Statistical Tests for Categorical Data

Bill Engels
Genetics Department
University of Wisconsin
Madison, WI 53706

13 February 1988

These programs allow you to do extremely powerful basic statistical tests on categorical data (data obtained by counting things).

I wrote these programs in order to analyze data from my own lab, but I added a standard Mac interface to make them available to other users. Weaker versions of some of these tests are available from commercial programs, but I found that I needed more power. I do not know of any commercial program on any system that does what these applications do.

All four programs have built-in documentation, so I will not go over them in detail here. I will simply provide a quick overview to indicate their functions. The programs have several features in common: (1) They avoid normal approximations, using only the actual distributions of the null hypotheses under test, (2) All computations are done in full SANE 80-bit precision, (3) Each program gives you the option of saving the full distribution of the statistic of interest to a text file, (4) They allow for text file input. Regular input from the keyboard is an option in Binomial Test and Fish6.

Binomial Test

If your data are thought to fit a binomial distribution, then you can use this program to (a) perform a one- or two-tail test of hypotheses concerning the parameter of the distribution, or (b) determine a confidence interval for the parameter. It can handle even very large numbers. The two-tail probabilities are obtained using the likelihood ratio to identify the "other" tail of the distribution. Thus the asymmetry of the binomial distribution is taken into account.

Fish6

Fisher's Exact Test is the best way to test for independence in a 2x2 contingency table. And Fish6 is the best way to do Fisher's Exact Test. It handles even very large tables with great precision and speed. Like Binomial Test, it orders possible outcomes on the basis of the likelihood ratio, thus giving you true two-tail tests.

Monte Carlo 2xN

Even a Mac (or a Cray for that matter) cannot compute fast enough to do the equivalent of Fisher's Exact Test for contingency tables much larger than 2x2. Therefore, if your need to test for independence in a table of N rows and 2 columns, and the expectations are too small to permit standard approximations, then you need to use a simulation ("Monte Carlo") procedure. Monte Carlo 2xN picks random 2xN tables with the same row and column sums as your observed table. It determines the probability of a table as "extreme" or more so than yours. The simulation is extremely fast, making this a practical alternative to standard but less reliable procedures.

Monte Carlo RxC

This is similar to Monte Carlo 2xN except it is generalized to allow an arbitrary number of rows (R) and columns (C). It can be used for 2xN tables, of course, but Monte Carlo 2xN is faster and more flexible for these cases.