

Traditional Measures of Diversity and Sensitivity of Power Entropies

Martin Horáček^{1,2}, Jana Zvárová^{1,2}

¹ Center of Biomedical Informatics, Institute of Computer Science AS CR, Prague, Czech Republic

² Institute of Hygiene and Epidemiology, First Faculty of Medicine of Charles University in Prague, Czech Republic

Abstract

Objectives: We dealt with the traditional measures of diversity and their sample estimates. We also studied a way to compare sensitivity to changes of different measures of diversity.

Methods: We proposed a new estimator of measures of diversity. We compared our estimator with three established estimators in a simulation study. We introduced a function called sensitivity to changes of a measure of diversity H and we described its basic characteristics.

Results: The proposed estimator compares favorably to other well established estimators. The sensitivity to changes has a clear interpretation and is easy to compute.

Conclusions: The sensitivity of measure of diversity to changes could be used to compare behavior of different measures of diversity and to select one or few that are the most suitable for a given problem.



Mgr. Martin Horáček

Keywords

Diversity, entropy, diversity estimates, sensitivity

Correspondence to:

Mgr. Martin Horáček

Center of Biomedical Informatics,
Institute of Computer Science AS CR
Address: Pod Vodárenskou věží 2, 182 07 Prague
E-mail: horacek@euromise.cz

EJBI 2011; 7(1):17–21

received: September 29, 2011

accepted: October 24, 2011

published: November 20, 2011

1 Introduction

In this paper, we dealt with functions that aim to capture the diversity of a given population. The diversity may relate e.g. to the genetic diversity - diversity of alleles of a chosen gene, to the species diversity in a chosen location, but also to a language or economic diversity. We were namely interested in the situation when the amount of diversity of the population depends solely on the probabilities p_i that an individual randomly sampled from the population contains the i -th out of r possible different mutually exclusive features. When these functions satisfy some additional requirements (described in the next paragraph) that are natural for a function that captures the diversity of a population, they are called the traditional measures of diversity.

More formally, traditional measure of diversity is a real functions H defined on the domain $\Delta^r = \{p = (p_1, \dots, p_r) : \sum_{i=1}^r p_i = 1, p_i \geq 0 \forall i\}$ that is

- nonnegative,
- symmetric with respect to permutations,
- minimal when one $p_i = 1$ (only one feature appears in the population)
- maximal when all $p_i \equiv 1/r$ (features are uniformly distributed),
- when one greater p_i increases at the expense of one smaller p_j , the value of $H(p)$ should not rise.

It may be useful to note that if a function $H : \Delta^r \rightarrow R^+$ is Schur-concave and nonnegative, it satisfies all five requirements (see [3]).

There are several frequently used traditional measures of diversity. We present some of them in the next section. Most of them are included or closely related to the f -entropies - a family of generalized entropies - proposed by Zvárová [1]. This family of entropies, their characteristics and how they can be used as measures of diversity is further studied i.e. in Zvárová, Vajda [2] and Horáček [3].

2 Traditional Measures of Diversity and Their Estimates

In this section, we introduce some of the most common traditional diversity measures like Simpson's index, Shannon's entropy, Rényi's entropy of order α , Hill's index and others. We introduce the topic of sample estimates of traditional measures of diversity and we develop a new type of estimator. Parts of sections 2 and 3 were published in the proceedings of the 7th Summer School on Computational Biology [4].

2.1 Examples of Traditional Measures of Diversity

The most often mentioned and used diversity measures include the number of features (e.g. alleles or species)

$$H_0(p) = \sum_{i=1}^r I_{(0,1]}(p_i) - 1$$

(where I denotes the identity function), the Simpson's index

$$H_2(p) = 1 - \sum_{i=1}^r p_i^2$$

and the Shannon's entropy

$$H_1(p) = - \sum_{i=1}^r p_i \ln p_i.$$

These three indices are generalized by the family of power entropies

$$H_\alpha(p) = (\alpha - 1)^{-1} \left(1 - \sum_{i=1}^r p_i^\alpha \right), \quad \text{when } \alpha > 0, \alpha \neq 1,$$

defined as limits when $\alpha = 0$ (identical to number of features) and $\alpha = 1$ (Shannon's entropy). When $\alpha = 2$, we get the Simpson's index.

