Confidence Intervals for Simpson's Diversity Index *

Zhiyi Zhang†
Department of Mathematics and Statistics
University of North Carolina at Charlotte
Email: zzhang@uncc.edu

May 2012

1 Introduction

Zhang and Zhou (2010) proposed a nonparametric estimator and derived a large sample confidence interval for Simpson's diversity index, as a special case in a family of diversity indices. (Interested readers may refer to the referenced paper for details.) However the important case of Simpson's index did not receive enough highlight in the presentation of the paper but rather was buried among other more general results. After many inquiries from scientists across a wide range of fields specifically about the notations and computation involved in the confidence interval for Simpson's diversity index, it is believed that a short tutorial on estimation of Simpson's index based on this paper is in order. The primary objective of this manuscript is to provide a step-by-step and easy-to-follow illustration of how a confidence interval for Simpson's index is calculated, by a simple exampleof Simpson's diversity index.

Before we dive into the illustration, an introduction to the basic notations is in order. Consider a multinomial probability distribution with countable categories indexed by a positive integer s, *i.e.*, $\{p_s\} = \{p_s; s \geq 1\}$ where p_s may be viewed as the proportion of s^{th} species in a population. Simpson (1949) defined a biodiversity index $\lambda = \sum_{s=1}^{S} p_s^2$ for a population with a finite number

AMS 2000 Subject Classifications. Primary 62f10, 62F12, 62G05, 62G20; secondary 62F15. *Keywords and phrases.* Simpson's biodiversity index, confidence intervals

†Author contact information: Tel: (704) 687-4549. Email: zzhang@uncc.edu

of species S, which has an equivalent form

$$\zeta_{1,1} = 1 - \lambda = \sum_{s=1}^{S} p_s q_s \tag{1.1}$$

where $q_s = 1 - p_s$. $\zeta_{1,1}$ assumes a value in $[0, 1)$ with a higher level of $\zeta_{1,1}$ indicating a more diverse population, and is widely used across many fields of study.

Remark 1.1. *In the subsequent text, we will illustrate the steps of producing a* $(1-\alpha)\times 100\%$ *confidence interval for* $\zeta_{1,1}$*. However it is clearly implied that, if a confidence interval for* $\zeta_{1,1}$ *is* (A, B)*, then a corresponding confidence interval for Simpson's index* λ *will be* $(1-B, 1-A)$*.*

Suppose an *iid* sample of size n is available and is summarized into frequencies $\{Y_s; s \geq 1\}$ where Y_s is the number of observations of the s^{th} species found in the sample. Let

$$\hat{p}_s = \frac{Y_s}{n}$$

be the sample relative frequency for the s^{th} species.

Let

$$\begin{aligned}
Z_{1,1} &= \tfrac{n}{n-1} \textstyle\sum \hat{p}_s(1-\hat{p}_s) \\
Z_{2,0} &= \tfrac{n}{n-1} \textstyle\sum 1_{[\hat{p}_s \geq 2/n]} \hat{p}_s(\hat{p}_s - 1/n) \\
Z_{3,0} &= \tfrac{n^2}{(n-1)(n-2)} \textstyle\sum 1_{[\hat{p}_s \geq 3/n]} \hat{p}_s(\hat{p}_s - 1/n)(\hat{p}_s - 2/n) \\
Z_{2,1} &= \tfrac{n^2}{(n-1)(n-2)} \textstyle\sum 1_{[\hat{p}_s \geq 2/n]} \hat{p}_s(\hat{p}_s - 1/n)(1-\hat{p}_s) \\
Z_{1,2} &= \tfrac{n^2}{(n-1)(n-2)} \textstyle\sum 1_{[\hat{p}_s \geq 1/n]} \hat{p}_s(1-\hat{p}_s)(1 - 1/n - \hat{p}_s)
\end{aligned} \tag{1.2}$$

where the summation $\sum$ is over all possible s.

Remark 1.2. *Note that, while the summation* $\sum$ *is over all possible* s*, the indicator function excludes some of the values of* s *so that the factors in the summand will be non-negative.*

Provided that the distribution $\{p_s\}$ is non-uniform. A $(1-\alpha)\times 100\%$ confidence interval for $\zeta_{1,1}$ is given by

$$Z_{1,1} \pm 2 z_{\alpha/2} \frac{\hat{\sigma}_1(1,1)}{\sqrt{n}} \tag{1.3}$$

where $\hat{\sigma}_1(1,1)$ is such that

$$4\hat{\sigma}_1^2(1,1) = Z_{1,2} - 2Z_{2,1} + Z_{3,0} - (Z_{1,1} - Z_{2,0})^2 \tag{1.4}$$

and $Z_{1,1}$, $Z_{1,2}$, $Z_{2,1}$, $Z_{2,0}$ and $Z_{3,0}$ are all given in (1.2).

2 A Step-by-Step Illustration

Consider an example of *iid* sample of butterflies from Region $\mathscr{A}$ given in Table 1. There are 13 different species observed in the sample of size $n = 147$.

