Review of the Statistical Techniques in Medical Sciences

[1]Okeh, U. M. and [2]Ugwu, A. C.
[1]Department of Industrial Mathematics and Applied Statistics, Ebonyi State University, Nigeria.
[2]Department of Radiology, Federal Medical Center, Ebonyi State, Nigeria.

Corresponding author: Ugwu, A. C. Department of Radiology, Federal Medical Center, Ebonyi State, Nigeria. Email: tonybullng@yahoo. Phone: +234 8076241297)

Abstract

Medical scientists often times are faced with the need to choose the right statistical technique for a given set of data. There are a number of criteria that should be a guide in making the right choice: the research questions, the category of the variables involved the size of the sample, the scale of measurement of the variable, the type of study design, among others. This article reviews some procedures which will be followed by any medical researcher in selecting the appropriate statistical techniques. Of course, all statistical techniques have certain underlying assumptions, which must be checked before the technique is applied.

Keywords: Variable, Prospective Studies, Retrospective Studies, Statistical significance

Introduction

Traditional scientific inquiry consists of four interrelated stages: problem definition, data gathering, data analysis, and data interpretation. Statistics are used in the data analysis and interpretation stages of an inquiry. There is nothing magical about statistics; however, like any tool, the appropriate use of statistics depends upon a basic knowledge of their capacities and limitations. Fortunately, one need not be a statistician to take advantage of this important tool of scientific inquiry. Statistical procedures fall loosely into three general categories: descriptive, associative and inferential (Anon, 2007). Descriptive statistics aims to provide meaningful and convenient techniques for describing features of data that are of interest. It is basically a device for organizing data and bringing into focus their essential characteristics for the purpose of conclusion. It deals with the collection, representation, calculation and processing, i.e. the summarization of data to make it more informative and comprehensive. It involves graphical and tabular approaches to describe, summarize and analyse the data. The primary function of descriptive statistics is to provide meaningful and convenient techniques for describing features of data that are of interest (Rastogi, 2006). Associative statistics seek to identify meaningful interrelationships between or among data. Addressing the question “Is there a relationship between salt intake and diastolic blood pressure among middle –age women?” is a problem definition suitable for analysis by associative statistics. Inferential statistics seek to assess the characteristics of a sample in order to make more general statements about the parent population, or about the relationship between different samples or populations. Addressing the question “Does a low sodium diet lower the diastolic blood pressure of middle–age women? ”Represents a problem definition suitable for inferential statistics (Anon, 2007).

Materials and Methods

Articles textbooks and the internet was consulted as the source of information *vis a vis* the use of statistics in medical sciences.

Results and Discussion

Statistical significance: Whereas descriptive statistics simply portray data, associative and inferential statistics test the likelihood or probability of a set of observations relative to chance. For this reason, associative and inferential statistical procedures provide both a statistical value (e.g. r, F, t) and a level of probability or “P” value. The “P” value simply represents the probability that the observed findings are a “chance” occurrence, i.e. due to random fluctuations or errors in sampling. A “P” value of 0.01, therefore, indicates that the probability is, I out of 100 that the observed finding is a chance event. Conversely, one could say with 99% confidence that the observed finding is “meaningful”- in the sense of whatever hypothesis was originally posed, obviously, from a statistical viewpoint, nothing is ever absolutely sure. Medical researchers and statisticians must therefore always accept the possibility that what they observe is not a true relationship but simply a chance aberration. However, in order to minimize the likelihood of being wrong, a researcher usually “presets” an acceptable level of probability for chance occurrences. Called the alpha level (designated by the Greek α). This acceptable error level is usually set at either 0.05 or 0.01, with the latter being the most conservative. Should the results of statistical analysis provide a value greater than the alpha level, e.g., P=0.10, the researcher would not be willing to claim that the findings were meaningful, or, in statistical parlance, “the findings were not statistically significant. “This simply means that the probability of the findings being a chance occurrence (in our example, I out of 10) is too high to have confidence in the results. On the other hand, when researchers use the term “statistically significant” the are simply saying that the probability of the occurrence being due to chance alone (the “P” value) is less than their preset error level (α). For example, if a researcher sets α at 0.05 (indicating that he is willing to accept being wrong 1 time out of 20) and the results of statistical analysis provided a “P” value of 0.02; he would portray the findings as “statistically significant”. This is because 0.02 is less than 0.05. Whether or not the findings have practical importance is, of course, another matter which one can be judged only by logical, as opposed to statistical analysis of the data.

