ISSN 0005-1179, Automation and Remote Control, 2020, Vol. 81, No. 2, pp. 272–284. © Pleiades Publishing, Ltd., 2020.
Russian Text © The Author(s), 2020, published in Avtomatika i Telemekhanika, 2020, No. 2, pp. 57–172.

INTELLECTUAL CONTROL SYSTEMS, DATA ANALYSIS

On Confidence Estimation Based on Quantitative Similarity Coefficients

I. V. Rodionov[*,**,a] **and A. N. Sozontov**[***,b]

[*] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
[**] *Steklov Mathematical Institute, Russian Academy of Sciences, Moscow, Russia*
[***] *Institute of Plant and Animal Ecology, Ural Branch of the Russian Academy of Sciences, Yekaterinburg, Russia*
e-mail: [a] *vecsell@gmail.com,* [b] *a.n.sozontov@gmail.com* Received April 16, 2019
Revised July 7, 2019
Accepted July 18, 2019

Abstract—We consider the problem of estimating the accuracy of quantitative similarity coefficients. For this purpose, we introduce a new concept of the similarity measure for the corresponding coefficient. We show that only frequency forms of quantitative similarity coefficients represent consistent estimates of their similarity measures. We obtain asymptotic confidence intervals for the Ružička and Bray–Curtis similarity measures based on the coefficients with the same names. We also propose a criterion for the homogeneity of two populations based on the above-mentioned coefficients.

Keywords: similarity coefficient, confidence estimation, homogeneity criterion, Bray–Curtis index, Jaccard index

DOI: 10.1134/S0005117920020101

1. INTRODUCTION

Similarity coefficients (SC), originally proposed by biologists, are widely used in chemistry, sociology, linguistics, jurisprudence, and other fields, as well as in methods of processing multidimensional data; in particular, they formed the foundation for some forms of cluster analysis. Currently, there exist several dozen to several hundred SC (see, for example, [1]), but statistical theory for describing SCs has virtually not been developed. For example, for different ratios between sample sizes most of the quantitative SCs used (see definitions below) are estimating, in essence, different values, while the data used to construct qualitative is often too small to make reliable statistical conclusions. Existing methods for assessing the accuracy of SCs either are based on extremely strong assumptions such as the uniform distribution of species in the population (this most often happens for qualitative SCs) or do not have a strict mathematical formalization at all. Several attempts have been made to construct bootstrap confidence intervals for some SCs. In this work, we obtain exact asymptotic confidence intervals for the most popular quantitative Bray–Curtis and Ružička similarity coefficients and propose a criterion for testing the homogeneity hypothesis of two populations based on the previous result.

Let X and Y be two descriptive sets [2, 3] that describe two compared samples, i.e., finite sets of species (types of objects) such that each species is associated with the number of its occurrences in the corresponding sample. In other words, we number the species found in the two considered samples from 1 to S and denote by X_i and Y_i the number of objects of type i in the first and second samples respectively. Denote by a the number of common species for the two sets in comparison;

by b and c, the number of unique species for the first and second sets, respectively:

$$a = \sum_{i=1}^{S} I(X_i \neq 0, Y_i \neq 0),$$
$$b = \sum_{i=1}^{S} I(X_i \neq 0, Y_i = 0),$$
$$c = \sum_{i=1}^{S} I(X_i = 0, Y_i \neq 0).$$

It is easy to see that $S = a + b + c$.

A similarity coefficient, or similarity index, of two populations $C(X, Y)$ is a dimensionless indicator that reflects the measure of proximity (similarity) of these populations X and Y. As a rule, similarity indices are considered in order to compare two populations; however, there are methods for finding similarities between three or more sets at the same time [4, 5]. In this work, we do not consider such comparison options and their corresponding CS. We call a CS *qualitative* if it depends only on a, b, and c, i.e., the values of such SC are affected only by the presence or absence of species in the compared populations. A similarity coefficient is called *quantitative* if the values $\{X_i\}_{i=1}^{S}$ and $\{Y_i\}_{i=1}^{S}$ are also used to construct it. We call quantitative SCs that depend only on the frequencies X_i/n and Y_i/m, $1 \leqslant i \leqslant S$, of the occurrence of species i in populations X and Y respectively, *frequency* CS. Here $n = \sum_{i=1}^{S} X_i$ and $m = \sum_{i=1}^{S} Y_i$. For any qualitative SC, one can consider its quantitative counterpart by replacing indicators $I(X_i \neq 0)$, $I(Y_i \neq 0)$, $i = 1, \ldots, S$, of the presence of species in a population with frequencies X_i/n and Y_i/m, $1 \leqslant i \leqslant S$. As we will show below, introducing other quantitative counterparts for qualitative SCs is not justified from the statistical point of view.

Let us discuss the general requirements that are usually imposed on similarity indices [6, 7]:

A1. Symmetry: $C(X, Y) = C(X, Y)$;

A2. Equality to zero for disjoint sets: $C(X, Y) = 0$, if $a = 0$;

A3. Equality to unity for identical populations: $C(X, Y) = 1$ if $a = b = c$ for qualitative KS and $X_i/n = Y_i/m \ \forall i,\ 1 \leqslant i \leqslant S$, for frequency KS;

A4. "Monotonicity" in the amount of similarity.

