BIO-STATISTICS: A BRIEF OVERVIEW

*Amit Sharma

Abstract

Research is an integral part of medical colleges. Students are often unaware of the importance of statistics in research work. The following article gives a brief overview of some of the common definitions, types and methods of statistics that they came across while working on a project. The purpose of this article is to help them understand and choose the best method applicable for the interpretation of the findings of their research work.

Keywords: Biostatistics; Medical students; Research; Interpretation.

Introduction

Medical field cannot survive without continuous research. The need and curiosity of indulging into research projects starts right from the day a student enters into medical college. In case of postgraduate students, at the time of doing their thesis/dissertation they come across the task of interpretation of the data of their findings, which requires the knowledge of the type of statistical test that should be applied. When examining the statistics section of an article, it is helpful to have a systematic approach. It is not essential to understand the exact workings and methodology of every statistical test encountered, but it is necessary to understand selected concepts such as parametric and nonparametric tests, correlation, and numerical versus categorical data. This working knowledge will allow spotting obvious irregularities in statistical analyses encountered. Various articles have been published to highlight the importance of statistics.[1-4] Greenhalgh[5,6] proposes asking the following questions, among others, as a first pass of any article: (1) Were the two groups evaluated for comparability at baseline? (2) Does the test chosen reflect the type of data presented (parametric vs nonparametric, categorical vs. numerical)? (3) Have the data been analyzed according to the original study protocol? (4) If an obscure test was used (essentially any test not mentioned in this review), was an explanation and a reference provided?

Corresponding author:
* Assistant Professor, Department of Forensic Medicine& Toxicology, Hamdard Institute of Medical Sciences & Research Delhi. India

Definitions

Mean

The mean is the arithmetic average of all the values in a set of data. It is often used to approximate the central tendency of a set of data but is very susceptible to outliers. For example, the mean income of Omaha, Nebraska, would likely be inordinately high because of Warren Buffett.

Median

The median is the middle value of a data set when the set is ordered chronologically. It is not influenced by outliers but does not give any

indication of the actual values of the numbers in a data set; it is not an average. It is therefore useful for small, highly skewed sets of numbers.

Distribution

When the values of a data set are plotted on a graph, the shape of the resultant curve defines the distribution. The normal (also called Gaussian) distribution is the classic bell curve, and in this distribution the mean = median (Figure 1).

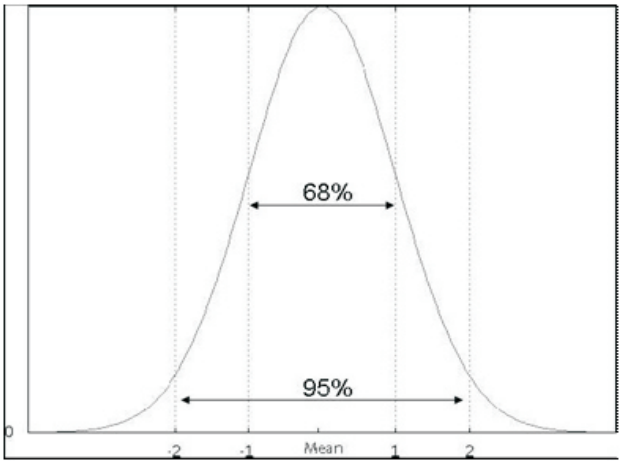

Figure 1: Normal (Gaussian) distribution.

Mean = 0 in this example. The 1 SD on either side of the mean encompasses 68% of all values under the curve, while 2 SD encompasses 95% of all values under the curve.

This allows several important mathematical assumptions to be made, allowing the use of statistical tests that are very sensitive. One of Fisher's assumptions states that the larger a sample size, the more closely the distribution will approximate a normal distribution. This is one mathematical reason why studies with large sample sizes are better.

Standard Deviation

The standard deviation (s or SD) denotes how far away from the mean an individual value lies. For a normally distributed data set, approximately 68% of the values will lie within 1 SD of the mean, and about 95% of the values will lie within 2 SDs (Figure 1).

