УДК 616-07(076.5)

С. 48–53 | Том VII. № 1. 2013

Корреляционный анализ данных: зонд в глубины скрытых механизмов взаимодействий

Г. П. Тихова

ООО «ИнтелТек Лаб», Петрозаводск

Correlation analysis: exploring hidden mechanics of relationships

G. P. Tikhova

IntelTeck Lab Ltd, Petrozavodsk

В статье подробно рассматривается понятие линейной зависимости между показателями и ее исследование методом корреляционного анализа. На примере конкретного клинического исследования обсуждается графическая и содержательная интерпретация значений коэффициентов корреляции, полученных в ходе статистической обработки экспериментальных данных.

The concept of linear relationship between parameters and its investigation using methods of correlation analysis are considered in present paper. Graphical and subject interpretation of correlation coefficient values obtained during statistical processing of experimental data is discussed on the base of clinical study.

Цель любой статистической обработки данных состоит в том, чтобы из огромного количества чисел вывести несколько интегральных величин, которые позволят судить о тенденциях или закономерностях изучаемого процесса. Еще лучше, если эти расчетные интегральные величины позволяют сделать какие-то качественные выводы в терминах уже той предметной области, в которой проводится исследование. Самое трудное, настоящее узкое место современных биомедицинских исследований – это правильное и оптимальное описание проблемы в терминах математической статистики (для того чтобы применить статистические методы, которые работают только в поле статистики), а затем корректно провести обратную процедуру – перевести результаты статистической обработки данных в термины предмета исследования. В связи с этим непросто проводить исследования с применением статистики в медицине и биологии, имеется в виду настоящие исследования, а не просто украшение научной работы парой статистических тестов. Поэтому не надо отчаиваться и расстраиваться, если не все сразу, слету понятно и хорошо интерпретируется. Это закономерно. Результаты исследования с применением статистики всегда требуют длительного обдумывания, они часто воспринимаются лучше в диаграммах и графиках, чем в числах, поэтому изложение ваших гипотез или заключений будет выглядеть нагляднее и понятнее, если полученные величины или исходные данные будут продублированы графическими изображениями. Кроме того, как правило, эти результаты не отвечают сразу на практические вопросы типа насколько больше/меньше, какова оптимальная доза или инфузионный объем и т. п. Чаще всего это описание скрытых механизмов взаимодействия тех показателей, которые мы можем регистрировать. Они (показатели) находятся на поверхности процесса, но именно их взаимодействие отражает глубинный механизм, который по-настоящему важен для понимания проблемы и ее практического решения. Важно подчеркнуть, что каждый показатель в отдельности не дает достаточного представления об этом, часто вообще никакого не дает, даже намека, а вот именно оценка их коллективного изменения, динамика их связей и зависимостей в ходе исследования может натолкнуть на потрясающие догадки. Все, что может сделать даже самый квалифицированный и опытный математик, это максимально приблизить интерпретацию полученных результатов к терминам

той медицинской проблемы, которая исследуется, а также попытаться максимально наглядно подать результаты статистической обработки. Клинический опыт и специальные знания конкретных патологий и особенностей развития клинического процесса, которыми обладает только врач-специалист, требуются для дальнейшей расшифровки полученных результатов и получения клинических выводов и предположений. Это фактически рождение нового знания, и это очень трудно. Если вы получили от математика или программной системы, развернутой на вашем компьютере, какие-то числа и не знаете, что дальше с ними делать, постарайтесь сформулировать четко, что вам непонятно, что именно неясно. Возможно, причина этого непонимания – просто недостаток информации или неудачное описание, замусоренное чужими терминами, значение которых вам неизвестно или не очень знакомо. Попытаемся в данной статье разъяснить смысл некоторых статистических понятий, которые, как нам кажется, могут вызывать затруднения в понимании, скорее даже в осознании некоторых результатов статистичекой обработки данных.

В этой статье мы поговорим о зависимостях между изучаемыми показателями, которые довольно часто используются в медицинских исследованиях, чтобы статистически достоверно подтвердить или опровергнуть причинно-следственные или ассоциативные связи, обнаруженные в ходе эксперимента или предполагаемые, исходя из опыта, наблюдений и теории.

