

В.П. Ильин

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ КОЛИЧЕСТВЕННЫХ ДАННЫХ В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

ФГБУ «Научный центр проблем здоровья семьи и репродукции человека» СО РАМН (Иркутск)

В работе приводятся методические особенности корреляционного анализа количественных переменных, касающихся условий применения, вычисления, трактовки и определения истинности или опосредованности парных связей при проведении медико-биологических исследований.

Ключевые слова: парная корреляция, теснота и степень объясненной связи, истинная и опосредованная корреляция

CORRELATION ANALYSIS QUANTITATIVE VARIABLES IN MEDICAL-BIOLOGICAL RESEARCHES

V.P. Iljin

Scientific Centre of Family Health and Human Reproduction Problems SB RAMS, Irkutsk

The methodical particularities of the using happen to in medical- biological correlation analysis quantitative variable, concerning conditions of the using, calculations, interpretations and determinations to true or mediated fresh relationships when undertaking medical-biological studies.

Key words: fresh correlation, tightness and degree explained relationship, true and mediated correlation

Конечная цель научного исследования, в том числе и медико-биологического, состоит в нахождении связей (зависимостей) между изучаемыми переменными. Философия науки учит, что не существует иного способа представления знания, кроме как в терминах зависимостей между количествами или качествами, выраженными какими-либо переменными. Исследование корреляций по существу и состоит в измерении таких зависимостей непосредственным образом [6].

1. КОРРЕЛЯЦИЯ И ЕЕ СВОЙСТВА

Корреляция является мерой связи между двумя или более переменными. В зависимости от типов шкалы [1–6], в которых замерены изучаемые переменные, корреляция может быть описана различными методами статистического анализа. Так для метрических и интервальных переменных – это корреляция Пирсона. При этом переменные должны быть замерены, по крайней мере, в интервальной шкале. Для переменных, замеренных в шкале наименований – это конкордация, ассоциация или сопряженность признаков. Для переменных, замеренных в порядковой шкале – это ранговые методы корреляции: Спирмена, Кендалла, гамма и другие. Кроме типа шкалы, в которой замерены переменные, на выбор метода корреляционного анализа влияет распределение изучаемых переменных.

Что такое корреляция? Корреляция – это согласованное изменение переменных. Более точно, корреляция – это согласованность распределений наблюдаемых переменных. Моделью корреляции является прямая линия, определяющая наличие и степень пропорциональности изучаемых переменных. При этом саму согласованность, возможно не всегда, можно выразить в явном виде. Корреляционные методы

особенно эффективны, когда связи не совершенны, так что каждому значению x вовсе не соответствует одно единственное значение y . Например, такими переменными могут быть рост, вес и объем талии у женщин, когда у женщин с одним ростом встречаются различные значения веса и объема талии.

Коэффициент корреляции изменяется от $-1,00$ до $1,00$. Значение корреляции « $-1,00$ » представляет собой вариант идеальной детерминированной отрицательной связи, значение « $+1,00$ » – является выражением идеальной детерминированной положительной корреляции. Значение $0,00$ представляет собой отсутствие корреляции.

Заметим, что при значениях корреляции более $0,95$ следует более внимательно рассмотреть первичные данные, относящиеся к наблюдаемым переменным. Значения корреляции большие величины $0,95$ на 5% уровне значимости являются неким аналогом детерминированного значения корреляции равной 1 . Что значит корреляция равная 1 ? Математически означает полную функциональную связь и отсутствие изменчивости (дисперсии) одновременно у двух показателей. Причин множество, но наиболее распространенных – две. Первая. Обе переменные детерминированы, т.е. константы и состоят из одинакового набора цифр. Второй – одна из переменных получена из второй с помощью арифметических операций (например, умножения, сложения или более сложных преобразований).