Selecting statistical procedures: There are literally hundreds of statistical procedures available to medical researchers. The choice of a statistical procedure depends on both the nature of the problem being addressed and the level (s) of measurement obtained during the data gathering stage of the inquiry. In selecting a statistical test, the most important question is “what is the main study hypothesis?” In some cases there is no hypothesis; the investigator just wants to “see what is there”. For example, in prevalence study, there is no hypothesis to test, and the size of the study is determined by how accurately the investigator wants to determine the prevalence. If there is no hypothesis, then there is no statistical test.

Problem definition and variables: “A question well asked is a question half answered” is a common axiom in disciplined inquiry. Carefully defining one’s problem via specific research questions or hypotheses provides a foundation for all that follows, including the choice of statistical tests. Carefully worded research questions or hypotheses guide the research by delineating which variables will be scrutinized and what relationships may exist between or among these variables. Clear operational definitions of these variables provide a sound basis for their measurement. In most research designs, there are three categories of variables (Anon).Independent variables are antecedent and presumed causal to an observed phenomenon. Dependent variables represent responses thought to be influenced by the independent variable. For example, in a controlled clinical trial of a new drug (the classical experimental design), the independent variable would be the subject’s group assignment (experimental group receives drug, control group gets placebo). The dependent variable might be the sought-after therapeutic effect of the drug. A third category of variables, called nuisance or intervening variables, represent factors which might alter the relationship between independent and dependent variables. In the drug trial previously described, possible intervening variables include patient age, metabolism, and route of administration. Because intervening variable can alter the relationship between independent and dependent variables, researchers must address their potential moderating effect. Although only the true experimental design can establish cause and effect relationships similar terminology applies to other research methods. In this context, factors which are logically antecedent to a phenomenon are termed independent variables, with the observed phenomenon itself being measured as the dependent variable. Lacking the inherent controls characterizing the true experimental design, however, no cause and effect relationship can or should be presumed (Rastogi, 2006).

Levels of measurement: Once categorized according to the above scheme (i.e., independent, dependent, intervening), variables must be measured. This of course, is the basis for data gathering. Data gathering employs measurement scale or sets of rules for quantifying and assigning values to a particular variable (a characteristic that varies from one observation to another in a group of observations or measurements). Some examples of variables include the heights, or weights of teenage girls in a given school, the ages of patients attending a health diagnostic clinic, genotypes; etc. A single observation or value of a given variable is often referred to as a variable. Typically, four levels of measurement apply to data gathering. Data levels may be characterized as nominal, ordinals interval and ratio (Oyeka, 1996).

Data on a nominal scale: These are variables measured qualitatively using certain properties they possess rather than quantitatively. Such variables are called attributes. Genotypes are commonly encountered biological attributes. Taxonomic categories also form a nominal classification scheme. Other examples include skin colour (black, brown, red or white), eye colour (blue or brown) and such dichotomies as male-female, fertile-sterile, well-sick, alive-dead. As will be seen, generally, only certain statistical techniques can be applied to nominal scale data and it is important to be able to identify such situation (Oyeka, 1996).

Data on ordinal scale: These are measurements carried out when observations not only differ from category to category but can also be rank-ordered according to some criterion. In this case interest may be in relative rather than in quantitative differences. One may refer to one entity as being longer, brighter than another; patients may be classified as unimproved, improved, much improved; four cranial sizes may be labeled 1, 2, 3 and 4 to denote their magnitudes relative to each other. In each of these examples members of any one category are all considered equal, and say, better, bigger, or lower than those in another category, which in turn bears a similar relationship to another category. For instance, a patient classified as improved is in a worse condition of health than a patient classified as much improved but in better health than a patient classified as unimproved. Numbers assigned to ordinal data merely help to order or rank the observations from higher to lowest (or vice versa). However, a great many statistical procedures are applicable to ordinal data (Oyeka, 1996).

Data on interval scale: Data is said to be measured on an interval scale whenever it is not only possible to order the measurements, but the distance or interval between any two measurements is also a known constant. Unlike the nominal and ordinal scales, the interval scale is truly quantitative. E.g. two common temperature scales: Celsius or centigrade (C), and Fahrenheit (F). The difference between say, 30^0C (86^0F) and 40^0C (104^0F) is equal to the difference between 50^0C (122^oF) and 60^0C (146^0F). Another example is time measurement. The interval between 9.00am (0900hrs) and 10.00am (1000hrs) is the same as the interval between 3.00pm (1500hrs) and 4.00pm (1600hrs). We do not, however, mean to imply that a temperature of 30^0C (86^0F) is half, as hot as a temperature of 60^0C (146^0F), or that one could take ratios of times of day, for example. In each case, the selected zero point is arbitrary and not a true zero in that the quantity being measured is not totally absent at this point (Oyeka, 1996).

