С.Г. Светуньков

О ВЗАИМОСВЯЗИ ЯВЛЕНИЙ, ДАННЫЕ О КОТОРЫХ ИЗМЕРЯЮТСЯ В НОМИНАЛЬНОЙ ШКАЛЕ

Одним из важнейших разделов математической статистики является раздел, называемый корреляционным анализом.

Задачей этого раздела является выявление взаимосвязи между двумя и более случайными переменными и оценка степени этой взаимосвязи.

К сожалению, эта задача до сих пор не имеет удовлетворительного решения, хотя корреляционному анализу в математической статистике уделяется много внимания. Конечно же, основной причиной того, что корреляционный анализ слабо справляется со своей задачей, является сложность этой задачи.

Действительно, очень редко когда исследователь сталкивается с этой задачей в ситуации, когда перед ним находятся данные о двух факторах, которые изолированы от влияния других факторов. В таком случае задача выявления степени коррелированности факторов решается довольно успешно.

Но в подавляющем большинстве случаев исследователь сталкивается с ситуацией, когда ему приходится анализировать множественные взаимосвязи и при этом ему сложно априорно оценить то, какие факторы действительно являются взаимосвязанными, а у каких факторов такой взаимосвязи нет или она опосредована: «Из того, что значения коэффициента корреляции r_{xy} высоки, нельзя вывести ни одного из следующих утверждений:

1) *y* зависит от *x*;

2) *x* зависит от *y*;

3) *x* и *y* совместно зависят от какой-то третьей переменной» [2, с. 44]

Довольно часто исследователи сталкиваются с ситуацией, когда абсолютно не связанные друг с другом факторы имеют высокое значение коэффициента парной корреляции. Это явление получило название «ложная корреляция». И предложил этот термин сам К. Пирсон, который и обосновал формулу коэффициента парной корреляции. Для демонстрации этого явления К. Пирсон показал, что если два независимых друг от друга фактора x_{1i} и x_{2i} имеют общий знаменатель x_{3i}, то между рядами $\{x_{1i}/x_{3i}\}$ и $\{x_{1i}/x_{3i}\}$ будет высчитываться коэффициент парной корреляции, далёкий от нуля, и свидетельствующий о наличии линейной взаимосвязи между факторами [1, с. 719].

Если при анализе корреляции между данными, измеренными в метрической - самой точной шкале измерения данных, - возникают сложности с вычислением корреляции, то что говорить об выявлении и оценке степени коррелированности рядов, данные о которых измеряются в шкале низшего уровня – номинальной шкале? Здесь ситуация ещё сложнее.

Поскольку математические действия с данными, измеренными в номинальной шкале, невозможны, работают с количеством наблюдений за этими данными. При этом чаще всего говорят не о данных, а о признаках, поскольку именно это свойство и фиксируется в ходе наблюдения за объектами.

Как понимать корреляцию между такими признаками? Общее правило

таково. Если рост количества наблюдений при изменении одного признака соответствует росту или уменьшению количества наблюдений другого признака, то говорят о наличии между ними корреляции. Если же изменения количества наблюдений за одним признаком не сопровождаются изменением количества наблюдений другого признака, то корреляции нет.

Для расчёта корреляции данных, измеренных в метрических шкалах, их сводят в специальные таблицы, которые называют таблицами сопряжённости.

Табл. 1.
Общий вид таблицы сопряжённости

Значение признаков	x_0	x_1	**Итог**
y_0	a	b	$a+b$
y_1	c	d	$c+d$
Итог	$a+c$	$b+d$	$N=a+b+c+d$

Чаще всего используют три коэффициента, которые используют приведённые в таблице сопряжённости числа:

1) коэффициент ассоциации Юла:

$$Q = \frac{ad - bc}{ad + bc}, \qquad (1)$$

2) коэффициент коллигиации Юла:

$$Q_k = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}, \qquad (2)$$

4) коэффициент сопряжённости Пирсона:

$$\varphi = \frac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}. \qquad (3)$$

Мы не будем рассматривать логику формирования каждого из этих коэффициентов. Оставим это для более подробного исследования. Отметим только, что эти коэффициенты в ряде случаев не справляются с задачей корреляционного анализа.

Приведём два таких случая, когда эти три коэффициента неправильно

выявляют корреляцию между признаками.

