О. В. САРМАНОВ

О ЛОЖНОЙ КОРРЕЛЯЦИИ МЕЖДУ СЛУЧАЙНЫМИ ВЕЛИЧИНАМИ

Введение

Так называемая ложная корреляция возникает при следующих обстоятельствах: подлежит изучению корреляционная зависимость между случайными величинами $x_1, x_2, \ldots, x_n$, однако условия наблюдений не позволяют непосредственно найти значения этих величин, наблюдению доступны некоторые другие величины: $\xi_1, \xi_2, \ldots, \xi_n$.

Величиной ξ_i, зависящей от x_i, $i=1, 2, \ldots, n$, заменяют величину x_i, в частности вместо изучения пары величин x_i, x_j изучают пару ξ_i, ξ_j и по этому изучению пытаются сделать заключение об определенных свойствах пары x_i, x_j.

Но величины ξ_i обычно зависят не только от x_i, но либо от других величин x_j, $j \neq i$, либо от некоторых новых величин $l, m, \ldots$, не зависящих от x_i и x_j.

В наиболее общем случае, представляющем практический интерес,

$$\xi_i = f(x_i;\ x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, l, m, \ldots) \qquad (1)$$

$$i = 1, 2, \ldots, n.$$

Наличие в выражениях (1) для ξ_i общих аргументов, отличных от x_i, создает добавочную зависимость между ξ_i, не обусловленную зависимостью между x_i; эту добавочную зависимость мы и называем ложной корреляцией.

Приведем наиболее важные примеры возникновения ложной корреляции. Пусть по результатам анализов горных пород, сложных химических соединений, почв или биологических объектов мы хотим изучить зависимость между компонентами $x_1, x_2, \ldots, x_n$; мы их будем считать неотрицательными случайными величинами. Если проба (или образец), взятая для анализа, обладает свойством репрезентативности, а это всегда будет предполагаться, то результаты анализов дают величины $\xi_1, \xi_2, \ldots, \xi_n$, пропорциональные неизвестным значениям

$x_1, x_2, \ldots, x_n$, с одинаковым коэффициентом пропорциональности l:

$$\xi_i = lx_i, \quad i = 1, 2, \ldots, n; \tag{2}$$

l — новая положительная случайная величина, например, отношение веса пробы, взятой для анализа, к весу подлежащей изучению части земной коры или рудного тела. Значения l обычно не зависят от значений x_i, поэтому в дальнейшем мы всегда будем считать, что l не зависит от величин $x_1, x_2, \ldots, x_n$.

Наличие в формуле (2) общего неизвестного множителя l является источником ложной корреляции между ξ_i и ξ_j, не обусловленной зависимостью между x_i и x_j. От множителя l избавляются переходом к процентным величинам:

$$z_i = \frac{100\, x_i}{x_1 + x_2 + \ldots + x_n}, \quad i = 1, 2, \ldots, n. \tag{3}$$

С точки зрения теории корреляции, величины (3) гораздо сложнее, чем величины (2), зависят от x_i, ибо множитель $\frac{100}{x_1 + x_2 + \ldots + x_n}$, появившийся вместо множителя l, теперь зависит от всех изучаемых величин и, как мы покажем, вносит гораздо большие искажения в корреляционную зависимость между z_i, по сравнению с искажением, вызванным множителем l.

Кроме формул (2) и (3), являющихся простейшими частными случаями (1), мы рассмотрим еще искажение зависимости, вызываемое общим слагаемым m:

$$\eta_i = y_i + m, \quad i = 1, 2, \ldots, n, \tag{4}$$

где m не зависит от величины $y_1, y_2, \ldots, y_n$. Логарифмирование положительных ξ_i в формуле (2) приводит к формуле вида (4).

В этой статье рассматривается лишь простейшая мера линейной зависимости, а именно обычный коэффициент корреляции, причем рассматриваются следующие вопросы. Как отличается коэффициент корреляции между величинами ξ_i и ξ_j или z_i и z_j от интересующего исследователя коэффициента корреляции между x_i и x_j? Можно ли, оперируя с величинами вида (2)—(4), восстановить коэффициент корреляции между x_i и x_j? Если последняя задача находит решение, то мы будем говорить, что ложная корреляция исключается.