Data on ratio scale: This is a measurement scale having a constant interval size and a true zero point. This is, of course, equivalent to an interval scale data set with a true zero point. Examples are number of leaves on a plant, lengths, heights, weights, rates, volumes, ratios, etc. Many statistical procedures are available for use with data on ratio scale of measurement provided that all the necessary assumptions are met.

Whenever possible, data should be gathered at the highest level. The higher level of precision provided by interval and ratio data allows for more powerful statistical testing. Moreover, high level data easily can be converted to lower levels, i.e. ordinal or nominal. The reverse is not true (Oyeka,1996). Data on any of these measurement scales can be either continuous or discrete. A set of data is said to be continuous if it can conceivably assume any numerical value within any two points on a continuum, for example, height, or age of a plant or person. In contrast, a discrete set of data, also termed discontinuous or meristic data is one that can assume a finite or countable number of numerical values, for example, numbers of colonies of micro-organisms or animals. A variable is said to be a continuous variable if its data are continuous and a discrete variable if its data are discrete.

Choice of the statistical technique: To choose the appropriate statistical test, first categorize your variables as independent and dependent (intervening or nuisance variables are usually treated as additional independent variables). Next, determine the number of independent and dependent variables in the study. Finally, determine the level of measurement (nominal, ordinal or interval) applied to each relevant variable. Then use the table below to determine which statistical test or tests might be appropriate.

Table 1: Choice of statistical test from paired or matched observations

Variable	Test
Nominal	MeNemar's Test
Ordinal (ordered categories)	Wilcoxon
Quantitative(discrete or non normal)	Wilcoxon
Quantitative (normal*)	Paired T-test

Source: *Campbell, 1993.*

It is helpful to decide the input variables and the outcome variables. For examples, in clinical trial the input variable is the type of treatment (a nominal variable) and the outcome may be some clinical measure perhaps normally distributed. If a set of data (value of the population) is normally distributed, the range should be within mean ± 3SD (3 sigma rule). The required test is then the T-test. However, if the input variable is continuous, say a clinical score, and the outcome is nominal, say cured or not cured, logistic regression is the required analysis. A t -test in this case may help but would not give us what we require, namely the probability of a cure for a given value of the clinical score. As another example, suppose we have a cross sectional study in which we ask a random sample of people whether they think their general practitioner is doing a good job, on a five point scale, and we wish to ascertain whether women have a higher opinion of general practitioners than men have. The input variable is gender, which is nominal. The outcome variable is the five point ordinal scale. Each person's opinion is independent of the others, so we have independent data. From here we know we should use a $\chi 2$ test for trend, or a Mann-Whitney U test (with correction for ties). Note, however, if some people share a general practitioner and others do not, then the data are not independent and a more sophisticated analysis is called for.

Note that these tables should be considered as guides only and each case should be considered on its merits.

a. If data are censored
b. The kruskal-Wallis test is used for comparing ordinal or non-normal variables for more than two groups, and is a generalization of the Mann–Whitney U test. Kruskal –Wallis is alternative nonparametric procedure for one – factor analysis and it is used when the assumptions of the F- test are not satisfied and when the assumptions are satisfied. Kruskal-Wallis test is based on a test statistic denoted as H, computed from ranks determined for pooled sample observations rather than from the observations themselves (Kruskal-Wallis, 1952).
c. Analysis of variance is a general technique, and one version (one way analysis of variance) is used to compare normally distributed variables for more than two groups, and is the parametric equivalent of the kruskal-Wallis test (Kruskal-Wallis, 1952).
d. If the outcome variable is the dependent variable, then provided the residuals (see) are plausibly normal, then the distribution of the independent variable is not important.
e. There are a number of advanced techniques, such as poisson regression, for dealing with these situations. However, they require certain assumptions and it is often easier to either dichotomize the outcome variable or treat it as continuous.