Во-первых, что такое линейная зависимость между показателями и почему с ней все так носятся и стараются обнаружить именно линейность во взаимосвязи исследуемых признаков? Линейная зависимость между двумя переменными – это зависимость, которая выражается уравнением вида $Y = \mathbf{a}X + \mathbf{b}$, где **a** и **b** – коэффициенты, совершенно любые числа. На графике такая зависимость отображается в виде прямой линии, коэффициенты **a** и **b** однозначно определяют ее положение относительно декартовой системы координат XY. Но не в этом самая главная прелесть линейной зависимости для исследователя. Основное ее преимущество в том, что при изменении X на единицу, Y всегда будет изменяться на одну и ту же величину. Например, если $Y = 2X + 5$, то при изменении X с 2 на 3, Y изменится на 2 (с 9 на 11), и при скачке X с 5 на 6, приращение Y будет по-прежнему 2. Иными словами при увеличении X на 1, Y всегда будет прирастать на 2. Обратите внимание, что 2 – это тот самый коэффициент, который стоит при X в формуле, показывающей, как можно из X рассчитать Y. Эта пропорция сохраняется для любых значений X. Для сравнения можно привести пример нелинейной зависимости, например, самой простой: $Y = X^2$. В этом случае при изменении X на 1, Y будет прирастать по-разному, в зависимости от начального значения X. При изменении X с 1 на 2, Y возрастет с 1 до 4, т. е. изменится на 3, а при скачке X с 3 до 4, Y изменится с 9 до 16, т. е. уже на 7. В этом случае никакой пропорции нет, и мы не можем сказать ничего определенного об изменении Y при вариации X, кроме того, что они будут изменяться в одну сторону. Эта зависимость гораздо сложнее. Для того чтобы нам знать об изменении Y столько же, сколько мы знаем об этом, имея линейную зависимость, необходимо гораздо больше информации, и она более привязана к частным случаям. Очень важно прочувствовать это кардинальное отличие линейной зависимости от всех остальных. Итак, если два показателя линейно связаны друг с другом, это означает, что они пропорциональны и сохраняют эту пропорциональность всегда, какие бы значения они ни принимали. Мы можем не знать точно, с какого на какое значение изменилось X, но, зная на сколько увеличилось X, мы всегда можем сказать, на сколько увеличится/уменьшится Y. Понятно, что, обнаружив среди наших показателей линейнозависимые, мы сможем прогнозировать их с меньшими затратами на всем интервале их значений, чем если мы имеем какие-то сложные нелинейные зависимости.

Теперь обратимся к коэффициентам корреляции. В нашем примере мы рассчитываем и интерпретируем парные коэффициенты корреляции. Чаще всего именно эти коэффициенты и приводятся в публикациях. (Есть еще и множественные коэффициенты той же линейной корреляции, но мы их не будем пока касаться). Итак, по данным двух показателей, полученным в ходе эксперимента или наблюдения, рассчитывается парный параметрический коэффициент корреляции (Пирсона). Что это такое и что он означает? Парный коэффициент корреляции Пирсона – это некое число из интервала от –1 до +1, отражающее, как принято говорить, степень или силу или тесноту линейной зависимости, подчеркнем еще раз, именно линейной зависимости, т. е. только зависимости типа $Y =$

aX + **b**, а не любой зависимости вообще. Только линейной и никакой другой. Это очень важно прочувствовать и осознать. Например, два показателя могут быть чрезвычайно сильно связаны зависимостью, которая выражается формулой $Y = X + X^2 + X^3$, но рассчитанный коэффициент корреляции будет при этом очень низким. Он покажет слабую зависимость, почему? А именно потому, что он покажет степень линейной зависимости, т. е. связи, описываемой формулой вида Y = **a**X +**b**, тогда как мы имеем дело с нелинейной связью, гораздо более сложной и формула у нее намного сложнее и очень мало напоминает ту, которую мы предполагаем. Крайне важно помнить, что если у нас коэффициент корреляции низкий, это свидетельствует лишь о слабости (или отсутствии) линейной зависимости, а не зависимости вообще. Зависимость нелинейная, любая другая, может иметь место и весьма сильная. Как определить в этом случае, есть такая зависимость или все же два показателя никак не связаны друг с другом, мы расскажем в следующей публикации.

А теперь второй важнейший момент, касающийся корреляции. Выше говорилось, что коэффициент корреляции – это число от –1 до +1, отражающее тесноту или силу линейной связи между показателями. А как понимать термин «теснота линейной связи», что такое теснота? Интуитивно ясно, что чем выше (по модулю) значение коэффициента, тем показатели более зависимы друг от друга, но все же, что именно измеряется этим числом, которое мы рассчитываем?

