BIOMETRICS 42, 171–176
March 1986

# Estimation of Simpson's Diversity When Counts Follow a Poisson Distribution

**N. I. Lyons and K. Hutcheson**

Department of Statistics and Institute of Ecology,
University of Georgia, Athens, Georgia 30602, U.S.A.

## SUMMARY

Moments of Simpson's index of diversity are derived assuming species frequencies follow underlying Poisson distributions. The Pearson curves are used to approximate 95% confidence limits on the population diversity. The results are compared to confidence interval estimation using the jackknife procedure in a simulation study for small samples.

## 1. Introduction

Most of the exact distributional properties available for Gini's (1912) index of diversity [more commonly referred to as Simpson's (1949) index] are derived assuming an underlying multinomial distribution of species frequencies, with simple random sampling. In one sense this is the least restrictive model in that it assumes only that individuals fall into $s$ categories, or species, at random with probabilities $p_i$, $i = 1, \ldots, s$. However, the model also assumes that the total sample size is fixed. Questions arise as to the validity of using these distributional properties in practical situations in which total sample size is random.

The fixed sample size restriction may be relaxed by assuming that each species frequency follows an underlying Poisson distribution with mean $\mu_i$, $i = 1, \ldots, s$. In fact, the joint distribution of independent species frequencies, conditioned on a fixed total sample size, is multinomial with $p_i = \mu_i/\sum \mu_i$, $i = 1, \ldots, s$. In this paper a method for computing the moments of the index under the Poisson assumptions is given and its asymptotic distribution is derived. The adequacy of the Pearson curve fit to the distribution of the index is examined with respect to confidence interval estimation. The results are used to assess the effect of incorrectly assuming a fixed sample size when in fact it is random. Quadrat sampling is introduced because it is more commonly used in practice, and it allows comparison of confidence interval estimation results with two other commonly used methods.

## 2. Moments

Suppose $q$ quadrats are placed randomly over a spatial area containing $s$ species. Let $n_{ij}$ represent the number of individuals of species $i$ in quadrat $j$. If individuals of a species are distributed at random over the area, then $n_{ij}$ has a Poisson distribution, say $P(\mu_i)$. If the $n_{ij}$ are independent, then the species counts, $n_i = \sum_j n_{ij}$, and the total count, $N = \sum_i n_i$, are $P(q\mu_i)$ and $P(q \sum \mu_i)$, respectively.

Stiteler and Patil (1969) point out that Simpson's (1949) diversity index

$$d_U = 1 - \sum n_i(n_i - 1)/[N(N - 1)],$$

is an MVU estimator of the population diversity, $1 - \sum p_i^2$, for any arbitrary distribution

*Key words:* Confidence intervals; Jackknife; Moments; Pearson curves.

of $N$ truncated at 0 and 1, provided the conditional distribution is multinomial. Assuming a Poisson distribution, they present tables of approximations to the expectations of functions of $N$ required for estimating the unconditional variance of $d_{\mathrm{U}}$ (although an error was made in simplification).

The biased version of Simpson's index $d_{\mathrm{B}} = 1 - \sum (n_i/N)^2$ is related to $d_{\mathrm{U}}$ by

$$d_{\mathrm{B}} = [(N-1)/N]d_{\mathrm{U}}.$$

Since $d_{\mathrm{U}}$ is unbiased, $\mathrm{E}(d_{\mathrm{U}} \mid N) = E(d_{\mathrm{U}})$, and the iterative property of conditional expectations can be used to obtain the $k$th moment of $d_{\mathrm{U}}$, expressed in terms of the conditional moments of $d_{\mathrm{B}}$, by

$$\mu_k(d_{\mathrm{U}}) = \mathrm{E}_n\left[\left(\frac{N}{N-1}\right)^k \mu_k(d_{\mathrm{B}} \mid N)\right]. \tag{1}$$

