)
8	Two Broken Arms in Cast
9	Unipolar Major Depression
10	Severe Continuous Migraine
11	Quadriplegia

The interval scale represents a much higher level of measurement than the ordinal scale. It possesses the properties of magnitude and equal interval between adjacent units but does not have an absolute zero point. Thus, the interval scale possesses the properties of ordinal scale and, in addition, has equal intervals between adjacent units. Equal intervals between adjacent units means that there are equal amounts of the variable being measured between adjacent units on the scale (Robert Pagano, 1994, p24). Temperature recorded by the clinical thermometer is a good example of an interval scale. Clinical thermometers are either calibrated with a Fahrenheit scale or a Celsius scale. Both of these are interval scales. The intervals in the scale are equally placed. Although the scales have an interval labeled zero, they do not have a true zero point. A reading of 20^{o}C is not twice as hot as 10^{o}C. Interval scales allow arithmetic operations such as addition, subtraction, but do not allow for multiplication, or division. However, differences between two interval scale measurements can be treated as ratio data. For example the change in temperature measured with the Celsius or Fahrenheit scale can be divided by the change in temperature in another instance measured with the same scale.

Figure-2:
Interval-Scale Thermometer



The next and highest, level of measurement is called a ratio scale. It has all the properties of an interval scale and, in addition, has an absolute zero point. Without an absolute zero point, it is not legitimate to do ratios with the scale readings. Since the ratio scale has an absolute zero point, ratios are permissible (hence the name ratio scale). A good example to illustrate the difference between interval and ratio scales is to compare the Celsius scale of temperature with the Kelvin scale. Zero on the Kelvin scale is an absolute zero i.e. complete absence of heat. Zero on the Celsius scale is the temperature at which water freezes. It is an arbitrary zero point that actually occurs at 273° Kelvin. The Celsius scale is an interval scale and the Kelvin scale is a ratio scale. The difference in heat between 8° and 9° is the same as between 99° and 100° whether the scale is Celsius or Kelvin. However, we can not do ratios with the Celsius scale. A reading of 20° Celsius is really 273° + 20° = 293° Kelvin and 10° Celsius is 273° + 10° = 283° Kelvin. Clearly 293° Kelvin would not be twice as hot 283°. A reading of 20° Kelvin would be twice as hot as 10° Kelvin. Examples of ratio scale measurements are, time, length, weight, age, frequency counts, etc. Ratio scales allow for all arithmetic operations such as addition, subtraction, multiplication, and division.

Measurement instruments bearing an interval or ratio scales have a least count. The least count of an instrument is the smallest unit on the scale. Thus all measurements made on a continuous variable are approximate. For example some of the analogue bathroom weighing scales have a least count of 0.5 Kg. Suppose the exact weight of a person has a fraction of less than 0.5 Kg. The weighing machine can not distinguish this quantity. The observer will record the weight to the nearest 0.5 Kg. Hence the real limits of a continuous variable would be (the recorded measurement $\pm$ 0.5 $\times$ least count). Suppose a study records

10 people's body weight as 64.5 Kg. The real limit of these weights would be 64.25 to 64.75 Kg.

A Clinical Scale with Least Count of lbs or 0.5 Kg.

[1] These weighing scales are known under a variety of names such as; Fitness scales, Personal scales, Bathroom scales, Professional Dial Scales, etc.

Data:

Table-1: Different measurements on literacy giving different types of data.

Nominal	Ordinal	Interval
High School Pass or Not	Just Literate	Years of Schooling
	Primary	
Literate or Not	Secondary School	
	High School	
	College	
	Post Graduate	

Measurement on a variable gives data. The data can be nominal, ordinal, interval or ratio. A variable that is intrinsically nominal, such as Sex, can only be measured by a nominal scale i.e. a classification system. An ordinal variable when measured with help of a nominal

scale yields nominal data. It can yield ordinal data, if the measurement scale is ordinal. Similarly an interval type quantitative variable yields nominal data when measured with nominal scale, can yield both nominal and ordinal data when measured with an ordinal scale and similarly can give nominal, ordinal or interval data when measured with an interval scale. Operationally speaking, data is the starting point of statistical analyses. Naturally many statistical texts introduce the concept of qualitative, quantitative data and distinguish between nominal, ordinal, interval or ratio data and then proceed to illustrate various statistical operations.

