International Journal of Sciences & Applied Research

www.ijsar.in

Biostatistics: From Conception to Execution

Amit Mittal*

Department of Community Medicine, Rajshree Medical research Institute Bareilly, India.

Correspondence Address: *Amit Mittal, Department of Community Medicine, Rajshree Medical research Institute Bareilly, India.

Abstract

Statistical analyses of biomedical research are obligatory to churn out any plausible explanation of the elaborate studies and expensive battery of tests performed. Statistics therefore, is common place in biomedical literature. Despite their wide use, even the simpler terms, analyses and explanations are sometimes misunderstood or misinterpreted by researchers who have either a limited or faulty knowledge of statistics. Surprisingly many are oblivious of the role of statistician in today's research set up. A sound understanding of the above is needed to foster closer and regular interactions between statisticians and other members of a research team for effective incorporation of the statistical elements into research and discussion of research output. An attempt is made to outline a few basics elements of statistics that are commonly found in biomedical literature with a word of advice on how to go about them.

Keywords: Non Parametric test, Parametric test, p value

Introduction

Statistics implies both data and statistical methods. It can be considered as an art as well as science. Statistics can neither prove nor disprove anything. It is just a tool. Statistics without scientific application has no roots. Thus statistics may be defined as the discipline concerned with the treatment of numerical data derived from group of individuals. These individuals may be human beings, animals, or other organisms. Statistical method has two major branches mainly descriptive and inferential. Descriptive statistics includes measure of central tendency and variability. This type of statistics is commonly used to summarize data about sociodemographic and clinical features whereas inferential statistics is used to express the level of certainty about estimates and includes hypothesis testing, standard error of mean and confidence interval. Biostatistics is a branch of statistics applied to biological or medical sciences. It covers applications and contributions not only from health, medicines and nutrition but also from fields such as genetics, biology, epidemiology and many others (Rao., 2007).

Types of data

In research, it is necessary to study certain characteristics in a group of subjects, such as age, sex, socioeconomic group etc. Each of these characteristics may vary from person to person and is referred to as a variable. The values taken by these variables are referred to as data. Data collected during a study may fall in to one of the following three types of data:

(i) Nominal (categories, attributes) e.g. sex (male, female), religion (Hindu, Muslim, Christian and others), blood groups (O, A, B and AB), Yes/no type (patient responded or not, cured or uncured, hypertensive or normal, smoker or non smoker) (Nanivadekar and Kannapam., 1991).

(ii) Ordinal (graded) e.g. severity of pain, itching or erythema may be graded as absent=0, mild=1, moderate=2, severe=3, socioeconomic status, degree of cigarette smoking (non smoker, ex-smoker, light smoker, heavy smoker) (Nanivadekar and Kannapam., 1991).

(iii) Interval/ratio type (measurements) e.g. age, height, blood glucose levels etc· (Nanivadekar and Kannapam., 1991).

Study design

The most important phase of any research is the planning and design phase, as a proper and complete study design constitutes the basis for healthy research. Errors, flaws and shortcomings, occurring in the planning stage can have a vast negative impact on the validity and reliability of research results, as they affect all subsequent stages of an investigation. Each study has some advantages and disadvantages. Randomized, controlled clinical trials are the most powerful designs possible in medical research. Biostatistician can help us design our experiments in such a way that unexpected variables or events are less likely to sabotage our study. It can also help us determine the no. of measurements necessary for the level of evidence desired and determine the right experimental design for desired comparisons. So in other sense, meeting with a statistician we embark on data collection is like practicing preventive medicine.

Compiling the data

At the stage of compiling of data, firstly the data source should be decided; afterwards the inclusion of the data source, completeness and reliability should be examined carefully (Sumbuloglu and Sumbuloglu., 2001). At the stage of compiling data, one of the most common errors occurs while using the data previously recorded and compiling them during secondary compiling. Researchers may not find the exact variable they will examine in the recordings. In that case the researchers may struggle to increase the number of data or try to change the structure of the data.

Data Manipulation

Manipulation of the data can distort and alter the data, leading to misinterpretation and misrepresentation of the experimental results. The best approach to making figures is to be mindful of how we design our experiments so that we do not have to manipulate our data after the fact. If our data are not correctly organized, it is better to return the experiment with all the samples to get the lanes correct and get the right exposure.

