© 2020 г. И.В. РОДИОНОВ, канд. физ.-мат. наук (vecsell@gmail.com)
(Институт проблем управления им. В.А. Трапезникова РАН, Москва;
Математический институт им. В.А. Стеклова, РАН, Москва),
А.Н. СОЗОНТОВ, канд. биол. наук (a.n.sozontov@gmail.com)
(Институт экологии растений и животных УрО РАН, Екатеринбург)

О ДОВЕРИТЕЛЬНОМ ОЦЕНИВАНИИ НА ОСНОВЕ КОЛИЧЕСТВЕННЫХ КОЭФФИЦИЕНТОВ СХОДСТВА[1]

Рассматривается задача оценивания точности количественных коэффициентов сходства. Для этого вводится новое понятие меры сходства соответствующего коэффициента. Показано, что состоятельными оценками своих мер сходства являются только частотные формы количественных коэффициентов сходства. Получены асимптотические доверительные интервалы для мер сходства Ружички и Брея–Кертиса на основе одноименных коэффициентов. Также предложен критерий однородности двух совокупностей на основе упомянутых коэффициентов.

Ключевые слова: коэффициент сходства, доверительное оценивание, критерий однородности, индекс Брея–Кертиса, индекс Жаккара.

DOI: 10.31857/S0005231020020105

1. Введение

Коэффициенты сходства (КС), изначально предложенные биологами, нашли широкое применение в химии, социологии, лингвистике, юриспруденции и т.д., а также в методах работы с многомерными данными, в частности они легли в основу некоторых форм кластерного анализа. В настоящее время насчитывается от нескольких десятков до нескольких сотен КС (см., например, [1]), однако статистическая теория, описывающая КС, практически не развита. Так, при разных соотношениях между объемами выборок большинство используемых количественных КС (определения см. далее), по сути, оценивают различные величины, тогда как данные, используемые для построения качественных КС, часто бывают достаточно бедны, чтобы делать надежные статистические выводы. Существующие методы оценивания точности КС либо основаны на чрезвычайно узких предположениях типа равномерной распределенности видов в совокупности (наиболее часто — для качественных КС), либо вообще не носят строгого математического характера. Предпринято также несколько попыток построить бутстрепные доверительные интервалы для некоторых КС. В настоящей статье получены точные асимптотические доверительные интервалы для наиболее популярных количественных коэффициентов сходства Брея–Кертиса и Ружички и предложен критерий проверки гипотезы однородности двух совокупностей на основе предыдущего результата.

[1] Разделы 1 и 2 статьи выполнены И.В. Радионовым за счет гранта Российского научного фонда (проект №19-11-00290) в Математическом институте им. В.А. Стеклова Российской академии наук.

Пусть X и Y — два дескриптивных множества [2, 3], описывающие две сравниваемые выборки, т.е. конечные множества видов (типов объектов) такие, что каждому виду сопоставлено количество его попаданий в соответствующую выборку. Иными словами, занумеруем виды, встретившиеся в двух исследуемых выборках, числами от 1 до S и обозначим через X_i и Y_i количество объектов i-го вида в первой и второй выборках соответственно. Обозначим через a количество общих видов для двух сравниваемых множеств, а через b и c — количество уникальных видов для первого и второго множества соответственно:

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

Легко видеть, что $S = a + b + c$.

Коэффициентом, или индексом, сходства двух совокупностей $C(X, Y)$ будем называть безразмерный показатель, отражающий меру близости (сходства) указанных совокупностей X и Y. Как правило, индексы сходства рассматриваются для сравнения двух совокупностей, однако существуют методы поиска сходства между тремя и более множествами одновременно [4, 5]. В настоящей статье такие варианты сравнения и соответствующие им КС не рассматриваются. Назовем КС *качественным*, если он зависит только от a, b и c, т.е. на значения таких КС влияет только наличие/отсутствие вида в сравниваемых совокупностях. Коэффициент сходства называется *количественным*, если для его построения используются величины $\{X_i\}_{i=1}^{S}$ и $\{Y_i\}_{i=1}^{S}$. Количественные КС, зависящие только от частот X_i/n и Y_i/m, $1 \leqslant i \leqslant S$, появлений вида i в совокупностях X и Y соответственно, будем называть *частотными*. Здесь $n = \sum_{i=1}^{S} X_i$ и $m = \sum_{i=1}^{S} Y_i$. Для любого качественного КС можно предложить его количественный аналог, заменив индикаторы $I(X_i \neq 0)$, $I(Y_i \neq 0)$, $i = 1, \ldots, S$, присутствия видов в совокупности на частоты X_i/n и Y_i/m, $1 \leqslant i \leqslant S$. Как будет показано далее, введение других количественных аналогов для качественных КС не оправдано со статистической точки зрения.