Another frequently mentioned and used indices include the γ -entropic function

$$H_{A,\gamma}(p) = (1 - \gamma)^{-1} \left[1 - \left(\sum_{i=1}^r p_i^{1/\gamma} \right)^\gamma \right], \quad \text{if } \gamma > 0, \gamma \neq 1,$$

Hill's index

$$H_{H,\alpha}(p) = \left(\sum_{i=1}^r p_i^\alpha \right)^{\frac{1}{1-\alpha}}, \quad \text{when } \alpha > 0, \alpha \neq 1$$

and Rényi's entropy of order α

$$H_{R,\alpha}(p) = (1 - \alpha)^{-1} \ln \left(\sum_{i=1}^r p_i^\alpha \right), \quad \text{when } \alpha > 0, \alpha \neq 1.$$

The introduced generalized parametric indices are handy in several ways. Namely they could be used to improve properties of some procedures based on the common Shannon's entropy. When the Shannon's entropy is replaced by a suitable parametric index, its variable parameter could be used to fine-tune the procedures. This is done e.g. in Andrade and Wang [5].

2.2 Sample Estimates of Traditional Measures of Diversity

Let $p = \{p_1, \dots, p_r\} \in \Delta^r$ be a vector of unknown probabilities p_i that an individual randomly chosen from a population has a feature of type A_i out of r possible features. In this situation, the estimate of measure of diversity $H(p)$ is usually done on the basis of relative frequencies $\hat{p}_n = (X_1/n, \dots, X_r/n) = (\hat{p}_1, \dots, \hat{p}_r)$ of features observed in a sample of n individuals selected from the population randomly with replacement. In that case, the distribution of the vector $X = (X_1, \dots, X_r)$ is multinomial $M(n, p)$. Several estimators that use the observed relative frequencies were suggested in the past. Their qualities, namely their bias and variance, respectively their mean squared error, may vary depending on the chosen diversity index and on the population in which they are used.

The most commonly used estimator, often called the "plug-in" estimator, consists in simply replacing the unknown probabilities p_i with the observed relative frequencies \hat{p}_i . However, despite \hat{p}_i is an unbiased estimate of p_i , the plug-in estimator is generally biased.

Sometimes, the bias could be easily corrected. For example, the mean value of the plug-in estimate of Simpson's index is

$$\begin{aligned} \mathbb{E}H_2(\hat{p}_n) &= 1 - n^{-2} \sum_{i=1}^r \mathbb{E}X_i^2 \\ &= 1 - n^{-2} \sum_{i=1}^r [\text{var}X_i + (\mathbb{E}X_i)^2] \\ &= 1 - n^{-2} \sum_{i=1}^r [np_i(1 - p_i) + n^2 p_i^2] \\ &= (1 - n^{-1}) H_2(p). \end{aligned}$$

Thus, the unbiased estimate of Simpson's index is

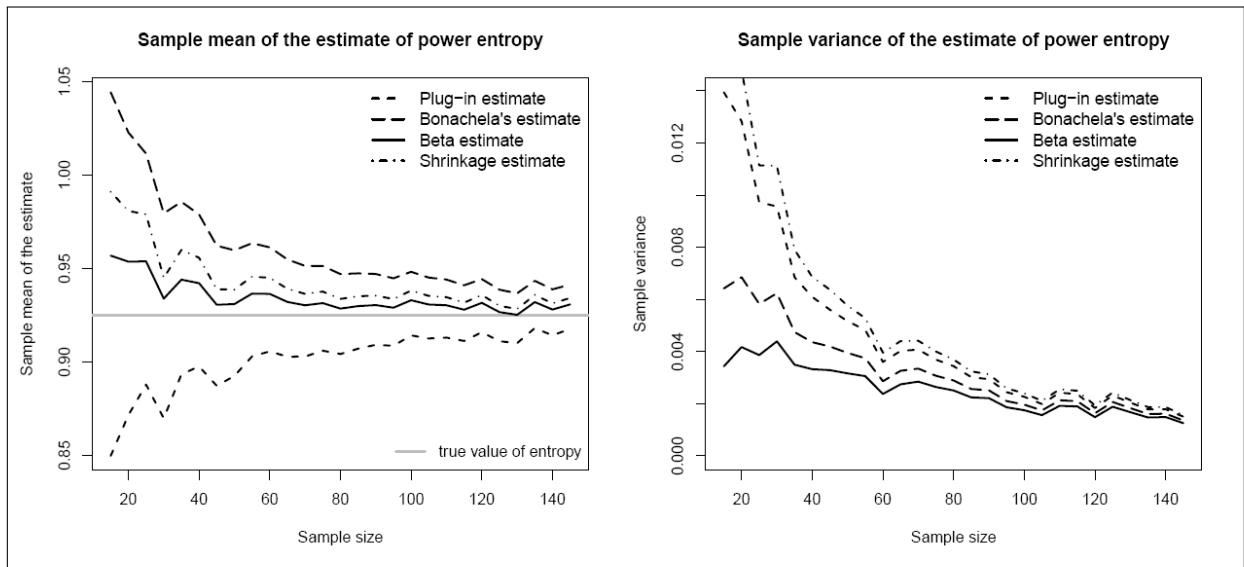


Figure 1: The sample mean and sample variance - estimates of power entropy $H_{3/2}$.