Для ответа на этот вопрос сформулируем его иначе. Что значит более или менее зависимы? Попробуем изобразить это графически. Начнем с предельного случая, который на практике никогда не встречается: показатели полностью и однозначно зависят (линейно!) только друг от друга, тогда на графике мы получим прямую линию, а значение парного коэффициента корреляции равным +1 (если с увеличением одного показателя второй пропорционально увеличивается, прямая пропорциональность) или –1 (если с увеличением одного показателя второй пропорционально уменьшается, обратная пропорциональность). Итак, если два показателя полностью (в математике это называется «функционально») зависят друг от друга, то на графике получается прямая линия, а по содержанию собственно сама линейная функция (рис. 1А). Теперь «размажем» немного прямую так, чтобы вокруг нее образовалось узкое плотное облако из точек, ведь в реальном эксперименте никогда не бывает так, чтобы все измерения легли строго по прямой линии, однако через это облако по середине можно провести прямую линию, как бы стержень, к которому наши измерения притягиваются (рис. 1Б). В этом случае говорят, что зависимость линейная, близкая к функциональной, а коэффициент корреляции обычно очень высок 0,8–0,9 по модулю. Растянем облако точек-измерений еще сильнее от центральной прямой (рис. 1В). Визуально оно еще сохраняет скопление около некоторой прямой линии, но плотность этого облака падает, снижается и коэффициент корреляции, примерно, 0,5–0,6 по модулю. Эту процедуру можно продолжать до полного исчезновения какой-либо возможности однозначно провести прямую через точки измерений, потому что облако из вытянутого в некотором направлении превратилось в совершенно круглое и никаких направлений в нем не наблюдается – коэффициент корреляции равен 0 (рис. 1Г).

Эти графики наглядно демонстрируют, что такое теснота линейной зависимости двух показателей. Это близость расположения их измерений к некоторой однозначно определяемой прямой

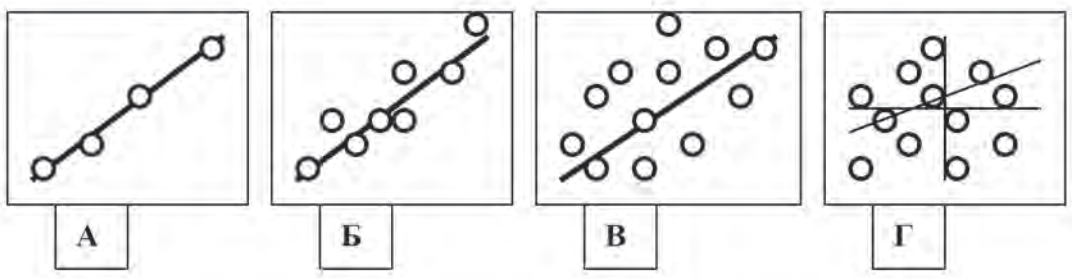

Рис. 1. Интерпретация значений парного коэффициента корреляции

линии, которая отражает их линейную зависимость друг от друга, если, конечно, такая линия и такая зависимость между ними существует. Чем плотнее облако группируется и вытягивается вдоль этой линии, тем выше коэффициент корреляции. Чем больше похоже становится облако точек-измерений на прямую линию, тем ближе коэффициент корреляции к 1 или –1. А что означает рассеивание точек все дальше от прямой? Что отталкивает их от «линии притяжения»? Причиной «рассеивания» облака является вмешательство в парную связь двух показателей сторонних факторов, сравнимых по силе с той связью, которая исследуется в этой паре. Чем меньше рассеивание, тем слабее сторонние факторы, чем рассеивание больше, тем сильнее другие внешние по отношению в исследуемой паре влияния.

В качестве примера рассмотрим небольшую часть исследования взаимосвязи азотистого баланса и его составляющих (суточного потребления азота и суточной потери азота) у пациентов с трансплантацией костного мозга [1]. Поскольку азотистый баланс (АБ) рассчитывается как разность между поступившим (N) и выделенным (MN) азотом за сутки, то он линейно зависит от обеих своих составляющих. Но если катаболическая фаза усиливается или снижается под воздействием развития патологии, течения послеоперационного периода и интенсивности нутриционной поддержки, то значения парных коэффициентов корреляции между АБ с его компонентами будут различными и покажут, в каких точках АБ более зависит от потерь азота, чем от поступления его в организм, а когда наоборот, нутриционная поддержка сможет оказать клинически значимое влияние на АБ и существенно снизить потерю азота.