Обсудим общие требования, которые, как правило, налагаются на индексы сходства [6, 7]:

- **A1.** Симметричность: $C(X, Y) = C(X, Y)$;
- **A2.** Равенство нулю для непересекающихся совокупностей: $C(X, Y) = 0$, если $a = 0$;
- **A3.** Равенство единице для совпадающих совокупностей: $C(X, Y) = 1$, если $a = b = c$ для качественных КС и $X_i/n = Y_i/m$ $\forall i$, $1 \leqslant i \leqslant S$, для частотных КС;
- **A4.** "Монотонность" по величине сходства.

Для качественных индексов сходства, в частности, последнее свойство означает следующее: если зафиксировать S и множество видов, то значение КС должно быть тем больше, чем больше значение a. Впрочем, далеко не все индексы сходства удовлетворяют условиям **A1**–**A4**, см. [1]. Далее, определим также меру сходства μ качественного КС как величину, равную этому

КС в случае, если бы вместо выборок КС вычислялся по генеральным совокупностям, из которых взяты данные выборки. Ясно, что при росте размера выборок к бесконечности качественный КС будет сходиться к своей мере сходства. Определим также меру сходства количественного КС как величину, получающуюся при замене X_i и Y_i в записи данного КС на вероятности p_i и q_i появления i-го вида из первой и второй генеральной совокупности соответственно. В разделе 2 будет показано, что количественные КС будут состоятельными оценками своих мер сходства тогда и только тогда, когда они являются частотными.

Необходимость сравнения множеств стояла перед биологами еще в XIX в., однако способы давать степени их (не)сходства количественную оценку появились лишь в начале XX в. По-видимому, самый первый КС, I_J, который до сих пор наиболее популярен среди индексов сходства, предложил швейцарский ботаник Поль Жаккар [8]. По своей сути I_J есть отношение мощности пересечения множеств видов в двух совокупностях к мощности их объединения,

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

и является качественным коэффициентом сходства. Его количественный аналог носит название коэффициента Ружички [9]

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

Другой популярный КС был предложен практически одновременно Л.Р. Дайсом в [10] и Т. Сёренсеном в [11]

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

количественная форма которого была предложена задолго до них Чекановским в [12], она также известна под названием индекса Брея–Кертиса [13]

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

Легко видеть, что индекс Жаккара выражается через индекс Сёренсена–Дайса как $I_J = I_{DS}/(2 - I_{DS})$. К настоящему времени предложены десятки качественных КС. Наряду с индексами Жаккара и Сёренсена–Дайса наиболее используемы индексы Охиаи I_O, Кульчинского I_K и Мориситы I_M, см. [7]. Все они монотонно возрастают от нуля до единицы в зависимости от количества общих видов и, по сути, отличаются лишь разной чувствительностью к малым и большим значениям a по сравнению с S.

Впервые попытка оценить точность индексов сходства была предпринята Сёренсеном в [11], однако его метод требует наличия не двух, а достаточно большого количества выборок, что не всегда осуществимо. Немалое количество публикаций посвящено доверительному оцениванию качественной меры сходства в предположении, что все виды в генеральной совокупности распределены одинаково, см. [14–17], что, очевидно, никогда не выполняется на практике. Однако стоит отметить, что если имеются лишь данные о наличии/отсутствии вида в выборках и отсутствуют сведения о количестве объектов каждого из видов в совокупности, то выбор любого другого распределения не является обоснованным. Кроме того, величины a, b и c сильно зависят от наличия редко встречающихся видов в выборке. При росте количества наблюдений соотношения между a, b и c могут существенно измениться, что препятствует точности статистического анализа качественных КС при малом и среднем количестве наблюдений. В связи с доверительным оцениванием качественной меры сходства отметим также публикацию [18], где предполагается, что распределение видов в генеральной совокупности является дискретным логнормальным, и публикацию [19], где принято предположение, что один доминантный вид встречается чаще остальных, которые уже имеют равную вероятность попадания в выборку.

Для построения доверительных интервалов для качественных мер сходства может быть полезен основанный на бутстрепе метод оценивания количества видов в генеральной совокупности по выборке из нее, развитый Чао в [20–22]. Так, в [22] в предположении, что индекс Жаккара меньше своей меры сходства, предложен доверительный интервал для меры сходства I_J, впрочем, без какого-либо математического обоснования. Несмотря на перспективность метода, Чао не удалось построить доверительный интервал для какого-либо качественного КС в общих предположениях. Авторам данной статьи неизвестны работы, где была бы рассмотрена задача построения доверительных интервалов для мер сходства на основе количественных КС.

