Abundance-Ubiquity Method (in the plasticity of bacterial communities paper)

J. D. Nulton

October 15, 2013

Abstract

Note: Some edited version of this material should be included in an appendix, if the abundance-ubiquity method is presented in the text.

The abundance-ubiquity figure in the text shows (1) an expectation curve and (2) a 0.01 significance cutoff curve. Both are based on the hypothesis H_0 as discussed in the text. The content below develops the theory that links the null-hypothesis H_0 and the test statistics $u = \textit{ubiquity}$ and $a = \textit{abundance}$ and gives the explicit origin of those curves.

Consider a large population of individuals classified into OTUs. Suppose a particular OTU, Q, has a relative abundance a within the population. Q will be considered fixed for purposes of the analysis. Suppose further that m samples of sizes $n_1, n_2, \ldots, n_m$ are taken from the population. Note that the assumption of one population embodies the null-hypothesis H_0 for the actual sampling experiment. Our goal is to derive an expression for the probability that Q will be represented by at least one individual in exactly k of the samples. In that event, as defined earlier, Q's ubiquity would be k/m. Let K be the random variable whose probability distribution we are seeking, i.e. let K count the number of samples in which Q shows up at least once. We will obtain K's distribution indirectly by first finding K's expectation and variance: $E[K]$ and $Var[K]$. Once these moments are obtained, the distribution is derived.

The random variable K is analyzed with the aid of the following events: E_j: "Q is observed at least once in the jth sample." Let X_j denote the random variable characteristic of E_j, i.e. X_j takes the value 1 if E_j occurs, and the value 0 otherwise. Then K is given by

$$K = \sum_{j=1}^{m} X_j. \tag{1}$$

This reduces the problem to the events E_j and the variables X_j. A randomly selected individual is Q with probability a. The probability that Q is *not* observed among n_j randomly selected individuals, i.e. that E_j does *not* occur, is $(1-a)^{n_j}$. Therefore the probability that E_j *does* occur is

$$p_j = 1 - (1-a)^{n_j}. \tag{2}$$

This is also the expectation of the variable X_j. Consequently we have

$$E[X_j] = p_j, \tag{3}$$

$$Var[X_j] = E[X_j^2] - E[X_j]^2 = E[X_j] - E[X_j]^2 = p_j(1-p_j). \tag{4}$$

and, by Eq. (1),

$$E[K] = \sum_{j=1}^{m} p_j, \qquad Var[K] = \sum_{j=1}^{m} p_j(1-p_j). \tag{5}$$

It remains to find the cumulative distribution for K, i.e. $Prob\{K \leq k\}$, where k is the number of samples in which Q showed up. This places Q in the distribution predicted by the H_0.

Consider the case in which all m samples have the same size n and $p_j = p$ for all samples. In that case $E[K] = mp$ and $Var[K] = mp(1-p)$, and it is easy to show that the distribution for K is binomial, and

$$Prob\{K \leq k\} = \sum_{i=0}^{k} \binom{m}{i} p^i (1-p)^{m-i}. \tag{6}$$

In the general case where the sample sizes are different, the best we can do is to match the moments to a binomial producing equations

$$m^* p^* = \sum_{j=1}^{m} p_j, \qquad m^* p^* (1-p^*) = \sum_{j=1}^{m} p_j(1-p_j), \tag{7}$$

which have the solutions

$$p^* = \frac{\sum_{j=1}^m p_j^2}{\sum_{j=1}^m p_j}, \qquad m^* = \frac{(\sum_{j=1}^m p_j)^2}{\sum_{j=1}^m p_j^2}. \tag{8}$$

The right of (6) is just $1 - f(p, k+1, m-k)$, where f denotes *the incomplete beta function*, whose integral representation [1] is defined for *non*-integer values of k and m. So finally a satisfactory approximation to the cumulative distribution can be written

$$Prob\{K \leq k\} \simeq f(p^*, k+1, m^* - k). \tag{9}$$

Finally, in terms of the quantities discussed here, the curves superimposed on the scattergram in the figure are obtained as follows. The coordinate system for the figure is (u, a), so the curves represent a relation between these coordinates.

The expectation curve is based on

$$u = E[K]/m, \tag{10}$$

where we see from (5) and (2) that $E[K]$ is ultimately a function of a.

The 0.01 significance curve is based on

$$Prob\{K \leq k\} = f(p^*, k+1, m^* - k) = 0.01, \tag{11}$$

where we see from (8) and (2) that p^* and m^* are ultimately functions of a, and $k = mu$.

References

[1] Feller, W., *An Introduction to Probability Theory and its Applications, Vol. 1*, Wiley (1950).