P Value

The P value represents the probability that the observed outcome was the result of chance. Arbitrarily in the scientific and medical communities, $P < .05$ has been chosen as the cutoff for "statistically significant." This means that there is a less than 5% possibility (1 chance in 20) that the observed result was from chance alone. A P value outside the significant range could indicate one of two things. Either there is no "real" difference between the two sets of data, or the sample size was too small to detect a difference between the two sets, even if it exists. You cannot tell from the P value which of these possibilities is at fault, however.

Confidence Interval

A confidence interval (CI) is a range of values within which it is fairly certain that the true value lies. This is based on the idea that if you were to repeat the exactly the same study on random groups of subjects multiple times, you would not get the exact same results each time. You would have a range. For the purposes of interpreting the medical literature, 95% CI is talked about frequently. What this represents mathematically, for a given statistical result, is that we can be 95% certain that the true value lies within the range denoted by the CI (ie, within 2 SD). The narrower the range of the CI, the closer any observed value is to the true value, and the more precise the result. Confidence intervals can be positive or negative numbers, and this has no implications. If the 95% CI includes zero as well

as values of practical importance, however, then the result is not statistically significant, regardless of the P value.[7]This applies to inferential statistics and is slightly different for odds ratios, which are not discussed in this review. Mathematically, zero means there is no difference between the sample and the true population value. Therefore, the possibility of no difference between the 2 groups has not been excluded, despite $P < .05$. It is also important to remember that such calculations serve only to determine the mathematical validity of results; they do not determine the clinical utility of said results. Determining such utility is a fascinating topic unto itself and is well beyond this brief overview of statistics.

Types of Data

Numbers are assigned to many different things in the annals of research, but they fall broadly into two categories, each amenable to different statistical tests.

Numerical Data

The number itself has relevance. Examples of numerical data are things that are directly measured as a number such as CD4 counts, low-density lipoprotein levels, blood alcohol content, tumorsize etc,.

Categorical Data

Numbers that are assigned to non-numerical values of interest represent categorical data. Examples of categorical data include the city of residence or the sex of study participants.

Types of statistical tests;

There are two broad categories of statistical tests.

Parametric Tests

Parametric tests assume that sample data come from a set with a particular distribution, typically from a normal distribution. Generally, this requires a large sample size. Because the distribution is known (mathematically, because the shape of the curve is known), parametric tests are able to examine absolute differences between individual values in a sample and are more powerful. They are able to identify smaller differences than are nonparametric tests and should be used whenever possible.

Nonparametric Tests

Nonparametric tests make no such assumptions about the distribution of originating data and therefore must ignore absolute values of data points and focus instead on ordinal properties (eg, which is smallest, which is most common). It is more difficult to demonstrate statistical significance with a nonparametric test (ie, the difference between the two groups must be larger) than with a parametric test.

Tests

Chi-square

Mathematically speaking, the chi-square (x^2) test is a nonparametric test. Practically, it requires large sample sizes and should not be used when numbers are less than 20. The chi-square test is used with categorical data, and actual tally numbers must be used, not percentages or means. It tests the distribution of two or more independent data sets compared with a theoretical distribution. The more alike the distributions are, the more related they are determined to be, and the larger the chi-square

value. A value of $x^2 = 0$ implies there is no relationship between the samples.

Fisher Exact Test

Analogous to the chi-square test, the Fisher exact test is a nonparametric test for categorical data but can be used in situations in which the chi-square test cannot, such as with small sample sizes. This test is used when comparing two data sets that create a contingency table and tests the association (contingency) between the two criteria. This is observed when each data set has a "yes/no" answer, such as tumor cells present or absent, blood cultures positive or negative, breast cancer specimens that are estrogen receptor/progesterone receptor positive or negative, and so forth.

Student t Test

The Student t test is likely the most widely used test for statistical significance and is a parametric test suitable for either numerical or categorical data. The test compares the means of two data sets to determine if they are equal; if they are, then no difference exists between the sets. It exists as both a paired and unpaired test. A paired test means that the same thing was measured on each subject twice. For example, you measured each subject's heart rate, administered a beta blocker, then measured each person's heart rate again and want to compare the difference before and after administration of the drug. If this was not done, use an unpaired test.