2. Основные результаты

2.1. Статистическая корректность количественных КС

Прежде всего покажем на основе индекса сходства Ружички (1), что количественные КС являются состоятельными оценками своих мер сходства только в случае $n/m \to 1$ при $n, m \to \infty$, где n и m, напомним, — размеры первой и второй совокупности соответственно. Рассмотрим в рамках исследуемой задачи две полиномиальных модели: в j-м испытании независимо от других испытаний появляется по одному объекту из каждой генеральной совокупности согласно распределениям вероятностей $\{p_i\}_{i\geqslant 1}$ и $\{q_i\}_{i\geqslant 1}$ соответственно, т.е. в j-м испытании i-й объект выпадает с вероятностями p_i и q_i для первой и второй группы соответственно. Обозначим случайные величины, соответствующие выпадению объекта определенного вида в первой и второй группе в j-м испытании как ξ_j и η_j. Таким образом, имеем

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

Тогда поскольку $\{I(\xi_j = i)\}_{j\geqslant 1}$ и $\{I(\eta_j = i)\}_{j\geqslant 1}$ — последовательности независимых одинаково распределенных случайных величин, то по усиленному закону больших чисел

$$\begin{aligned} &\frac{X_i}{n} \xrightarrow{\text{п.н.}} EI(\xi_j = i) = P(\xi_j = i) = p_i, \\ &\frac{Y_i}{m} \xrightarrow{\text{п.н.}} q_i \qquad (\text{п.н. — почти наверное}) \end{aligned} \tag{3}$$

при $n, m \to \infty$.

Вернемся к обсуждению индекса Ружички. Его мера сходства, очевидно, равна

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

Тогда при использовании теоремы о наследовании сходимостей, свойств сходимости почти наверное и соотношений (3) в случае $n/m \to 1$ при $n, m \to \infty$ элементарно показывается, что

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

С другой стороны, при $n/m = d$, $n, m \to \infty$ индекс Ружички почти наверное сходится к величине

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

отличной при $d \neq 1$ от μ_R, тогда как частотный аналог данного индекса

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

сходится почти наверное к μ_R в независимости от соотношения между m и n. Аналогичные рассуждения справедливы и для индекса Брея–Кертиса (2), и для других количественных индексов сходства, которые не являются частотными. Тем самым выводы о генеральной совокупности, полученные на основе количественных КС, не являющихся частотными, не могут считаться достоверными и статистически корректными.

Рассмотрим индекс сходства Ружички и построим доверительный интервал для его меры сходства. Преобразуем данный индекс следующим образом:

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

т.е. фактически индекс Ружички зависит только от $\sum_{i=1}^{S} \max(X_i, Y_i)$.

Предположим сначала, что $n = m$. Пусть $p_i > q_i$, тогда при $n \to \infty$ выполнено, что $P(X_i > Y_i) \to 1$ и $P(\max(X_i, Y_i) = X_i) \to 1$, т.е. при больших n выпадение i-го объекта для второй совокупности не будет влиять на значение индекса Ружички с вероятностью, близкой к единице. Определим A как множество номеров i, таких что $p_i > q_i$, $A = \{i : p_i \geqslant q_i\}$, и $B = \{i : q_i > p_i\}$. Определим также

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

Тогда мера сходства индекса Ружички переписывается в виде

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

Теорема. Пусть $n = m$. Тогда при $n \to \infty$

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

где

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

Доказательство теоремы. Определим

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

Эти случайные величины, очевидно, могут принимать значения $0, 1$ и 2. Заметим, что поскольку по условию количество видов S конечно, то

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

при $n \to \infty$. С другой стороны,

$$\sum_{i\in A} X_i + \sum_{i\in B} Y_i = \sum_{i\in A}\sum_{j=1}^{n} I(\xi_j = i) + \sum_{i\in B}\sum_{j=1}^{n} I(\eta_j = i) =$$

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

Тем самым для доказательства теоремы достаточно показать, что при $n \to \infty$

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

и

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

где статистика $2n/T_n - 1$ получена подстановкой суммы T_n в выражение (5) вместо $\sum_{i=1}^{S} \max(X_i, Y_i)$, и воспользоваться леммой Слуцкого.

Докажем сначала соотношение (8). Заметим, что $\{\zeta_j\}_{j=1}^{n}$ — независимые одинаково распределенные случайные величины, и найдем $E\zeta_1$ и $Var\zeta_1$. Имеем, что

$$E\zeta_1 = EI(\xi_1 = i, i \in A) + EI(\eta_1 = i, i \in B) =$$

$$= \sum_{i\in A} P(\xi_1 = i) + \sum_{i\in B} P(\eta_1 = i) = \sum_{i\in A} p_i + \sum_{i\in B} q_i = P + Q,$$

$$Var\zeta_1 = VarI(\xi_1 = i, i \in A) + VarI(\eta_1 = i, i \in B) = P(1-P) + Q(1-Q).$$

Из центральной предельной теоремы получаем, что

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

Применяя дельта-метод для функции $g(x) = 2/x$, находим, что

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

откуда и следует соотношение (8).

