Choosing and using statistical test

Introduction

Statistical tests are mathematical tools for analyzing quantitative data generated in a research study. The multitude of statistical tests makes a researcher difficult to remember which statistical test to use in which condition. There are various points which one needs to ponder upon while choosing a statistical test. These include the type of study design (which we discussed in the last issue), number of groups for comparison and type of data (i.e., continuous, dichotomous or categorical).

In the present article, we are going to discuss these points, but before that, let us go through some important concepts in statistics.

There are four main levels of measurement/types of data used in statistics. They have different degrees of usefulness in statistical research. Ratio measurements have both a meaningful zero value and the distances between different measurements defined; they provide the greatest flexibility in statistical methods that can be used for analyzing the data. Interval measurements have meaningful distances between measurements defined, but the zero value is arbitrary (as in the case with longitude and temperature measurements in Celsius or Fahrenheit). Ordinal measurements have imprecise differences between consecutive values, but have a meaningful order to those values. Nominal measurements have no meaningful rank order among values. Because variables conforming only to nominal or ordinal measurements cannot be reasonably measured numerically, sometimes they are grouped together as categorical variables, whereas ratio and interval measurements are grouped together as quantitative or continuous variables due to their numerical nature.

In the study of statistics, we focus on mathematical distributions for the sake of simplicity and relevance to the real world. Understanding these distributions will enable us to visualize the data easier and build models quicker. However, they cannot and do not replace the work of manual data collection and generating the actual data distribution. Distributions show what percentage of the data lies within a certain range. So, given a distribution and a set of values, we can determine the probability that the data will lie within a certain range. The same data may lead to different conclusions if they are interposed on different distributions. So, it is vital in all statistical analysis for data to be put onto the correct distribution.

P Value

Assume that there are data collected from two samples and that the means of the two samples are different. In this case, there are two possibilities: the samples really have different means or the other possibility is that the difference that is observed is a coincidence of random sampling. However, there is no way to confirm any of these possibilities.

All one can do is calculate the probabilities (known as “P” value in statistics) of observing a difference between sample means in an experiment of the studied sample size. The value of P ranges from zero to one. If the P value is small, then the difference is quite unlikely to be caused by random sampling, or in other words, the difference between the two samples is real. One has to decide this value in advance, i.e., at which smallest accepted value of P, the difference will be considered as real difference.

The statistical significance of a result is the probability that the observed relationship (e.g., between variables) or a difference (e.g., between means) in a sample occurred by pure chance (“luck of the draw”), and that in the population from which the sample was drawn, no such relationship or differences exist. Using less technical terms, we could say that the statistical significance of a result tells us something about the degree to which the result is “true” (in the sense of being “representative of the population”).More technically, the P value represents a decreasing index of the reliability of a result. The higher the P value, the less we can believe that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population. Specifically, the P value represents the probability of error that is involved in accepting our observed result as valid, i.e., as "representative of the population." For example, a P value of 0.05 (i.e., 1/20) indicates that there is a 5% probability that the relation between the variables found in our sample is a "fluke." In other words, assuming that in the population there was no relation between those variables, whatsoever, and we were repeating experiments such as ours one after another, we could expect that approximately in every 20 replications of the experiment there would be one in which the relation between the variables in question would be equal or stronger than in ours. In many areas of research, the P value of 0.05 is customarily treated as a "cut-off" error level.