Mathematical Statistics
Stockholm University

Diversity indices for infectious strains

Åke Svensson

Research Report 2002:5

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.matematik.su.se/matstat

Diversity indices for infectious strains

Åke Svensson
University of Stockholm

June 3, 2002

Abstract

A model for the spread of different strains of an infectious agent is used to study how diversity of the distribution of attacking infectious strains is reflected in clustering of infectious cases and in different diversity measures related to such clustering. Three measures are considered: the number of clusters, the mean cluster size and an index based the coefficient of variation of the observed cluster sizes. Monotonicity properties related to strain diversity, attack intensity, duration of the study and speed of transmission are derived.

1. Introduction

The relative abundance of different subgroups of a population is often referred to as the diversity of the population. Indices measuring diversity have been used in ecology (related to distribution of a population over different species in the flora or fauna), in genetics (related to the distribution of different alleles at the same locus), and in economics (related to the distribution of individuals over income classes).

The concept of diversity is of growing importance in the epidemiology of infectious diseases. Modern biotechnology now makes it possible, in many cases, to classify the infectious agents into groups with similar features, e.g. DNA fingerprints obtained through RFLP. This makes it possible to identify infections with a possible common source. It has proved to be a valuable complement to contact tracing which often is used to find chains of sexually transmitted infections.

A considerable number of studies have been published discussing clustering patterns obtained from DNA fingerprints of isolates from tuberculosis patients (cf. Godfrey-Faussett (1999) and Murray and Alland (2002)). The present study started from an interest to analyze such clusters. Tuberculosis is initiated be the reactivating of a dormant infection that initiates a further spread in the surrounding population. It is important to distinguish between transmission chains that consists of one primary case and the secondary cases that descends from it and clusters that contain all infections with the same distinguishable strain. Actual calculations of diversity indices has to be based on observations of clusters. We will be interested in the relation between diversity of the primary cases and diversity or the strains causing infections.

In section 2 the concept of diversity is formulated and in section 3 some diversity indices are defined and the connection between diversity and Schur convexity is explained. A model for the occurrence of transmission chains, which apply to more general situations than the spread of tuberculosis, is formulated in section 4. The model is used in section 5 to investigate how different measures of clustering depends on basic parameters.

In ecology indices that uses closeness of groups are referred to as measures of bio–diversity. The indices, studied in this paper, are defined from the relative abundances of different strains, e.g., we will not consider "closeness" the strains or of their DNA fingerprints). The focus is on diversity rather than on bio–diversity.

2. A partial ordering for diversity

The study of diversity is based on the distribution (or relative abundance) of members in a population in a finite (or at most countable finite) number of mutually exclusive groups. We will assume that p_i stands for the proportion of members in the population that belongs to group i. Let

$$S^n = \{p = (p_1, \ldots, p_n); p_i \geq 0 \quad \text{for all} \quad i = 1, \ldots, n, \quad \text{and} \quad \sum_{i=1}^{n} p_i = 1\}. \tag{2.1}$$

We will consider distribution vectors that belongs to $\cup_n S^n$. Let $(p_{(1)}, p_{(2)}, \ldots, p_{(n)})$ be a permutation of the elements of the vector p such that $p_{(1)} \geq p_{(2)} \geq p_{(3)} \ldots$. Since we are comparing the distributions it is of no importance to which groups the proportions are related.

A distribution p is said to majorize $\tilde{p}$, if

$$\sum_{i=1}^{k} p_{(i)} \geq \sum_{i=1}^{k} \tilde{p}_{(i)}, \tag{2.2}$$

for all $k = 1, 2, \ldots$. If this relation is satisfied we write $p \succ \tilde{p}$. When comparing the two distribution with unequal number of elements we expand vector with lower dimension by adding a suitable number of zeros.

It is natural to say, as has been suggested by Solomon (1979), that the distribution $\tilde{p}$ is more diverse than a distribution p if $p \succ \tilde{p}$. There is a equivalent, more abstract definition of the partial ordering given by the relation $\succ$. It says that $p \succ \tilde{p}$, if there exist a doubly stochastic nxn real matrix P such that

$$\tilde{p} = pP. \tag{2.3}$$

A proof can be found in Hardy et al. (1934) and Marshall and Olkin (1979).