given by

$$\hat{H}_2(\hat{p}_n) = n(n - 1)^{-1}H_2(\hat{p}_n).$$

However, it is often difficult or impossible to find a good correction of the plug-in estimate for other diversity measures. For example, it can be shown that Shannon's index does not have an unbiased estimate (Blyth [6]). Hence, several authors dealt with this problem and suggested more sophisticated estimators. We present an estimator proposed by Bonachela et al. [7] that is called the balanced estimator. We suggested a modification of this estimator that takes into account the likely distribution of values of p_i in the interval $[0, 1]$.

Let us assume that the diversity measure is in the form

$$H(p) = F\left(\sum_{i=1}^r h(p_i)\right),$$

where F and h are an arbitrary real continuous functions. This form includes all previously mentioned indices save the number of features.

Bonachela et al. proposed their estimator in the form

$$\hat{H}(X) = F\left(\sum_{i=1}^r \zeta(X_i)\right),$$

where the function ζ is chosen to minimize

$$\Phi_{\zeta}^2(p_i) = [E(\zeta(X_i) - h(p_i))]^2 + \text{var}(\zeta(X_i))$$

possibly weighed by a function $w(p_i)$ when we have some prior knowledge about the distribution of values $p_i \in [0, 1]$. This way, if we disregard the possible influence of the function F and the correlations, we are able to simultaneously reduce the variance and the square of bias of the estimate.

The weighted average error is then given by

$$\bar{\Phi}_{\zeta}^2(p_i) = \int_0^1 \Phi_{\zeta}^2(p_i)w(p_i)dp_i. \tag{1}$$

The necessary condition for minimality of the error is a zero value of the derivatives

$$\frac{\delta}{\delta\zeta(k)}\bar{\Phi}_{\zeta}^2(p_i) = 0, \quad k \in \{0, \dots, n\}.$$

Therefore, we chose such ζ that

$$\begin{aligned} &\frac{\delta}{\delta\zeta(k)} \int_0^1 [h^2(p_i) - 2h(p_i) \sum_{j=0}^n P(X_i = j)\zeta(j) + \\ &+ \sum_{j=0}^n P(X_i = j)\zeta^2(j)]w(p_i)dp_i = 0 \end{aligned}$$

which can be simplified to

$$\int_0^1 [\zeta(k)P(X_i = k) - h(p_i)P(X_i = k)]w(p_i)dp_i = 0.$$

Since

$$P(X_i = k) = \binom{n}{k} p_i^k (1 - p_i)^{n-k},$$

we got

$$\zeta(k) = \frac{\int_0^1 h(p_i)w(p_i)\binom{n}{k}p_i^k(1-p_i)^{n-k}dp_i}{\int_0^1 w(p_i)\binom{n}{k}p_i^k(1-p_i)^{n-k}dp_i}. \tag{2}$$

Bonachela et al. [7] derived the form of balanced estimator for Shannon's and power entropies with the weight function equal to 1 on the whole interval $[0, 1]$ of possible values of p_i .

However, if we have zero prior knowledge about the values of the components of vector p , it is natural not to prefer any point from Δ^r . Vector p can be thus viewed

as a realization of a random vector $Y = (Y_1, \dots, Y_r)$ that has a uniform distribution on Δ^r . The weight function could then be chosen proportional to the expected value of the components p_i , i.e. as a marginal density of random variable Y_i .

This marginal density is proportional to

$$f(y_1) \propto \int_0^{1-y_1} \dots \int_0^{1-y_1-\dots-y_{r-2}} dy_{r-1} \dots dy_2$$

$$= \frac{(1-y_1)^{r-2}}{(r-2)!},$$

which is (outside a multiplicative constant) a density of Beta distribution $B(1, r-1)$.

We found the ζ function that minimizes vaha with $w(p_i)$ chosen as $(1-p_i)^{r-2}$ and we derived the corresponding estimator. We called this estimator a β -estimator. We describe the derivation of the β -estimator for Shannon's entropy, i.e. when $h(p_i) = -p_i \ln p_i$ and $F(x) = x$. The symbols Γ , B and Ψ denote the Gamma, Beta and Digamma functions, respectively.