Докажем теперь соотношение (9). Используя (6) и (7), получаем при $n \to \infty$, что

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

Поскольку по закону больших чисел $T_n/n \to P+Q$ по вероятности при $n \to \infty$, то из соотношения (10) следует, что $\sum_{i=1}^{S} \max(X_i, Y_i)/n$ также стремится по вероятности к $P+Q$. Разделив левую часть (10) на $T_n \sum_{i=1}^{S} \max(X_i, Y_i)/(2n^2)$, имеем по лемме Слуцкого, что

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

откуда и следует (9). Тем самым, доказательство теоремы закончено.

Обратимся теперь к изучению частотного аналога индекса Ружички C'_R (4). Следующее утверждение говорит, что асимптотическое поведение C'_R несколько отличается от асимптотического поведения индекса Ружички при $n = m \to \infty$.

Предложение 1. При $n, m \to \infty$ и $n/m \to d > 0$ выполнено, что

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

где

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

Доказательство предложения 1. Пользуясь тем, что $\sum_{i\in A} X_i$ и $\sum_{i\in B} Y_i$ представимы как суммы независимых одинаково распределенных случайных величин и независимы, из центральной предельной теоремы получаем, что

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

и

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

при $n \to \infty$ и $m \to \infty$ соответственно, откуда в условиях предложения с помощью леммы Слуцкого имеем, что

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

Дальнейшее доказательство повторяет с легкими изменениями соответствующие шаги доказательства теоремы.

Рассмотрим теперь частотную форму индекса Брея–Кертиса (2),

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

и исследуем ее асимптотическое поведение. Стоит заметить, что частотная форма широко используемого индекса Охиаи $I_O = a/\sqrt{(a+b)(a+c)}$ также равна C'_{BC}. Из соотношения (11) и из модификации формулы (10) для $n \neq m$ вытекает следующее утверждение.

Предложение 2. В условиях предложения 1 выполнено, что

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

где $\mu_{BC} = (2 - P - Q)$ *— мера сходства Брея–Кертиса и* $V_{BC}(d) = P(1-P) + dQ(1-Q)$.

2.3. Доверительное оценивание количественных мер сходства

Чтобы построить доверительные интервалы для мер сходства Ружички и Брея–Кертиса, необходимо оценить асимптотическую дисперсию одноименных частотных КС. Рассмотрим случайное множество номеров $A' = \{i : X_i/n > Y_i/m\}$. Покажем, что

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

Действительно,

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

по закону больших чисел. Снова пользуясь законом больших чисел и свойствами сходимости по вероятности, имеем, что

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

а сходимость $\widehat{Q}$ к Q по вероятности доказывается аналогично.

Используя теорему о наследовании сходимостей и свойства сходимости по вероятности, легко показать, что при $n, m \to \infty$ и $n/m \to d$ выполнено, что

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

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

Тем самым из леммы Слуцкого в тех же условиях получаем, что

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

что позволяет выписать асимптотические доверительные интервалы для мер сходства Ружички и Брея–Кертиса: с вероятностью, стремящейся к $1-\alpha$, мера сходства Ружички принадлежит интервалу

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

где $u_{1-\alpha/2}$ — квантиль уровня $1-\alpha/2$ стандартного нормального распределения; доверительный интервал для меры сходства Брея–Кертиса выписывается аналогично.

2.4. Проверка гипотезы однородности с помощью КС

Наконец, рассмотрим задачу проверки гипотезы об однородности сравниваемых совокупностей. На первый взгляд, применение двухвыборочного критерия хи-квадрат Пирсона решает данную задачу, однако критерий имеет определенные условия применимости, которым получаемые на практике совокупности не всегда удовлетворяют. Так, часто в сравниваемых группах видов бывает несколько доминирующих видов, а все остальные виды встречаются в количестве 1–2 экземпляров, что препятствует прямому применению критерия хи-квадрат. С другой стороны, проверить гипотезу однородности двух совокупностей при помощи коэффициентов сходства можно всегда.

Итак, будем проверять гипотезу $H_0:\mu_R=1$. В случае совпадающих генеральных совокупностей мера сходства двух этих совокупностей равняется единице, если для соответствующего КС выполнено свойство **A3**, поэтому вместо μ_R в определение гипотезы можно подставить любую другую меру сходства, количественную или качественную, для которой выполняется данное свойство. Построим критерий проверки H_0 с помощью частотного индекса Ружички C'_R. Поскольку данный индекс всегда меньше либо равен 1, предложим такой критерий:

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

где $u_{1-\alpha}$ — $(1-\alpha)$-квантиль $N(0,1)$. Из (12) следует, что данный критерий будет иметь асимптотический уровень значимости α. С помощью частотного индекса Брея–Кертиса C'_{BC} аналогичным образом можно предложить еще один критерий проверки H_0:

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

он также имеет асимптотический уровень значимости α.