The first four conditional moments of $d_{\mathrm{B}}$ are given by Lyons and Hutcheson (1979). Computation of $\mathrm{E}[N^{k-r}/(N-1)^k]$ for values of $k = 1, 2, 3, 4$, and $r = 1, \ldots, 7$, is required. This expectation for the 0–1 truncated distribution of $N$ with parameter $\lambda = q \sum \mu_i$ is given by

$$\mathrm{E}[N^{k-r}/(N-1)^k] = \frac{1}{C} e^{-\lambda} \sum_{N=2}^{\infty} \frac{\lambda^N}{N!} N^{k-r}/(N-1)^k,$$

where $C = 1 - \Pr(N=0) - \Pr(N=1)$. Since $N^{k-r}/(N-1)^k$ is decreasing in $N$, the series may be truncated at some value, say $N_0$, for approximations as follows:

$$\mathrm{E}[N^{k-r}/(N-1)^k] \approx \frac{1}{C} e^{-\lambda} \sum_{N=2}^{N_0} \frac{\lambda^N}{N!} N^{k-r}/(N-1)^k, \tag{2}$$

and the truncation error, $e_0$, can be bounded by

$$0 < e_0 < N_0^{k-r}/(N_0-1)^k[1 - \Pr(2 \leq N \leq N_0)].$$

Since the distribution of $N$ depends only on the quantity $q \sum \mu_i$, the moments are invariant to scale in the $\mu_i$'s in the sense that, if the $\mu_i$'s are increased by a common factor, $q$ may be reduced by the same factor without affecting the moments of $d_{\mathrm{U}}$. In practice this implies that if quadrat size is doubled, the number of quadrats may be cut in half and $d_{\mathrm{U}}$ has the same distributional properties.

## 3. Asymptotic Results

The asymptotic distributions of $d_{\mathrm{U}}$ and $d_{\mathrm{B}}$ are the same since $d_{\mathrm{U}} = [N/(N-1)]d_{\mathrm{B}}$ and the coefficient $N/(N-1)$ converges in probability to 1. Using Taylor series expansions of functions of asymptotically normal random variables (see, e.g., Serfling, 1980, p. 124), one can show that $d_{\mathrm{U}}$ and $d_{\mathrm{B}}$ are asymptotically

$$\mathrm{N}\left(1 - \sum_{i=1}^{s}\left(\frac{\mu_i}{\sum \mu_i}\right)^2, \frac{1}{q}\sum_{i=1}^{s}\left[\frac{4\mu_i(\mu_i \sum \mu_i - \sum \mu_i^2)^2}{(\sum \mu_i)^6}\right]\right)$$

in general. If $\mu_i = \mu$ for all $i$, $d_{\mathrm{U}}$ and $d_{\mathrm{B}}$ are asymptotically $[1/(s^2\mu)]\chi^2_{s-1}$. These are the same limits obtained for the multinomial case with $N$ fixed.

Both indices can be expressed as functions of $s$ $U$-statistics, each of order $m = 1$. Therefore, if $\sum \mu_i$ is held fixed and $q$ is allowed to tend to infinity, the results of Callaert and Janssen (1978) can be used to show that the rate of convergence to normality for general underlying distributions is of order $O(q^{-1/2})$. In practice this implies that the error incurred in using the normal distribution for approximating probabilities concerning these indices would be on the order of magnitude of $1/\sqrt{q}$ for a large number of quadrats.

Values of skewness, $\beta_1$, and kurtosis, $\beta_2$, were computed using equations (1), (2), and the results of Lyons and Hutcheson (1979), to study the actual rate of convergence to the asymptotic distribution. Three models for the $\mu_i$'s were used: the equiprobable model, the broken stick model, and a model based on the geometric series. Without loss of generality, the value of $\sum \mu_i$ was set to 1. In all three cases $\beta_1$ and $\beta_2$ for the Poisson model differ only slightly from those of the multinomial fixed sample size model for $q > 20$. For the equiprobable case with $2 \leqslant s \leqslant 100$, the chi-square approximation is good for small $q$ ($q \sum \mu_i > 85$). With the broken stick model, the rate of convergence to normality slows as $s$ increases, but is good for smaller $s$. For this model larger values of $s$ result in flatter distributions of abundances, approaching the equiprobable case. The equiprobable case tends to the $\chi^2$ limiting distribution, but the index is still in the normal asymptotic region as long as the $\mu_i$'s are not exactly equal.