Пример первый. Данные табл. 2 подобраны таким образом, что на переход от признака x_0 к признаку x_1 первый сопряжённый признак y_0 реагирует, а второй сопряжённый признак y_1 не реагирует. Это может быть, например, когда x_0 – это просмотр по телевизору футбольных матчей, а x_1 – просмотр по телевизору хоккейных матчей. При этом второй признак – это пол зрителя передачи: y_0 – мужской пол, y_1 - женский пол.

Из таблицы видно, что мужчины предпочитают смотреть хоккей, а женщинам – всё равно что смотреть, лишь бы в квартире был какой-нибудь шум.

Табл. 2.
Пример данных, при котором коэффициенты (1) - (3) отражают отсутствующую корреляцию

Значение признаков	x_0	x_1	**Итог**
y_0	10	200	*210*
y_1	20	20	*40*
Итог	*30*	*220*	*N=250*

Какие результаты покажут эти три коэффициента? Вот они:

1) коэффициент ассоциации Юла равен Q= -0,905 (очень сильная корреляция),

2) коэффициент коллигиации Юла Q_k= -0,635 (заметная корреляция),

3) коэффициент сопряжённости Пирсона φ = -0,510 (заметная корреляция).

В рассматриваемом примере корреляции между видом спортивных соревнований, транслируемых по телевизору, и полом зрителя нет. Следовало бы рассмотреть другую зависимость, а именно – между видом трансляции (спортивная передача или развлекательная передача) и полом зрителя. В этом случае взаимосвязь между признаками есть и вычисление корреляции должно эту взаимосвязь выявить.

Пример второй. Данные табл. 3 подобраны иначе - на переход от признака x_0 к признаку x_1 первый сопряжённый признак y_0 реагирует увеличением количества наблюдений, также как и второй признак y_1. Но второй признак увеличивает количество наблюдений в два раза.

На эту очевидную взаимосвязь между признаками все три коэффициента реагируют одинаково плохо:

1) коэффициент ассоциации Юла равен Q= 0,

2) коэффициент коллигиации Юла Q_k= 0,

3) коэффициент сопряжённости Пирсона φ = 0.

Табл. 3.
Пример данных, при котором коэффициенты (1) - (3) не выявляют существующую корреляцию

Значение признаков	x_0	x_1	**Итог**
y_0	10	30	*40*
y_1	20	60	*80*
Итог	*30*	*90*	*N=120*

Исследователь, доверившись этим показателям, сделает вывод о том, что между признаками *x* и *y* никакой корреляции нет, а это не так.

Следовательно, актуальной является задача по поиску новых подходов и методов выявления и оценивания степени корреляции между номинальными данными.

Покажем одну из таких возможностей.

Прежде всего, обратим внимание на то, что рассматриваемые таблицы сопряжённости могут быть представлены графически в трёхмерном пространстве. Осями этого пространства являются признаки *x* и *y*, а также количество наблюдаемых появлений каждого признака *n* (рисунок 1).

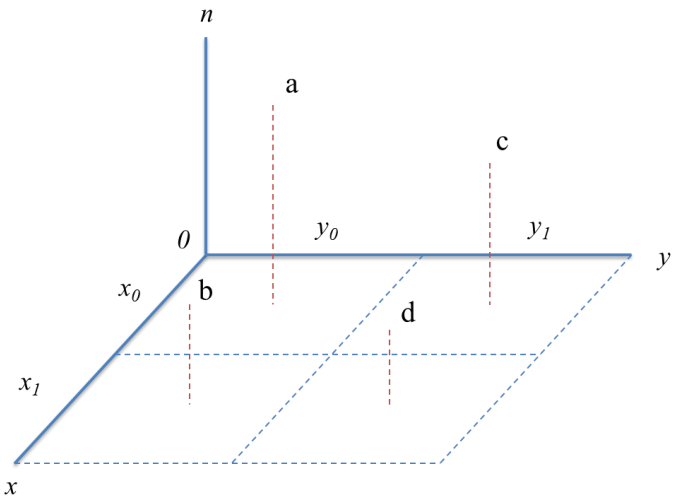

Рисунок 1. Таблица сопряжённости признаков в графической интерпретации

Четыре точки в трёхмерном пространстве проецируются на каждую из составляющих пространство плоскости. Интерес представляют плоскости *n0y* и *n0x*. Первая плоскость будет изображена так:

