Cornell University
Cornell Statistical Consulting Unit

StatNews #01

Sparse Contingency Tables
May 1996
Updated May 2012

With data obtained from both laboratory experiments and surveys, one is often confronted with frequency (or contingency) tables representing the occurrences of combinations of two categorical variables. A next step often involves testing the hypothesis of no association between the row variable and the column variable or testing the equality of proportions.

For example, one might be interested in the number of fish of different species caught in three different lakes.

	SALMON	**TROUT**	**BASS**
KEUKA LAKE	0	2	0
SENECA LAKE	0	4	0
CAYUGA LAKE	13	5	1

When requesting a Pearson chi-square test with such a table, to test the independence of the lake and bacteria variables, you might have been disconcerted by the warning message: "77.8% of the cells have expected counts less than 5. Chi-square may not be a valid test."

The reason for this message is that asymptotic methods, which invoke the Central Limit Theorem, are not reliable when the data are sparse, skewed or heavily tied. Previously a solution to this problem was found by collapsing categories of variables or by transforming the problematic variables. These methods are not ideal since they entail a loss of information that in some cases is of major interest.

Now we can rely on exact methods which will calculate p-values by constructing a reference set of all possible outcomes in which the exact null probability of each outcome is known. Then summing up the probabilities of those outcomes in the reference set that are at least as extreme as the one observed, will produce the exact p-value.

In the example above the asymptotic p-value equals 0.0244 whereas the exact method yields a p-value equal to 0.0861. The difference in the two values might in some cases lead you to quite different interpretations. Most standard statistical packages offer now some exact tests. In addition, specialized software is available to implement exact p-values and confidence intervals for a wide range of tests. StatXact is the best known package. It is very easy to use and allows you to obtain your desired answer in less than 5 minutes.

Along the same line we also have LogXact which will present, besides maximum likelihood estimation of the parameters in logistic regression, exact inference which will yield exact p-values and confidence intervals for the effect of covariates. Exact logistic regression is also now available in some commonly used statistical software packages (e.g. Stata and SAS). See also newsletter #82 for another method to fit logistic regression with sparse data.

If you experience a need for these methods, please contact the Cornell Statistical Consulting Unit.

Author: Francoise Vermeylen (fmv1@cornell.edu)

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).