$Species(s)$	$Frequency(y)$
A	53
B	51
C	13
D	8
E	5
F	5
G	3
H	3
I	2
J	1
K	1
L	1
M	1

Table 1: Butterfly Data

2.1 Step 1: Data Preparation

Using the data in Table 1, the following six columns are added to the data table

- $\hat{p}_s = y_s/n$,
- $\hat{p}_s(1 - \hat{p}_s)$,
- $\hat{p}_s(\hat{p}_s - 1/n)$,
- $\hat{p}_s(\hat{p}_s - 1/n)(\hat{p}_s - 2/n)$,
- $\hat{p}_s(\hat{p}_s - 1/n)(1 - \hat{p}_s)$, and

- $\hat{p}_s(1-\hat{p}_s)(1-\hat{p}_s-1/n)$.

s	y_s	$\hat{p}_s$	$\hat{p}_s(1-\hat{p}_s)$	$\hat{p}_s(\hat{p}_s-1/n)$
A	53	0.3605	0.2306	0.1275
B	51	0.3469	0.2266	0.1180
C	13	0.0884	0.0806	0.0072
D	8	0.0544	0.0515	0.0026
E	5	0.0340	0.0329	0.0009
F	5	0.0340	0.0329	0.0009
G	3	0.0204	0.0200	0.0003
H	3	0.0204	0.0200	0.0003
I	2	0.0136	0.0134	9.26×10^{-5}
J	1	0.0068	0.0068	0.0000
K	1	0.0068	0.0068	0.0000
L	1	0.0068	0.0068	0.0000
M	1	0.0068	0.0068	0.0000

s	y_s	$\hat{p}_s(\hat{p}_s-1/n)(\hat{p}_s-2/n)$	$\hat{p}_s(\hat{p}_s-1/n)(1-\hat{p}_s)$	$\hat{p}_s(1-\hat{p}_s)(1-\hat{p}_s-1/n)$
A	53	0.0442	0.0816	0.1459
B	51	0.0393	0.0771	0.1464
C	13	0.0005	0.0066	0.0729
D	8	0.0001	0.0025	0.0483
E	5	1.89×10^{-5}	0.0009	0.0315
F	5	1.89×10^{-5}	0.0009	0.0315
G	3	1.89×10^{-6}	0.0003	0.0194
H	3	1.89×10^{-6}	0.0003	0.0194
I	2	0.0000	9.13×10^{-5}	0.0131
J	1	0.0000	0.0000	0.0067
K	1	0.0000	0.0000	0.0067
L	1	0.0000	0.0000	0.0067
M	1	0.0000	0.0000	0.0067

2.2 Step 2: Find $Z_{1,1}$

Noting the sum of values in the $\hat{p}_s(1-\hat{p}_s)$ column is 0.7353 and $n = 147$, by (1.2),

$$Z_{1,1} = \frac{147}{146} \times 0.7353 = 0.7404$$

2.3 Step 3: Find $Z_{2,0}$

Noting the sum of values in the $\hat{p}_s(\hat{p}_s - 1/n)$ column is 0.2579 and $n = 147$, by (1.2),

$$Z_{2,0} = \frac{147}{146} \times 0.2579 = 0.2596$$

2.4 Step 4: Find $Z_{3,0}$

Noting the sum of values in the $\hat{p}_s(\hat{p}_s - 1/n)(\hat{p}_s - 2/n)$ column is 0.0843 and $n = 147$, by (1.2),

$$Z_{3,0} = \frac{147^2}{146 \times 145} \times 0.0843 = 0.0860$$

2.5 Step 5: Find $Z_{2,1}$

Noting the sum of values in the $\hat{p}_s(\hat{p}_s - 1/n)(1 - \hat{p}_s)$ column is 0.1701 and $n = 147$, by (1.2),

$$Z_{2,1} = \frac{147^2}{146 \times 145} \times 0.1701 = 0.1736$$

2.6 Step 6: Find $Z_{1,2}$

Noting the sum of values in the $\hat{p}_s(1 - \hat{p}_s)(1 - \hat{p}_s - 1/n)$ column is 0.5553 and $n = 147$, by (1.2),

$$Z_{1,2} = \frac{147^2}{146 \times 145} \times 0.5553 = 0.5668$$

2.7 Step 7: Find $\hat{\sigma}_1(1,1)$

Using (1.4),

$$\begin{aligned} 4\hat{\sigma}_1^2(1,1) &= Z_{1,2} - 2Z_{2,1} + Z_{3,0} - (Z_{1,1} - Z_{2,0})^2 \\ &= 0.5668 - 2 \times 0.1736 + 0.0860 - (0.7404 - 0.2596)^2 = 0.0744 \end{aligned}$$

and therefore

$$\hat{\sigma}_1^2(1,1) = 0.0186 \quad or \quad \hat{\sigma}_1(1,1) = 0.1364.$$

2.8 Step 8: Find a 95% Confidence Interval for $\zeta_{1,1}$

Using (1.3) and $z_{\alpha/2} = 1.96$, the desired confidence interval is

$$Z_{1,1} \pm 2z_{\alpha/2}\frac{\hat{\sigma}_1(1,1)}{\sqrt{n}} = 0.7404 \pm 2 \times 1.96 \times \frac{0.1364}{\sqrt{147}} = 0.7404 \pm 0.0441 = (0.6963, 0.7845)$$

2.9 Step 9: Find a 95% Confidence Interval for Simpson's Index, λ

Noting Remark 1.1, the desired confidence interval is

$$(1 - 0.7845, 1 - 0.6963) = (0.2155, 0.3037).$$

References

[1] Simpson, E.H. (1949), *Measurement of Diversity*, Nature, Vol. 163, p.688.

[2] Zhang, Z. & Zhou, J. (2010). Re-parameterization of multinomial distributions and diversity indices, Journal of Statistical Planning and Inference, 140, pp. 1731-1738.