We will be interested in diversity indices, which are functions of the relative abundance distribution. It is natural to require that a diversity index should have a greater value for a more diverse relative abundance distribution than for a less. There is a mathematical term for this requirement. A function ϕ is said to be Schur–convex if

$$\phi(p) \geq \phi(\tilde{p})$$

whenever $p \succ \tilde{p}$. cf. Marshall and Olkin (1979) and Baczkowski et al. (1998). Thus a diversity index should be Schur–concave.

If the diversity index has the form

$$\phi(p) = \sum_i \psi(p_i),$$

where ψ is a continuous function it follows from results in Hardy et al. (1934) and Marshall and Olkin (1979) that a necessary condition for ϕ to be Schur–concave is that the function ψ is concave on the interval $[0, 1/2[$. A sufficient condition is that it is concave on the interval $[0, 1]$. In case ψ is not continuous characterizations are given by Ng (1998).

In case the diversity index has a more complicated structure there are still rather simple characterizations guaranteeing Schur–concavity. If the function ϕ is continuously differentiable then it is Schur–concave on an interval if and only if

$$\phi_{(i)}(z) = \partial\phi(z)/\partial z_{(i)}$$

is increasing in i for all $z_{(1)} \geq z_{(2)} \geq \ldots$. This result is due to Schur (1923) and Ostrowski (1952) (cf Marshall and Olkin (1979)).

3. Diversity indices

Several diversity indices has been suggested in the literature. We will here only mention a few. A commonly used index is the Gini–Simpson index (cf Gini (1912) and Simpson (1949)) which is defined as

$$G(p) = 1 - \sum_i p_i^2. \tag{3.1}$$

The function G is obviously Schur-concave. In genetics this index occurs under the name of heterozygosity (cf Sham (1998)). $G(p)$ is the probability that two individuals chosen at random, independently of each other, does not belong to the same group. There are several equivalent versions of this index, e.g.,

$$D(p) = -\ln(G(p)).$$

The Shannon–Wiener index, is derived from information theory, and is defined as

$$H(p) = -\sum p_i \ln(p_i). \tag{3.2}$$

The function G is obviously Schur–concave.

A general diversity index suggested by Good has the form

$$H_{\alpha,\beta}(p) = \sum p_i^\alpha \{-\ln(p_i)\}^\beta, \tag{3.3}$$

where α and β are positive integers (cf. Good (1953) and Good (1982)), or more general for positive real numbers (cf. Baczkowski et al. (1998)). $H_{2,0}$ gives the Gini–Simpson index and $H_{1,1}$ gives the Shannon–Wiener index. Baczkowski et al. (1998) investigates for which values of α and β the function $H_{\alpha,\beta}$ is Schur–concave.

In ecology there is a set of indices that are based on the species richness that is observed in a random sample. Assume that there are S members in a population distributed over N species according to the distribution given by p. The expected number of species observed in a random sample (taken without replacement) of size s equal

$$N - \sum_{j=1}^{N} \binom{S(1-p_j)}{s} / \binom{S}{s} \approx N - \sum_{j=1}^{N} (1-p_j)^s. \tag{3.4}$$

The approximation is valid when the sampling fraction is small. These indices are called rarefaction diversity (cf. Hurlbert (1971) and Heck et al. (1975)) and are Schur–concave.

4. A model for transmission chains

We will assume that the population under consideration is exposed to attacks of different strains of infections agents. If an infection "attacks" the population it may start spreading within the population and a transmission chain will develop. The starting impulse of the chain may be that a dormant infections becomes infectious (as in the case of tuberculosis), that an infectious individual enters the population, or that a member of the population comes in contact with an external source of infection. Depending on the situation and the population the chain may or may not include a primary case.

We will consider infections that occurs in a population of size P during the time of study. The observations are assumed to be gathered during the time interval $[0, T)$.

