Об одном подходе к вычислению ранговой корреляции. Часть I

Б. Ф. Мельников.

Аннотация—**При вычислении ранговой корреляции возникает та же самая ситуация, что и в некоторых других предметных областях: у исследователей возникло мнение, что новые возможные построения в этой предметной области либо невозможны (всё уже сделано), либо не нужны. В статье мы пытаемся показать, что возможны и дальнейшие теоретические разработки, которые могут привести к достаточно интересным практическим результатам. Мы предлагаем свой вариант вычисления коэффициента ранговой корреляции, который – в его самом простом виде – можно считать «располагающимся между» коэффициентами Кендалла и Спирмена. Конкретнее, мы используем не только условия совпадения результатов предикатов, определяющих порядок элементов соответствующих пар (как при вычислении коэффициента Кендалла), и не только обобщение общего варианта вычисления парной корреляции на случай ранговой (как при вычислении коэффициента Спирмена) – а оба этих приёма вместе. Полученные формулы мы применяем для вычисления разных вариантов коэффициентов ранговой корреляции в нескольких различных предметных областях – и рассуждениями пытаемся обосновать то, что предлагаемый нами критерий достаточно удачен. Также мы предлагаем варианты обобщения нашего критерия – причём мы считаем, что эти возможные обобщения не совпадают с обобщениями, приводимыми в классических монографиях (но и не противоречат им, а дополняют их).**

Ключевые слова—**коэффициент корреляции, ранговая корреляция Спирмена, ранговая корреляция Кендалла, новый вариант коэффициента ранговой корреляции.**

I. ВВЕДЕНИЕ

В самых разных исследованиях, относящихся к различным предметным областям, большую роль играет вычисление коэффициента корреляции между некоторыми заданными случайными величинами – и вряд ли в этой фразе есть что-то новое. Однако, по мнению автора, здесь возникает та же самая ситуация, что и в некоторых других предметных областях: в связи с многочисленными теоретическими публикациями и ещё более многочисленными примерами практического применения у исследователей возникло мнение, что в этой области всё уже сделано, что к имеющимся формулам (конечно же, вполне приемлемым) надо относиться как к «истине в последней инстанции» [1]. В статье мы пытаемся показать, что возможны и дальнейшие теоретические разработки – причём даже «в самых стартовых» определениях – которые могут привести к достаточно интересным практическим результатам.

В статье мы рассматриваем *ранговую корреляцию* – её можно считать частным случаем парной корреляции. (При этом, по-видимому, можно всегда говорить просто «корреляция» – употребляемые в статье термины согласованы с [1], [2] и другими источниками.) Ранговая корреляция – это корреляция двух случайных величин, заданных несколькими соответствующими друг другу парами значений этих величин; при этом в общем случае для вычисления коэффициента корреляции рассматриваемые случайные величины, конечно, могут быть заданы и как непрерывные, т. е. как две функции одного и того же аргумента, с помощью которых при необходимости можно сформировать пары значений.

Итак, в настоящей статье мы пытаемся показать, что возможны дальнейшие *теоретические* разработки, которые могут привести к достаточно интересным *практическим* результатам. Для этого мы предлагаем *свой вариант* вычисления коэффициента ранговой корреляции, который – в его самом простом виде – можно считать «располагающимся между» коэффициентами Кендалла и Спирмена (подробнее об этом «расположении» см. далее). Конкретнее, мы используем:

- не только условия совпадения результатов предикатов, определяющих порядок элементов в соответствующих парах (как при вычислении коэффициента Кендалла),
- и не только обобщение общего варианта вычисления парной корреляции на случай корреляции ранговой (как при вычислении коэффициента Спирмена) –

а оба этих приёма вместе. Полученные формулы мы применяем для вычисления разных вариантов коэффициентов ранговой корреляции в нескольких различных предметных областях – и *рассуждениями* пытаемся обосновать то, что предлагаемый нами критерий достаточно удачен. Также мы предлагаем варианты обобщения нашего критерия – причём мы считаем, что эти возможные обобщения не совпадают с обобщениями, приводимыми в классических монографиях (но и не противоречат им, а дополняют их).

Приведём содержание части I статьи по разделам.

В разделе II описаны стандартные классические подходы к вычислению коэффициента ранговой корреляции: обычная формула коэффициента корреляции, а также коэффициенты ранговой корреляции Спирмена и Кендалла (последний приводится в двух вариантах, оба по сравнению с классикой немного модифицированы). При-

Статья получена 17 августа 2024 г

Борис Феликсович Мельников, Университет МГУ – ППИ в Шэньчжэне (bormel@mail.ru, bormel@smbu.edu.cn).

[1] На эту тему приведём такой пример. В нескольких работах автора, посвящённых анализу близости ДНК-цепочек, аналогичная мысль высказывалась по поводу вычисленной ранее (им же и многими другими авторами) близости между некоторыми геномами. Например, «кто-то как-то когда-то» посчитал, что близость между геномами человека и шимпанзе составляет 99% – и это значение цитируется во множестве последующих публикаций, в том числе – в научно-популярной литературе. Однако есть же и другие методы подсчёта близости геномов, которые могут дать иные численные результаты …

водимые обозначения в целом согласованы с известными монографиями.

Название раздела III – «Плюсы и минусы обычных вариантов вычисления ранговой корреляции»; обычными мы считаем такие варианты:

- «стандартный» метод – т. е. просто обычный коэффициент парной корреляции;
- коэффициент Спирмена;
- коэффициент Кендалла, немного модифицированный; основная идея Кендалла сохраняется, т. е. мы считаем число обменов, необходимых для выстраивания элементов второй последовательности в том же порядке, в котором идут элементы последовательности первой;
- коэффициент Кендалла, сильно модифицированный; изменение нужно из-за того, что при большом количестве точных совпадений в значениях рассматриваемых случайных величин коэффициент Кендалла даёт не очень адекватные результаты – поэтому для этого сильно модифицированного коэффициента мы при любом варанте совпадения условно считаем значение этого коэффициента для рассматриваемой пары пар значений равным 0[2].