Parameters:

A parameter quantifies a characteristic of the population. It is the true, but usually unknown, state of nature (i.e. the universe of study) about which we want to make an inference. Recall that population or the study universe can be real or conceptual. In fact, the study population or universe we deal with in inferential statistics are mostly conceptual rather than real. Appropriate conceptualization of the population is the fundamental research design act required before data can be analysed to draw any conclusion. Suppose we are interested to know about a population, we have that population physically available to us, and do not have any time or resource constraint to measure any characteristic of interest. Then we would not need to make any inference. We would simply measure the characteristic of interest and compute the parameters. But ideal measurement systems do not exist. Any measurement system will have some margin of error. So, at the least, we have to make some inference about the size of this error, to arrive at the exact parameter value. Since we can not know with certainty the size of the error, we can not compute with certainty the value of the parameter. The only opportunity available to us is to estimate the parameter value. More commonly, however, we conceive of study universes, all elements of which may not physically exist for measurement to happen. Suppose we are interested to assess the mortality risk to which infants born and brought up in a given village are exposed. Here the mortality risk is a state of nature. It is the variable to which people living in the village are exposed. We know that infant mortality risk is non negative non zero in any area. Theoretically, we can not have a situation where infant mortality risk can be said to be zero. Even if all infants actually born in the area survive until their first birth day, we can not say that infant mortality risk is zero. The risk of death is always there, howsoever small it may be! That is clearly our current understanding of the state of nature as far as biological life is concerned.

Table-2: Probability of observing no infant death during the course of a year in communities with different population size living under different levels of infant mortality risk (IMR).

Population Size	IMR-> Births	10	30	50	70
		Probability of at least one death in a year			
500	10	0.90	0.74	0.60	0.48
1000	20	0.82	0.54	0.36	0.23
2000	40	0.67	0.30	0.13	0.05
3000	60	0.55	0.16	0.05	0.01
4000	80	0.45	0.09	0.02	0.00
5000	100	0.37	0.05	0.01	0.00
10000	200	0.13	0.00	0.00	0.00
20000	400	0.02	0.00	0.00	0.00
30000	600	0.00	0.00	0.00	0.00
40000	800	0.00	0.00	0.00	0.00

[1] IMR is shown as number of deaths / 1000 child birth.
[2] Crude birth rate is assumed to remain constant at 20 births / 1000 population.
[3] The probabilities have been calculated using Lotus 123 @Binomial(Births,1,IMR,2)

The above table shows results from a thought experiment. It computes the probability of observing at least one infant death during the course of a year in a hypothetical community of different population size and experiencing different infant mortality risks. The infant mortality risk (IMR) is the population parameter. Four population parameter values are chosen. Remember that a population parameter is in truth a fixed value. It does not change since the population parameter is essentially a descriptive measure of the study population. The four population parameters imply four different populations. The table shows that for a village with 1000 population and truly experiencing infant mortality risk of 30 infant deaths / 1000 live births, the probability of observing a single infant death during the course of a year is less than half. Suppose we happen to study such a village, observe 20 child births here during the course of a year, and treat those child births as the population of our study. For any given year the chance that we will not encounter any infant death is a little more than half. Considering that chance is usually lumpy (Abelson, 1995 p17-38) one may not come across a single infant death for two to three years in a row. By the same logic, one may come across upto three infant deaths in a year given the same infant morality risk of 30 infant deaths / 1000 live births acting on the same village with 1000 persons. Of course, the probability of such an event would be very low, some where around 0.0183 implying that such an event may occur, on average, once in 55 years. In the first scenario, if we assume our study universe to consist of all infants born in this hypothetical village of 1000 people, we would conclude that the infants in this village are not exposed to any mortality risk. In the second scenario of 3 observed infant deaths, we might conclude that the infant mortality risk in this village is 150 infant deaths / 1000 live births! Note that we started with an assumption that the true IMR for this village to be 30 infant deaths / 1000 live births. Instead, visualizing a larger hypothetical population of which the 20 infants born in the village during a year is a sample, allows us to view the sample statistic to be an estimate the true infant mortality risk. Even then, the second scenario with three observed deaths will give us a point estimate of 150 infant deaths / 1000 live births. But this time, since we have assumed the 20 births to be a sample from a large number of potential births, we will calculate statistical confidence limits and give out a range of values that straddles the point estimate of 150. Also note that the population of the village is different from the study population. Here the study population consists of infants born or likely to be born in the village. The only link of the village

population to the study population is that the village population contribute to the sample by giving birth to infants. We are using the same word population to mean two different concepts. Firstly the village population means actual human beings living in a village. The study population here is conceived to consist of the infants actually born as well as the infants that could have born in the village. A less confusing term would be the study universe, which is synonymous with study population.