3. Моделирование

Цель данного раздела — демонстрация асимптотических свойств предложенных доверительных интервалов и критериев однородности, основанных на КС. Рассмотрим сначала задачу доверительного оценивания меры сходства Брея–Кертиса двух генеральных совокупностей с 10 видами в каждой, подчиняющихся распределениям $\{p_i\}_{i=1}^{10}$ и $\{q_i\}_{i=1}^{10}$ соответственно. Для моделирования были выбраны усеченные распределения Пуассона с параметрами 3,5 (темные столбцы) и 5 (светлые столбцы) соответственно, гистограммы соответствующих распределений $\mathcal{P} = \{p_i\}_{i=1}^{10}$ и $\mathcal{Q} = \{q_i\}_{i=1}^{10}$ представлены на рис. 1.

На рис. 2 выведены графики верхней и нижней границ доверительных интервалов (штриховые линии) уровня доверия $\alpha = 0{,}95$ для меры сходства Брея–Кертиса распределений $\mathcal{P}$ и $\mathcal{Q}$ и график значений частотного индекса сходства Брея–Кертиса C'_{BC} (сплошная линия) в зависимости от n. На рисунке слева $n = 2m$, на рисунке справа $n = m/3$. Истинное значение меры Брея–Кертиса для данных распределений равно $\mu_{BC} = 0{,}76$ с точностью до третьего знака после запятой (штрихпунктирная линия).

Легко видеть, что построенные доверительные интервалы по своему поведению мало отличаются от асимптотических доверительных интервалов, построенных на основе стандартного условия асимптотической нормальности оценки. В частности, при возрастании размеров выборок ширина доверительных интервалов уменьшается, а истинное значение меры сходства практически всегда попадает в доверительный интервал.

Приведенные на рис. 3 графики доверительных границ для меры сходства Ружички обладают теми же свойствами, что и доверительные границы для меры сходства Брея–Кертиса. Здесь доверительные границы и сами значения частотного коэффициента Ружички C'_R построены для двух разных пар

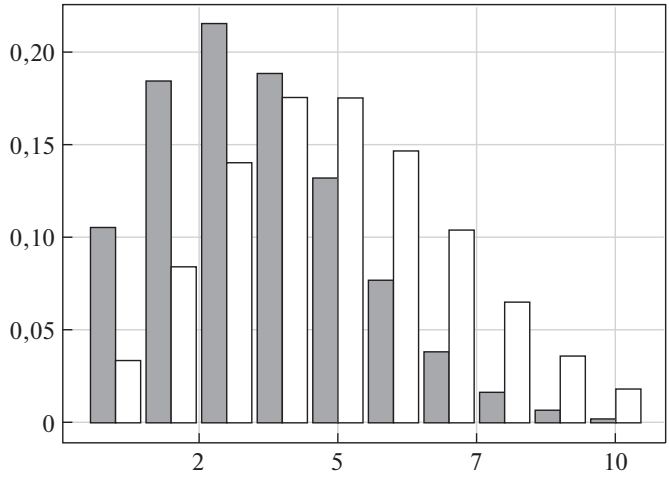

Рис. 1. Гистограмма распределений $\{p_i\}_{i=1}^{10}$ (темные столбцы для $\lambda = 3{,}5$) и $\{q_i\}_{i=1}^{10}$ (светлые столбцы для $\lambda = 5$).