После вычисления коэффициентов корреляции для всего множества пар пар значений мы – в тех случаях, где это требуется – усредняем полученные значения.

В этом же разделе III мы, среди прочего, отмечаем, что даты публикаций определений коэффициентов показывают «движение по степени важности», в сторону убывания, *самих значений* элементов последовательностей – и, соответственно, увеличения важности значений предиката, сравнивающего эти значения.

Также в этом разделе мы рассматриваем несколько похожих вариантов для специально подобранного примера двух последовательностей пар; все эти варианты мы в статье называем «нулевым примером». Для каждого из этих вариантов мы приводим посчитанные значения коэффициентов ранговой корреляции – и даже минимальное полученное нами значение (для разных вариантов его вычисления) превышает 0.75; однако мы такое значение для рассмотренных примеров считаем слишком большим – и будем далее его «исправлять» в оставшейся части статьи.

В разделе IV мы предлагаем свой вариант вычисления ранговой корреляции. В нём мы рассматриваем множество пар пар случайных величин, определяем формулу коэффициента для пары пар таких значений, а итоговое значение корреляции получается усреднением всех полученных значений коэффициенов.

Разделе V – продолжение рассмотрения «нулевого примера» и его модификаций для нашего варианта вычисления ранговой корреляции. В нём мы показываем, что на специально подобранных примерах наш вариант может быть более адекватен – и, более того, «неадекватные» значения других вариантов могут располагаться *с разных сторон* от ожидаемого (естественного, адекватного) значения.

В разделе VI мы рассмотрим только описание многочисленных предметных областей, в которых, *по нашему мнению*, предлагаемый вариант вычисления ранговой корреляции даёт на уже рассмотренных примерах более адекватные результаты. При этом мы в части I приводим только краткое их описание – а *подробно примеры практического применения* предлагаемого коэффициента и полученные результаты – в сравнении с результатами для других вариантов коэффициентов – *будут приведены в части II*. Также в части II мы рассмотрим предлагаемое обобщение нашего варианта вычисления ранговой корреляции.

II. ПРЕДВАРИТЕЛЬНЫЕ СВЕ́ДЕНИЯ. КЛАССИЧЕСКИЕ ПОДХОДЫ К ВЫЧИСЛЕНИЮ КОЭФФИЦИЕНТА РАНГОВОЙ КОРРЕЛЯЦИИ

В этом разделе мы приведём обычные статистические характеристики, которые будут использованы в настоящей статье. Как мы уже отмечали, обозначения в целом согласованы с известными монографиями [1], [2] – но иногда мы будем использовать обозначения «более математические», например, *не будем* использовать записи вроде MXY и.т. п.

Две рассматривыемые случайные величины будут в этом разделе обозначаться X и Y; их наблюдаемые реальные реализации обозначаются таким же образом соответствующими нижними индексами, т. е.,

$$X_i \quad \text{и} \quad Y_i \quad \text{для} \quad i = 1, 2, \ldots, N.$$

Теперь сформулируем *обычное определение корреляции*: коэффициент корреляции может быть вычислен по формуле

$$R(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \cdot \sigma_Y},$$

где

$$\operatorname{cov}(X, Y) = M_{X \cdot Y} - M_X \cdot M_Y.$$

Ниже в наших таблицах и фрагментах программ этот вариант коэффициента *будет иметь номер* 0.

Далее сформулируем определение *«немного модифицированного» коэффициента корреляции Кендалла*[3]. Для этого сначала определим количество *несоответствий* (это число можно также назвать «коэффициентом энтропии»): каждое *несоответствие* ((i, j)-несоответствие) возникает в том случае, когда для некоторых $i \neq j$, выполнено следующее:

$$X_i > X_j \quad \text{но} \quad Y_i < Y_j. \tag{1}$$

Будем обозначать общее количество таких несоответствий записью $\mathrm{entr}(X, Y)$, или просто E – если при этом не возникнет неопределённости.

Поскольку максимально возможное количество таких несоответствий равно $\frac{N \cdot (N-1)}{2}$, мы определим *«немного модифицированный» коэффициент корреляции Кендалла* как

$$1 - \frac{4 \cdot E}{N \cdot (N-1)};$$

это значение равно:

[2] Для обычного варианта вычисления коэффициента Кендалла мы в этом случае должны считать его равным 1 – что, конечно, неестественно.

[3] Стандартное определение приводить не будем – в связи со следующим фактом. Коэффициент корреляции, вычисленный для обычного варианта Кендалла во-первых и этой нашей интерпретацией во-вторых (т. е., можнро сказать, «корреляция между корреляциями») – всегда равен 1. Иными словами, если для двух пары последовательностей первая даёт больший коэффициент согласно обычному определению, то и вторая даст больший коэффициент – и наоборот.

- 1 при отсутствии несоответствий,
- и -1 в случае их максимально возможного числа.

Мы видим, что основная идея Кендалла сохраняется, т. е. можно сказать, что мы считаем число обменов, необходимых для выстраивания элементов второй последовательности в том же порядке, в котором идут элементы последовательности первой. Ниже в наших таблицах и фрагментах программ этот вариант коэффициента *будет иметь номер* **2**.

Отметим ещё, что то же самое значение может быть также вычислено следующим образом. Для пары (i, j) коэффициент энтропии мы вычисляем по той же самой формуле (1), а далее считаем сумму всех таких коэффициентов и делим полученный результат на значение $\frac{N \cdot (N-1)}{2}$, которое ранее нами уже было использовано. Такой вариант будет важен впоследствии, при рассмотрении обобщения вариантов корреляции.