For qualitative similarity indices, in particular, the last property means the following: if we fix S and the set of species, then the value of CS should increase with the value of a. However, far from all similarity indices satisfy conditions **A1**–**A4**; see [1]. Next, let us also define the similarity measure μ of a qualitative SC as the value equal to this SC if, instead of the samples, the SC was computed by general populations from which the sample data was taken. It is clear that with sample size tending to infinity, a qualitative SC will converge to its similarity measure. We also define the similarity measure of the quantitative SC as the value obtained by replacing X_i and Y_i in the formula for this SC with probabilities p_i and q_i of species i occurring in the first and second populations respectively. In Section 2, we will show that quantitative SCs will represent consistent estimates of their similarity measures if and only if they are frequency SCs.

The need for comparing sets faced biologists as early as the XIX century. However, methods that allow to quantify their degree of (dis)similarity appeared only at the beginning of the XX century. Apparently, the very first CS, I_J, which still remains the most popular among similarity indices, was proposed by the Swiss botanist Paul Jaccard [8]. In essence, I_J is the ratio of the size of intersection of sets of species in two populations to the size of their union,

$$I_J = \frac{a}{a + b + c},$$

and it is a qualitative similarity coefficient. Its quantitative analogue is called the Ružička coefficient [9]:

$$C_R = \frac{\sum\limits_{i=1}^{S} \min(X_i, Y_i)}{\sum\limits_{i=1}^{S} \max(X_i, Y_i)}. \tag{1}$$

Another popular CS was offered almost simultaneously by L.R. Dice in [10] and T. Sørensen in [11]:

$$I_{DS} = \frac{2a}{2a + b + c},$$

whose quantitative form was proposed long before them by Chekanovsky in [12]; it is also known as the Bray–Curtis index [13]

$$C_{BC} = \frac{\sum\limits_{i=1}^{S} 2\min(X_i, Y_i)}{\sum\limits_{i=1}^{S} X_i + \sum\limits_{i=1}^{S} Y_i}. \tag{2}$$

It is easy to see that the Jaccard index can be expressed via the Sørensen–Dice index as $I_J = I_{DS}/(2 - I_{DS})$. By now, dozens of qualitative SC have been proposed. Along with the Jaccard and Sørensen–Dice indices, the most often used are Ochiai indices I_O, Kulczynski's I_K, and Morisita's I_M, see [7]. All of them monotonically increase from zero to one depending on the number of common species and, in fact, differ only in their sensitivity to small and large values of a compared to S.

For the first time, an attempt to evaluate the accuracy of similarity indices was made by Sørensen in [11], but his method requires not two but a sufficiently large number of samples, which is not always feasible. A considerable number of publications have been devoted to confidence estimation of the qualitative similarity measure under the assumption that all species in the general population are distributed equally, see [14–17],which, obviously, never holds in practice. However, it is worth noting that if we only have data on the presence or absence of a species in the samples and there is no information on the number of objects of each of the species in the population, then the choice of any other distribution is not justified. In addition, the values a, b, and c strongly depend on the presence of rare species in the sample. As the number of observations increases, the relations between a, b, and c can change significantly, which hinders the accuracy of statistical analysis for qualitative SCs with small or average number of observations. In relation to confidence estimation of the qualitative similarity measure we also note the work [18], where it is assumed that the distribution of species in the general population is discrete lognormal, and the work [19], where it is assumed that one dominant species occurs more often than others that, in turn, have an equal probability of occurring in a sample.

To construct confidence intervals for qualitative similarity measures, Chao in [20–22] developed a useful bootstrapping-based method for estimating the number of species in the general population based on a sample. In particular, the work [22] constructs, under the assumption that the Jaccard index is less than its similarity measure, a confidence interval for the similarity measure I_J; it does so, however, without any mathematical justification. Although the method is promising, Chao was not able to construct a confidence interval for any qualitative SC under general assumptions. We are not aware of any work that considers the problem of constructing confidence intervals for similarity measures based on quantitative SCs.

2. MAIN RESULTS

2.1. Statistical Correctness of Quantitative SC

First of all, we show on the basis of the Ružička similarity index (1) that quantitative SCs represent consistent estimates of their similarity measures only in case when $n/m \to 1$ for $n, m \to \infty$; recall that n and m are the sizes of the first and second population respectively. Consider two polynomial models in the framework of the considered problem: in trial j, one object from each general population occurs according to the probability distributions $\{p_i\}_{i\geqslant 1}$ and $\{q_i\}_{i\geqslant 1}$ respectively independently of other trials, i.e., in trial j object i occurs with probabilities p_i and q_i for the first and second group respectively. We denote random variables corresponding to the occurrence of an object of a certain type in the first and second group in trial j as ξ_j and η_j respectively. Thus, we have

$$X_i = \sum_{j=1}^{n} I(\xi_j = i), \quad Y_i = \sum_{j=1}^{m} I(\eta_j = i).$$

Then, since $\{I(\xi_j = i)\}_{j\geqslant 1}$ and $\{I(\eta_j = i)\}_{j\geqslant 1}$ are sequences of independent identically distributed random variables, according to the strengthened law of large numbers

$$\begin{aligned} &\frac{X_i}{n} \xrightarrow{\text{a.s.}} EI(\xi_j = i) = P(\xi_j = i) = p_i, \\ &\frac{Y_i}{m} \xrightarrow{\text{a.s.}} q_i \qquad \text{(a.s.—almost surely)} \end{aligned} \tag{3}$$

for $n, m \to \infty$.