Hypothesis

The primary object of statistical analysis is to find out whether the effect produced by a compound under study is genuine and is not due to chance. Hence the analysis usually attaches a test of statistical significance. First step in such a test is to state the null hypothesis. In this we make assumption that there exist no differences between the two groups. Alternative hypothesis states that there is a difference between two groups.

(1) Type I error (False positive)

It is the probability of finding a difference; when no such difference actually exists, which results in the acceptance of an inactive compound as an active compound. Such an error which is not unusual may be tolerated because in subsequent trials, the compound will reveal itself as inactive and thus finally rejected.

(2) Type II error (False negative)

It is the probability of inability to detect the difference when it actually exists, thus resulting in the rejection of an active compound as an inactive. This error is more serious than type I error because once we labeled the compound as inactive, there is possibility that nobody will try it again. Thus an active compound will be lost (Lang., 2004).

(3) Power of study

Power of study is very important while calculation of sample size. It can be calculated after calculation of study called as posteriori power calculation. This is very important to know whether study had enough power to pick up the difference if it existed. Any study to be scientifically sound should have at least 80% power. If power of study is less than 80% and the difference between groups is not significant, then we can say that difference between groups could not be detected, rather than "no difference" between the groups. If we increase the power of study, then sample size also increases. It is always better to decide power of study at initial level of research.

Level of significance

If the probability of an event is high, we say it is not rare but if the probability of an event is low, we say it is rare. In biostatistics a rare event is called significant, whereas a non rare event is called non significant. The p value at which we regard an event as enough to be regarded as significant is called significance level. In medical research most commonly p value less than 0.05 is considered as significant otherwise non significant level (Mahajan, 2010).

Importance of sample size in medical research

Sample is a fraction of the universe. Studying the universe is the best parameter. But when it is possible to achieve the same result by taking fraction of the universe, a sample is taken. Hence an adequate sample size is of prime importance in biomedical studies. If the sample size is too small, the study may fail to detect a true difference that actually exists. By convention the type II error should be 0.2 or 20% or less. Then the power of study will be 0.8 or 80% or more. If the sample size is too large, it may be concluded that even a very small difference is statistically significant but, in actuality this difference is not clinically significant. We conclude that a difference exists but actually it does not exist. This is known as type I error, by convention this error should be 0.05 or 5 % or less. A very large sample size also increases the cost and causes delay in completion of research project (Zodpey, 2004).

Factors influencing sample size

Prevalence of particular event –If the prevalence is high, small sample can be taken and vice-versa. If prevalence is not known then it can be obtained by a pilot study.

Calculation of sample size

Calculation of sample size plays a key role while doing any research. Before calculation of sample size, following five points are to be considered very carefully. First of all we have to assess the minimum expected difference between the groups. Then we have to find out the S.D. of the variables. Now set the level of significance (generally set as p<0.05) and power of study. After deciding all these parameters, we have to select the formula from computer programmes to obtain the sample size. Various software are available free of cost for calculation of sample size and power of study.

Sampling

A statistician is often dealing with inferential methods, it is important that the characteristics of a small sample be a true indication of the characteristics of the population about which he may make generalized conclusions. For the sample to be representative of the population it must be obtained in such a way that no bias enters in to its selection. For greatest likelihood of this being the case every individual in the population must have the same chance of being selected for the sample. There are many methods of sampling, and the choice of particular method is up to the discretion of the researcher. For all sampling methods particularly simple random sampling is used unconsciously. To represent the population by the sample, determining the subjects that will be included in the sample is the next stage that requires attention after deciding the appropriate sampling technique (Kan., 1998; Tokol., 2000). So the criterion of the selection must be clearly determined. One of the most common errors in selection of the subject is collecting the units by different researchers who are not in the research group. Selection type of the units to the sample is also defined due to the research subject. The misrepresentation of non probability sampling as random sampling has important implications (Williamson., 2003). Random sampling methods are used when a sample of subjects is selected from a population of possible subjects in observational studies, such as cohort, case control and cross sectional studies (Dawson and Trap, 2001). Especially in the studies that are made in order to get the knowledge about the population, probability sampling is inevitable. But in some cases researchers make mistakes by not using probability techniques. So constructing probability sampling or non probability sampling due to the research subject should be examined very carefully. For that a biostatistician plays a big role.