Different strains of infectious agents that will, independently, start transmission chains according to Poisson process in time with the intensity πp_i respectively. It is natural to assume that the attack intensity π depends on the size of the population under study. A simple relation is that π is proportional to the size of the population, i.e., $\pi = \lambda P$.

The number of individuals in the population that belongs to a transmission chain will increase in time. Let X_t be this number t time units after the primary attack. The stochastic processes describing the development of the transmission chains are assumed to be independent and identically distributed. Denote the Laplace transform of X_t by

$$L_t(s) = \mathrm{E}\left(\exp\{-sX_t\}\right).$$

Given that m transmission chains are started during the study the starting times will be distributed as m independent random variables which are uniformly distributed over the interval (0,T). The number of cases in the chains will have the distribution of m independent random variables with Laplace transforms

$$Q_T(s) = \frac{1}{T} \int_0^T L_t(s) dt. \tag{4.1}$$

We will assume that we can distinguish between strains (fingerprints) but that we are not able to determine which transmission chain a case belongs to. Thus, the analysis have to be based on observed clusters and observed cluster sizes. The number of cases in a cluster with fingerprints in accordance with strain i has Laplace transform:

$$\exp\{T\pi p_i\,(Q_T(s)-1)\}. \tag{4.2}$$

5. Diversity indices derived from clusters

We will here study some statistical functions based on observations on clusters and their sizes. In particular, we will be interested in how the expected values of these statistics are influenced by the transmission of the strains. The formation of transmission chains and clusters depends, according to the model, on several parameters. Our main interest is related to the diversity distribution of the attacking infectious strains given by $p = (p_1, p_2, \ldots)$. In particular, we will discuss monotonicity properties related to the partial ordering of the abundance distribution p.

The other components of the model are the attack rate π, the time of the study T, and the distribution, $\mathcal{L}(X.)$, of the stochastic process X_t that describes the spread of a transmission chain. These auxiliary parameters will in the following be collected in the vector $\theta = (\pi, T, \mathcal{L}(X.))$. We will also be interested in monotonicity properties related to θ.

The parameters π and T are real numbers and have the corresponding natural ordering. There is a natural partial ordering of the processes describing the development of the transmission chains. We will say that the transmission process $X.$ develops faster than the chain $Y.$, or that $X. \geq Y.$, if $\mathrm{P}(X_t \leq k) \leq \mathrm{P}(Y_t \leq k)$ for all t and all k. These inequalities implies that the Laplace transform of X_t is smaller than the Laplace transform of Y_t. By constructing a convenient sample space it is always possible to find a (Skorohod) representation such that $X_t = Y_t + \tilde{Y}_t$ where $\tilde{Y}_t$ is a non-negative random variable.

5.1 *The number of clusters*

The number of infections with strain i is denoted by K_i. This means that the total number of observed infections is $\sum_i K_i$.

A first concern may be the number of clusters, i.e., the number of different strains that are observed during the study. Let

$$N = \sum_i \mathrm{I}(K_i \geq 1). \tag{5.1}$$

This number equals the number of clusters.

The probability that a strain, or a collection of strains, which has the proportion q is observed during the study, i.e.,

$$I_\theta(q) = 1 - \exp\left\{\pi T q (Q_T(\infty) - 1)\right\}. \tag{5.2}$$

Analogously with the rarefaction index we can consider the expected number of strains that are active under the course of the study, i.e. the number of clusters observed. Simple calculations yield that the expected number of clusters observed is

$$R_\theta(p) = \mathrm{E}(N) = \sum_i \mathrm{I}_\theta(p_i) = \sum_i \left(1 - \exp\{\pi T p_i (Q_T(\infty) - 1)\}\right). \tag{5.3}$$

For a fixed θ the function R_θ is well defined and Schur–concave. Thus, it can serve as a diversity index. The expected number of clusters will be larger the more diverse the strain distribution is.

However, it is important to note that the function are also monotone in the other parameters. The expected number of clusters increases in π and T. It is also larger the faster the development of the transmission chain is.