Let us go back to the discussion of the Ružička index. Its similarity measure is obviously equal to

$$\mu_R = \frac{\sum\limits_{i=1}^{S} \min(p_i, q_i)}{\sum\limits_{i=1}^{S} \max(p_i, q_i)}.$$

Then, using the theorem on the inheritance of convergences, properties of almost sure convergence, and relations (3), in case when $n/m \to 1$ for $n, m \to \infty$ it is elementary to show that

$$C_R = \frac{\sum\limits_{i=1}^{S} \min\left(\frac{X_i}{n}, \frac{Y_i}{m}\frac{m}{n}\right)}{\sum\limits_{i=1}^{S} \max\left(\frac{X_i}{n}, \frac{Y_i}{m}\frac{m}{n}\right)} \xrightarrow{\text{a.s.}} \mu_R.$$

On the other hand, for $n/m = d$, $n, m \to \infty$, the Ružička index almost surely converges to

$$\mu_R(d) = \frac{\sum\limits_{i=1}^{S} \min(dp_i, q_i)}{\sum\limits_{i=1}^{S} \max(dp_i, q_i)},$$

which for $d \neq 1$ is different from μ_R, whereas the frequency counterpart of this index

$$C'_R = \frac{\sum\limits_{i=1}^{S} \min\left(\frac{X_i}{n}, \frac{Y_i}{m}\right)}{\sum\limits_{i=1}^{S} \max\left(\frac{X_i}{n}, \frac{Y_i}{m}\right)} \tag{4}$$

converges almost surely to μ_R regardless of the relation between m and n. Similar reasoning is also valid for the Bray–Curtis similarity index (2), and for other quantitative similarity indices that are

not frequency indices. Thus, conclusions about the general population obtained on the basis of quantitative SCs that are not frequency SCs cannot be considered reliable and statistically correct.

2.2. Asymptotic Normality of Quantitative SC

Let us consider the Ružička similarity index and construct a confidence interval for its similarity measure. We transform this index as follows:

$$C_R = \frac{\sum\limits_{i=1}^{S} \min(X_i, Y_i)}{\sum\limits_{i=1}^{S} \max(X_i, Y_i)} = \frac{n+m-\sum\limits_{i=1}^{S} \max(X_i, Y_i)}{\sum\limits_{i=1}^{S} \max(X_i, Y_i)} = \frac{n+m}{\sum\limits_{i=1}^{S} \max(X_i, Y_i)} - 1, \tag{5}$$

i.e. in fact, the Ružička index depends only on $\sum_{i=1}^{S} \max(X_i, Y_i)$.

First, suppose that $n = m$. Let $p_i > q_i$; then for $n \to \infty$ it holds that $P(X_i > Y_i) \to 1$ and $P(\max(X_i, Y_i) = X_i) \to 1$, i.e., for large n the occurrence of object i for the second population will not affect the value of the Ružička index with probability close to one. We define A as the set of numbers i such that $p_i > q_i$, $A = \{i : p_i \geqslant q_i\}$, and $B = \{i : q_i > p_i\}$. We also define

$$P = \sum_{i \in A} p_i, \quad Q = \sum_{i \in B} q_i.$$

Then the similarity measure of the Ružička index can be rewritten in the form

$$\mu_R = \frac{2 - \sum\limits_{i=1}^{S} \max(p_i, q_i)}{\sum\limits_{i=1}^{S} \max(p_i, q_i)} = \frac{2}{P+Q} - 1.$$

Theorem. *Let $n = m$. Then for $n \to \infty$*

$$\sqrt{n}(C_R - \mu_R) \xrightarrow{d} N\left(0, V_R\right),$$

where

$$V_R = \frac{4(P(1-P) + Q(1-Q))}{(P+Q)^4}.$$

Proof of Theorem. Let

$$\zeta_j = I(\xi_j = i, i \in A) + I(\eta_j = i, i \in B), \quad j \geqslant 1.$$

These random variables can obviously take the values $0, 1$, and 2. Note that since by assumption the number of species S is finite, then

$$P\left(\sum_{i=1}^{S} \max(X_i, Y_i) = \sum_{i \in A} X_i + \sum_{i \in B} Y_i\right) \to 1 \tag{6}$$

as $n \to \infty$. On the other hand,

$$\begin{aligned} \sum_{i \in A} X_i + \sum_{i \in B} Y_i &= \sum_{i \in A} \sum_{j=1}^{n} I(\xi_j = i) + \sum_{i \in B} \sum_{j=1}^{n} I(\eta_j = i) \\ &= \sum_{j=1}^{n} (I(\xi_j = i, i \in A) + I(\eta_j = i, i \in B)) = \sum_{j=1}^{n} \zeta_j =: T_n. \end{aligned} \tag{7}$$

Thus, to prove the theorem, it suffices to show that for $n \to \infty$

$$\sqrt{n}\left(\left(\frac{2n}{T_n} - 1\right) - \mu_R\right) \xrightarrow{d} N\left(0, \frac{4(P(1-P) + Q(1-Q))}{(P+Q)^4}\right) \tag{8}$$

and

$$\sqrt{n}\left(C_R - \left(\frac{2n}{T_n} - 1\right)\right) \xrightarrow{d} 0, \tag{9}$$

where statistics $2n/T_n - 1$ is obtained by substituting the sum T_n in the expression (5) instead of $\sum_{i=1}^{S} \max(X_i, Y_i)$, and then use Slutsky's lemma.