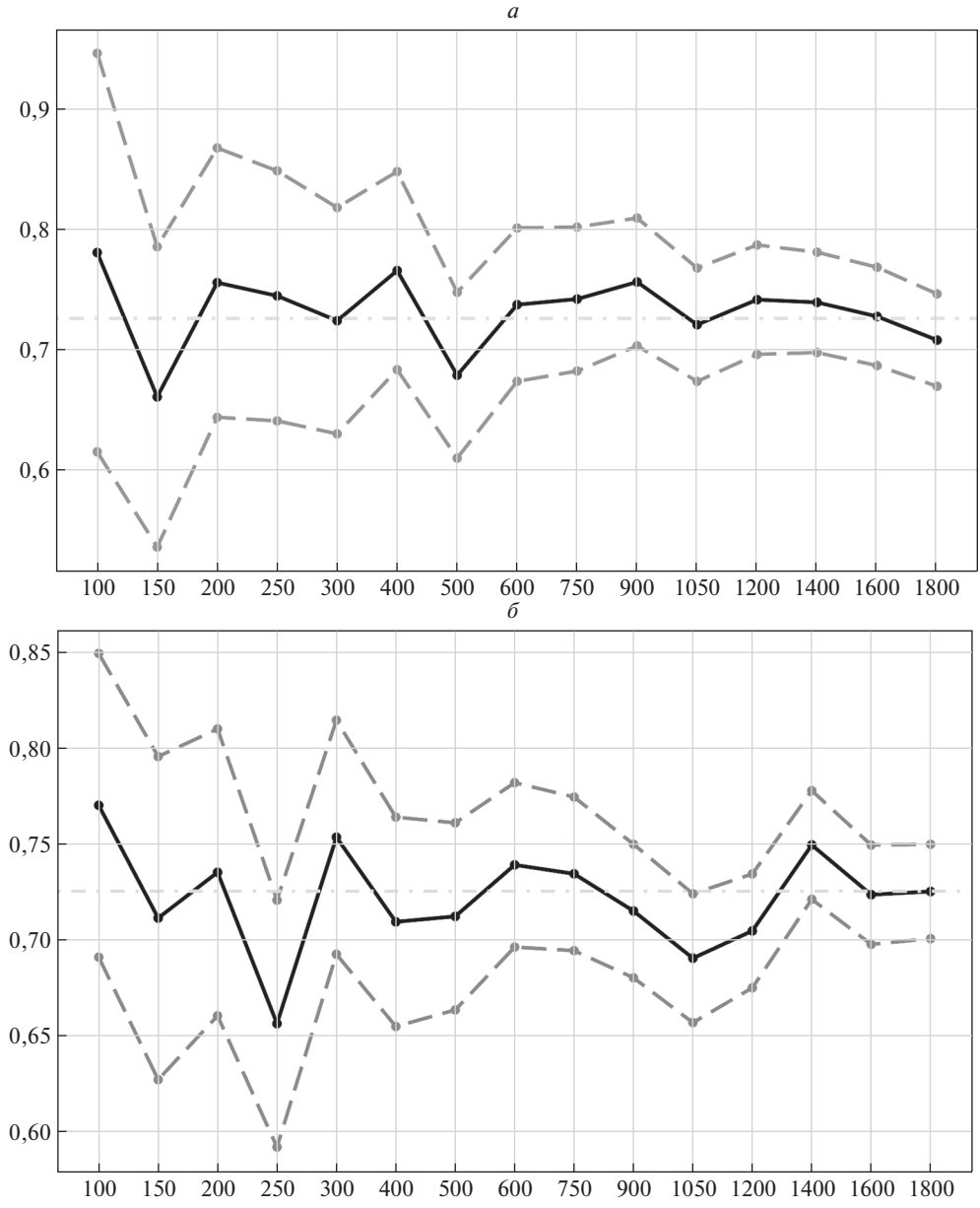

Рис. 2. Доверительные границы для меры сходства Брея–Кертиса и значение частотного индекса Брея–Кертиса в зависимости от n ($а$ — $n = 2m$, $б$ — $n = m/3$).

дискретных распределений, гистограммы которых выведены на соответствующих графиках, в зависимости от n при $n = m$.

Наконец, обратимся к задаче проверки однородности двух совокупностей с помощью коэффициентов сходства. На рис. 4 представлена верхняя граница из критерия (13) при $\alpha = 0{,}025$ (штриховая линия) для меры сходства Брея–