5.2 *The clustering index*

The observed mean cluster size equals:

$$M = \frac{\sum_i K_i}{\sum_i \mathrm{I}(K_i \geq 1)}. \tag{5.4}$$

Heuristically it seems clear that the expected mean cluster size will be smaller the more diverse the strain distribution is. In epidemiological literature on clusters of tuberculosis it is common to consider

$$C_{n-1} = 1 - 1/M = \frac{\sum_i [K_i - \mathrm{I}(K_i \geq 1)]}{\sum_i K_i}. \tag{5.5}$$

There is a one-to-one correspondence between M and C_{n-1}. If all clusters have a unique primary case C_{n-1} is the proportion of all cases that are secondary cases. Simple calculation yields that the expected value of C_{n-1}, given that there are at least one observed case, is

$$C_\theta(p) = \mathrm{E}(C_{n-1}) = 1 - \frac{\sum_i [Z_\theta(0) - (1 - \mathrm{I}_\theta(p_i)) Z_\theta(p_i)]}{\mathrm{I}_\theta(1)}. \tag{5.6}$$

where

$$Z_\theta(q) = \int_0^\infty \exp(\pi T (1-q)(Q_T(s) - 1)) ds.$$

To prove this we consider the functions

$$r_j(s) = \mathrm{E}\left(\frac{\mathrm{I}(K_j \geq 1)}{\sum_i K_i} \exp(\{-s \sum_i K_i\} \mid \sum_i K_i \geq 1 \right).$$

Differentiating with respect to s we find that

$$\begin{aligned} r'_j(s) &= -\mathrm{E}\left(\mathrm{I}(K_j \geq 1)\exp\{-s\sum_i K_i)\} \mid \sum_i K_i \geq 1\right) \\ &= -\mathrm{E}\{\exp(-s\sum_i K_i\} \mid \sum_i K_i \geq 1) \\ &+ \mathrm{P}(K_j = 0)\mathrm{E}\{\exp(-s\sum_{i\neq j} K_i\} \mid \sum_{i\neq j} K_i \geq 1). \end{aligned} \tag{5.7}$$

The functions r_j can be obtained by solving this differential equation with the boundary value $r_j(\infty) = 0$.

Obviously $C_\theta(p) = \sum_i r_i(0)$ is Schur-convex for fixed value of θ. This means that the expected value of C_{n-1} is larger for a less diverse strain distribution than for a more diverse distribution. Thus the inequality is reversed compared to the index that counts the number of observed strains.

The simple inequality

$$\frac{\sum_i K_i + 1}{\sum_i \mathrm{I}(K_i \geq 1) + 1} \leq \frac{\sum_i K_i}{\sum_i \mathrm{I}(K_i \geq 1)} \leq \frac{\sum_i K_i + 1}{\sum_i \mathrm{I}(K_i \geq 1)}$$

implies that the the mean cluster size (and C_{n-1}) decreases when a new cluster is formed and increases if a new case with a strain that is alreadey observed occurs. There can thus not exist any general monotonicity property related to the parameters π and T. Trivially M and C_{n-1} are increasing in the partial ordering defined above for the transmission processes $X_\cdot$.

5.3 *The quadratic index*

It seems natural to define an index that is based on the variability on the cluster sizes. We will thus define the quadratic index:

$$V = 1 - \frac{\sum_i K_i^2}{(\sum_i K_i)^2}. \tag{5.8}$$

Let

$$v_j(s) = \mathrm{E}\left(\frac{K_j^2}{(\sum_i K_i)^2}\exp\{-s\sum_i K_i\} \mid \sum_i K_i \geq 1\right).$$

This function will solve the differential equation

$$\begin{aligned} v''_j(s) &= \mathrm{E}\left(K_j^2\exp\{-s\sum_i K_i\} \mid \sum_i K_i \geq 1\right) \\ &= \frac{[\pi T p_i Q''_T(s) + (\pi T p_i Q'_T(s))^2]\exp\{\pi T(Q_T(s) - 1)\}}{\mathrm{I}_\theta(1)} \end{aligned}$$

with the boundary conditions $v_j'(\infty) = v_j''(\infty) = 0$.