First we prove relation (8). Note that $\{\zeta_j\}_{j=1}^{n}$ are independent identically distributed random variables, and we will find $E\zeta_1$ and $Var\zeta_1$. We have that

$$\begin{gathered} E\zeta_1 = EI(\xi_1 = i, i \in A) + EI(\eta_1 = i, i \in B) \\ = \sum_{i \in A} P(\xi_1 = i) + \sum_{i \in B} P(\eta_1 = i) = \sum_{i \in A} p_i + \sum_{i \in B} q_i = P + Q, \\ Var\zeta_1 = VarI(\xi_1 = i, i \in A) + VarI(\eta_1 = i, i \in B) = P(1-P) + Q(1-Q). \end{gathered}$$

The central limit theorem implies that

$$\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n} \zeta_j - (P+Q)\right) \xrightarrow{d} N\left(0, P(1-P) + Q(1-Q)\right), \quad n \to \infty.$$

Using the delta method for the function $g(x) = 2/x$, we get that

$$\sqrt{n}\left(\frac{2n}{\sum\limits_{j=1}^{n} \zeta_j} - \frac{2}{P+Q}\right) \xrightarrow{d} N\left(0, \frac{4(P(1-P) + Q(1-Q))}{(P+Q)^4}\right), \quad n \to \infty,$$

which implies relation (8).

Let us now prove relation (9). Using (6) and (7), we get for $n \to \infty$ that

$$\sqrt{n}\left(\frac{\sum\limits_{i=1}^{S} \max(X_i, Y_i)}{n} - \frac{T_n}{n}\right) \xrightarrow{d} 0. \tag{10}$$

Since by the law of large numbers $T_n/n \to P + Q$ in probability for $n \to \infty$, then relation (10) implies that $\sum_{i=1}^{S} \max(X_i, Y_i)/n$ also tends in probability to $P + Q$. Dividing the left-hand side (10) by $T_n \sum_{i=1}^{S} \max(X_i, Y_i)/(2n^2)$, we have by Slutsky's lemma that

$$\sqrt{n}\left(\frac{2n}{T_n} - \frac{2n}{\sum\limits_{i=1}^{S} \max(X_i, Y_i)}\right) \xrightarrow{d} 0,$$

which implies (9). This completes the proof of the theorem.

We now turn to studying the frequency counterpart of Ružička's index C'_R (4). The following statement says that the asymptotic behavior of C'_R is different from the asymptotic behavior of the Ružička index at $n = m \to \infty$.

Proposition 1. *For $n, m \to \infty$ and $n/m \to d > 0$ it holds that*

$$\sqrt{n}(C'_R - \mu_R) \xrightarrow{d} N\left(0, V_R(d)\right),$$

where

$$V_R(d) = \frac{4(P(1-P) + dQ(1-Q))}{(P+Q)^4}.$$

Proof of Proposition 1. Using the fact that $\sum_{i\in A} X_i$ and $\sum_{i\in B} Y_i$ can be represented as sums of independent identically distributed random variables and are independent, we obtain from the central limit theorem that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i\in A} X_i - P\right) \xrightarrow{d} N(0, P(1-P))$$

and

$$\sqrt{m}\left(\frac{1}{m}\sum_{i\in B} Y_i - Q\right) \xrightarrow{d} N(0, Q(1-Q))$$

with $n\to\infty$ and $m\to\infty$ respectively, which under the assumptions of the proposition by Slutsky's lemma implies that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i\in A} X_i + \frac{1}{m}\sum_{i\in B} Y_i - (P+Q)\right) \xrightarrow{d} N(0, (P(1-P) + dQ(1-Q))). \tag{11}$$

The rest of the proof repeats, with slight changes, the corresponding steps of the proof of the theorem.

Now let us consider the frequency form of the Bray–Curtis index (2),

$$C'_{BC} = \sum_{i=1}^{S} \min(X_i/n, Y_i/m) = 2 - \sum_{i=1}^{S} \max(X_i/n, Y_i/m),$$

and examine its asymptotic behavior. Note that the frequency form of the widely used Ochiai index $I_O = a/\sqrt{(a+b)(a+c)}$ is also equal to C'_{BC}. Relation (11) and a modification of formula (10) for $n \neq m$ the following statement.

Proposition 2. *Under the assumptions of proposition 1 it holds that*

$$\sqrt{n}(C'_{BC} - \mu_{BC}) \xrightarrow{d} N\left(0, V_{BC}(d)\right),$$

where $\mu_{BC} = (2-P-Q)$ *is the Bray–Curtis similarity measure and* $V_{BC}(d) = P(1-P) + dQ(1-Q)$.

2.3. Confidence Estimates of Quantitative Similarity Measures

In order to construct confidence intervals for Ružička and Bray–Curtis similarity measures, we need to estimate the asymptotic variance of the corresponding frequency SCs. Consider a random set of numbers $A' = \{i : X_i/n > Y_i/m\}$. Let us show that

$$\widehat{P} := \frac{1}{n}\sum_{i\in A'} X_i \xrightarrow{P} P, \quad \widehat{Q} := \frac{1}{m}\sum_{i\in \overline{A'}} Y_i \xrightarrow{P} Q, \quad n, m \to \infty.$$

Indeed,

$$P\left(\sum_{i\in A'} X_i \neq \sum_{i\in A} X_i\right) \leqslant \sum_{i\in A} P(X_i/n < Y_i/m) + \sum_{i\in B} P(X_i/n \geqslant Y_i/m) \to 0$$

by the law of large numbers. Using the law of large numbers and the properties of convergence in probability again, we have

$$\widehat{P} = \left(\frac{1}{n}\sum_{i\in A'} X_i - \frac{1}{n}\sum_{i\in A} X_i\right) + \frac{1}{n}\sum_{i\in A} X_i \xrightarrow{P} P, \quad n, m \to \infty,$$

and the convergence of $\widehat{Q}$ to Q in probability can be proved similarly.