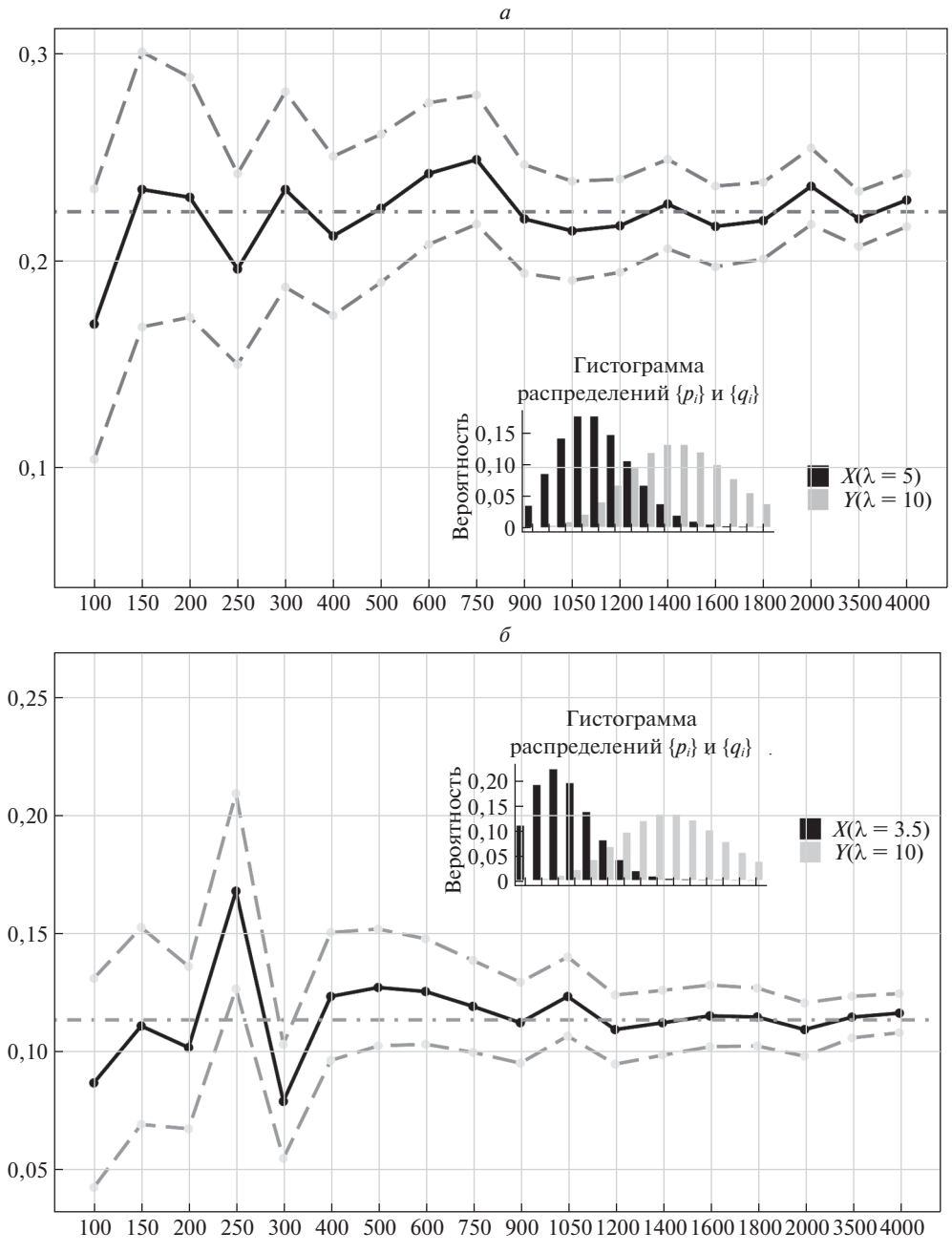

Рис. 3. Доверительные границы для меры сходства Ружички и значения частотного индекса Ружички для выборок X и Y из усеченных распределений Пуассона с параметрами *а* – (5, 10) и *б* – (3,5, 10) в зависимости от n, $m = n$.

Кертиса (равной в данном случае единице) для выборок из одной генеральной совокупности из 10 видов в зависимости от n при $n = m$, гистограмму распределения тоже см. на рисунке. Также на графике выведена нижняя граница

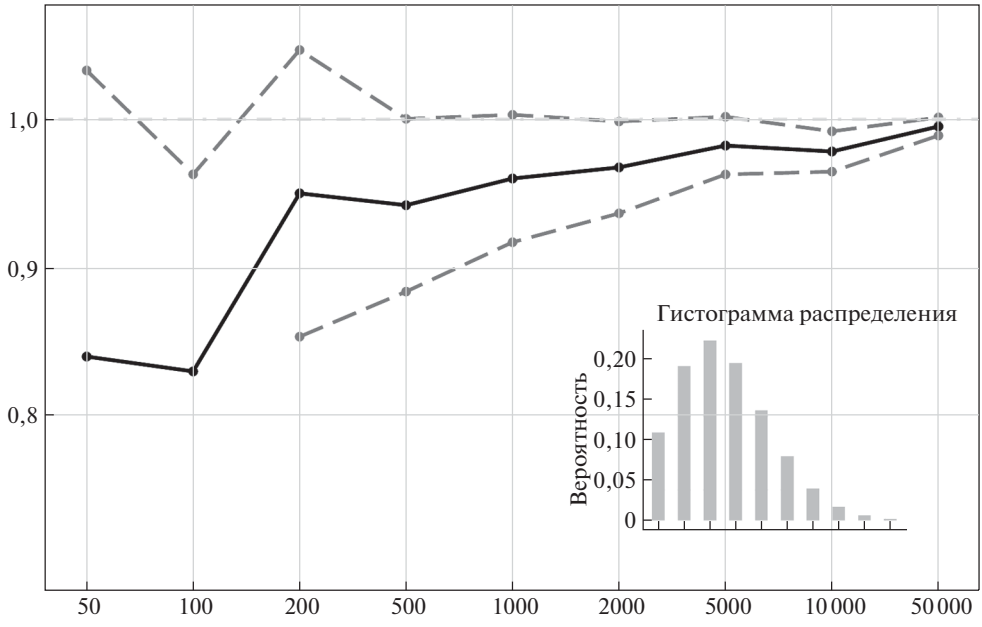

Рис. 4. Доверительные границы для меры сходства Брея–Кертиса в зависимости от n в случае выборок из одной генеральной совокупности.