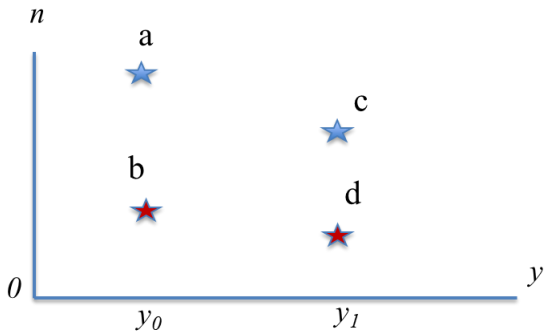

Рисунок 2. Проекции точек таблицы сопряжённости на плоскость n0y

Как известно, через две точки можно провести одну и только одну прямую. Проведём на плоскости рисунка 2 через точки *a* и *c* одну прямую, а через точки *b* и *d* – другую. Получим:

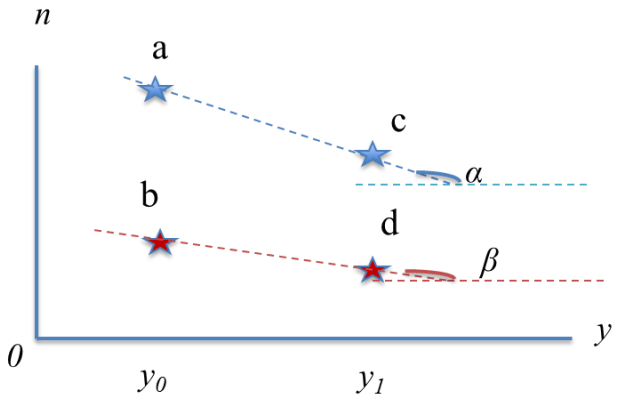

Рисунок 3. Прямые на плоскости n0y

Тангенсы углов каждой из прямых считаются элементарно: $tg\alpha=(a-c)$, $tg\beta=(b-d)$.

Аналогично и на второй плоскости *n0x* через проекции точек также можно провести две другие прямые:

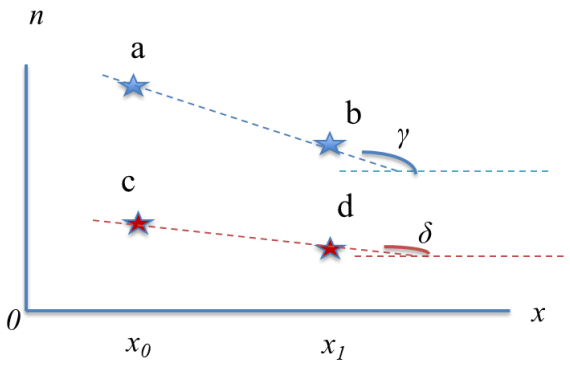

Рисунок 4. Прямые на плоскости n0y

И их углы наклона вычисляются элементарно: $tg\gamma=(a-b)$, $tg\delta=(c-d)$.

Если хотя бы один из признаков не реагирует на изменение сопряжённого признака и остаётся постоянным, то прямая линия, проходящая через проекции таких точек, будет характеризоваться тем, что тангенс такого угла будет равен нулю. Такую ситуацию мы определили как отсутствие

корреляции. Поэтому в таком случае произведение тангенсов всех четырёх углов будет равно нулю, а во всех остальных случаях оно будет больше нуля. Произведение четырёх тангенсов тяжело интерпретировать, а вот средняя геометрическая этих тангенсов имеет простое толкование. Поэтому следует использовать среднюю геометрическую, для которой:

$$\sqrt[1/4]{(a-c)\cdot(b-d)\cdot(a-b)\cdot(c-d)} \geq 0. \quad (4)$$

Первое условие для измерения наличия или отсутствия корреляции между признаками средняя геометрическая выполняет: когда корреляции нет, она равна нулю. Но все коэффициенты, в той или иной степени используемые для вычисления степени корреляции, не должны по модулю превышать единицу, а средняя геометрическая (4) может быть какой угодно. Поэтому средняя геометрическая должна быть отмасштабирована так, чтобы она не превышала единицу. Для этого следует отмасштабировать числа в таблице сопряжённости *a, b, c, d,* разделив каждое из них на максимальное значение из этих чисел, то есть, преобразовать таблицу сопряжённости 1 к новому виду:

Табл. 2.
Отмасштабированная таблица сопряжённости

Значение признаков	x_0	x_1
y_0	*a/max(a,b,c,d)*	*b/ max(a,b,c,d)*
y_1	*c/ max(a,b,c,d)*	*d/ max(a,b,c,d)*

