Int J Emerg Med (2008) 1:197–199
DOI 10.1007/s12245-008-0061-z

CLINICAL RESEARCH CAPSULE

Understanding statistical tests in the medical literature: which test should I use?

M. Fernanda Bellolio · Luis A. Serrano · Latha G. Stead

Received: 30 July 2008 / Accepted: 31 July 2008 / Published online: 25 September 2008

When writing or reading articles, one should be aware whether the statistical tests performed were appropriate for the type of data collected and used, thereby avoiding misleading conclusions. The goal of all statistical tests is to determine whether two (or more) variables are associated with one another or independent from each other at the population level.

In this issue of IJEM, our Clinical Research Capsule reviews the most common tests used in published literature and some of the pitfalls associated with their use. This article is intended for non-statisticians.

One of the first things to keep in mind is the *type* of data and outcomes the author wants to measure and correlate. In order to do this, one must define the variables of the study. Are we looking at a continuous variable, one that can be quantified on an infinite scale, such as temperature or age, or is it a categorical variable, one that has to be grouped in classes. Categorical variables can be nominal or ordinal. Nominal variables are data that can be counted, but not ordered or measured. Nominal data can be further broken down into dichotomous (e.g., dead or alive) or have several categories (e.g., blood type). Ordinal data are numerical values that have a natural order and thus can be ranked and ordered. However, the distance between two values on an ordinal scale may not represent an equal degree of difference. For example, the modified Rankin score is a measure of outcome after stroke where a value of 0 is no functional deficit, and a value of 6 is dead. The difference between 0 and 1 is slight; however, the difference between a 2 and a 3 is very significant, as it distinguishes being functionally independent (2) versus dependent (3). Other examples of ordinal variables include birth order and pain severity scales.

The second important thing to keep in mind is how the results are *distributed*. Do they follow a "bell curve" (also called a Gaussian distribution), similar to biological phenomena and exam grading techniques, or do the results tend to cluster resulting in a skewed distribution? (Fig. 1). With normally distributed data, mean and SD are reported. For skewed data, median and interquartile ranges are reported.

Sometimes, it may be desirable to "normalize" skewed data. This is known as transformation. When data are skewed, the commonly applied transformations are $1/x$, log (x), and sqrt(x), exponentiating, squaring, or cubing x, where the x's are the data values.

Also to be considered is whether the data are *matched*—meaning are sample subjects or data points related to one another, or are they independent?

Once these three questions are answered, one is able to choose to the appropriate statistical test and, thus, decipher if an inappropriate test is used.

For two dichotomous or binary variables, one will be able to build a 2×2 table. If the data follow a normal

None of the authors have financial disclosures to make. All of the authors have seen and approved the manuscript and take full responsibility.

M. F. Bellolio (✉) · L. A. Serrano · L. G. Stead
Division of Research. Department of Emergency Medicine,
Mayo Clinic College of Medicine,
200 First Street SW, Generose Bld. G-410,
Rochester, MN 55905, USA
e-mail: bellolioavaria.mariafernanda@mayo.edu

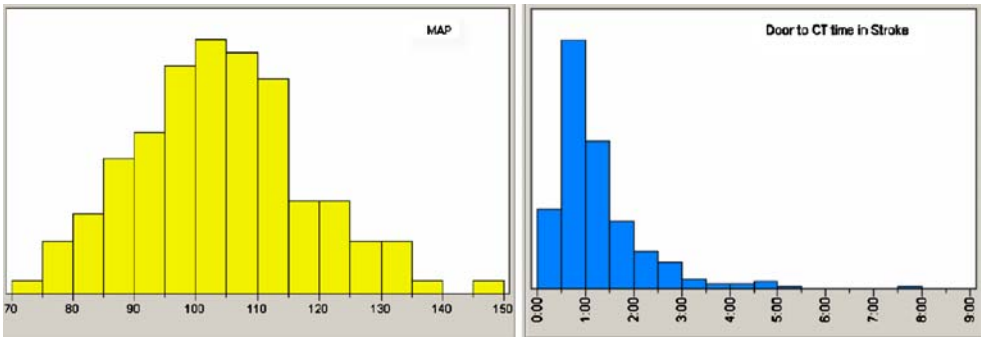

Fig. 1 *Left panel* demonstrates a normal distribution of mean arterial pressure (MAP) in patients with acute ischemic stroke; *right panel* demonstrate skewed distribution of door to computed tomography time in patients with acute ischemic stroke