Несмотря на частое употребление *на практике* коэффициента корреляция Кендалла [4], в разных публикациях приводятся разные *варианты критики* этого критерия; при этом мы считаем наиболее важным такой недостаток: *при большом количестве совпадений в значениях рассматриваемых случайных величин коэффициент Кендалла даёт не очень адекватные результаты*. Именно поэтому мы будем рассматривать ещё и дальнейшую его модификацию – назовём её *«сильно модифицированным» коэффициентом корреляции Кендалла*. Его наиболее удобно определять для пар пар значений случайных величин X и Y – аналогично тому, как это было сделано выше, при втором варианте вычисления обычного коэффициента Кендалла.

Ниже в наших таблицах и фрагментах программ этот вариант коэффициента *будет иметь номер* **3**. Фрагмент компьютерной программы на Си++ для вариантов 2 и 3 приведён на рис. 1 – при этом автор надеется, что текст программы полностью ясен и, кроме того, он может служить комментарием к приведённым выше определениям.

```
if (nReg==2) return
    (pairOne.GetA()-pairOne.GetB())*(pairTwo.GetA()-pairTwo.GetB()) < 0 ?
        -1 : 1;
    // -1 if the pair is "incorrect" and +1 if it is "correct"
else if (nReg==3) { // a more complicated version of the previous one:
    // we take into account the equality of 0 in one of the pairs
    double rOne = pairOne.GetA()-pairTwo.GetA();
    if (::IsNulla(rOne)) return 0;
    double rTwo = pairOne.GetB()-pairTwo.GetB();
    if (::IsNulla(rTwo)) return 0;
    return (rOne*rTwo) < 0 ? -1 : 1;
}
```

Рис. 1. Часть текста функции для вычисления двух вариантов модифицированного коэффициента корреляции Кендалла

Перейдём к *коэффициенту корреляции Спирмена*. Он вычисляется обычным образом, т. е.

$$\frac{\sum_{i=1}^{n} (x_i - M_X) \cdot (y_i - M_Y)}{\sqrt{n \cdot \sigma_X \cdot \sigma_Y}};$$

эта формула получена путём эквивалентного изменения соответствующих формул из [1], [2]. Ниже в наших таблицах и фрагментах программ этот вариант коэффициента *будет иметь номер* **1**.

[4] При этом сто́ит отметить, что коэффициент корреляция Спирмена (о котором далее), по-видимому, в практических задачах используется чаще. Косвенный аргумент – такой: поисковики по запросу «коэффициент корреляция Спирмена» дают примерно в 2.5 – 3 раза больше результатов, чем по запросу «коэффициент корреляция Кендалла».

Заранее отметим, что в разделе IV будет приведён *наш вариант* вычисления ранговой корреляции; в наших дальнейших таблицах и фрагментах программ этот, предлагаемый нами вариант коэффициента *будет иметь номер* **4**.

III. ПЛЮСЫ И МИНУСЫ ОБЫЧНЫХ ВАРИАНТОВ ВЫЧИСЛЕНИЯ РАНГОВОЙ КОРРЕЛЯЦИИ

Как уже было отмечено во введении и в разделе II, ранговую корреляцию можно считать и самым простым способом, т. е. как обычный коэффициент корреляции. По мнению автора настоящей статьи, даты публикаций определений коэффициентов [5] показывают «движение по степени важности», в сторону убывания, *самих значений* элементов последовательностей, для которых коэффициенты вычисляются: эти значения наиболее важны для общего метода вычисления корреляции, менее важны для коэффициента Спирмена и не важны для коэффициента Кендалла: в последнем случае важен только относительный порядок этих значений. С этой точки зрения, коэффициент Спирмена, по-видимому, чаще всего описывает ситуации наиболее адекватно: например, нужно как-то отражать малую разность между значениями одного ряда – чего не делается в случае коэффициента Кендалла.

Таким образом, можно считать, что среди различных алгоритмов вычисления ранговой корреляции обычный коэффициент корреляции «находится на одном полюсе», а коэффициент Кендалла – «на противоположном». Коэффициент Спирмена «находится между ними», и при этом, по мнению автора настоящей статьи, он «существенно ближе» к обычному коэффициенту корреляции.

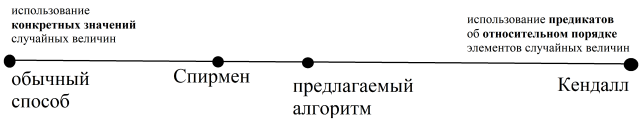

Рис. 2. «Расположение» вариантов вычисления ранговой корреляции

И, немного забегая вперёд, отметим следующее:

- во-первых, предлагаемый нами вариант вычисления коэффициента корреляции также будет находиться «между» – при этом «ближе к полюсу Кендалла» (чем вариант Спирмена);
- во-вторых, коэффициент Спирмена фактически один, он не подлежит никакой модификации – в то время как предлагаемый нами алгоритм вычисления коэффициента корреляции фактически даёт множество различных способов, которые различаются рассмотрением двух заданных пар случайных величин.

Специально отметим, что это пока не «критика общего способа» вычисления ранговой корреляции – см. [1, стр. 346] – а лишь возможный взгляд на описание предлагаемого нами варианта. А такую «критику общего способа» мы рассмотрим в части II настоящей статьи.

По поводу сказанного в этом разделе выше рассмотрим следующий *интересный пример* – и заранее отметим, что он подобран «вручную». На основании его рассмотрения

[5] Математические обоснования общего метода корреляции были даны О. Браве в 1846 году, коэффициент Ч. Спирмена был предложен в 1904 году, а коэффициент М. Кендалла – в 1938 году.