Parameters are usually represented by Greek alphabets. This is a convention. For example the Greek letter μ is traditionally used to denote arithmetic mean of a variable. The Greek letter σ is used to represent population variance of the variable. Europe's ancient philosophy developed in the Graeco-Roman world. Ancient Roman philosophy sprang from the Greek tradition. Most of the philosophical writings were in the Greek language. Of all the branches of human knowledge, philosophy involves a lot of reflection and thinking about the world. The tradition of using Greek alphabets to denote population parameters appears to reinforce the idea that they refer to state of nature, and the truth about a variable, that we may not ever know with complete certainty.

Sample:

A sample is a subset of the study universe.

Statistic:

A statistic is any function of the data. Data invariable come from samples. Thus statistic is a number calculated on sample data that quantifies a characteristic of the sample. A sample statistic gives information about a corresponding population parameter. For example, the sample mean for a sample of data would enable us to estimate the population mean.

References:

Abelson Robert P. Statistics as Principled Argument. Hillsdale, NJ, USA//Hove UK: Lawrence Erlbaum Associates, Publishers; 1995.

Agresti Alan. Categorical data analysis. New York: John Wiley and Sons; 1990.

Argyrous George. Statistics for Social and Health Research. With a Guide to SPSS. New Delhi, London, Thousand Oaks: SAGE Publications; 2000.

Bailey Kenneth D. Typologies and taxonomies. An introduction to classification techniques. Newburry Park//London//New Delhi: Sage; 1994.

Mahapatra Prasanta, Srilatha S., Sridhar P. A patient satisfaction survey in public hospitals. Journal. Academy of Hospital Administration 2001 Jul-2001 Dec 31;13(2):11-5.

Pagano Robert R. Understanding statistics in the behavioral sciences. Fourth edition. Minneapolis-St Paul: West publishing co.; 1994.

Pedhazur Elazar J.; Liora Pedhazur Schmelkin. Measurement, Design, and Analysis: An Integrated Approach. Hillsdale, New Jersey//Hove//London: Lawrence Erlbaum Associates; 1991.

Tiryakian E.A. Typologies. in: Sills D.L., Editor. International encyclopedia of the social sciences. Vol 16.1968. pp. 177-86.

Exercises

1. Identify the study population and sample in the following situations:
   i. Patient Satisfaction Survey: The AP Vaidya Vidhana Parishad (APVVP) manages a large network of public hospitals in Andhra Pradesh. A Patient Satisfaction Surveys (PSS) was conducted by the Institute in the year 1999. All of the 19 District Hospitals, and all of the six Area Hospitals were covered by the survey. The study team wanted to gather a sample of about 50 inpatients from each hospital. The survey team visited the hospital with short notice to the hospital authorities. All patients who had completed three days of stay in the hospital were listed. If the number exceeded 50 a simple random sample was taken to select 50 patients. If the list was equal to 50, all in the list were interviewed. If less than 50, the team recruited additional patients as they completed three day stay in the hospital. A patient satisfaction questionnaire was administered to the patient or an attendant.
   ii. Study on the Structure and Dynamics of Private Health Sector in Andhra Pradesh (SDPHAP): This study sought to understand the structure and dynamics of the private health sector in Andhra Pradesh, in order to provide insights for meaningful policy intervention. Three areas in Andhra Pradesh were selected for the study. The area around the state capital namely Hyderabad - Ranga Reddy districts was chosen purposively. In addition, Visakhapatnam district was chosen from the economically developed districts and Warangal was selected from the economically backward districts. Thus the sample areas are; (a) Hyderabad - Ranga Reddy, (b) Visakhapatnam, and (c) Warangal districts. Altogether 256 health care institutions were surveyed consisting of 150 HCIs in the private and non profit sector and 106 HCIs in the public sector. Three types of HCIs were studied, namely (a) large hospitals, (b) small hospitals and (c) clinics or primary health centres.
2. Following is an extract from the IHS Director's report to the eighth annual general meeting (2000-2001), on 06 December, 2001. Identify the descriptive and / or inferential role of the statistics cited by the Institute Director.