§ 1. О корреляции между процентными величинами

То обстоятельство, что процентные расчеты, т. е. переход от величин вида (2) к величинам вида (3), вносит искажение корреляционной зависимости, неоднократно отмечалось в литературе [1], [2], [3]. Искажение особенно очевидно, если $n = 2$. Процентные содержания двух единственных компонентов всегда связаны функциональной

зависимостью вида $z_2 = 100 - z_1$, и коэффициент корреляции между z_1 и z_2 всегда равен -1, каковы бы ни были x_1 и x_2; последние могут быть независимыми или иметь положительный коффициент корреляции.

Как показывает нижеследующая теорема, и при сколь угодно большом числе компонентов x_i, вполне возможны случаи, когда коэффициент корреляции между z_i и z_j не дает никакой информации о коэффициенте корреляции между x_i и x_j.

Теорема 1 (о ложной корреляции). *Если случайные величины* $x_1, x_2, \ldots, x_n$ *положительны, одинаково распределены и находятся в симметричной корреляционной зависимости* (в симметричной множественной корреляции), *то процентные величины* $\frac{100\,x_i}{x_1 + x_2 + \ldots + x_n}$ *тоже находятся в симметричной корреляции с постоянным коэффициентом корреляции* $R(z_i, z_j) = -\frac{1}{n-1}$, $i \neq j$, *который не зависит от законов распределения* x_i *и от коэффициента корреляции* $R(x_i, x_j)$. *В частности,* x_i *могут быть независимыми одинаково распределенными величинами.*

Доказательство. Отбросив множитель 100, не влияющий на коэффициент корреляции, рассмотрим величины

$$z_i = \frac{x_i}{x_1 + x_2 + \ldots + x_n}, \quad i = 1, 2, \ldots, n. \tag{5}$$

В силу симметрии, все z_i имеют одинаковые законы распределения, в частности все их моменты не зависят от i, по этой же причине, коэффициент корреляции $R(z_i, z_j)$ не зависит от i и j.

Так как

$$\sum_{i=1}^{n} z_i = 1, \tag{6}$$

то $M \sum_{i=1}^{n} z_i = nMz_i = 1$, откуда

$$Mz_i = \frac{1}{n}, \quad i = 1, 2, \ldots, n. \tag{7}$$

С другой стороны, в силу (6), дисперсия $\sum_{i=1}^{n} z_i$ равна нулю:

$$0 = D \sum_{i=1}^{n} z_i = \sum_{i=1}^{n} Dz_i + 2 \sum_{i<j} \operatorname{cov} z_i z_j = nDz_i + n(n-1) \operatorname{cov} z_i z_j,$$

откуда (при $Dz_i \neq 0$)

$$R(z_i, z_j) = \frac{\operatorname{cov} z_i z_j}{Dz_i} = -\frac{1}{n-1}, \tag{8}$$

что и требовалось доказать.

Так как ежегодно появляется множество работ, в которых постоянно оперируют с процентными величинами, и, установив между

ними корреляционную зависимость, пытаются без всяких изменений перенести ее на исходные величины x_i, то доказанная теорема имеет большое практическое значение. Теорема 1 показывает, что в ряде случаев, когда компоненты равноправны или примерно равноправны, по найденному коэффициенту корреляции между z_i и z_j принципиально невозможно судить о коэффициенте корреляции между x_i и x_j.

Непосредственным обобщением доказанной теоремы является

Теорема 2. *Пусть* $x_1, x_2, \ldots, x_n$ *положительны и находятся в симметричной корреляционной зависимости, обозначим через* A *сумму* $k \geqslant 1$ *величин вида* (5), *а через* B — *сумму каких-нибудь других* $l \geqslant 1$ *величин того же вида,* $k+l \leqslant n$, *тогда коэффициент корреляции между этими суммами не зависит от законов распределения* x_i, *от коэффициента корреляции между ними и равен*

$$R(A, B) = -\sqrt{\frac{kl}{(n-k)(n-l)}}. \tag{9}$$

Доказательство. Обозначим через C сумму слагаемых вида (5), которые не входят в A и B (их может и не быть), тогда $A+B+C=1$ и, следовательно, $D(A+B+C)=0$ или

$$DA+DB+DC+2[\operatorname{cov} AB+\operatorname{cov} AC+\operatorname{cov} BC]=0; \tag{10}$$