мы далее попытаемся объяснить, почему именно предлагаемый нами вариант подсчёта парной корреляции мы считаем наиболее адекватным – хотя, конечно, подобную степень адекватности каждый исследователь должен определять самостоятельно для себя. Итак, рассмотрим две такие последовательности:

```
0  0  1  1  0  3  6  3  6  3  6  6
0  2  2  3  1  3  5  4  7  5  8  6
```

(пары берутся согласно порядка элементов).

После вычислений мы для этого примера получаем следующие варианты коэффициентов корреляции [6]:

```
corr-0: 0.884
corr-1: 0.940
corr-2: 1
corr-3: 0.758
```

(про номера коэффициентов говорилось выше).

Рассмотрим важные комментарии, связанные с каждым из рассчитанных значений вариантов.

Во-первых заметим, что значения `corr-0` и `corr-1` очень близки; это легко объяснить тем, что коэффициент ранговой корреляции Спирмена определяется близко к обычному («неранговому») коэффициенту корреляции.

Во-вторых объясним, почему значение `corr-2` равно 1; самое простое объяснение заключается в следующем. Определим частичный порядок согласно значениям первых элементов пар, а в случае их равенства – по вторым элементам. В рассматриваемом примере мы можем расположить все пары в таком порядке, то есть при этом получается линейный порядок. Согласно определению коэффициента Кендалла этого достаточно, чтобы итоговое значение было равно 1.

Конечно, в этом примере считать значение корреляции равным 1 неестественно. Это значение улучшается при применении нашей модификации коэффициента Кендалла, в которой, согласно приведенным выше определениям, две пары с одним равным элементом не считаются коррелированными. Полученное при этом значение 0.758, конечно, более естественное, и очень важно, что оно значительно меньше значений, полученных в соответствии с алгоритмами для `corr-0` и `corr-1`.

В-третьих, для некоторых практических задач полезно применение *нормализации*; таковой обычно является линейное отображение значений случайных величин (обычно обеих, изредка одной из них) – независимо от второй случайной величины – в новый отрезок, обычно [0, 1], где 0 соответствует минимальному значению рассматриваемой случайной величины, а 1 – максимальному её значению. При этом, например, для обоих вариантов коэффициента Кендалла подобная нормализация не приводит к изменению посчитанного итогового значения [7], однако в задачах, рассмотренных в [3], [4], [5], подобная нормализация часто важна.

Интересно отметить следующее обстоятельство. Действительно, можно заранее предположить, что результаты вычисления коэффициентов корреляции не изменятся в результате проведения нормализации, и в большинстве случаев это просто следует из определений. Однако менее очевидно, что при специальном варианте «растягивания» исходных данных, а именно замене `3` на `100`, `4` на `101`, `6` на `200`, `7` на `201` и т. д., результаты вычисления коэффициентов корреляции также практически не изменятся. Действительно, рассмотрим такие входные данные (ср. с приведёнными выше):

```
0 0 1   1 0 100 200 100 200 100 200 200
0 2 2 100 1 100 102 110 201 102 202 200
```

При этом результаты расчетов *немного* изменяются, но их общее соотношение остается примерно таким же, как и раньше:

```
corr-0: 0.930
corr-1: 0.942
corr-2: 1
corr-3: 0.758
```

Практически ничего не меняется в случае проведения описанной выше нормализации: результаты *в точности совпадают* с теми, которые относились к первому рассмотрению этого «нулевого» примера; впрочем, это легко объяснимо.

Итак, даже минимальное полученное нами значение коэффициента корреляции (для разных вариантов его вычисления) превышает 0.75. Это, конечно, более естественно, чем 1, но и такое значение для рассмотренных примеров мы считаем слишком большим [8] – *и будем далее его «исправлять»*.

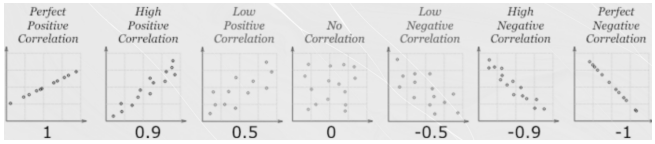

Рис. 3. Варианты значений коэффициента корреляции

IV. ПРЕДЛАГАЕМЫЙ ВАРИАНТ ВЫЧИСЛЕНИЯ РАНГОВОЙ КОРРЕЛЯЦИИ

Итак, в этом разделе мы предлагаем *свой вариант* вычисления ранговой корреляции. В нём мы рассматриваем множество пар пар случайных величин, определяем формулу коэффициента для пары пар таких значений, а итоговое значение корреляции получается усреднением всех полученных значений коэффициенов. И, как тоже было отмечено, мы, как и в случае коэффициента Спирмена, будем находиться «между» обычным вариантом вычисления парной корреляции и коэффициентом Кендалла – но при этом отличие от коэффициента Спирмена существенное.

[6] Здесь и ниже мы, как правило, производим округление до 3 значащих десятичных цифр.

[7] Для «обычной» корреляции и коэффициента Спирмена – как правило, приводят к «небольшим» изменениям. Мы не будем обсуждать, что в данном случае означает «небольшие».

[8] По-видимому, нужно согласиться, что корреляция между последовательностями есть – но она «едва заметна». Поэтому «по классике» (`https://www.codecamp.ru/blog/what-is-a-weak-correlation/` и мн. др.) здесь адекватным было бы значение порядка 0.3 – 0.4: «как правило, коэффициент корреляции между 0.25 и 0.5 считается «слабой» корреляцией между двумя переменными», см. рис. 3.

По этому поводу приведём также текст с одного из известных сайтов: «Всё на самом деле просто. Можно учиться в игровой форме, в этом поможет сайт `https://www.guessthecorrelation.com/`. Это веб-игра, в которой даётся случайное поле рассеивания величин, и нужно определить корреляцию «на глаз». … После 10 уровней можно отличить корреляцию 0.2 от 0.3 с первого взгляда».