Диаграммы рассеяний, изображенные на плоскости, наглядно показывают зависимость между 2 переменными. Так, если на оси X отложить значения одной переменной, а на оси Y – другой, то расположение точек на плоскости представляет собой корреляцию между переменными x и y . Чем больше расположение точек напоминает круг, тем меньше корреляция (значение

коэффициента корреляции r ближе к 0). Чем больше расположение точек представляет прямую линию, тем больше (теснее) корреляция между переменными (значение коэффициента корреляции r ближе к 1).

На рис. 1, заимствованном из документации к пакету Statistica, приведена линейная корреляция Пирсона в вариантах: статистически существенной положительной корреляции $r = +0,90$ (GROUP 3, 4), отсутствующей $r = 0,00$ (GROUP 1) и статистически незначимой корреляции $r = +0,30$ (GROUP 2). Очевидно, что при $r = +0,90$ расположение множества точек ближе к прямой линии, в случае $r = +0,60$ более растянуто по второй оси эллипса, и при $r = +0,00$ расположение точек отличается от линии и по расположению больше напоминает круг, чем вытянутый эллипс. Соотношение осей эллипса и показывает выраженность корреляции.

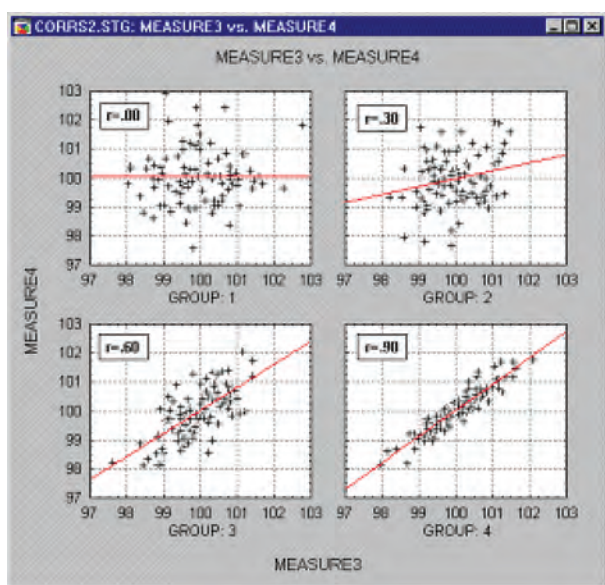


Рис. 1. Линейная корреляция Пирсона.

Предполагается, что обе переменные измерены в одинаковой шкале, хотя бы интервальной, что может определять степень «пропорциональности» двух переменных друг другу. Отметим, что значение корреляции не зависит от используемых конкретных единиц измерения, например, соотношения роста и веса. Коэффициент корреляции будет одинаковым независимо от того, сантиметры, метры, килограммы или граммы используются в качестве единиц измерения. Качественной характеристикой тесноты связи служит размытость облака точек вокруг линии корреляции. Более тесной связи соответствует более тесная группировка данных вокруг линии. Прямая пропорциональная зависимость означает, что линейная зависимость является статистически значимой, положительной и прямая линия корреляции направлена вверх. Это значит, что с ростом одной переменной растет и вторая. Обратная пропорциональная зависимость означает, что линейная зависимость является статистически значимой, отрицательной, прямая линия корреляции направлена вниз и с ростом одной переменной убывает вторая.

2. ОСНОВНЫЕ СВОЙСТВА КОРРЕЛЯЦИИ [1, 2, 3, 6]

1. Коэффициент парной корреляции вычисляется для количественных признаков, представленных в метрической шкале, как минимум интервальной.
2. Коэффициент корреляции симметричен, т.е. не изменяется, если X и Y поменять местами.
3. Коэффициент корреляции является величиной безразмерной.
4. Коэффициент корреляции не изменяется при изменении единиц измерения признаков X и Y .

Содержательная интерпретация коэффициента корреляции вытекает из определения свойств корреляции и заключается в описании одно направленности (разнонаправленности) и тесноты связи (пропорциональности). Однако его квадрат – R^2 , называемый коэффициентом детерминации, выражается в долях единицы или в % и показывает насколько изменения зависимой переменной объясняются изменениями независимой, или в статистических терминах, коэффициент детерминации описывает долю дисперсии (разброса) одной переменной, согласованно изменяющейся с дисперсией (разбросом) другой переменной.