разделив (10) на $\sqrt{DA \cdot DB}$, найдем $R(A, B)$:

$$R(A,B)=-\frac{1}{2}\sqrt{\frac{DA}{DB}}-\frac{1}{2}\sqrt{\frac{DB}{DA}}-\frac{1}{2}\frac{DC}{\sqrt{DADB}}-\frac{\operatorname{cov} AC+\operatorname{cov} BC}{\sqrt{DADB}}; \tag{11}$$

в силу равенства дисперсий и ковариаций величин z_i, получим

$$\frac{DA}{DB}=\frac{kDz_i+k(k-1)\operatorname{cov} z_iz_j}{lDz_i+l(l-1)\operatorname{cov} z_iz_j}=\frac{k+k(k-1)R(z_i, z_j)}{l+l(l-1)R(z_i, z_j)}=$$

$$=\frac{k+k(k-1)\left(-\frac{1}{n-1}\right)}{l+l(l-1)\left(-\frac{1}{n-1}\right)}=\frac{k(n-k)}{l(n-l)}; \tag{12}$$

аналогично, пользуясь равенством (8), найдем

$$\sqrt{\frac{DA}{DB}}+\sqrt{\frac{DB}{DA}}+\frac{DC}{\sqrt{DADB}}=\frac{(l+k)(n-k-l)+k(n-k)+l(n-l)}{\sqrt{kl(n-k)(n-l)}}, \tag{13}$$

$$\left.\begin{aligned}\frac{\operatorname{cov} AC}{\sqrt{DADB}}&=-\frac{k(n-k-l)}{\sqrt{kl(n-k)(n-l)}}\\ \frac{\operatorname{cov} BC}{\sqrt{DADB}}&=-\frac{l(n-k-l)}{\sqrt{kl(n-k)(n-l)}}\end{aligned}\right\} \tag{14}$$

Подставляя (13) и (14) в (11), получим для $R(A, B)$ выражение (9).

Равенства (7) и (8) показывают, что первые моменты и коэффициенты корреляции симметрично связанных величин z_i зависят только от числа этих величин, а не от законов распределения исходных

величин x_i. Вторые моменты Mz_i^2 и смешанные моменты Mz_iz_j уже зависят от закона совместного распределения величин x_i и могут меняться в пределах, устанавливаемых следующей теоремой.

Теорема 3. *Вторые моменты величин* (5), *построенных с помощью положительных, симметрично связанных величин* x_i, $i=1, 2, \ldots, n$, *удовлетворяют неравенствам*

$$\frac{1}{n^2} \leqslant Mz_i^2 \leqslant \frac{1}{n}, \quad i=1, 2, \ldots, n, \tag{15}$$

$$0 \leqslant Mz_iz_j \leqslant \frac{1}{n^2}, \quad i \neq j. \tag{16}$$

Доказательство. Так как величины z_i неотрицательны, а $Mz_i = \frac{1}{n}$, то левые части неравенств (15) и (16) очевидны. При выводе (8) мы пользовались соотношением $nDz_i + n(n-1)\operatorname{cov} z_iz_j = 0$, которое после сокращения на n и использования равенства (7) приводится к виду

$$Mz_i^2 + (n-1)\,Mz_iz_j = \frac{1}{n}, \quad i \neq j. \tag{17}$$

Заменяя в равенстве (17) смешанный момент Mz_iz_j его нижней границей, т. е. нулем, получим правую часть неравенства (15), а заменяя Mz_i^2 его нижней границей $\frac{1}{n^2}$, получим правую часть (16), что завершает доказательство всех утверждений теоремы.