В предлагаемом варианте вычисления ранговой корреляции мы, подобно методам (2) и (3), рассматриваем множество пар пар: первая пара – это X_i и X_j (для двух реализаций случайной величины X), а вторая – Y_i и Y_j (с теми же индексами для Y). Аналогично методам (2) и (3), каждое значение корреляции для одной пары пар должно быть в диапазоне от -1 до 1 (с обычным смыслом этих значений), а итоговое значение корреляции будет получено *усреднением* всех полученных значений (в наших обозначениях всего будет $\frac{N\cdot(N-1)}{2}$ пар).

Для каждой из этих пар мы вычисляем значение, показанное далее на рис. 4. На нём значения X_i и X_j приведены слева, а значения Y_i и Y_j – справа. Важно, что $X_i \leqslant X_j$ и $Y_i \leqslant Y_j$ (иначе мы изменяем *порядок* элементов в «нарушенной» паре, при этом также *изменяя знак ответа*), а также $X_j - X_i \leqslant Y_j - Y_i$ (иначе мы изменяем *имена* переменных, *не изменяя* знак ответа).

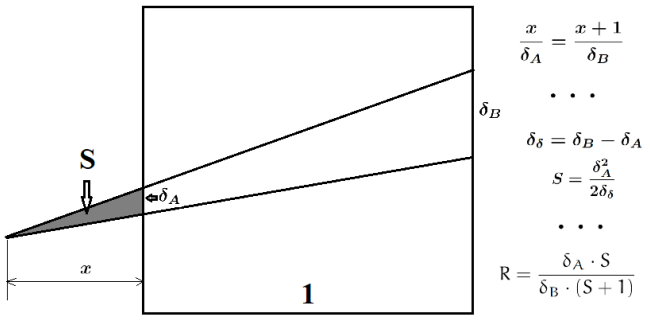

Рис. 4. Предлагаемый вариант вычисления ранговой корреляции

Ответом мы полагаем значение

$$R = \frac{\delta_A \cdot S}{\delta_B \cdot (S+1)}, \quad \text{где} \quad S = \frac{\delta_A^2}{2\delta_\delta} \quad \text{и} \quad \delta_\delta = \delta_B - \delta_A .$$

Соответствующий мини-алгоритм на языке Си++ (для одного из этих вариантов) приведён на рис. 5:

```
bool bOrder = true; // by default, the correct order is in both pairs
double A1 = pairOne.GetA(), B1 = pairOne.GetB(),
       A2 = pairTwo.GetA(), B2 = pairTwo.GetB();
if (A1<A2) { Swap(A1,A2); Swap(B1,B2); bOrder = !bOrder; }
if (B1<B2) { Swap(B1,B2); bOrder = !bOrder; }
// we obtained A1>=A2, B1>=B2,
// and if !bOrder then we make the negative answer
double deltaA = A1-A2, deltaB = B1-B2;
if (deltaA>deltaB) { Swap(A1,B1); Swap(A2,B2); Swap(deltaA,deltaB); }
// we obtained deltaA<=deltaB,
// but we do not change bOrder here!
if (::IsNulla(deltaA)) return (bOrder ? deltaB : -deltaB);
double deltadelta = deltaB-deltaA;
if (::IsNulla(deltadelta)) return 0.0;
double double Return = (deltaA*S)/deltaB*(S+1.0);
return (bOrder ? Return : -Return);
```

Рис. 5. Часть текста функции для вычисления предлагаемого в статье варианта коэффициента корреляции

Стоит отметить, что в предварительных версиях мы проводили тестирования по приведённым выше формулам, а также по двум другим:

- во-первых, по более простой формуле: для того же самого случая $A_1 > A_2$ и $B_1 > B_2$ полагаем $R = \frac{\delta_A}{\delta_B}$;
- во-вторых [9], умножая приведённое выше значение R основного варианта на $\frac{\delta_A+\delta_B}{2}$.

[9] Мы здесь приводим формулу для случая, когда предварительно была проведена нормализация, причём именно по описанному выше мини-алгоритму. Более общий случай (без предварительной нормализации) не намного сложнее.

Однако после ряда вычислительных экспериментов мы пришли к выводу, что формулы, приведённые в основном тексте статьи, более удачны [10].

Далее приведены примеры нашей версии парной корреляции для некоторых конкретных пар значений. Подписи к приведенным выше рисункам показывают, наблюдаем ли мы сильную, среднюю или малую корреляцию – включая рисунки для вырожденных случаев.

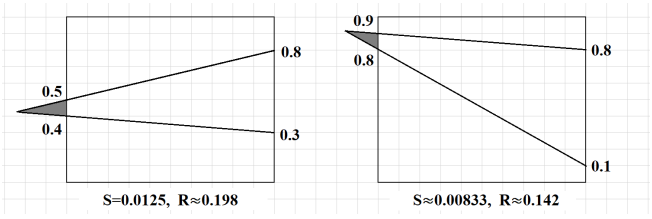

Рис. 6. Пример расчета значения в случае «малой» величины корреляции

Во-первых, рассмотрим рис. 6. Оба приведённых на нём примера соответствуют одному и тому же порядку попарного расположения элементов (как и все дальнейшие рисунки – в противном случае мы меняем знак ответа), но в то же время в одной из последовательностей [11] разница в значениях этих элементов намного меньше, чем в другой. Как и ожидалось, значение корреляции положительное, но очень малое.

