Appendix to "2^k Contingency Tables in Ecology", by E. C. Pielou

APPENDIX: STATISTICAL TESTS OF SIGNIFICANCE

D. S. Robson
Cornell University

BU-351-M January, 1971

Abstract

E. C. Pielou has proposed a new measure of association between the species occurring in a sample of N quadrats. When occurrence is measured by presence or absence the data may be displayed in an N × k matrix of zeroes and ones, the i^{th} row sum s_i then denoting the number of different species occurring in the i^{th} quadrat, and the j^{th} column sum n_j denoting the number of quadrats in which the j^{th} species occurs. If the k species occur mutually independently then the conditional variance of the row sum s_i, given the column sums $n_j = Np_j$, $j=1,\cdots,k$, becomes

$$\mathrm{Var}(s) = \sum_{j=1}^{k} p_j(1-p_j)$$

and Pielou has proposed that the difference

$$\frac{\nu(s)}{N} = \frac{1}{2}[m_2(s)-\mathrm{Var}(s)]$$

between the observed row variance

$$m_2(s) = \frac{1}{N}\left[\sum_{i=1}^{N} s_i^2 - \frac{1}{N}\left(\sum_{1}^{N} s_i\right)^2\right]$$

and the variance predicted under the independence hypothesis be taken as a measure of dependence. The reduction in ν/N obtained by deleting a specified subset (Y) of species from the data matrix was further proposed as a statistic for testing the relation between Y and the remaining (X) species.

Biometrics Unit, Cornell University, Ithaca, New York.

This reduction in ν/N is algebraically expressible as

$$\frac{\nu(x+y)}{N} - \frac{\nu(x)}{N} = \frac{\nu(y)}{N} + m_{11}(x,y)$$

where $\nu(y)/N$ measures association <u>within</u> the Y-class and the product-moment $m_{11}(x,y)$ measures association <u>between</u> the Y-class and the X-class. Separate, independent, large sample tests of these two different aspects of Pielou's null hypotheses are here derived in the form:

$$\frac{m_2(y)}{\mathrm{Var}(y)}(N-1) \sim \chi^2_{(N-1)\text{d.f.}} \qquad r^2_{xy}(N-1) \sim \chi^2_{1\ \text{d.f.}} \quad .$$

Appendix to "2^k Contingency Tables in Ecology", by E. C. Pielou

APPENDIX: STATISTICAL TESTS OF SIGNIFICANCE

D. S. Robson
Cornell University

BU-351-M January, 1971

If the collection of k species is partitioned into two classes, say the X-class and the Y-class, then the total number of species s_i appearing in the i^{th} quadrat is expressible as the sum $s_i = x_i + y_i$ of the numbers in the two classes. The statistic ν/N calculated from the complete data matrix, say $\nu(x+y)/N$, is then related to the corresponding statistics of the X- and Y-data matrices by the equation

$$\frac{\nu(x+y)}{N} = \frac{\nu(x)}{N} + \frac{\nu(y)}{N} + m_{11}(x,y)$$

where the product moment is defined by

$$m_{11}(x,y) = \frac{1}{N}\left[\Sigma xy - \frac{1}{N}(\Sigma x)(\Sigma y)\right] \quad .$$

The reduction resulting from deletion of one class of species, say the Y-class,

$$\frac{\nu(x+y)}{N} - \frac{\nu(x)}{N} = \frac{\nu(y)}{N} + m_{11}(x,y)$$

will thus have expectation zero if each species in the Y-class is not only independent of all species in the X-class, implying $m_{11}(x,y)$ has mean zero, but also independent of all other species in the Y-class, implying $\nu(y)/N$ has mean zero. A hypothesis of no expected reduction is thus a composite of two null hypotheses, each of which is separately and independently testable.