Итак, в точке исследования Т1 (табл. 1) значение корреляции АБ с суточным потреблением азота и АБ с суточными потерями азота приблизительно одинаково. Это означает, что эти два компонента вносят одинаковый по значимости вклад в формирование величины азотистого баланса. Но в точке Т2 (табл. 2) корреляция между АБ и потреблением азота равна нулю, тогда как значение коэффициента корреляции между АБ и потерями азота возрастает до 0,87, что свидетельствует, что на этом этапе исследования потери азота почти полностью определяли значение азотистого баланса и потребление азота вносило в эту величину несравнимо меньший вклад.

Таблица 1. **Парные коэффициенты корреляции между АБ, суточным потреблением (N) и суточной потерей (MN) азота в точке исследования Т**

Точка 1	N	MN	АБ
N	1		
MN	0,47	1	
АБ	0,49	-0,54	1

Таблица 2. **Парные коэффициенты корреляции между АБ, суточным потреблением (N) и суточной потерей (MN) азота в точке исследования Т2**

Точка 2	N	MN	АБ
N	1		
MN	0,42	1	
АБ		-0,87	1

На графике это развитие событий видно особенно ярко (рис. 2). При переходе от Т1 к Т2 облако точек, отражающих связь между АБ и потреблением азота, расплывается в круг и становится совершенно бесформенным, значит в связь АБ – N мощно вмешиваются сторонние силы, большую часть из которых составляет потеря азота, что видно, если сравнить графики первого и второго столбцов на рис. 2. Тогда как облако АБ – N в точке Т2 расплывается в круг, облако АБ – MN сжимается и вытягивается вдоль прямой линии, демонстрируя почти полное отсутствие вмешательства каких-либо факторов в сильнейшую парную зависимость АБ-суточные потери азота. Это означает, что на этапе Т2 азотистый баланс в наибольшей степени определяется потерями азота. Дальнейшие содержательные выводы можно делать из этого факта. Так, можно проанализировать и сравнить все остальные графики и корреляции, и динамика процесса предстанет достаточно глубоко и всесторонне, а из нее можно сделать уже и практические выводы или предположения.

Есть еще один важнейший момент, связанный с интерпретацией и сравнением коэффициентов корреляции. Поскольку получение в исследуемой паре того или иного значения корреляции по своей природе процесс вероятностный, то необходимо отделить различие этих значений, обусловленное стохастической природой данных, и различие, причиной которого является факторы воздействия,