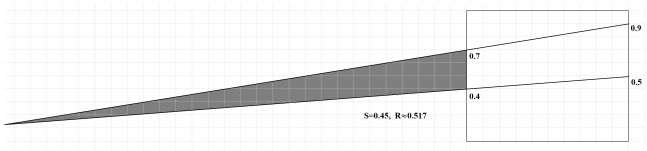

Рис. 7. Пример расчета значения в случае «большой» величины корреляции

Во-вторых, рассмотрим рис. 7. Он соответствует случаю, когда разница между одинаковыми значениями намного больше. Как и ожидалось, значение корреляции больше, чем 0.5.

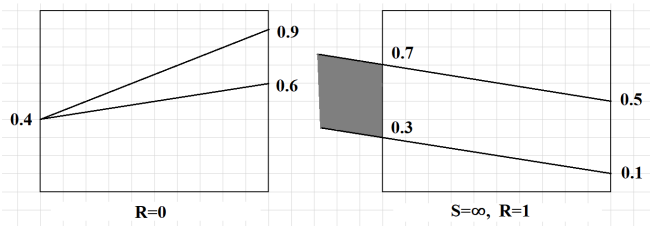

Рис. 8. Примеры для вырожденных случаев

В-третьих, рассмотрим два вырожденных случая, см. рис. 8. Здесь стóит отметить, что для варианта, приведённого на правом рисунке, значение R может существенно различаться для трёх приведённых выше значений – но подробности мы в настоящей статье рассматривать не будем.

[10] Подробности могут представлять некоторый интерес, и, возможно, мы обсудим их в одной из следующих публикаций. Пока же мы будем использовать один из этих вариантов – причём *не будем конкретизировать, какой именно*, это не очень принципиально.

[11] Не «в одной из пар», это разные вещи.

В заключение раздела повторим, что общую ранговую корреляцию будет считать усреднённой суммой всех возможных пар пар значений. Поэтому следующий раздел мы начнём с конкретных результатов вычислений для «примера номер 0»; сразу отметим, что эти результаты можно проверить «вручную», то есть без вычислений на компьютере.

V. О НЕКОТОРЫХ СПЕЦИАЛЬНО ПОДОБРАННЫХ ПРИМЕРАХ

Продолжим рассмотрение «нулевого» примера. Его первая часть (до замены `3` на `100` и т. д.) даёт

```
corr-4: 0.375
```

а вторая часть (после такой замены) –

```
corr-4: 0.298
```

Очень важно следующее: *мы считаем эти значения совершенно адекватными!*

Следующий «сделанный вручную» пример также был получен с помощью некоторой модификации первого варианта «нулевого примера» – умножением исходных значений обеих последовательностей этого варианта на 1000 и добавлением после этого умножения значений от 1 до 12 для первой последовательности и наоборот, от 12 до 1 с шагом −1 для второй последовательности. Такие небольшие добавки к существенным изменениям коэффициентов не привели, вот полученные значения после проведения нормализации:

```
corr-0: 0.939
corr-1: 0.930
corr-2: 0.788
corr-3: 0.788
corr-4: 0.342
```

(точное равенство 1 для более простого варианта коэффициента Кендалла по понятным причинам теперь не достигается).

Обобщая можно сказать, что для всех исходных модификаций этого «нулевого примера» выполняются следующие *ожидаемые* факты:

- значения `corr-0` и `corr-1` очень близки;
- все значения кроме последнего, в частности `corr-2` и `corr-3`, очень велики и вряд ли могут быть названы адекватными;
- а значение `corr-4` таковым *всегда* является.

Последний пример, приведённый далее, также сделан «вручную», в нём способ формирования обеих последовательностей понятен на основе их рассмотрения:

```
1001  1006
1002  1005
1003  1004
1004  1003
1005  1002
1006  1001
2001  2006
2002  2005
2003  2004
2004  2003
2005  2002
2006  2001
```

(для наглядности мы здесь располагаем две последовательности не по строкам, как ранее, а по столбцам).

Этот пример, по-видимому, наиболее интересен! Он показывает не только и не столько неадекватность классических способов вычисления ранговой корреляции, сколько тот факт, что такие неадекватные значения могут располагаться *с разных сторон* от ожидаемого (естественного, адекватного) значения, вычисляемого с помощью алгоритма `corr-4`. После проведения нормализации и вычислений получены следующие значения:

```
corr-0: 1.000
corr-1: 0.510
corr-2: 0.091
corr-3: 0.091
corr-4: 0.536
```

Специально отметим, что значение `corr-0` равно 1 только при огруглении: разница наблюдается в 5-й десятичной цифре после запятой. Но существенно более важным мы считаем очень большую разницу значений `corr-0` и `corr-1` (не появлявшуюся выше) – и, по нашему мнению, *именно для исследования подобных примеров и была разработана Спирменом его теория*. Также несложно «почти без вычислений» объяснить значение (примерно 0.1), получаемое обоими версиями алгоритма Кендалла: мы должны выполнить все возможные перестановки «внутри» первых 6 элементов (их всего $\frac{6\cdot5}{2} = 15$ штук), плюс «внутри» вторых 6 элементов (их тоже 15 штук); общее же максимальное число возможных перестановок для 12 элементов равно $\frac{12\cdot11}{2} = 66$ – что превышает число посчитанных нами необходимых перестановок немногим более, чем в 2 раза (превышение ровно в 2 раза дало бы нулевую корреляцию).

И, как и ранее, «наше» значение `corr-4`, а также близкое к нему значение `corr-1`, мы считаем *наиболее адекватными*.

VI. КРАТКО О НЕКОТОРЫХ ПРИМЕРАХ ПРАКТИЧЕСКОГО ПРИМЕНЕНИЯ

Как уже было сказано во введении, подробно примеры применения предлагаемого коэффициента ранговой корреляции и полученные при этом результаты – в сравнении с результатами для других вариантов коэффициентов – будут приведены в части II настоящей статьи. В этом разделе мы рассмотрим только краткое описание многочисленных (и очень разных) предметных областей. Эти области мы старались упорядочить «по возрастанию их сложности» – причём под сложностью мы имеем в виду не столько сами эти области, сколько вспомогательные алгоритмы, необходимые для конкретного применения разных вариантов вычисления ранговой корреляции.