First, we replaced $h(p_i)$ and $w(p_i)$ with the appropriate forms and calculated the integral in the denominator of equation zeta1

$$\zeta(X_i) = \frac{\int_0^1 h(p_i) p_i^{X_i} (1-p_i)^{n-X_i+r-2} dp_i}{B(X_i+1, n-X_i+r-1)}.$$

The partial derivative of the Beta function satisfies

$$\frac{\delta}{\delta x} B(x, y) = B(x, y) [\Psi(x) - \Psi(x+y)],$$

and the numerator can be expressed as

$$\int_0^1 h(p_i) p_i^{X_i} (1-p_i)^{n-X_i+r-2} dp_i$$

$$= - \int_0^1 p_i \ln(p_i) p_i^{X_i} (1-p_i)^{n-X_i+r-2} dp_i$$

$$= - \lim_{\alpha \rightarrow 0} \int_0^1 \frac{p_i^\alpha - 1}{\alpha} p_i^{X_i+1} (1-p_i)^{n-X_i+r-2} dp_i$$

$$= \lim_{\alpha \rightarrow 0} \frac{1}{\alpha} [B(X_i+2, n-X_i+r-1) - B(X_i+\alpha+2, n-X_i+r-1)]$$

$$= B(X_i+2, n-X_i+r-1) [\Psi(n+r+1) - \Psi(X_i+2)].$$

Therefore, the ζ function follows

$$\zeta(X_i) = \frac{X_i+1}{n+r} [\Psi(n+r+1) - \Psi(X_i+2)]$$

$$= \frac{X_i+1}{n+r} \sum_{k=X_i+2}^{n+r} \frac{1}{k}$$

and the β -estimator of Shannon's entropy is

$$\hat{H}_1(X) = \sum_{i=1}^r \frac{X_i+1}{n+r} \sum_{k=X_i+2}^{n+r} \frac{1}{k}.$$

The β -estimator for power entropies, whose satisfy $F(x) = x$ and $h(p_i) = (\alpha-1)^{-1} (p_i - p_i^\alpha)$, could be derived in a similar manner. With the weight function chosen as $w(p_i) = (1-p_i)^{r-2}$, the β -estimator of power entropies satisfies

$$\hat{H}_\alpha(X) = (\alpha-1)^{-1} \left[1 - \sum_{i=1}^r \frac{B(n+r, \alpha)}{B(X_i+1, \alpha)} \right].$$

In Fig. 1 we can see a comparison of the β -estimator, Bonachela's original balanced estimator, the plug-in estimator and the James-Stein shrinkage estimator [8] in a population with 6 possible different features distributed as $p = (24/50, 11/50, 9/50, 3/50, 2/50, 1/50)$. The figures show the sample mean and sample variance computed out of 300 trials. The figures were done in R [9]. Another comparison of sample variance and absolute values of sample bias is presented in Table 1. This time, we compared estimators of $H_{1/2}$ and $H_{5/2}$ when $p_1 = (13, 9, 2, 2, 1)/27$, $p_2 = (12, 8, 5, 3, 3, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)/46$ and sample size is $n = 50$.

Table 1: Absolute sample bias and sample variance of estimates.

		$H_{1/2}$		$H_{5/2}$	
		bias	var	bias	var
p_1	plug-in	0.0843	0.0339	0.0094	0.0010
	balanced	0.0059	0.0143	0.0169	0.0008
	beta	0.0052	0.0135	0.0009	0.0006
	shrink	0.0354	0.0211	0.0056	0.0010
p_2	plug-in	0.7476	0.1486	0.0075	0.0002
	balanced	0.2867	0.0281	0.1673	0.0002
	beta	0.1288	0.0218	0.0038	0.0001
	shrink	0.0651	0.0817	0.0052	0.0002

3 Sensitivity to Changes

The indices used to measure diversity differ more or less in their qualities. Their characteristic that is frequently of interest is the rate of the change in value of diversity measure connected to changes in frequencies of a given feature. Several authors dealt with this problem, namely Boyle et al. [10], who were interested mostly in the empirical results, and Izsak [11], who tried to construct a sensitivity measure on a theoretical background. On the suggestion of I. Vajda, we propose a sensitivity measure that is easier to compute and has a clearer interpretation, compared to the Izsak's sensitivity.

Define the sensitivity of diversity measure H to changes in the j -th feature as

$$S_H(p|j) = \lim_{\epsilon \rightarrow 0} \frac{H(p_{j,\epsilon}) - H(p)}{\epsilon H(p)}$$

where

$$p_{j,\epsilon} = \frac{(p_1, \dots, p_{j-1}, p_j + \epsilon p_j, p_{j+1}, \dots, p_r)}{1 + \epsilon p_j}$$

This way, the sensitivity of the measure $H(p)$ to changes in p_j is defined as (a limit form of)

$$\frac{\text{relative change of } H}{\text{relative change of } p_j}$$

and reflects the ratio between relative changes of $H(p)$ and p_j when given p_j alters by a small margin.