Замечание. Зависимость Mz_i^2 от закона совместного распределения x_i можно показать на примерах. Рассмотрим две симметрично связанные положительные величины x_i и x_j; пусть каждая из них принимает лишь k значений $a_1, a_2, \ldots, a_k$, и пусть $p_{ij} = P\{x_1 = a_i, x_2 = a_j\}$. Тогда

$$Mz_1^2 = \sum_{i=1}^{k} \frac{a_i^2}{(a_i + a_i)^2}\, p_{ii} + \sum_{i<j} \left[\frac{a_i^2}{(a_i + a_j)^2} + \frac{a_j^2}{(a_j + a_i)^2} \right] p_{ij} =$$

$$= \frac{1}{4} \sum_{i=1}^{k} p_{ii} + \frac{1}{2} \sum_{i<j} p_{ij}\, \frac{2a_i^2 + 2a_j^2}{(a_i + a_j)^2} =$$

$$= \frac{1}{4} \sum_{i=1}^{k} p_{ii} + \frac{1}{2} \sum_{i<j} p_{ij}\, \frac{(a_i + a_j)^2 + (a_i - a_j)^2}{(a_i + a_j)^2} =$$

$$= \frac{1}{4} + \frac{1}{2} \sum_{i<j} p_{ij} \left(\frac{a_i - a_j}{a_i + a_j} \right)^2 \geqslant \frac{1}{4}. \tag{18}$$

Так как $\left(\frac{a_i - a_j}{a_i + a_j}\right)^2 \leqslant 1$, то

$$Mz_i^2 \leqslant \frac{1}{4} + \frac{1}{2} \sum_{i<j} p_{ij} = \frac{1}{4} + \frac{1}{4} \left(1 - \sum_{i=1}^{k} p_{ii} \right) \leqslant \frac{1}{2}.$$

Оценка сверху $Mz_i^2 \leqslant \frac{1}{2}$ $(n=2)$ не может быть понижена, так как можно положить $p_{ii}=0$, $i=1, 2, \ldots, k$, с другой стороны, взяв последовательность $a_1 < a_2 < \ldots < a_k$ такой, что a_i неограниченно растут, причем $\frac{a_{i-1}}{a_i} \to 0$, $i=2, 3, \ldots, k$, мы все дроби $\left(\frac{a_j - a_i}{a_i + a_j}\right)^2$ сделаем сколь угодно близкими к единице.

Сделаем еще несколько замечаний о средних значениях дисперсий и ковариаций величин (5), в общем несимметричном случае.

Пусть теперь $x_1, x_2, \ldots, x_n$ — произвольно связанные положительные величины, тогда как и раньше можно определить величины z_i формулой (5), причем $\sum_{i=1}^{n} z_i = 1$, поэтому $D\sum_{i=1}^{n} z_i = 0$,

$$0 = \sum_{i=1}^{n} Dz_i + \sum_{i \neq j} \operatorname{cov} z_i z_j = n\overline{Dz_i} + n(n-1)\,\overline{\operatorname{cov} z_i z_j},$$

откуда

$$\frac{\overline{\operatorname{cov} z_i z_j}}{\overline{Dz_i}} = -\frac{1}{n-1}, \tag{19}$$

где $\overline{Dz_i}$ есть среднее арифметическое из n дисперсий, а $\overline{\operatorname{cov} z_i z_j}$ — среднее арифметическое из $n(n-1)$ ковариаций z_i и z_j при $i \neq j$.

Если мы предположим, что все дисперсии z_i одинаковы,

$$Dz_1 = Dz_2 = \ldots = Dz_n, \tag{20}$$

то

$$\overline{R(z_i, z_j)} = -\frac{1}{n-1}, \tag{21}$$

где $\overline{R(z_i, z_j)}$ есть средний коэффициент корреляции между величинами z_i и z_j, $i \neq j$.

Выражение (21), верное при условии (20), указывается Чейзом в его рукописи [4].

Если же, кроме того, равны все ковариации z_i, z_j, мы приходим к равенству (8). Теорема 1 выясняет достаточные условия, накладываемые не на z_i, а на исходные величины x_i, при которых равенство (8) будет иметь место.

§ 2. Исключение ложной корреляции, вызванной процентными расчетами

В предыдущем параграфе было показано, что в симметричном случае, когда все компоненты x_i равноправны, по коэффициенту корреляции между процентными величинами z_i вообще нельзя судить о коэффициенте корреляции между исходными величинами. Если же имеются добавочные сведения об особенностях отдельных компонен-

тов x_i, то в ряде случаев ложная корреляция может быть полностью устранена.