симметричного доверительного интервала (т.е. уровень доверия данного интервала равен 0,95) и значения частотного индекса Брея–Кертиса (сплошная линия), с помощью которого и построены данные доверительные интервалы. Заметим, что доверительные интервалы для разных значений n не всегда содержат единицу, т.е. при малых n теоретический уровень значимости критерия (13), по-видимому, является заниженным. Это связано с тем, что в данном случае коэффициент сходства Брея–Кертиса всегда ниже своей меры сходства из-за особенностей определения КС.

4. Заключение

В статье рассмотрена проблема оценивания точности количественных коэффициентов сходства. Исчерпывающий обзор публикаций, касающихся этой темы, показал, что до сих пор не было предложено удовлетворительного алгоритма, имеющего строгое обоснование, позволяющего найти границы доверительного интервала для меры сходства какого-либо КС и/или оценить его дисперсию. В настоящей статье авторы предлагают способ построения асимптотических доверительных интервалов для мер сходства двух наиболее часто применяемых КС — Брея–Кертиса и Ружички. Используя приведенный метод, можно получить доверительные интервалы для мер сходства любых других частотных КС. Остаются открытыми вопросы о степени чувствительности различных КС и о взаимоотношении асимптотических и бутстрепных доверительных интервалов, которые будут рассмотрены авторами в дальнейшем.

СПИСОК ЛИТЕРАТУРЫ

1. *Cha S.-H.* Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions // Int. J. Math. Model. Meth. Appl. Sci. 2007. V. 1. No. 4. P. 300–307.
2. *Семкин Б.И.* Дескриптивные множества и их приложения // Исследования систем. 1. Сложные системы. Владивосток: 1973. С. 83–94.
3. *Semkin B.I.* The Axiomatic Approach to Introducing Measures for Ordering and Classification of Descriptive Sets // Patt. Recogn. Image Anal. 2011. V. 21. No. 2. P. 164–166.
4. *Diserud O.H., Ødegaard F.* A Multiple-Site Similarity Measures // Biol. Lett. 2007. V. 3. No. 1. P. 20–22.
5. *Baselga A., Jimenez-Valverde A., Niccolini G.* A Multiple-Site Similarity Measure Independent of Richness // Biol. Lett. 2007. V. 3. No. 6. P. 642–645.
6. *Cheetham A.H., Hazel J.E.* Binary (Presence-Absence) Similarity Coefficients // J. Paleontol. 1969. V. 43. No. 5. P. 1130–1136.
7. *Песенко Ю.А.* Принципы и методы количественного анализа в фаунистических исследованиях. М.: Наука, 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. V. 37. P. 241–272.
9. *Ružička M.* Anwendung mathematisch-statistischer Methoden in der Geobotanik (Synthetische Bearbeitung von Aufnahmen) // Biológia, Bratisl. 1958. V. 13. P. 647–661.
10. *Dice L.R.* Measures of the Amount of Ecologic Association between Species // Ecology. 1945. V. 26. No. 3. P. 297–302.
11. *Sörensen 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. P. 1–34.
12. *Czekanowski J.* Zur differential Diagnose der Neandertalgruppe // Korrespbl. Dtsch. Ges. Anthropol. 1909. Bd 40. S. 44–47.
13. *Bray J.R., Curtis J.T.* An Ordination of Upland Forest Communities of Southern Wisconsin // Ecol. Monogr. 1957. V. 27. P. 325–349.
14. *Glime J.M., Clemons R.M.* Species Diversity of Stream Insects on Fontinalis Spp. Compared to Diversity on Artificial Substrates // Ecology. 1972. V. 53. No. 3. P. 458–464.
15. *Li X., Dubes R.C.* A Probabilistic Measure of Similarity for Binary Data in Pattern Recognition // Patt. Recogn. 1989. V. 22. No. 4. P. 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. V. 39. P. 143–156.
17. *Baroni-Urbani C., Buser M.W.* Similarity of Binary Data // Syst. Zool. 1976. V. 25. No. 3. P. 251–259.
18. *Engen S., Grøtan V., Sæther B.-E.* Estimating Similarity of Communities: a Parametric Approach to Spatio-Temporal Analysis of Species Diversity // Ecography. 2011. V. 34. No. 2. P. 220–231.
19. *McCormick W.P., Lyons N.I., Hutcheson K.* Distributional Properties of Jaccard's Index of Similarity // Commun. Statist. Theor. Meth. 1992. V. 21. No. 1. P. 51–68.

20. *Chao A.* Estimating the Population Size for Capture-Recapture Data with Unequal Catchability // Biometrics. 1987. V. 43. No. 4. P. 783–791.

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

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

Статья представлена к публикации членом редколлегии Е.Я. Рубиновичем.

Поступила в редакцию 16.04.2019

После доработки 07.07.2019

Принята к публикации 18.07.2019