Wilcoxon Signed Rank Test and Mann-Whitney U Test

The Wilcoxon signed rank test and Mann-Whitney U test are nonparametric analogs of the paired and unpaired t tests, respectively, in many situations. There are specific scenarios in which other tests are used as nonparametric analogs of the Student t test, such as when examining survival time or when examining categorical data, which are beyond the scope of this review. If these scenarios are encountered, it is advisable to examine closely the references listed in the study to determine the reason for the choice of test used. These tests analyze whether the median of the two data sets is equal or if the sample sets are drawn from the same population. As with all nonparametric tests, they have less power than the parametric counterpart (Student t test) but can be used with small samples or non-normally distributed data.

Analysis of Variance

Analysis of variance (ANOVA or F test) is a generalization of the Student t test (or Wilcoxon or Mann-Whitney U test) when three or more data sets are being compared. There are both parametric and nonparametric analyses of variance referred to as ANOVA by sum of squares or ANOVA by rank, respectively.

Pearson Product Moment Correlation Coefficient

Perhaps the most misused and misunderstood of all statistical analyses is the Pearson product moment correlation coefficient test. Fundamentally, correlation is a departure from independence of two or more variables. This can be in the form of any relationship, positive or negative. Correlation does not equal causation, nor does it imply causation; it merely records the fact that two or more variables are not

completely independent of one another. Pearson coefficient (r) is a parametric test that can be used with numerical or categorical data, but it is only meaningful under very select circumstances. For it to be a valid measure, all of the following four criteria must be met. (1) The data must be normally distributed. (2) The two data sets must be independent of one another. One value should not automatically vary with the other. For example, number of drinks consumed and blood alcohol level are not independent of one another; you must drink alcohol to change your blood alcohol level. (3) Only a single pair of measurements should be made on each subject. (4) Every r value calculated should be accompanied by a P value and/or CI.2

Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient test (rs or r) is the nonparametric counterpart to the Pearson coefficient and is a good option when all of the criteria cannot be met to calculate a meaningful r value and numerical data are being examined. As a nonparametric test, the degree of departure from independence will have to be greater to reach significance using this test than it would with the Pearson test. It is also more laborious to calculate, although in the modern era of statistical software this is largely irrelevant.

Regression Analysis

Regression quantifies the numerical relationship between two variables that are correlated. Essentially, it creates an equation wherein if one variable is known, the other can be estimated. This process, called extrapolation, should be done cautiously, and likewise interpreted cautiously. The numerous reasons for this are beyond the scope of this introduction.

Conclusion

A full time statistician is usually there in almost all medical colleges to help them in this regard but most of the time he/she would be so much overburdened to give sufficient time for every student to explain the importance of different types of tests available. The sole purpose of this article is to make the medical students aware about biostatistics.

References

1. Guyatt G, Jaeschke R, Heddle N, et al. Basic statistics for clinicians: 1. hypothesis testing. CMAJ. 1995;152(1):27-32.
2. Guyatt G, Jaeschke R, Heddle N, et al. Basic statistics for clinicians: 2. interpreting study results: confidence intervals. CMAJ. 1995;152(2):169-173
3. Guyatt G, Jaeschke R, Heddle N, et al. Basic statistics for clinicians: 3. assessing the effects of treatment: measures of association. CMAJ. 1995;152(3):351-357.
4. Guyatt G, Jaeschke R, Heddle N, et al. Basic statistics for clinicians: 4. correlation and regression. CMAJ. 1995;152(4):497-504.
5. Greenhalgh T. How to read a paper: statistics for the nonstatistician, I: different types of data need different statistical tests [erratum appears in BMJ. 1997;315(7109):675]. BMJ.1997;315:364-366.
6. Greenhalgh T. How to read a paper: statistics for the nonstatistician, II: "significant" relations and their pitfalls. BMJ. 1997;315:422-425.
7. Blume J, Peipert JF. What your statistician never told you about P-values. J Am AssocGynecolLaparosc. 2003;10(4):439-444.