For the geometric series the $\mu_i$'s are determined by

$$\mu_i = \delta^{i-1} \Big/ \sum_{k=1}^{s} \delta^{k-1}, \quad i = 1, \ldots, s.$$

In this model $\mu_{i+1} = \delta\mu_i$, $i = 1, \ldots, s - 1$, where $0 < \delta \leqslant 1$. Controlling the value of $\delta$ allows the distribution of the $\mu_i$'s to range continuously from the equiprobable case ($\delta = 1$) to one with many rare species ($\delta$ close to 0). In this model increasing $s$ beyond 15 does not change the values of $\beta_1$ and $\beta_2$ significantly. This is due to the fact that the means of the rarest species are practically negligible as soon as $\delta$ departs from 1 for large $s$, and is also a reflection of the relative insensitivity of Simpson's index to rare species. Convergence to normality is fast for $\delta$ small ($0 < \delta \leqslant .4$) and slows steadily as $\delta$ increases, regardless of $s$. For the same number of quadrats the normal approximation is better for skewed distributions of species abundances than for nearly equal abundances.

## 4. Confidence Interval Estimation

Given either exact or estimated values of $\beta_1$ and $\beta_2$, the Pearson curves can be used to approximate the distributions of $d_{\mathrm{U}}$ and $d_{\mathrm{B}}$. A simulation study was conducted to determine the adequacy of the Pearson curves to approximate 95% confidence limits on $1 - \sum p_i^2$ using only the geometric model. The limits were obtained by substituting the conditional maximum likelihood estimates, $n_i/N$, for $p_i$, and the observed sample size, $N$, for $q \sum \mu_i$ in expressions for the moments.

The percentages of confidence limits covering $1 - \sum p_i^2$ were recorded. In addition, the percentages of observed values of $d_{\mathrm{U}}$ and $d_{\mathrm{B}}$ falling between the 2.5 and 97.5 percentage points of the selected Pearson distribution based on the true values of $\beta_1$ and $\beta_2$ were tabulated.

For comparison the jackknife procedure [suggested by Zahl (1977)], Odum's (1971) procedure, and the multinomial approximation were used. For Odum's procedure, values of $d_{\mathrm{U}}$ and $d_{\mathrm{B}}$ are computed on each quadrat and the variance of these values is used to estimate the variance of the index based on pooled quadrat data. The $t$-distribution with $q - 1$ degrees of freedom was assumed for the jackknife and Odum methods. The Pearson curves were used for the multinomial approximation using the moments in Lyons and Hutcheson (1979).

Preliminary results indicated that the central 95% probability limits based on the true values of $\beta_1$ and $\beta_2$ differed only in the third significant digit for the multinomial and Poisson models for $q \sum \mu_i$ greater than 20. The value of $s$ had little effect beyond $s = 10$. In no case was the biased estimator as good as $d_{\mathrm{U}}$, even with the jackknife procedure.

Subsequently, 500 iterations were made using the geometric model at each of four values of $q \sum \mu_i$ (5, 10, 15, 20), four values of $\delta$ (.2, .5, .8, 1.0), and three values of $q$ (3, 5, 10) for $s = 5$. Only the index $d_U$ was considered. The computation was quite time-consuming because each iteration required four calls to a sequence of subroutines (Bouver and Bargmann, Technical Report No. 110, Department of Statistics, University of Georgia, 1975) to determine and evaluate percentage points of the appropriate Pearson curve.

## 5. Conclusion

For the Poisson model, the Pearson distribution limits based on true values of $\beta_1$ and $\beta_2$ were shifted too far to the left of the actual distribution for all values of $\delta$ with $q \sum \mu_i = 5$. For $q \sum \mu_i > 5$ the limits were satisfactory for $\delta = 1.0$ and $\delta = .2$. There was one rejection for $\delta = .5$ with $q \sum \mu_i = 15$ ($.025 < P\text{-value} < .05$) and one for $\delta = .8$ and $q \sum \mu_i = 10$ ($.025 < P\text{-value} < .05$). As expected, incorrectly assuming a fixed sample size (multinomial model) resulted in limits which were narrower in all cases. The error was significant for nearly all combinations of values of $\delta$ and $q \sum \mu_i$. For both Odum's and the jackknife methods, the estimate of variance was zero in some cases for small sample sizes. These cases were counted as falling in the appropriate tail areas. For the multinomial approximation a few samples resulted in estimated values of skewness and kurtosis falling out of the Pearson range. These cases were not counted in the simulation, resulting in slightly fewer than 500 iterations being recorded.