Outliers

Sometimes when we analyze the data one value is very extreme from the others. Such value is referred as outliers. This could be due to two reasons. Firstly the value obtained may be due to chance; in that case, we should keep that value in final analysis as the value is from the same distribution. Secondly it may be due to mistakes; in such cases these values should be deleted, to avoid invalid results.

How to choose an appropriate statistical test

The following table is a good representation for selecting a statistical test.

	Non parametric test		**Parametric test**
	Nominal data	**Ordinal data**	**Ordinal, interval, ratio data**
One Group	Chi square goodness of fit	Wilcoxon signed rank test	One group t test
Two unrelated groups	Chi square	Wilcoxon rank sum test, Mann Whitney test	Student’s t test
Two related groups	Mc Nemar’s test	Wilcoxon signed rank test	Paired student’s t test
K unrelated groups	Chi square test	Kruskal Wallis , One way analysis of variance	ANOVA
K related groups		Friedman’s matched samples	ANOVA with repeated measurements

Statistical software

If the researchers do not consult a statistician and if they do not have adequate statistics knowledge, one of the most common errors is the error sourced from the statistical software which makes the statistical analysis easier. There are many statistical software available for carrying out the analysis; One of the most popular software is SPSS. But it has some drawbacks because it has not covered the following topics such as Power Analysis, Time Series Analysis, Sample Size Calculations, Direct computations from contingency tables, Limited Programming options, Lacking in effective Control Charts presentations, Simulation, Vital statistic, Index numbers, Missing Monte Carlo Markov Chain feature, Optimal designs, Missing data techniques, Only few Design options, Limited non-parametric statistics, Only few distributions, Reproduction of graphics, GAM (Generalized Additive Models), Data Mining Techniques. For that we have other Statistical packages such as MINITAB**,** Epi-Info**,** STATGRAPHICS**,** SYSTAT**,** S-PLUS**,** BMDP**,** SAS**,** STATA**,** R**,** MATHEMATICA**,** STATISTICA etc. (Murphy, 2004)

The Graph pad instat software is also very popular as its demo version can be freely downloaded from the website www.graphpad.com.

Conclusion

The role of statistics or a statistician in biomedical research should start at a very early planning stage of an experiment or a clinical trial to establish the design and size of a study that will ensure a good prospect of detecting effects of clinical or scientific value. Statistics analyses data to make inferences applicable in a wider population. Advanced research work demands applied statistical methods which include the formulation and testing of mathematical models to make relevant inferences from observed data. Application of such advanced methods require a clear understanding of the purpose of research and the anticipated goals before any conclusions are arrived at based upon the statistical tools employed. Better structured study programs should build the basic understanding in biomedical scientists and medical graduates and help seamless exchange of ideas between the statistician and the biomedical researcher.

References

Dawson B, Trap RG. Basic and clinical biostatistics 3rd ed. Lange medical books/ Mc Graw- Hill International editions, 2001:21,71,107.

Kan I. Biostatistics. Bursa: Uludag University Publication, 1998: 55.

Lang T. Twenty statistical errors even you can find in biomedical research article. Croat Med J. 2004; 45:361-70.

Mahajan BK. Methods in biostatistics, 7th ed. New Delhi: Jaypee Brothers Medical Publisher Ltd.; 2010. Sample variability and significance; PP.104-16.

Murphy JR. Statistical errors in immunologic research. J.Allergy Clin Immunol 2004;114: 1259-63.

Nanivadekar AS, Kannappan AR. Statistics for clinicians 3,4,7. J. Assoc Physicians India 1991; 39: 194-8, 222; 273-7,281; 549-53. .

Rao KV. What is Statistics? What is Biostatistics? In: Rao KV, editor. Biostatistics: A mannual of statistical methods for use in health, nutrition and anthropology. 2nd ed.. New Delhi: Jaypee Brothers Medical Publisher Ltd; 2007. pp.1-4.

Sumbuloglu K, Sumbuloglu V. Consciously use of biostatistics principles and methods in scientific researches. Pfizer Ltd. Co. 2002: 7-40.

Tokol T. Marketing research. Bursa: Vipas Co. 2000:19.

Williamson G.R. Misrepresenting random sampling? A systematic review of research papers in the journal of advanced nursing. J. Adv. Nurs 2003; 44 (3):278-88.

Zodpey SP. Sample size and power analysis in medical research. Indian J Dermatol Venereol Leprol 2004; 70:123-8.