3.1 Sensitivity of Power Entropies

The derivation of the formula for sensitivity of power entropies was done in Horáček [3]. If all $p_i > 0$, the sensitivity of power entropies satisfies

$$S_{H_\alpha}(p|j) = \alpha \frac{\sum_{i=1}^r p_i^{\alpha-1} (p_i - \delta_{ij})}{1 - \sum_{i=1}^r p_i^\alpha},$$

if $\alpha \neq 1$ and

$$S_{H_1}(p|j) = \frac{\sum_{i=1}^r (p_i - \delta_{ij}) \ln p_i}{\sum_{i=1}^r p_i \ln p_i}.$$

A comparison of the sensitivity in a population with $p = (24/50, 11/50, 9/50, 3/50, 2/50, 1/50)$ is shown in Fig. 2.

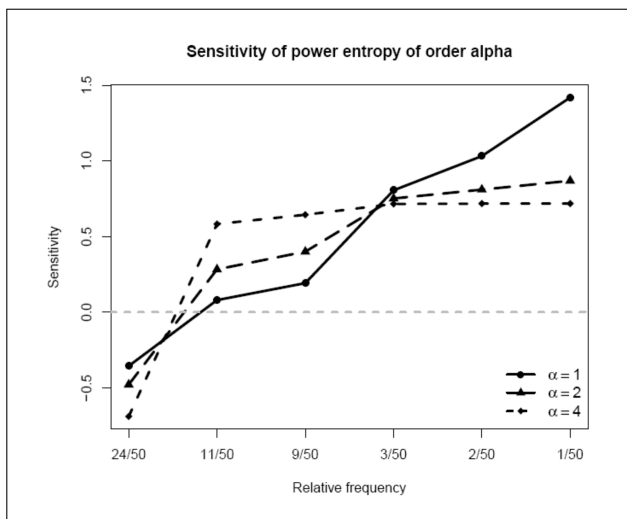


Figure 2: Comparison of the sensitivity to changes - power entropies.

We can see that with decreasing α , the power entropies are more sensitive to the fluctuations in the features that are rare in the population. If we look for example at the sensitivity of Shannon's entropy, say a 10% increase in $p_5 = 1/25$ would result in about 10% increase in $H_1(p)$, while a 10% increase in $p_1 = 24/50$ would result in about 3% decrease of $H(p)$ and a small change in $p_2 = 11/50$ wouldn't likely change the $H(p)$ much at all.

Acknowledgements

This work was supported by the project 1M06014 of the Ministry of Education, Youth and Sports of the Czech Republic and by the project SVV-2011-262514 of Charles University in Prague.

References

- [1] Zvárová J.: On Measures of Statistical Dependence. Časopis pro pěstování matematiky 1974; 99: 15–29
- [2] Zvárová J., Vajda I.: On Genetic Information, Diversity and Distance. Methods of Inform. in Medicine 2006; 2: 173–179
- [3] Horáček, M.: Measures of biodiversity and their applications. Master thesis, Charles university, Prague, supervisor J. Zvárová 2009
- [4] Horáček, M., Zvárová J.: Traditional Measures of Diversity, Their Estimates and Sensitivity to Changes. Proceedings of the 7th Summer School on Computational Biology. 2011; 73–81.
- [5] Andrade, M. de, Wang, X.: Entropy Based Genetic Association Tests and Gene-Gene Interaction Tests. Statistical Applications in Genetics and Molecular Biology 2011; 10: Iss. 1, Article 38
- [6] Blyth, C. R.: Note on estimating information. Annals of Math. Stat. 1959; 30: 71–79
- [7] Bonachela, J. A., Hinrichsen, H., Muñoz, M. A.: Entropy estimates of small data sets. J. of Phys. A: Math. and Theor. 2008; 41: 1–9
- [8] Hausser, J., Strimmer, K.: Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. Journal of Machine Learning Research 2009; 10: 1469-1484.
- [9] R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org> 2011
- [10] Boyle, T. P., Smillie, G. M., Anderson, J. C., and Beeson, D. R.: A sensitivity analysis of nine diversity and seven similarity indices. Research Journal Water Pollution Control Federation 1990; 62: 749–762
- [11] Izsak, J.: Sensitivity Profiles of Diversity Indices. Biom. J. 1996; 38: 921–930