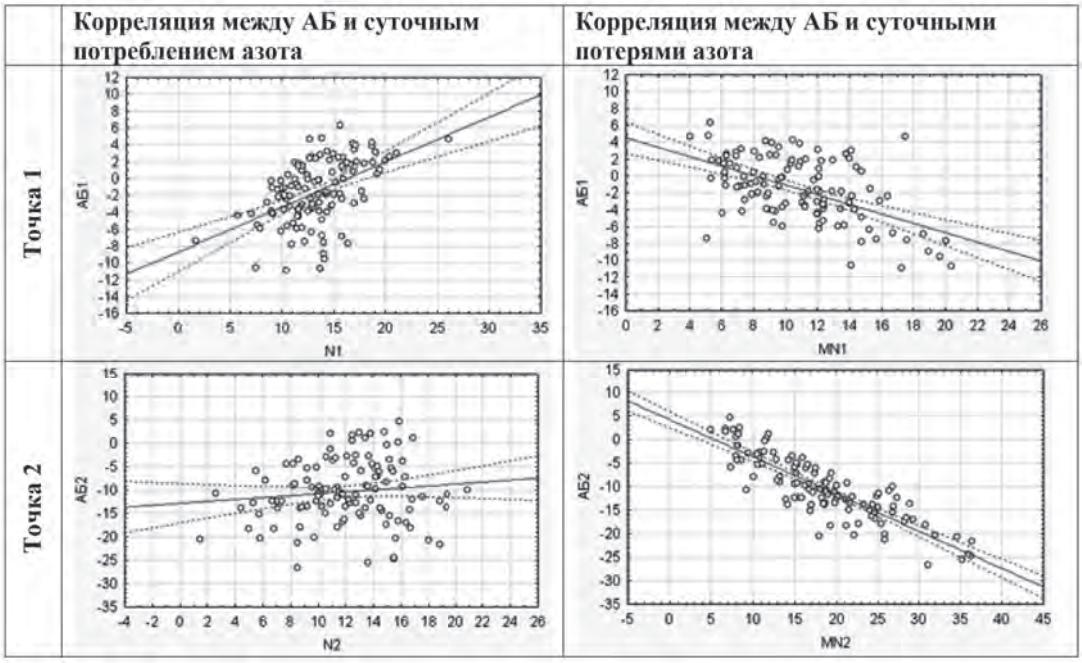

Рис. 2. Графическое отображение силы линейной связи между парами показателей, выраженной значениями парных коэффициентов корреляции

т. е. истинное, реальное различие (в статистике это называется «статистически значимое/достоверное различие»). Для этого вместе с коэффициентом корреляции рассчитывается его ошибка, которая и показывает, насколько можно доверять полученному значению (насколько оно «случайно» и обусловлено конкретной выборкой). Эта ошибка определяет статистическую значимость коэффициента корреляции, т. е. тот факт, что полученное значение статистически значимо/достоверно отличается от нулевого. Статистической значимости все равно чему равен коэффициент корреляции: 0,9 или 0,1. Статистическая значимость указывает лишь на то, что это значение достоверно отличается от нуля и это обусловлено природой изучаемого процесса, а не случайными вариациями данных. Иными словами, статистическая значимость подтверждает, что какая-то степень зависимости есть. Но реально это может быть степень зависимости равная 0,7 или 0,2. Для определения статистической значимости конкретная величина корреляции не имеет значения, эта значимость рассчитывается по специальному алгоритму, и отвечает на вопрос: мог бы получиться 0, если бы исходные данные случайно чуть-чуть изменились, или нет? Если корреляция статистически значима, значит ответ – нет, ноль получиться от случайной флуктуации данных не может. Все, на этом сфера компетенции математической статистики заканчивается и дальше уже начинается собственно территория исследователя. Что такое корреляция, равная 0,2? Это значит, грубо говоря, лишь 4% всего изменения одного показателя может зависеть от изменения второго. Много это или мало? Просто ничтожно мало. Стоит ли принимать во внимание такую зависимость, причем линейную? Как правило, ее игнорируют, особенно если имеют место более сильные корреляции. Обычно корреляции ниже 0,3 считаются пренебрежимо малыми и в рассмотрение не принимаются. Такие корреляции называются клинически незначимыми. Иными словами, они отличны от нуля, но они не имеют клинического значения, а только засоряют выделения существенных факторов влияния. Вообще человеческое сознание устроено так, что при сравнении ему хочется от чисел перейти к каким-то качественным градациям. При сравнении корреляции тоже удобнее сравнивать не просто числа, а несколько степеней зависимости, поэтому обычно разбивают весь интервал клинически значимых значений корреляции на три градации: слабая, средняя и сильная зависимость. Конкретные пороги, вообще говоря, устанавливаются исследователем, но в наших

Таблица 3. **Разбиение интервала значений коэффициента корреляции на качественные уровни связи**

Интервал значений коэффициента корреляции	Сила линейной связи
<0,3	Зависимость клинически незначима
[0,3; 0,45)	Слабая зависимость
[0,45; 0,65)	Средняя зависимость
[0,65; 1,0]	Сильная зависимость

исследованиях мы всегда придерживаем примерно одного и того же разбиения (табл. 3).

Этот метод, во-первых, позволяет отсеять незначительные изменения корреляции (внутри одной градации), которые, скорее всего, обусловлены случайной природой данных, и обратить внимание на существенную трансформацию связи (когда градации меняются, особенно если значение коэффициента корреляции перескакивает через одну градацию). Такая трансформация связи, безусловно, имеет причиной кардинальные перемены в состоянии исследуемого процесса. Для наглядности эти же результаты могут быть отражены на диаграммах с разной толщиной дуг, связывающих исследуемые показатели в различных точках исследования.

Литература

1. *Matushevskaya V., Fedorenko D., Melnichenko V., Tikhova G.* Changes in protein intake – nitrogen balance relationship in patients undergoing BMT. Clinical Nutrition. 2012; 7 (1): 232.