Research design: In many ways the design of a study is more important than the analysis. A badly designed study can never be retrieved, whereas a

Table 2: Choice of statistical test for independent observations

	Nominal	Categorical (>2 Categories)	Ordinal	Out Come Quantitative Discrete	Variable Quantitative Non-Normal	Quantitative Normal
Nominal	X^2 or Fisher's	X^2	X^2 trend or Mann – Whitney	Mann –Whitney	Mann –Whitney or log –rank (a)	Student's t test
Categorical (>2 categories)	X^2	X^2	Kruskal- Wallis (b)	Kruskal - Wallis (b)	Kruskal- wallis (b)	Analysis of variance(c)
Input variable ordinal (ordered categories)	X^2_ trend or Mann-Whitney	(e)	Spearman rank	Mann –Whitney	Spearman rank	Spearman rank or linear regression (d)
Quantitative discrete	Logistic regression	(e)	(e)	Spearman rank	Spearman rank	Spearman rank or linear regression (d)
Quantitative non-normal	Logistic regression	(e)	(e)	(e)	Plot data and Pearson or Spearman rank	Plot data and Pearson or spearman rank and linear regression
Quantitative normal	Logistic regression	(e)	(e)	(e)	Linear regression(d)	Pearson and linear regression

Source: *Campbell, 1993*

poorly analyzed one can usually be reanalyzed. Consideration of design is also important because the design of a study will govern how the data are to be analysed. Most medical studies consider an input, which may be a medical intervention or exposure to a potentially toxic compound, and an output, which is some measure of health that the intervention is supposed to affect. The simplest way to size studies is with reference to the time sequence in which the input and output are studied. Here are some of the study designs:- The most powerful studies are prospective studies, and the paradigm for these is the randomized controlled trial (Pocock,1982). In this subjects with a disease are randomized to one of two (or more) treatments, one of which may be a control treatment. In other words, this study is characterized by the identification of the two study samples (treatment and control) on the basis of the presence (A) or absence (Ā) of the antecedent factor and by estimating for both samples the proportion developing the disease or condition under study. It therefore means that the importance of randomization is that we know in the long run treatment groups will be balanced in known and unknown prognostic factors. It is important that the treatments are concurrent meaning that the active and control treatments occur in the same period of time.

A parallel group design is one in which treatment and control are allocated to different individuals, To allow for the therapeutic effect of simply being given treatment, the control may consist of a placebo, an inert substance that is physically identical to the active compound. If possible a study should be double-blinded- neither the investigator nor the subject being aware of what treatment the subject is undergoing. Sometime it is impossible to blind the subject, for example when the treatment is some form of health education, but often it is possible to ensure that the people evaluating the outcome are unaware of the treatment. A matched design comes about when randomization is between matched pairs. E.g. that between different parts of a patient's body. A crossover study is one in which two or more treatments are applied sequentially to the same subject. The advantage is that each subjects then acts as his/her own control and so fewer subjects may be required. The main disadvantage is that there may be a carry over effect in that the action of the second treatment is affected by the first treatment. An example of a crossover trial is when different dosages of bran are compared within the same individual (Senn, 1992). One of the major threats to validity of a clinical trial is compliance. Patients are likely to drop out of trials if the treatment is unpleasant, and often fail to take medication as prescribed. It is usual to adopt a pragmatic approach and analyze by intention to treat (analyzing the study by the treatment that the subject was assigned to and not the one they actually look for). The alternative is to analyze per protocol or on study. Drop outs should of course be reported by treatment group. A checklist for writing reports on clinical trials is available (Gardner, 1986).

A quasi experimental design is one in which treatment allocation is not random. An example of this type of design is that in which injuries are compared in two dropping zones (Armitage, 1994). This is subject to potential bias in that the reason why a person is allocated to particular dropping zone may be related to their risk of a sprained ankle.

A cohort study is one in which subjects, initially disease free, are followed up one a period of time. Some will be exposed to some risk factor, for example cigarette smoking. The outcome may be death and we may be interested in relating the risk factor to a particular cause of death. Clearly, these have to be large, long term studies and tend to be costly to carry out. If records have been kept routinely in the past then a historical cohort study may be carried out. Here, the cohort is all cases of appendicitis admitted over a given period and a sample of the records could be inspected retrospectively. A typical example would be to look at birth weight records and relate birth weight to disease in later life.