Первая область. В работах [6], [7] нами рассматривались разные варианты вычисления расстояний между геномами и специальные оценки (т. н. badness) этих вариантов для рассматриваемых матриц таких расстояний [12]. В простом варианте этой задачи мы получали пары последовательностей таких вариантов badness для разных

[12] Сами значения badness вычисляются с помощью той же ранговой корреляции (об этом будет подробнее сказано ниже, в пятой предметной области) – поэтому можно сказать, что здесь мы выполняем вычисления «ранговой корреляции для ранговой корреляции».

матриц – и считали корреляцию между этими последовательностями. Понятно, что в оптимальном случае коэффициент корреляции здесь должен быть равен 1 – то есть можно считать, что вспомогательные алгоритмы этой задачи направлены на увеличение получаемых коэффициентов корреляции. Несмотря на большой объём выполненных вычислительных экспериментов, мы считаем эту предметную область самой простой, поскольку доступных в Интернете алгоритмов вычисления расстояний между геномами немного, в упромянутых статьях их всего 4 [13] – соответственно, длины последовательностей получаются равными лишь $\frac{4\cdot 3}{2} = 6$.

Вторая область. Сами несколько вариантов вышеописанных алгоритмов для вычисления значений badness были описаны в нашей более ранней статье [8]. Эти значения badness вычислялись на основе рассмотрения всех треугольников матрицы расстояний – и всего мы рассматривали 6 вариантов алгоритмов для badness. В настоящей работе при генерации многих случайных треугольников (1000 штук) мы по разным алгоритмам ранговой корреляции сравнивали порядки треугольников в соответствующих последовательностях.

Третья область [14]. При изучении графиков показателей для распространения фемтосекундного лазерного излучения в среде, легированной золотыми наностержнями, а также в линейной среде, возникло следующее предположение. Можно измерить степень нелинейности, вычисляя распределения т. н. чирп-сигналов в линейной и нелинейной средах. Проверку сделанного предположения мы осуществляем с помощью алгоритмов ранговой корреляции.

Четвёртая область. Результаты чемпионатов СССР по футболу в 1969–1989 годах (можно условно считать, что в каждом из них участвовало по 16 команд). Мы считаем коэффициенты корреляции между расположением команд для двух последовательных годов (две заменяемые, «вылетевшие» команды условно отождествляем с двумя заменёнными) – как и в других примерах, для разных алгоритмов подсчёта этих коэффициентов. Посчитанные 20 значений мы для каждого алгоритма корреляции рассматриваем как варианты реализации случайной величины – и стандартные статистические характеристики полученных случайных величин представляют определённый интерес.

Пятая область. Для населения разных стран – корреляция между значениями IQ и некоторыми антронометрическими характеристиками; все необходимые данные взяты нами из Интернета [15]. В этой области интересно существенное изменение результатов при рассмотрении специально выбранного подмножества стран (конкретно – при исключении из рассмотрения стран европейских).

Шестая область. В отличие от области второй, где исследовались случайно сгенерированные треугольники (для нескольких вариантов badness), здесь для одного варианта badness (который мы в настоящее время считаем наиболее адекватным) мы исследуем конкретные треугольники для конкретных геномов и конкретных алгоритмов подсчёта расстояний между геномами (а именно, алгоритмами Нидлмана – Вунша и Джаро – Винклера). В части I статьи [9] приведены входные данные (по ним, среди прочего, можно определить сами треугольники и соответствующие им значения badness); результаты вычислений будут приведены в представленной к публикации части II той статьи.

Как уже было отмечено, анонсированные здесь примеры применения предложенного коэффициента ранговой корреляции и полученные результаты – в сравнении с результатами для других вариантов коэффициентов – будут приведены в части II настоящей статьи; мы в ней будем использовать приведённые здесь «номера» предметных областей. Также в части II мы рассмотрим предлагаемое *обобщение* нашего варианта вычисления ранговой корреляции.

БЛАГОДАРНОСТИ

Настоящая работа была частично поддержана грантом научной программы китайских университетов “Higher Education Stability Support Program” (раздел “Shenzhen 2022 – Science, Technology and Innovation Commission of Shenzhen Municipality”) – 深圳市 2022 年高等院校稳定支持计划资助项目.

Список литературы

[1] Лагутин М.Б. Наглядная математическая статистика. – М.: БИНОМ. Лаборатория знаний. 2012. 472 с.

[2] Wasserman L. All of statistics: a concise course in statistical inference. – Springer Science & Business Media. 2013. 442 p.

[3] Melnikov B.: Heuristics in programming of nondeterministic games. Programming and Computer Software, 2001, vol. 27, no. 5, pp. 277–288, DOI: 10.1023/ A:1012345111076.

[4] Melnikov B., Radionov A., Gumayunov V.: Some special heuristics for discrete optimization problems. ICEIS 2006 – 8th International Conference on Enterprise Information Systems, Proceedings, 2006, AIDSS, pp. 360–364, ISBN: 9728865422, 978-972886542-9.

[5] Melnikov B., Radionov A., Moseev A., Melnikova E.: Some specific heuristics for situation clustering problems. ICSOFT 2006 – 1st International Conference on Software and Data Technologies, Proceedings, 2006, no. 2, pp. 272–279, ISBN: 9728865694, 978-972886569-6.

[6] Melnikov B., Trenina M., Kochergin A.: On one problem of reconstructing matrix distances between chains of DNA. IFAC-Papers, 2018, vol. 51 (32), pp. 378–383, DOI: 10.1016/j.ifacol.2018.11.413.