1. Пусть среди величин x_i имеется одна постоянная величина, например, $x_1 = \text{const}$. Тогда вместо процентных величин z_i и z_j, $i > 1$, $j > 1$, $i \neq j$, рассмотрим отношения $\frac{z_i}{z_1}$ и $\frac{z_j}{z_1}$; легко показать, что коэффициент корреляции между ними равен искомому коэффициенту корреляции между x_i и x_j:

$$R\left(\frac{z_i}{z_1}, \frac{z_j}{z_1}\right) = R\left(\frac{x_i}{x_1}, \frac{x_j}{x_1}\right) = R(x_i, x_j), \tag{22}$$

так как, по условию,

$$x_1 = \text{const}. \tag{23}$$

В этом случае ложная корреляция, вызванная процентными расчетами, полностью исключается.

2. Пусть среди величин x_i имеется два компонента x_1 и x_2, независимые между собой, и пусть от них не зависят по крайней мере два компонента x_i и x_j; i; $j \neq 1$; 2; $i \neq j$. В этом случае тоже возможно восстановить истинное значение $R(x_i, x_j)$, оперируя исключительно с одними процентными величинами.

Для любой положительной случайной величины ξ введем обозначение

$$\Delta\xi = \frac{M\xi^2}{(M\xi)^2} \geqslant 1, \tag{24}$$

а для пары положительных величин ξ и η — обозначение

$$\Delta(\xi\eta) = \Delta\xi\eta = \frac{M(\xi\eta)}{M\xi M\eta}. \tag{25}$$

Тогда коэффициент корреляции между ними выражается следующим образом:

$$R(\xi, \eta) = \frac{\Delta\xi\eta - 1}{\sqrt{(\Delta\xi - 1)(\Delta\eta - 1)}}. \tag{26}$$

Возвращаясь к рассматриваемому случаю, найдем четыре величины

$$\Delta\frac{z_i}{z_1}, \quad \Delta\frac{z_j}{z_2}, \quad \Delta\frac{z_i}{z_2}, \quad \Delta\frac{z_j}{z_1}$$

и три коэффициента корреляции

$$R_0 = R\left(\frac{z_i}{z_1}, \frac{z_j}{z_2}\right); \quad R_1 = R\left(\frac{z_i}{z_1}, \frac{z_j}{z_1}\right); \quad R = R\left(\frac{z_i}{z_2}, \frac{z_j}{z_2}\right);$$

с их помощью вычисляются две вспомогательные величины:

$$u_1=\frac{1+R_1\sqrt{\left(\Delta\frac{z_i}{z_1}-1\right)\left(\Delta\frac{z_j}{z_1}-1\right)}}{1+R_0\sqrt{\left(\Delta\frac{z_i}{z_1}-1\right)\left(\Delta\frac{z_j}{z_2}-1\right)}};$$

$$u_2=\frac{1+R_2\sqrt{\left(\Delta\frac{z_i}{z_2}-1\right)\left(\Delta\frac{z_j}{z_2}-1\right)}}{1+R_0\sqrt{\left(\Delta\frac{z_i}{z_1}-1\right)\left(\Delta\frac{z_j}{z_2}-1\right)}}, \tag{27}$$

через которые, как показано в работе [3], $R(x_i, x_j)$ выражается следующим образом:

$$R(x_i, x_j)=\frac{R_0\sqrt{\left(\Delta\frac{z_i}{z_1}-1\right)\left(\Delta\frac{z_j}{z_2}-1\right)}}{\sqrt{\left(u_1\Delta\frac{z_i}{z_1}-1\right)\left(u_2\Delta\frac{z_j}{z_2}-1\right)}}. \tag{28}$$

Заметим, что если $x_1=\text{const}$, $x_2=\text{const}$, то

$$\Delta\frac{z_i}{z_1}=\Delta\frac{z_i}{z_2};\quad \Delta\frac{z_j}{z_1}=\Delta\frac{z_j}{z_2};\quad u_1=u_2=1,\quad R_1=R_2=R_0$$

и $R(x_i, x_j)=R_0$.