These studies differ in essence from retrospective studies, which start with diseased subjects and then examine possible exposure. Such case control studies are commonly undertaken as a preliminary investigation, because they are relatively quick and inexpensive. The comparison of the blood pressure in farmers and printers is an example of a case control study. It is retrospective because we argued from the blood pressure to the occupation and did not start out with subjects assigned to occupation. There are many confounding factors in case control studies. For example, does occupational stress cause high blood pressure, or do people prone to high blood pressure choose stressful occupation? A particular problem is recall bias, in that the cases, with the disease, are more motivated to recall apparently trivial episodes in the past than controls, who are disease free.

Cross sectional studies are common and include surveys, laboratory experiments and studies to examine the prevalence of a disease. Studies validating instruments and questionnaires are also cross sectional studies. The study of urinary concentration of lead in children and the study of the relationship between height and pulmonary anatomical dead space are also cross sectional studies.

Data sample size: One of the most common questions asked of a statistician about design is the number of patients to include. It is an important question, because if a study is too small it will not be able to answer the question posed, and would be a waste of time and money. It could also be deemed unethical because patients may be put at risk with no apparent benefit. However, studies should not be too large because resources would be wasted if fewer patients would have sufficed. The sample size depends on four critical quantities: the type I and type II error rates α and β, the variability of the data δ^2, and the effect size d. In a trial the effect size is the amount by which we would expect the two treatments to differ, or is the difference that would be clinically worthwhile. Usually α and β are fixed at 5% and 20% (or 10%) respectively. A simple formula for a two group parallel trial with a continuous outcome is that the required sample size per group is given by $n = 16\delta^2/d2$ for two sided α of 5% and β of 20%. For example, in a trial to reduce blood pressure, if a clinically worthwhile effect for diastolic blood pressure is 5mm Hg and the between subjects standard deviation is 10mm Hg, we would require $n = 16x100/25 = 64$ patients per group in the study. The sample size goes up as the square of the standard deviation of the data (the variance) and goes down inversely as the square of the effect size. Doubling the effect size reduces the sample size by four- it is much easier to detect large effects! In practice, the sample size is often fixed by other criteria, such as finance or resources, and the formula is used to determine a realistic effect size. If this is too large, then the study will have to be abandoned or increased in size. Also sample size guidelines for several of the simple statistical techniques, as well as references, are given in details (khamis, 1988).

Conclusion: Generally, any medical research involves interrelated stages of scientific inquiry which includes problem definition, data gathering, data analysis and data interpretation. If a research question is to be investigated and the investigation involves data, then several statistical issues need to be addressed. One of the first considerations is outlining the purposes of the statistical data analysis, categorizing the statistical procedures and identifications of the statistical significance. Next is categorizing the variable or data as independent and dependent, determining the number of independent and dependent variables in the study as well as determining the measurement scale of the data. Care must also be taken in identifying the type of study design involved. Of course, the size of the sample for statistical analysis has to be determined using appropriate method. This article has focused on the review of the various steps involved in choosing statistical technique which depends on both the nature of the problem being addressed and the measurement scale of data. While the information in this article will help guide any medical scientist toward the proper statistical analysis, it is nevertheless recommended that a statistician be consulted early in the research project so as to ensure the highest statistical standards.

Acknowledgements

The authors appreciate the helpful comments and suggestions of Prof. I. C. A Oyeka, Department of Applied Statistics, Nnamdi Azikiwe University Awka, Anambra State, Nigeria.

References

Anon(2007).The Role of Statistical Analysis in R esearch, http://www.umdnj.edu/id_sweb/sh ared /statslct.htm.

Armitage, P. and Berry, G. (1994). *Statistical Methods in Medical Research.* Blackwell Scientific Publication. Oxford.

Campbell, M. J. and Machin, D. (1993). *Medical Statistics: A commonsense Appro ach,* 2nd edition. John Wiley and Sons, Chichester.

Gardner, M. J., Machin, D. and Campbell, M. J. (1986). The use of checklists in assessing the statistical content of medical studies. *BMJ,* 292: 810-12.

Khamis, H. J. (1988). Statistics refresher: II Choice of sample size. *JDMS,* 4:176 – 183.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one Criterion analysis of variance.

J. Amer. Statistics Association, 47: 383 – 621.

Oyeka, CA.(1996). An Introduction to Applied Statistical Methods in the Sciences. Modern Printers, Nigeria.

Pocock SJ.(1982). *Clinical trials: A practical Approach,* John Wiley and Sons, Chichester.

Rastogi, V. B. (2006).*Fundamentals of Biostatistics.* Ane Books, India.

Senn, S. J. (1992). *The Design and Analysis of Crossover Trials.* John Wiley and Sons, Chichester.