distribution, the most common test will be Chi-square test. It is used to compare the proportion of subjects in two groups, and verify the independence of each other. For example, if a study about a certain treatment obtains data that shows that it reduces mortality more than placebo for a given disease, one would like to know if the results are true or merely a coincidence. Therefore, we perform a Chi-square test and obtain the *p* value. One limitation of the Chi-square testing is that its distribution breaks down as the frequencies decrease. If in one of the cells of your table there are five or less observations, the data is considered skewed. In this case, you need to use Fisher's exact test, specifically designed for small samples.

For two continuous variables (e.g., respiratory rate versus age), one can use linear regression or correlation. Linear regression allows us to predict the outcome for a particular value of the predictor. Correlation help us measure the association and direction between the variables.

In cases where the outcome or dependent variable (*Y* axis) is continuous (e.g., high blood pressure) and the independent variable (*X* axis) is binary (e.g., smoking yes/no); the distribution of the dependent variable will guide one in using (1) parametric tests [*t* test and analysis of variance (ANOVA)] for normally distributed data, or (2) nonparametric tests (Wilcoxon/Kruskal–Wallis or rank sum tests) for skewed data (Table 1).

In cases where the outcome or dependent variable (*Y* axis) is binary and the independent variable (*X* axis) is continuous, one should use logistic regression analyses.

When the data is matched (e.g., before and after measurement of a variable in the same patient), the appropriate test would be the McNemar test.

There are many sources of errors when selecting a statistical test. The first involves sources of bias. These are conditions or circumstances which affect the external validity of statistical results. The second are errors in methodology, which can lead to inaccurate or invalid results. The third are interpretation of results or how statistical results are applied to real world issues.

Common pitfalls:

1. Reporting the skewed data with mean and SD. Normally distributed data should be reported with mean (average) and SD or confidence intervals, and skewed data should be reported with median and interquartile ranges.
2. The study's overall statistical analyses were not performed to reject the null hypothesis.
3. If the investigator constructs a loose protocol and allows the experimenters to vary how they conduct the experimental procedures or interviews with different subjects, it is likely that the results of the experiment will be misleading.
4. The decision not to use the data was made after inspection of the results and without a predetermined rationale.
5. After an overall analysis had failed to reject the null hypothesis, the investigators perform a large number of new statistical tests on the data.

Table 1 Statistical test suggested

Parameter	Two independent samples	Two paired/ matched samples	Continuous and multiple predictors
Continuous outcome	2-sample *t* test	Paired *t* test	Linear regression
	Wilcoxon's rank sum test	Wilcoxon's rank test	ANOVA
Binary or categorical outcome	*Z* test and Chi-square test	McNemar's test	Logistic regression
	Relative risk	Sign test	
	Odds ratio		
	Fisher's exact test		

6. Do not take account of changing levels of significance when many statistical tests were performed on a single set of data (for example, perform 20 comparisons with one set of data)
7. Data omission—i.e., all the data in the analyses are not included (deleting patients with "inconvenient" results)
8. Conclusions cannot be derived form the results of the study
9. Reporting only *p* values. The mean, median, SD, confidence interval, relative risk, odds ratio, etc. should all be reported, to allow the reader to critique for him/herself the validity of the results. Look at magnitudes rather than *p* values.
10. Causal inference. Observational studies are very limited in their ability to show causal relationships. We will require a multifaceted approach to the research use of chronologically structured designs (placing variables in the roles of antecedents and outcomes) and ability of replication, to come to any conclusions regarding causality.
11. Precision and accuracy. Precision refers to how finely an estimate is specified; whereas accuracy refers to how close an estimate is to the true value. Estimates can be precise without being accurate.

There seems to be an erroneous notion that you can prove anything with statistics. However, this is only true if you use them *improperly.* Many times the data is overlooked, or the statistical test is not correctly selected. Always keep in mind that the simpler the experiment, the better will be its execution, and the more likely will one be able to see the "real truth" and what the results actually mean.

Conflicts of interest None.