The confidence limits based on estimating $\beta_1$ and $\beta_2$ from the generated data were too wide for all values of $\delta$ with $q \sum \mu_i = 5$. For the three cases $\delta = .5$, .8, and 1.0, the limits were shifted too far to the left for all values of $q \sum \mu_i$. The multinomial approximation provided limits which were also shifted too far to the left and too narrow. In these three cases the jackknife procedure produced the best limits for $q \sum \mu_i > 5$ in the sense of percent coverage and location; however, it tended toward overcoverage in the flatter distributions ($\delta = .8$, 1.0) and undercoverage for $\delta = .5$. Odum's method in almost all cases produced extremely narrow limits due to underestimation of variance. This caused the limits to vary considerably with respect to shift from case to case.

The results were the best overall for $\delta = .2$. This case is presented as an example in Table 1. For this table $q \sum \mu_i$ was extended to include 50 and 100 to illustrate that the four methods differ very little for large sample sizes. The percent coverage is given in decimal form. The entries in parentheses indicate numbers of confidence intervals falling outside (above/below) the population value of Simpson's index.

The values of the $p_i$'s are very skewed. Since Simpson's index is rather insensitive to rare species, the model for $\delta = .2$ is essentially reduced to a two- or three-species model. The Pearson curve approximation was poor for small to moderate mean sample sizes ($5 \leqslant q \sum \mu_i \leqslant 50$). Only when $q \sum \mu_i$ reached 100 did the procedure do well. Then the true values of $\beta_1$ and $\beta_2$ are 0.017 and 2.984, respectively, for the Poisson model, and 0.017 and 2.951 for the fixed sample size approximation. Thus, the normal distribution can be assumed with mean and variance estimated using Simpson's fixed sample size results. The jackknife procedure did well in this case for as few as three quadrats with $q \sum \mu_i = 10$. The adequacy of the results for the jackknife appear to depend more on the mean sample size per quadrat, than on the number of quadrats.

Clearly the sample sizes for this simulation are extremely small per quadrat, and erratic behavior is expected. For reasonably large sample sizes the fixed sample multinomial model may as well be assumed; however, for moderate sample sizes the jackknife provides the best alternative. The only advantage in using the Pearson curve approximation appears to be that quadrat sampling is not necessary. In the Poisson model the effect of subsampling

**Table 1**

*Percent coverage of* 95% *confidence intervals with* $s = 5$, $\delta = .2$, *and selected values of* $q$ *and* $q \sum \mu_i$ (500 *iterations*);
$p_1 = .800$, $p_2 = .160$, $p_3 = .032$, $p_4 = .006$, $p_5 = .001$