"People are gradually recognising the bibliographic niche being cultivated by the IHS library. Although our library is small, it has some collections in the area of health economics, health system research etc. not easily available elsewhere in Hyderabad. As of January 2000 we had 25 associate members. Currently we have 37 associate members including one life associate member. Most of these memberships are taken to access the Institutes library services. Currently the library services about 323 retrievals per month."

3. Indicate which of the following represent a variable and which a constant:
   i. The number of letters in the Devanagari script.
   ii. The number of letters in various languages of the world.
   iii. The number of hours in a day.
   iv. The time at which you eat dinner.
   v. The number of students admitted every year to the same academic program in which you are enrolled.
   vi. The amount of sleep you get each night.
   vii. Body weight of people in your class.
   viii. The number of playing cards in a pack.

4. Following is an extract of selected questions from the NFHS-2 Women's questionnaire[5]. NFHS-2 instructions about codes, if any, to be assigned to the responses are also given. Identify the type of variable (qualitative, quantitative, etc.) implied by respective questions.

No.	Questions and Filters	Coding Categories	
107	What is your marital status?	Currently Married	1
		Married but Gauna not performed	2
		Separated	3
		Deserted	4
		Divorced	5
		Widowed	6
		Never Married	7
119	Can you read and write?	Yes .....1 No 2	
201	Have you ever given birth?	Yes	1
		No	2
230	Are you pregnant now?	Yes	1
		No	2
		Unsure	8
231	How many months pregnant are you?	Months	
902	Respondent's Haemoglobin level G/DL		

5. Survey of water taps: A survey of water taps in a public building asks the following question for each tap. Identify the variable types and measurement scales.
   i. TapId: Alphanumeric identification given by you to uniquely identify each tap.
   ii. Location: Area where located. For example; Corridor, Patio, Gents Toilet -1, etc.
   iii. Date: Date of the survey.
   iv. Time: Time of the survey.
   v. Yield: Open the tap and observe for water flow and report; Dry, Trickle, or Flow.
   vi. Leaking: Applicable only if Yield # Dry. Open the tap and let water flow for a few seconds. Then close the tap. Observe if flow stops completely (Leaking = False) or there is some leak (Leaking = True).
   vii. Positioning: Consider the purpose for which the tap has been provided? Give your assessment about positioning of the tap from the perspective of the intended use. For example, a tap meant for hand wash should be is the tap positioning appropriate for its intended use? If yes, then Positioning = Y else, Positioning = N.

[5] The first National Family Health Survey (NFHS-1) in India was conducted during 1992-93. The second survey (NFHS-2) was conducted during 1998-99. The principal objective of NFHS is to provide state and national level estimates fertility, the practice of family planning, infant and child mortality, maternal and child health, and the utilization of health services provided to mothers and children. The NFHS in India is in many respects similar in scope to the Demographic and Health Surveys (DHS) elsewhere in the world. NFHS-2 India report has been published by International Institute for Population Sciences (IIPS); ORC Macro; Roy TK, et al. National Family Health Survey 1998-99 (NFHS-2). India. Mumbai (Bombay): International Institute for Population Sciences (IIPS); 2000 Oct.

viii. Design: Consider the purpose for which the tap has been provided? Give your assessment about design of the tap from the perspective of the intended use. For example, a tap in operation theatre area for hand wash and scrubbing should have an elbow operated lever to open and close the tap. A tap serving a sink should have a long enough neck, so that a person does not have to bend a lot while washing things in the sink. Thus ask; is the tap design appropriate for its intended use? If yes, then Design = Y else, Design = N.

ix. Tapwork: Assess if the tap needs any work. Choose one of the following: Repair, Replace, Reposition, Redesign, Reposition And Redesign, Connect (with water source), or None.