The expected value of $1 - V$ is the sum of $v_j(0)$ over all strains. Simple manipulations yield that

$$S_\theta(p) = \mathrm{E}(V) = G(p)H(\theta), \tag{5.9}$$

where G is the Gini-Simpson index and H is a function of θ. In fact, $H(\theta) = J(0)$ where J is the solution of of the differential equations

$$\begin{aligned} J''(s) &= (\pi T Q_T'(s))^2 \exp\{\pi T(Q_T(s) - 1)\}/\mathrm{I}_\theta(1), \\ J(\infty) &= 0 \\ J'(\infty) &= 0. \end{aligned}$$

This implies that

$$H(\theta) = J(0) = \frac{1}{\mathrm{I}_\theta(1)} \int_0^\infty s(\pi T Q_T'(s))^2 \exp\{\pi T(Q_T(s) - 1)\} ds. \tag{5.10}$$

$S_\theta(p)$ is Schur-concave for a fixed value of θ. This means that the expected value is larger for a more diverse strain distribution than for a less diverse distribution.

The dependence on the elements of the parameter θ is more complex. In general it is increasing in π and T for small values of πT and decreasing for large values. This is partly due to the fact that new cases when πT is small tends to be the formation of new clusters which will increase the value of V and changes when πT is large tends to be addition to already formed clusters, which will tend to decrease V if the clusters are large.

Acknowledgements

This work has been funded by The Bank of Sweden Tercentenary Foundation.

References

Baczkowski, A., Joanes, D. and Shamia, G. (1998). Range of validity of α and β for a generalized diversity index $\mathrm{H}(\alpha, \beta)$ due to Good. *Mathematical Biosciences* **148**, 115–128.

Gini, C. (1912). Variabilita e mutabilita. *Studi Economico-Giuridici della facolta de Giuisprudenz dell. Universita de Cagliari III, parte II* .

Godfrey-Faussett, P. (1999). Interpretation of cluster studies of tuberculosis. *Lancet* **353**, 427–8.

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237.

Good, I. (1982). Comment on the paper by G.P. Patil, C. Taillie. *J. Am. Statist. Assoc.* **77**, 285.

Hardy, G., Littlewood, J. and Pólya, G. (1934). *Inequalities.* Cambridge University Press, Cambridge.

Heck, A., van Belle, G. and Simberloff, D. (1975). Explicit Calculation of the rerefaction diversity measurement ant the determination of sufficient sample size. *Ecology* **56**, 1459–1461.

Hurlbert, S. (1971). The Nonconcept of Species Diversity: A Critique and Alternative Parameters. *Ecology* **52**, 577–586.

Marshall, A. and Olkin, I. (1979). *Inequalities: Theory of majorization and its Applications.* Academic Press, London.

Murray, M. and Alland, D. (2002). Methodological problems in the molecular epidemiology of tuberculosis. *American Journal of Epidemiology* **155**, 565–571.

Ng, C. (1998). Functions generating Schur–convex sums. In Gabriel, J., Lefévre, C. and Picard, P., editors, *International Series of Numerical Mathematics*, volume 80 of *International Series of Numerical Mathematics.* Birkhäuser Verlag, Basel.

Ostrowski, A. (1952). Sur quelques applications des fonctions convexes et concaves au sens de I. Schur. *J. math. Pures Appl.* **31**, 253–292.

Schur, I. (1923). Über eine Klasse von Mittelbildungen mit Anwendungen de Determinanten. *Theorie Sitzungsber. Berlin. Math.Gesellschaft* **22**, 9–20.

Sham, P. (1998). *Statistics in Human Genetics.* Arnold, London.

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

Solomon, D. (1979). A comparative approach to species diversity. In Grassle, J., Patil, G., Smith, W. and Taillie, C., editors, *Ecological Diversity in Theory and Practise.* International Cooperative Publishing House, Fairland.