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
recieved: 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 $\Delta^r = \{p = (p_1, \ldots, p_r) : \sum_{i=1}^{r} p_i = 1,\ p_i \geq 0\, \forall\, i\}$ 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 $H : \Delta^r \to R^+$ 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, Renyi'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

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

Hill's index

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

and Rényi's entropy of order α

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

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 $p = \{p_1, \ldots, p_r\} \in \Delta^r$ 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 $\hat{p}_n = (X_1/n, \ldots, X_r/n) = (\hat{p}_i, \ldots, \hat{p}_r)$ of features observed in a sample of n individuals selected from the population randomly with replacement. In that case, the distribution of the vector $X = (X_1, \ldots, X_r)$ is multinomial $\mathrm{M}(n, p)$. 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

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

Thus, the unbiased estimate of Simpson's index is where

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

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.

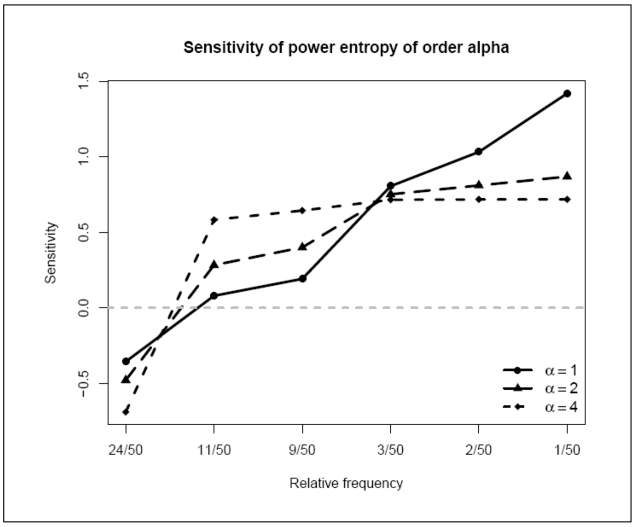

Figure 2: Comparison of the sensitivity to changes - power entropies.

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