В таком случае ни один из тангенсов углов не будет превышать единицу, и средняя геометрическая (4) будет лежать в пределах от нуля до единицы. Но поскольку средняя геометрическая чётного числа членов всегда не отрицательна, то при её использовании как некоторого инструмента для измерения степени корреляции не определяется направление взаимосвязи – прямая или обратная. Для того, чтобы придать коэффициенту это свойство, следует умножить (4) на знак произведения тангенсов:

$$S_{gs} = знак((a-c)\cdot(b-d)\cdot(a-b)\cdot(c-d))\sqrt[1/4]{(a-c)\cdot(b-d)\cdot(a-b)\cdot(c-d)}. \quad (5)$$

Например, в MS Excel есть встроенная функция, определяющая знак вычисляемого числа. В том случае, когда исследователь пользуется иным программным продуктом, то знак произведения тангенсов можно вычислить, разделив произведение тангенсов на произведение модулей тангенсов. Тогда искомый коэффициент будет иметь такой вид:

$$S_{gs} = \left(\frac{a-c}{|a-c|}\cdot\frac{b-d}{|b-d|}\cdot\frac{a-b}{|a-b|}\cdot\frac{c-d}{|c-d|}\right)\sqrt[1/4]{(a-c)\cdot(b-d)\cdot(a-b)\cdot(c-d)}. \quad (6)$$

Проверим как работает предложенный коэффициент на двух вышеприведённых примерах, когда существующие коэффициенты демонстрировали свою неточность.

Приведём результаты ниже:

Табл. 3.
Сравнительный анализ расчёта коэффициентов корреляции

Наименование коэффициента	Значение коэффициента по данным таблицы 2 (отсутствие взаимосвязи)	Значение коэффициента по данным таблицы 3 (наличие взаимосвязи)
Коэффициент ассоциации Юла	-0,905	0
Коэффициент коллигиации Юла	-0,635	0
Коэффициент Пирсона	-0,510	0
Коэффициент S_{gs}	0	0,369

Как видно из приведённых примеров, новый коэффициент показывает наличие связи там, где она есть и сигнализирует об отсутствии связи там, где её нет.

Теперь необходимо понять: как интерпретировать значения коэффициента, лежащего по модулю в промежутке от нуля до единицы.

Исходя из того, что тангенс – функция нелинейная, следует согласиться с тем, что градация, принятая для большинства коэффициентов, измеряющих степень взаимосвязи между случайными факторами, здесь не приемлема.

Действительно, при интерпретации значений различных корреляционных коэффициентов исходят из линейной зависимости между значениями коэффициентов и степенью тесноты связи. Выделяют несколько градаций степени взаимосвязи, например: сильная, средняя, слабая и её отсутствие. Всего четыре уровня. Тогда разделив отрезок изменения модуля какого-либо корреляционного коэффициента на четыре отрезка, дают каждому отрезку соответствующую интерпретацию:

- при значениях модуля коэффициента, лежащего в пределах от нуля до 0,25 говорят об отсутствии связи;

- при значениях модуля коэффициента, лежащего в пределах от 0,25до 0,5 говорят о слабой взаимосвязи;

- при его значениях, лежащих в пределах от 0,5 до 0,75 говорят о среднем уровне корреляции;

- при его изменениях выше 0,75 говорят о сильной зависимости.

Коэффициент (4) представляет собой среднее геометрическое тангенсов углов наклона и аргументом здесь выступают именно углы наклона. Поскольку (4) по модулю меняется в пределах от нуля до единицы, то аргумент (угол наклона) следует рассматривать в пределах от 0 до $\pi/4$. Четырём уровням

степени корреляционной взаимосвязи соответствуют четыре значения аргумента: π/16; 2π/16; 3π/16; 4π/16.

Теперь можно дать рекомендации о том, как интерпретировать силу связи в зависимости от значений коэффициента S_{gs}:

$|S_{gs}| < 0{,}199$ – отсутствие взаимосвязи между рассматриваемыми признаками;

$0{,}199 \leq |S_{gs}| < 0{,}414$ – слабая корреляционная взаимосвязь;

$0{,}414 \leq |S_{gs}| < 0{,}668$ – корреляционная взаимосвязь среднего уровня;

$0{,}668 \leq |S_{gs}|$ – сильная корреляционная взаимосвязь.

ЛИТЕРАТУРА

1. *Дружинин Н.К. О ложной корреляции // Экономика и математические методы, 1984, т. XX, вып. 4. С. 719 – 725.*
2. *Кейн Э. Экономическая статистика и эконометрия: Введение в количественный экономический анализ / Пер. с англ. Р. Мошкович, С. Николаенко, А. Шмидта. Москва: Статистика, 1977. Вып. 2. 1977. 228 с.*