[7] Melnikov B., Trenina M., Melnikova E.: Different approaches to solving the problem of reconstructing the distance matrix between DNA chains. Communications in Computer and Information Science (CCIS), 2020, vol. 1201, pp. 211–223, DOI: 10.1007/978-3-030-46895-8_17.

[8] Melnikov B., Pivneva S., Trifonov M.: Various algorithms, calculating distances of DNA sequences, and some computational recommendations for use such algorithms, 2017, Information Technology and Nanotechnology, Proceedings, DOI:10.18287/1613-0073-2017-1902-43-50.

[9] Ли Цзямянь, Му Цзинъюань, Мельников Б.: Об одном подходе к реализации алгоритмов Нидлмана – Вунша и Джаро – Винклера и их применении в корреляционном анализе сходства митохондриальных ДНК обезьян. Часть I. Общее описание работы, 2024, International Journal of Open Information Technologies, vol. 12, no. 9, pp. 1–10. ISSN: 2307-8162.

Борис Феликсович МЕЛЬНИКОВ,
профессор Университета МГУ – ППИ в Шэньчжэне
(http://szmsubit.ru/),
email$_1$: bormel@smbu.edu.cn,
email$_2$: bormel@mail.ru,
mathnet.ru: personid=27967,
elibrary.ru: authorid=15715,
scopus.com: authorId=55954040300,
ORCID: orcidID=0000-0002-6765-6800.

[13] В других статьях мы пользовались ещё 2 другими алгоритмами, но по некоторым причинапм их не удалось применить к этой задаче.

[14] Соответствующая статья автора с соавтором T. Lysak принята на конференцию 13th International Conference on Mathematical Modeling in Physical Sciences, Kalamata, Greece, September 30 – October 3, 2024.

[15] Сайт https://worldpopulationreview.com/.

About one approach to calculating the rank correlation. Part I

Boris Melnikov

***Abstract*—When calculating the rank correlation, the same situation arises as in some other subject areas: researchers have the opinion that new possible constructions in this subject area are either impossible (everything has already been done), or are not needed. In the article, we try to show that further theoretical developments are possible, which can lead to quite interesting practical results. We propose our own version of calculating the rank correlation coefficient, which (in its simplest form) can be considered "located between" the Kendall and Spearman coefficients. More specifically, we use not only the conditions for the coincidence of the results of predicates that determine the order of the elements of the corresponding pairs (as in calculating the Kendall coefficient), and not only a generalization of the general version of calculating the pair correlation in the case of rank (as in calculating the Spearman coefficient), but both of these techniques together. We use the obtained formulas to calculate different variants of rank correlation coefficients in several different subject areas, and try to justify by reasoning that the criterion we propose is quite successful. We also propose options for generalizing our criterion, and we believe that these possible generalizations do not coincide with the generalizations given in classical monographs (but they do not contradict them, but complement them).**

***Keywords*—correlation coefficient, Spearman's rank correlation, Kendall's rank correlation, a new version of the rank correlation coefficient.**

References

[1] Lagutin M. Visual mathematical statistics. – Moscow: BINOM. Laboratoriya znaniy. 2012. 472 p. (In Russian.)

[2] Wasserman L. All of statistics: a concise course in statistical inference. – Springer Science & Business Media. 2013. 442 p.

[3] Melnikov B.: Heuristics in programming of nondeterministic games. Programming and Computer Software, 2001, vol. 27, no. 5, pp. 277–288, DOI: 10.1023/ A:1012345111076.

[4] Melnikov B., Radionov A., Gumayunov V.: Some special heuristics for discrete optimization problems. ICEIS 2006 – 8th International Conference on Enterprise Information Systems, Proceedings, 2006, AIDSS, pp. 360–364, ISBN: 9728865422, 978-972886542-9.

[5] Melnikov B., Radionov A., Moseev A., Melnikova E.: Some specific heuristics for situation clustering problems. ICSOFT 2006 – 1st International Conference on Software and Data Technologies, Proceedings, 2006, no. 2, pp. 272–279, ISBN: 9728865694, 978-972886569-6.

[6] Melnikov B., Trenina M., Kochergin A.: On one problem of reconstructing matrix distances between chains of DNA. IFAC-Papers, 2018, vol. 51 (32), pp. 378–383, DOI: 10.1016/j.ifacol.2018.11.413.

[7] Melnikov B., Trenina M., Melnikova E.: Different approaches to solving the problem of reconstructing the distance matrix between DNA chains. Communications in Computer and Information Science (CCIS), 2020, vol. 1201, pp. 211–223, DOI: 10.1007/978-3-030-46895-8_17.

[8] Melnikov B., Pivneva S., Trifonov M.: Various algorithms, calculating distances of DNA sequences, and some computational recommendations for use such algorithms, 2017, Information Technology and Nanotechnology, Proceedings, DOI:10.18287/1613-0073-2017-1902-43-50.

[9] Li Jiamian, Mu Jingyuan, Melnikov B.: On an approach to the implementation of the Needleman – Wunsch and Jaro – Winkler algorithms and their application in the correlation analysis of the similarity of mitochondrial DNA of monkeys. Part I, 2024, International Journal of Open Information Technologies, vol. 12, no. 9, pp. 1–10. ISSN: 2307-8162. (In Russian.)

Boris MELNIKOV,
Professor of Shenzhen MSU – BIT University, China
(http://szmsubit.ru/),
email$_1$: bormel@smbu.edu.cn,
email$_2$: bormel@mail.ru,
mathnet.ru: personid=27967,
elibrary.ru: authorid=15715,
scopus.com: authorId=55954040300,
ORCID: orcidID=0000-0002-6765-6800.