Using the theorem on the inheritance of convergences and properties of convergence in probability, it is easy to show that for $n, m \to \infty$ and $n/m \to d$ it holds that

$$\widehat{V_R(d)} := \frac{4(\widehat{P}(1-\widehat{P}) + n\widehat{Q}(1-\widehat{Q})/m)}{(\widehat{P}+\widehat{Q})^4} \xrightarrow{P} V_R(d),$$
$$\widehat{V_{BC}(d)} := \widehat{P}(1-\widehat{P}) + n\widehat{Q}(1-\widehat{Q})/m \xrightarrow{P} V_{BC}(d).$$

Thus, under the same assumptions we obtain from Slutsky's lemma that

$$\sqrt{n}(\widehat{V_R(d)})^{-1/2}(C'_R - \mu_R) \xrightarrow{d} N(0,1) \quad \text{and} \quad \sqrt{n}(\widehat{V_{BC}(d)})^{-1/2}(C'_{BC} - \mu_{BC}) \xrightarrow{d} N(0,1), \tag{12}$$

which allows us to write asymptotic confidence intervals for Ružička and Bray–Curtis similarity measures: with probability tending to $1-\alpha$ the Ružička similarity measure belongs to the interval

$$\left(C'_R - u_{1-\alpha/2}(\widehat{V_R(d)})^{1/2}n^{-1/2}; C'_R + u_{1-\alpha/2}(\widehat{V_R(d)})^{1/2}n^{-1/2}\right),$$

where $u_{1-\alpha/2}$ is the quantile of level $1-\alpha/2$ of the standard normal distribution; the confidence interval for the Bray–Curtis similarity measure can be written similarly.

2.4. Testing the Homogeneity Hypothesis Using CS

Finally, we consider the problem of testing the homogeneity hypothesis for the compared populations. At first glance, the use of the Pearson chi-square two-sample criterion solves this problem. However, this criterion has certain applicability conditions, which populations obtained in practice do not always satisfy. In particular, often compared groups of species contain several dominant species, and all other species are found in the amount of 1–2 specimens, which prevents the direct use of the chi-square criterion. On the other hand, it is always possible to test the homogeneity hypothesis for two populations using similarity coefficients.

Thus, we will test hypothesis $H_0 : \mu_R = 1$. In case of coinciding general populations, the similarity measure of these two populations is equal to one if property **A3** is fulfilled for the corresponding SC, so instead of μ_R in the definition of the hypothesis one can substitute any other similarity measure, quantitative or qualitative, that satisfies this property. We will construct a test criterion H_0 using the frequency index Ružička C'_R. Since this index is always ≤ 1, we propose the following criterion:

$$\text{if } C'_R + u_{1-\alpha}(\widehat{V_R(d)})^{1/2}n^{-1/2} < 1, \text{ then reject } H_0,$$

where $u_{1-\alpha}$ is the quantile of level $(1-\alpha)$ of $N(0,1)$. It follows from (12) that this criterion will have asymptotic significance level α. Using the Bray–Curtis frequency index C'_{BC}, one can similarly propose another criterion for testing H_0 :

$$\text{if } C'_{BC} + u_{1-\alpha}(\widehat{V_{BC}(d)})^{1/2}n^{-1/2} < 1, \text{ then reject } H_0; \tag{13}$$

it also has asymptotic significance level α.

3. MODELING

The purpose of this section is to demonstrate the asymptotic properties of the proposed confidence intervals and homogeneity criteria based on SCs. Let us first consider the problem of confidence estimation for the Bray–Curtis similarity measure of two populations with 10 species in each, obeying the distributions $\{p_i\}_{i=1}^{10}$ and $\{q_i\}_{i=1}^{10}$ respectively. For the simulation, we chose truncated Poisson distributions with parameters 3.5 (dark columns) and 5 (light columns) respectively; histograms of the corresponding distributions $\mathcal{P} = \{p_i\}_{i=1}^{10}$ and $\mathcal{Q} = \{q_i\}_{i=1}^{10}$ are shown on Fig. 1.

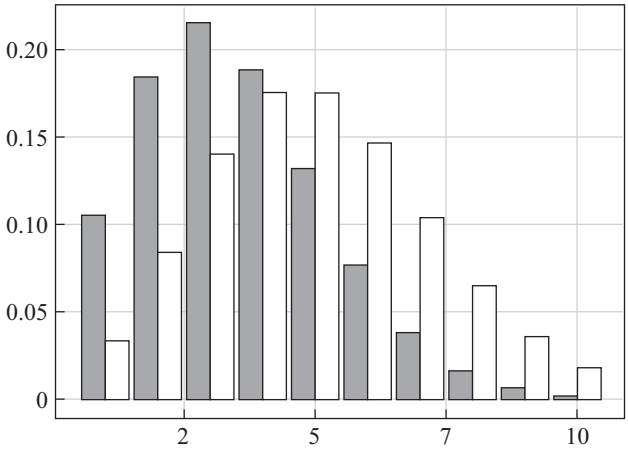

Fig. 1. Histogram of distributions $\{p_i\}_{i=1}^{10}$ (dark columns for $\lambda = 3.5$) and $\{q_i\}_{i=1}^{10}$ (light columns for $\lambda = 5$).