| | | Quadrats ($q$) | | |
|---|---|---|---|---|
| $q \sum \mu_i$ | Method | 3 | 5 | 10 |
| 5 | Poisson | 0.986 (0/7)* | 1.000 (0/0)* | 0.996 (2/0)* |
| | Mult. | 0.724 (0/71)* | 0.842 (3/0)* | 0.864 (2/0)* |
| | JK | 0.938 (28/3)* | 0.968 (16/0)* | 0.926 (37/0)* |
| | Odum | 0.848 (16/60)* | 0.656 (0/172)* | 0.074 (0/436)* |
| 10 | Poisson | 0.938 (0/31)* | 0.986 (6/1)* | 0.996 (2/0)* |
| | Mult. | 0.848 (0/73)* | 0.974 (11/1)* | 0.982 (4/1)* |
| | JK | 0.956 (11/11) | 0.940 (30/0)* | 0.960 (20/0)* |
| | Odum | 0.926 (28/9)* | 0.878 (2/59)* | 0.344 (0/328)* |
| 15 | Poisson | 0.940 (0/30)* | 0.960 (8/12) | 0.982 (5/4)* |
| | Mult. | 0.922 (0/39)* | 0.940 (12/18) | 0.974 (7/6)* |
| | JK | 0.962 (9/10) | 0.938 (30/1)* | 0.952 (21/3)* |
| | Odum | 0.960 (17/3)* | 0.928 (2/34)* | 0.580 (0/210)* |
| 20 | Poisson | 0.944 (0/28)* | 0.948 (9/17) | 0.946 (7/20)* |
| | Mult. | 0.914 (0/43)* | 0.942 (11/18) | 0.928 (9/27)* |
| | JK | 0.964 (12/6) | 0.978 (10/1)* | 0.932 (22/12)* |
| | Odum | 0.958 (19/2)* | 0.938 (9/22)* | 0.730 (0/135)* |
| 50 | Poisson | 0.936 (11/21)* | 0.938 (10/21)* | 0.932 (10/24)* |
| | Mult. | 0.936 (11/21)* | 0.936 (10/22)* | 0.930 (11/24)* |
| | JK | 0.948 (18/8) | 0.950 (10/15) | 0.946 (14/13) |
| | Odum | 0.940 (22/8)* | 0.944 (18/10) | 0.928 (11/25)* |
| 100 | Poisson | 0.948 (11/15) | 0.954 (7/16) | 0.930 (10/25)* |
| | Mult. | 0.948 (11/15) | 0.948 (9/17) | 0.926 (11/26)* |
| | JK | 0.948 (13/13) | 0.948 (12/14) | 0.936 (13/19) |
| | Odum | 0.944 (15/13) | 0.950 (12/13) | 0.932 (20/14) |

* Indicates observed counts significantly different from expected ($\alpha = .05$) using a standard $\chi^2$ test.

can be accomplished by randomly dividing a sample into subsamples. Then the jackknife appears to provide the best alternative for small sample sizes from populations with skewed species abundances for as few as three quadrats, provided the mean sample size per quadrat is reasonably large. It should be pointed out, however, that the species frequencies are randomly distributed and independent. Heltshe and Forrester (1985) study the behavior of the jackknife procedure using quadrat sampling generating clumped populations and different levels of skewed species distributions. In these situations they observed definite overcoverage with 95% confidence limits.

## ACKNOWLEDGEMENTS

This research was supported by NSF Grant No. DEB 82-01564 through the Institute of Ecology at the University of Georgia. The authors wish to thank the referees for their suggestions and comments.

## RÉSUMÉ

On calcule les moments de l'indice de Simpson sous l'hypothèse que les fréquences des différentes espèces suivent des distributions de Poisson. On utilise les courbes de Pearson pour approcher les limites de confiance à 95% de la diversité de la population. Ces résultats sont confrontés à l'estimation du mîeme intervalle de confiance simulé par une procédure jackknife dans le cas de petits exemples.

## References

Callaert, H. and Janssen, P. (1978). The Berry–Esseen theorem for U-statistics. *Annals of Statistics* **6,** 417–421.

Gini, C. (1912). Variabilité e mutabilitá. *Studi Economico-Giuridici Fac. Giurisprudenza Univ. Cagliari, A,* **III,** Parte II, 3–159.

Heltshe, J. F. and Forrester, N. E. (1985). Statistical evaluation of the jackknife estimate of diversity when using quadrat sampling. *Ecology* **66,** 107–111.

Lyons, N. I. and Hutcheson, K. (1979). Distributional properties of Simpson's index of diversity. *Communications in Statistics, Series A* **8,** 569–574.

Odum, E. P. (1971). *Ecology.* New York: Holt, Reinhart & Winston.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

Simpson, E. H. (1949). Measurement of diversity. *Nature* **163,** 688.

Stiteler, W. M. and Patil, G. P. (1969). Variance-to-mean ratio and Morisita's index as a measure of spatial patterns in ecological populations. In *Statistical Ecology: Spatial Patterns and Statistical Distributions,* J. F. Grassle, G. P. Patil, W. Smith, and C. Taillie (eds). Satellite Program in Statistical Ecology. Burtonsville, Maryland: International Cooperative Publishing House.

Zahl, S. (1977). Jackknifing an index of diversity. *Ecology* **58,** 907–913.

*Received February* 1984; *revised June and November* 1985.