§ 3. Исключение ложной корреляции, вызванной общим множителем

Двумя частными случаями, приведенными в предыдущем параграфе, и исчерпываются пока все изученные случаи исключения ложной корреляции между процентными величинами. Обратимся теперь к ряду величин $\xi_i=lx_i$, $i=1, 2, \ldots, n$, непосредственно полученных из наблюдений, и предположим, что не известная нам случайная величина l не зависит от x_i.

Согласно формуле (26), для нахождения коэффициента корреляции $R=R(\xi_i, \xi_j)$ нам нужно вычислить три величины:

$$\left.\begin{aligned}\Delta\xi_i\xi_j&=\frac{Ml^2x_ix_j}{Mlx_iMlx_j}=\Delta l\Delta x_ix_j,\\ \Delta\xi_i&=\frac{Ml^2x_i^2}{(Mlx_i)^2}=\Delta l\Delta x_i;\quad \Delta\xi_j=\Delta l\Delta x_j,\end{aligned}\right\} \tag{29}$$

откуда

$$R=\frac{\Delta l\Delta x_ix_j-1}{\sqrt{(\Delta l\Delta x_i-1)(\Delta l\Delta x_j-1)}}. \tag{30}$$

Нас же интересует коэффициент корреляции $\rho=R(x_i, x_j)$, равный

$$\rho=\frac{\Delta x_ix_j-1}{\sqrt{(\Delta x_i-1)(\Delta x_j-1)}}. \tag{31}$$

Из формул (30), (31) и неравенства (24) вытекает

Теорема 4. *Если коэффициент корреляции между величинами* $\xi_i = lx_i$ *и* $\xi_j = lx_j$ *отрицателен или равен нулю, то и коэффициент корреляции между* x_i *и* x_j *отрицателен или равен нулю.*

Доказательство. В самом деле, если $\rho > 0$, то $\Delta x_i x_j > 1$, но $\Delta l \geqslant 1$, поэтому $R > 0$, причем при $R = 0$ равенство $\rho = 0$ возможно только если $l = \text{const}$, так как только в этом случае $\Delta l = 1$.

Итак, наличие отрицательной связи между ξ_i и ξ_j всегда указывает на отрицательную связь между x_i и x_j.

Подставляя в (30) выражение $\Delta x_i x_j$ из (31), получим

$$\rho = R\sqrt{\left(1 + \frac{\alpha}{\Delta x_i - 1}\right)\left(1 + \frac{\alpha}{\Delta x_j - 1}\right)} - \frac{\alpha}{\sqrt{(\Delta x_i - 1)(\Delta x_j - 1)}}, \tag{32}$$

где

$$\alpha = \frac{\Delta l - 1}{\Delta l} \geqslant 0. \tag{33}$$

Если α мало по сравнению с $\Delta x_i - 1$ и $\Delta x_j - 1$, то можно пользоваться приближенной формулой, вытекающей из равенства (32):

$$\rho \cong R + R\left(\frac{1}{2}\frac{\alpha}{\Delta x_i - 1} + \frac{1}{2}\frac{\alpha}{\Delta x_j - 1}\right) - \frac{\alpha}{\sqrt{(\Delta x_i - 1)(\Delta x_j - 1)}}. \tag{32'}$$

Если же предположить, что среди величин x_i имеется одна пара независимых друг от друга (от всех прочих величин они могут зависеть как угодно), то ложная корреляция, вызванная общим множителем l, полностью исключается.

Пусть известно, что x_1 и x_2 независимы или даже только некоррелированы между собой. Найдем коэффициент корреляции между x_i и x_j; $i, j > 1$; $i \neq j$. В силу некоррелированности x_1 и x_2 получим $\Delta\xi_1\xi_2 = \Delta l \Delta x_1 x_2 = \Delta l$; итак,

$$\Delta l = \Delta\xi_1\xi_2 \tag{34}$$

находится из наблюдений.