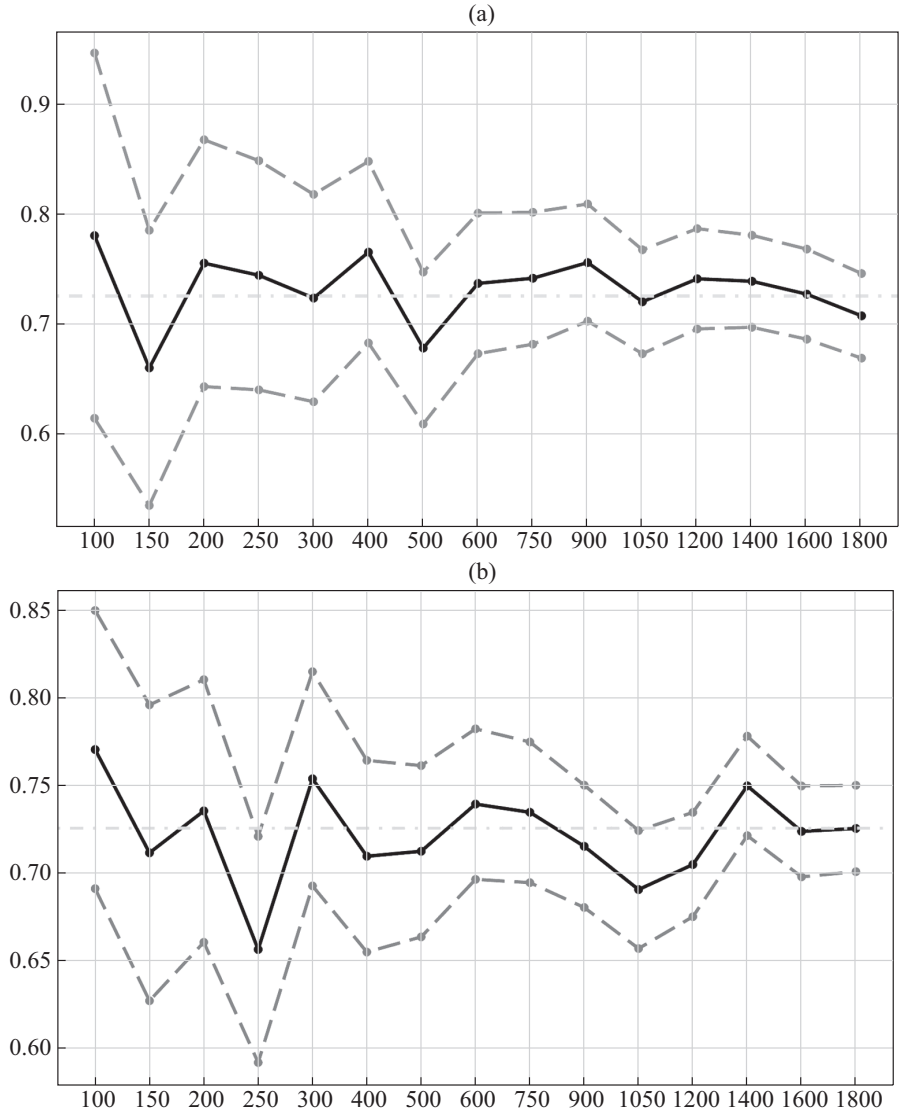

Fig. 2. Confidence intervals for the Bray-Curtis similarity measure and the value of the Bray–Curtis frequency index depending on n (a — $n = 2m$; b — $n = m/3$).