Если две переменные функционально линейно зависимы (точки на диаграмме рассеяния лежат на одной прямой), то можно сказать, что изменение одной из них полностью объясняется изменением другой, а это как раз тот случай, когда коэффициент детерминации равен 100 % (при этом коэффициент корреляции может быть равен как 1, так и -1).

Если две переменные линейно независимы (метод наименьших квадратов дает горизонтальную прямую), то распределения переменных независимы. Или в более простых терминах ни одна из переменных в своих изменениях никоим образом не определяет другую. В остальных случаях коэффициент детерминации указывает, какая часть дисперсии (изменений) одной переменной объясняется дисперсией другой переменной, т.е. насколько изменения являются согласованными.

Чем выше по модулю (по абсолютной величине) значение коэффициента корреляции, тем сильнее связь между признаками. Принято считать (это верно для выборок достаточно больших размеров при наличии существенной связи), что коэффициенты корреляции, которые по модулю больше 0,7, говорят о сильной связи (при этом коэффициенты детерминации > 50 %), т.е. изменения одной переменной связаны с дисперсией (изменениями) другой переменной более, чем наполовину. Коэффициенты корреляции, которые по модулю меньше 0,7, но больше 0,5 характеризуют связи средней силы (при этом коэффициенты детерминации меньше 50 %, но больше 25 %). Коэффициенты корреляции меньше 0,5 характеризуют слабые связи, в этом случае коэффициенты детерминации меньше 25 %.

В пакете Statistica вычисляется матрица корреляции для заданного набора переменных, у которой число строк и столбцов равно числу переменных, а на пересечении каждой строки и столбца стоит коэффициент корреляции соответствующей пары признаков. Матрица корреляции является симметричной, т.е. значения выше и ниже диагонали равны между собой.

Таблица 1

Пример матрицы корреляций для шести переменных, описывающих взаимосвязи гормонов у женщин при первичном бесплодии

Переменная	Корреляции значимые на уровне $p = 0,05$, $N = 53$ (Построчное удаление ПД)					
	кортизол	ТТГ	Т3	Т3св	Т4	Т4св
Кортизол	1,00	0,16	0,29	0,25	0,35	0,15
ТТГ	0,16	1,00	0,08	0,31	-0,27	-0,08
Т3	0,29	0,08	1,00	0,23	0,14	0,15
Т3св	0,25	0,31	0,23	1,00	0,05	0,22
Т4	0,35	-0,27	0,14	0,05	1,00	0,47
Т4св	0,15	-0,08	0,15	0,22	0,47	1,00

В таблице 1 приведен пример матрицы корреляций для шести переменных, описывающих взаимосвязи гормонов у женщин при первичном бесплодии. В этой матрице 36 коэффициентов корреляции. По диагонали 6 коэффициентов равных 1,00 – это корреляция каждой переменной сама с собой и по пять коэффициентов по парной корреляции с другими переменными, расположенными выше и ниже диагонали. Статистически значимые на 5 % уровне корреляции в матрице отмечены красным цветом.

Из таблицы видно, что статистически значимых связей всего 5 и все они являются слабыми, их значения меньше величины 0,5, согласованного изменения дисперсий менее 25 % (слабая связь). Связь ТТГ и Т4 – отрицательная, с ростом одной убывает другая. Остальные связи – положительные, что характеризует их однонаправленное изменение: с ростом одной слабо растет и другая.

Не всякий коэффициент корреляции является статистически значимым. В чем смысл понятия статистической значимости коэффициента корреляции? Как правило, при проведении исследований коэффициент корреляции вычисляется на основе выборочной группы с определенным числом наблюдений, выражающих взаимосвязи между переменными наблюдаемой группы, а не генеральной совокупности. Конечной целью исследований является перенос выборочных зависимостей на генеральную совокупность. Требуется оценить корреляцию в генеральной совокупности по значению выборочного коэффициента корреляции. Например, получено ненулевое значение выборочного коэффициента корреляции. Является ли оно отражением взаимосвязи между признаками генеральной совокупности или это только специфика данной выборки?