Аналогично, из формул (29) находим

$$\Delta x_i = \frac{\Delta\xi_i}{\Delta l} = \frac{\Delta\xi_i}{\Delta\xi_1\xi_2}; \quad \Delta x_j = \frac{\Delta\xi_j}{\Delta\xi_1\xi_2}, \tag{35}$$

$$\Delta x_i x_j = \frac{\Delta\xi_i\xi_j}{\Delta l} = \frac{\Delta\xi_i\xi_j}{\Delta\xi_1\xi_2}. \tag{36}$$

Подставляя (35) и (36) в (31), найдем

$$\rho = R(x_i, x_j) = \frac{\Delta\xi_i\xi_j - \Delta\xi_1\xi_2}{\sqrt{(\Delta\xi_i - \Delta\xi_1\xi_2)(\Delta\xi_j - \Delta\xi_1\xi_2)}}, \tag{37}$$

т. е., действительно, значение ρ выражается через величины ξ_1, ξ_2, ξ_i, ξ_j, доступные непосредственным наблюдениям.

Замечание. Если относительно x_i и x_j известно только, что они имеют одинаковые параметры $\Delta x_i = \Delta x_j = \Delta x$, то формулы (30)

и (31) упрощаются и связь между R и ρ делается более прозрачной. Тогда

$$R=\frac{\Delta l\Delta x_i x_j-1}{\Delta l\Delta x-1},\quad \rho=\frac{\Delta x_i x_j-1}{\Delta x-1},$$

откуда

$$\rho=R-(1-R)\frac{\Delta l-1}{\Delta x-1}\frac{1}{\Delta l}=R-(1-R)\frac{\alpha}{\Delta x-1}. \tag{38}$$

Если $R\leqslant 0$, то $\rho<0$, если $R=1$, то $\rho=1$. Поэтому как при R, близких к 1, так и при малых $\frac{\alpha}{\Delta x-1}$ верна приближенная формула $\rho\approx R$.

§ 4. Исключение ложной корреляции, вызванной общим слагаемым

Пусть наблюдениям доступны величины

$$\eta_1=y_1+m,\quad \eta_2=y_2+m,\ \ldots,\ \eta_n=y_n+m, \tag{4'}$$

где m не зависит от $y_1, y_2, \ldots, y_n$.

Наличие общего слагаемого m вносит ложную корреляцию и коэффициент корреляции между η_i и η_j не равен коэффициенту корреляции между y_i и y_j.

Верны, однако, следующие формулы:

$$M\eta_i=My_i+M_m;\quad D\eta_i=Dy_i+Dm,\qquad i=1,\ 2,\ \ldots,\ n. \tag{39}$$

Кроме того,

$$\operatorname{cov}\eta_i\eta_j=M\eta_i\eta_j-M\eta_iM\eta_j=$$
$$=M(y_i+m)(y_j+m)-(My_i+Mm)(My_j+Mm)=\operatorname{cov}y_iy_j+Dm, \tag{40}$$

$$\rho=R(y_i,\ y_j)=\frac{\operatorname{cov}y_iy_j}{\sqrt{Dy_iDy_j}}=\frac{\operatorname{cov}\eta_i\eta_j-Dm}{\sqrt{(D\eta_i-Dm)(D\eta_j-Dm)}}. \tag{41}$$

Поэтому, если коэффициент корреляции между η_i и η_j отрицателен, то и коэффициент корреляции между y_i и y_j тоже отрицателен.

Укажем один случай, в котором можно пользоваться приближенной формулой вида $\rho\approx R=R(\eta_i,\ \eta_j)$.

Пусть среди случайных величин $\eta_1, \eta_2, \ldots, \eta_n$ имеется одна величина, например η_1, дисперсия которой по крайней мере в k раз меньше дисперсий $D\eta_i$ и $D\eta_j$ двух других величин η_i и η_j, т. е.

$$\left.\begin{aligned}D\eta_1&\leqslant\tfrac{1}{k}D\eta_i,\\ D\eta_1&\leqslant\tfrac{1}{k}D\eta_j,\end{aligned}\right\} \tag{42}$$

или

$$Dy_1+Dm\leqslant\tfrac{1}{k}D\eta_i,$$
$$Dy_1+Dm\leqslant\tfrac{1}{k}D\eta_j,$$

или

$$\left.\begin{aligned} Dm &< \tfrac{1}{k}\, D\eta_i, \\ Dm &< \tfrac{1}{k}\, D\eta_j; \end{aligned}\right\} \tag{43}$$

отсюда

$$\rho = R(y_i, y_j) = \frac{R(\eta_i, \eta_j) - \dfrac{Dm}{\sqrt{D\eta_i D\eta_j}}}{\sqrt{\left(1 - \dfrac{Dm}{D\eta_i}\right)\left(1 - \dfrac{Dm}{D\eta_j}\right)}} = \frac{R - \theta_{ij}\dfrac{1}{k}}{\sqrt{\left(1 - \theta_i \dfrac{1}{k}\right)\left(1 - \theta_j \dfrac{1}{k}\right)}}, \tag{44}$$

где θ_i, θ_j, θ_{ij} лежат между нулем и единицей. Иначе говоря, погрешность приближенной формулы вида $\rho \approx R$ имеет порядок $\frac{1}{k}$.