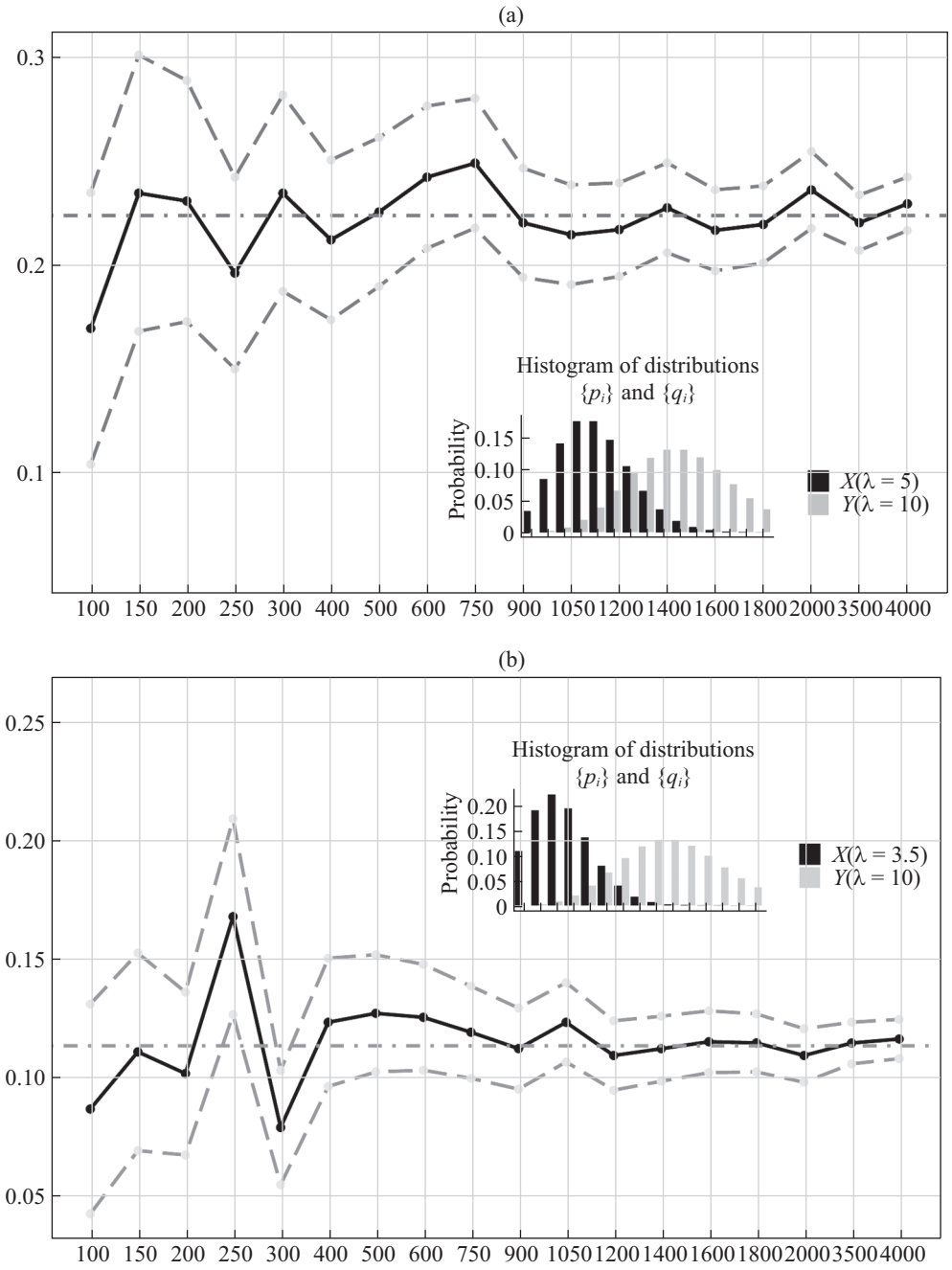

Fig. 3. Confidence intervals for Ružička similarity measure and the value of the Ružička frequency index depending on n, $m = n$.

Figure 2 shows the plots of the upper and lower boundaries of confidence intervals (dashed lines) for confidence level $\alpha = 0.95$ for the Bray–Curtis similarity measure for the distributions $\mathcal{P}$ and $\mathcal{Q}$ and the plot of the frequency index of the Bray–Curtis similarity index C'_{BC} (solid line) depending on n. On the left, $n = 2m$; on the right, $n = m/3$. The true value of the Bray–Curtis measure for these distributions is $\mu_{BC} = 0.76$ with accuracy up to the third decimal place (dash-dot line).

It is easy to see that the constructed confidence intervals differ little in their behavior from the asymptotic confidence intervals constructed on the basis of the standard condition for the asymptotic normality of the estimate. In particular, with increasing sample sizes, the width of

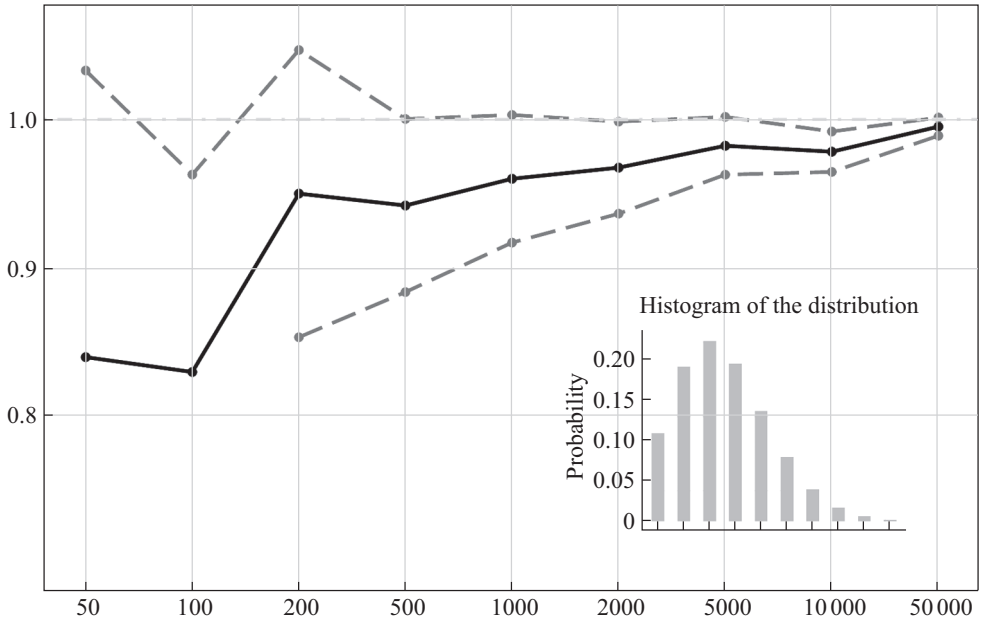

Fig. 4. Confidence intervals for the Bray–Curtis similarity measure depending on n in case of samples from the same population.

confidence intervals decreases, and the true value of the similarity measure almost always falls into the confidence interval.

Figure 3 shows plots of confidence interval boundaries for the Ružička similarity measure; they have the same properties as confidence intervals for the Bray–Curtis similarity measure. Here, the confidence intervals and the values of the Ružička frequency coefficient C'_R themselves are constructed for two different pairs of discrete distributions, whose histograms are displayed on the corresponding plots, depending on n with $n = m$.

Finally, we turn to the problem of testing the homogeneity of two populations using similarity coefficients. Figure 4 shows the upper bound from criterion (13) at $\alpha = 0.025$ (dashed line) for the Bray–Curtis similarity measure (equal to one in this case) for samples from a single general population of 10 species, depending on n for $n = m$; the distribution histogram is also shown in the figure. The plot also shows the lower boundary of the symmetric confidence interval (that is, the confidence level of this interval is 0.95) and values of the Bray–Curtis frequency index (solid line), with which these confidence intervals have been constructed. Note that confidence intervals for different values of n do not always contain one, i.e., for small n the theoretical level of significance of criterion (13) is apparently underestimated. This is due to the fact that in this case, the Bray–Curtis similarity coefficient is always lower than its similarity measure due to the characteristic features of the SC's definition.