If the Y-class is independent of the X-class then m_{11} has mean zero even when the expectation is conditioned on one of the order statistics, say

$$\{y\} = \{y_i | i=1,2,\cdots,N\} = \text{unordered collection of N y's .}$$

Thus,

$$E_{H_o}[m_{11}(x,y)|(x),\{y\}] = 0$$

where the average is taken over the N! equally likely permutations of (y). For example, if the Y-class contains a single species then $y_i = 0$ or 1 and

$$m_{11}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x}_N)y_i = \frac{n}{N}(\bar{x}_n-\bar{x}_N)$$

where $\bar{x}_n$ is the mean value of x in those n quadrats containing the Y-species. The permutation distribution of m_{11} in this case is, in effect, the distribution of a mean of a sample of size n drawn randomly and without replacement from a finite population of size N. The normal approximation to this distribution gives

$$\frac{\left[\frac{n}{N}(\bar{x}_n-\bar{x}_N)\right]^2}{\left(\frac{n}{N}\right)^2\frac{N-n}{n(N-1)}m_2(x)} \sim \chi^2_{1\ \text{d.f.}} \quad .$$

In the more general case we have

$$\text{Var}\,[m_{11}(x,y)|(x),\{y\}] = \frac{m_2(x)m_2(y)}{N-1}$$

and

$$\frac{[m_{11}(x,y)]^2}{m_2(x)m_2(y)}(N-1) = r^2_{xy}(N-1) \sim \chi^2_{1\ \text{d.f.}} \quad .$$

If the X- and Y-classes both contain a large number of species then the distribution of r^2_{xy} derived from bivariate (independent) normal theory might provide a better approximation than this chi-square distribution on one degree of freedom, but when one class is small the above approximation is the better of the two.

The statistical significance of $\nu(y)/N$ for the k_Y species in the Y-class can be tested in a 2^{k_Y} contingency table when k_Y is small, and can be tested by appealing to the Central Limit Theorem when k_Y is large. A row total y_i (see Table 1) is the sum of k_Y indicator variables, say

$$y_i = \sum_{j=1}^{k_Y} \delta_{ij}$$

which are independent under the null hypothesis, and hence as k_Y gets large the probability distribution of y_i approaches a normal distribution. The same may be said of the column totals

$$n_j = \sum_{i=1}^{N} \delta_{ij}$$

as N gets large, and hence the conditional distribution of the row totals, given the column totals, must approach multivariate normality. Since the conditional moments are

$$E_{H_o}(y_i|n_1,\cdots,n_{k_Y}) = \sum_{j=1}^{k_Y} E_{H_o}(\delta_{ij}|n_j) = \sum_{j=1}^{k_Y} \frac{n_j}{N} = \sum_{j=1}^{k_Y} p_j$$

$$\mathrm{Var}_{H_o}(y_i|n_1,\cdots,n_{k_Y}) = \sum_{j=1}^{k_Y} \mathrm{Var}_{H_o}(\delta_{ij}|n_j) = \sum_{j=1}^{k_Y} p_j(1-p_j)$$

$$\mathrm{Cov}_{H_0}(y_i, y_{i'} | n_1, \cdots, n_{k_Y}) = \sum_{j=1}^{k_Y} \mathrm{Cov}(\delta_{ij}, \delta_{i'j} | n_j) = \frac{-1}{N-1} \sum_{j=1}^{k_Y} p_j(1-p_j)$$

then the quadratic form of any N-1 y's, say $\underset{\sim}{y} = (y_1, \cdots, y_{N-1})$, reduces to

$$(\underset{\sim}{y} - E(\underset{\sim}{y}|\underset{\sim}{n}))V^{-1}(\underset{\sim}{y} - E(\underset{\sim}{y}|\underset{\sim}{n}))' = \frac{m_2(y)}{\sum_{j=1}^{k_Y} p_j(1-p_j)} (N-1)$$

where V is the conditional covariance matrix of $\underset{\sim}{y}$. For large k_Y this test statistic

$$\frac{m_2(y)}{\mathrm{Var}(y)}(N-1) = \left[1 + \frac{2}{\mathrm{Var}(y)} \frac{\nu(y)}{N}\right](N-1) \sim \chi^2_{N-1}$$

is therefore approximately chi-square distributed on N-1 degrees of freedom, and is asymptotically independent of $\chi^2_1 \sim r^2_{xy}(N-1)$ since the distribution of the latter is conditional on $\{y\}$ and hence on $m_2(y)$. These two chi-squares could therefore be added to provide a test of the composite null hypothesis, but the separate tests of the two facets of this null hypothesis are more informative.