Как и в предыдущем параграфе, если имеется пара некоррелированных компонентов, например y_1 и y_2, то ложная корреляция может быть полностью исключена (см. работу [5]).

Пусть y_1 и y_2 некоррелированы, тогда $\operatorname{cov} y_1 y_2 = 0$ и из формулы (40) можно найти неизвестную дисперсию

$$Dm = \operatorname{cov} \eta_1 \eta_2; \tag{45}$$

после этого в правой части (41) все величины станут известными и мы получим:

$$\rho = R(y_i, y_j) = \frac{\operatorname{cov} \eta_i \eta_j - \operatorname{cov} \eta_1 \eta_2}{\sqrt{(D\eta_i - \operatorname{cov} \eta_1 \eta_2)\,(D\eta_j - \operatorname{cov} \eta_1 \eta_2)}}. \tag{46}$$

Замечание. Случай исключения ложной корреляции, вызванной общим множителем, рассмотренный в предыдущем параграфе, сводится к только что рассмотренному случаю логарифмированием. Если $\xi_i = l x_i$, то, обозначив $\eta_i = \lg \xi_i$, $y_i = \lg x_i$, $m = \lg l$, получим величины вида $\eta_j = y_i + m$ с общим слагаемым. Если есть основание считать, что x_i и x_j связаны логарифмически нормальной корреляцией, т. е., что $y_i = \lg x_i$ и $y_j = \lg x_j$ находятся в нормальной корреляции, то логарифмирование весьма желательно. Заметим, что если мы нашли коэффициент корреляции ρ между логарифмами и корреляция между ними нормальна, то коэффициент корреляции r между исходными величинами выражается формулой:

$$r = R(x_i, x_j) = \frac{e^{\rho} - 1}{e - 1}, \tag{47}$$

где $\rho = R(\lg x_i, \lg x_j)$. Когда ρ возрастает от -1 до $+1$, r возрастает от $-\frac{1}{e} \approx -0{,}36788$ до 1, при $\rho = 0$ и $r = 0$.

При положительных ρ максимальное расхождение между ρ и r получается при $\rho = \ln(e-1) \approx 0{,}5413$, и оно равно

$$\sup_{0 \leqslant \rho \leqslant 1} (\rho - r) = \ln(e-1) - \frac{e-2}{e-1} \approx 0{,}1232. \tag{48}$$

Заметим, наконец, что ρ есть максимальный коэффициент корреляции (см. [6]) между x_i и x_j, поэтому он по абсолютной величине больше (или равен) соответствующего r, и из равенства $\rho = 0$ следует независимость x_i и x_j, а не только некоррелированность, следующая из равенства $r = 0$.

ЛИТЕРАТУРА

1. F. A. Chayes. A petrographic criterion for the possible replacement origin of rocks. Am. Journ. Sci., 246, 413—425, 1948.
2. А. Б. Вистелиус. Минеральные ассоциации и характерные парагенезисы апт-сенеманской терригенной толщи Закаспия. Доклады АН СССР, 97, № 3, 503—506, 1954.
3. О. В. Сарманов и А. Б. Вистелиус. О корреляции между процентными величинами. Доклады АН СССР, 126, № 1, 22—25, 1959.
4. F. A. Chayes. Numerical Correlation and Petrographic variation.
5. О. В. Сарманов и Ю. Т. Медведев. О ложной корреляции между случайными величинами. Теория вероятностей и ее применения. Резюме доклада.
6. О. В. Сарманов. Максимальный коэффициент корреляции (симметричный случай). Доклады АН СССР, 120, № 4, 715—718, 1958.