4. CONCLUSION

In this work, we have considered the problem of estimating the accuracy of quantitative similarity coefficients. An exhaustive review of publications on this topic has shown that so far no satisfactory algorithm has been proposed that would have a rigorous justification and would let one find the boundaries of the confidence interval for a similarity measure of any SC and/or estimate its variance. In this work, we have proposed a method for constructing asymptotic confidence intervals for similarity measures of the two most commonly used CS, Bray–Curtis and Ružička. Using the above method, one can obtain confidence intervals for similarity measures of any other frequency SC.

There remain open questions about the degree of sensitivity for various SCs and the relationship of asymptotic and bootstrap confidence intervals, which we will consider in future work.

ACKNOWLEDGMENTS

The work of I.V. Radionov in Sections 1 and 2 was supported by the Russian Science Foundation, project no. 19-11-00290 provided by the Steklov Mathematical Institute of the RAS.

REFERENCES

1. Cha, S.-H., Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions, *Int. J. Math. Model. Meth. Appl. Sci*, 2007, vol. 1, no. 4, pp. 300–307.
2. Semkin, B.I., Descriptive Sets and Their Applications, in *Issledovaniya sistem. 1. Slozhnye sistemy* (Systems Research. 1. Complex Systems), Vladivostok, 1973, pp. 83–94.
3. Semkin, B.I., The Axiomatic Approach to Introducing Measures for Ordering and Classification of Descriptive Sets, *Patt. Recogn. Image Anal*, 2011, vol. 21, no. 2, pp. 164–166.
4. Diserud, O.H. and Ødegaard F., A Multiple-Site Similarity Measures, *Biol. Lett*, 2007, vol. 3, no. 1, pp. 20–22.
5. Baselga, A., Jimenez-Valverde, A., and Niccolini, G., A Multiple-Site Similarity Measure Independent of Richness, *Biol. Lett*, 2007, vol. 3, no. 6, pp. 642–645.
6. Cheetham, A.H. and Hazel, J.E., Binary (Presence-Absence) Similarity Coefficients, *J. Paleontol.*, 1969, vol. 43, no. 5, pp. 1130–1136.
7. Pesenko, Yu.A., *Printsipy i metody kolichestvennogo analiza v faunisticheskikh issledovaniyakh* (Principles and Methods of Quantitative Analysis in Fauna Studies), Moscow: Nauka, 1982.
8. Jaccard, P., Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines, *Bulletin de la Société Vaudoise des Sciences Naturelles*, 1901, vol. 37, pp. 241–272.
9. Ružička, M., Anwendung mathematisch-statistischer Methoden in der Geobotanik (Synthetische Bearbeitung von Aufnahmen), *Biológia*, Bratisl., 1958, vol. 13, pp. 647–661.
10. Dice, L.R., Measures of the Amount of Ecologic Association between Species, *Ecology*, 1945, vol. 26, no. 3, pp. 297–302.
11. Sørensen, T., A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content, *Kongelige Danske Videnskabernes Selskab. Biol. krifter*, 1948, Bd. V, no. 4, pp. 1–34.
12. Czekanowski, J., Zur differential Diagnose der Neandertalgruppe, *Korrespbl. Dtsch. Ges. Anthropol.*, 1909, Bd. 40, S. 44–47.
13. Bray, J.R. and Curtis, J.T., An Ordination of Upland Forest Communities of Southern Wisconsin, *Ecol. Monogr.*, 1957, vol. 27, pp. 325–349.
14. Glime, J.M. and Clemons, R.M., Species Diversity of Stream Insects on Fontinalis Spp. Compared to Diversity on Artificial Substrates, *Ecology*, 1972, vol. 53, no. 3, pp. 458–464.
15. Li, X. and Dubes, R.C., A Probabilistic Measure of Similarity for Binary Data in Pattern Recognition, *Patt. Recogn.*, 1989, vol. 22, no. 4, pp. 397–409.
16. Bolton, H.C., On the Mathematical Significance of the Similarity Index of Ochiai as a Measure for Biogeographical Habitats, *Aust. J. Zool.*, 1991, vol. 39, pp. 143–156.
17. Baroni-Urbani, C. and Buser, M.W., Similarity of Binary Data, *Syst. Zool.*, 1976, vol. 25, no. 3, pp. 251–259.
18. Engen, S., Grøtan, V., and Sæther, B.-E., Estimating Similarity of Communities: A Parametric Approach to Spatio-Temporal Analysis of Species Diversity, *Ecography*, 2011, vol. 34, no. 2, pp. 220–231.

19. McCormick, W.P., Lyons, N.I., and Hutcheson, K., Distributional Properties of Jaccard's Index of Similarity, *Commun. Statist. Theor. Meth.*, 1992, vol. 21, no. 1, pp. 51–68.

20. Chao, A., Estimating the Population Size for Capture-Recapture Data with Unequal Catchability, *Biometrics*, 1987, vol. 43, no. 4, pp. 783–791.

21. Chao, A., Hwang, W.-H., Chen, Y.-C., and Kuo, C.-Y., Estimating the Number of Shared Species in Two Communities, *Statist. Sinica*, 2000, vol. 10, pp. 227–246.

22. Chao, A., Chazdon, R.L., Colwell, R.K., and Shen, T.J., A New Statistical Approach for Assessing Similarity of Species Composition with Incidence and Abundance Data, *Ecol. Lett.*, 2005, vol. 8, pp. 148–159.

This paper was recommended for publication by E.Ya. Rubinovich, a member of the Editorial Board