Для оценки значения коэффициента корреляции в генеральной совокупности по выборочным данным необходимо построить 95% доверительный интервал, основанный на ошибке коэффициента корреляции, умноженной на величину параметра t , зависящего от числа единиц наблюдения в выборочных группах (числа степеней свободы). Коэффициент корреляции генеральной совокупности будет с 95% вероятностью лежать в построенном доверительном интервале значения выборочного коэффициента корреляции. Затем необходимо провести проверку,

попадет ли нулевое значение в этот доверительный интервал. Если ноль не попадет в доверительный интервал, то это означает, что с 95 % долей вероятности в генеральной совокупности не может быть нулевого значения коэффициента корреляции, т.е. связь между признаками существует и в генеральной совокупности. В таком случае коэффициент корреляции является статистически значимым.

Если ноль попадет в вычисленный доверительный интервал, то с высокой вероятностью (95 % долей) в генеральной совокупности может оказаться нулевая корреляция, т.е. отсутствие связи. В этом случае коэффициент корреляции является статистически незначимым.

Еще раз подчеркнем, что величина коэффициента корреляции еще не гарантирует его значимости: даже большой по абсолютной величине коэффициент может оказаться статистически незначимым при малом объеме выборки, а небольшой коэффициент при большом объеме выборки – значимым.

Как упоминалось ранее, корреляция представляет собой линейную зависимость между распределениями двух переменных. Коэффициент корреляции – это мера прямой или обратной пропорциональности между распределениями двух переменных. Он чувствителен к связи только в том случае, если эта связь является монотонной, т.е. не меняет направления по мере увеличения значений одной из переменных.

Основными показателями корреляции являются сила, направление и надежность (достоверность) связи. Сила связи определяется по абсолютной величине корреляции (меняется от 0 до 1). Направление связи определяется по знаку коэффициента корреляции: положительный коэффициент – связь прямая; отрицательный – связь обратная. Надежность связи определяется p -уровнем статистической значимости (чем меньше p -уровень, тем выше статистическая значимость или достоверность связи).

3. УСЛОВИЯ ПРИМЕНЕНИЯ ПАРНОГО КОРРЕЛЯЦИОННОГО АНАЛИЗА [1, 2, 3, 6]

1. Переменные измерены в количественной (интервальной или метрической) шкале на одной и той же выборке объектов (так называемая внутригрупповая характеристика).

2. Связь между переменными является монотонной (не меняет направления по мере увеличения значений одной из переменных).

3. Распределения изучаемых переменных – нормальное, либо близкое к нормальному.

Основная проверяемая статистическая гипотеза в отношении коэффициентов корреляции является ненаправленной и содержит утверждение о равенстве корреляции нулю в генеральной совокупности $H_0: R_{xy} = 0$. При отклонении нулевой гипотезы принимается альтернативная гипотеза $H_1: R_{xy} \neq 0$ о наличии положительной, либо отрицательной корреляции в зависимости от знака вычисленного коэффициента корреляции.

4. ИНТЕРПРЕТАЦИЯ ЗНАЧЕНИЯ КОРРЕЛЯЦИЙ [1, 2, 3, 6]

Если по результатам статистической проверки $H_0: R = 0$ не отклоняется на уровне α , то содержательный вывод: связь между переменными x и y не обнаружена. Если $H_0: R = 0$ отклоняется на уровне α , то содержательный вывод: обнаружена положительная (отрицательная) связь между переменными x и y .

Что влияет на p -уровень значимости корреляции? Статистическая значимость коэффициента корреляции тем выше, чем больше его абсолютная величина (при одном и том же объеме выборки), и чем больше объем выборки (при одном и том же значении корреляции), и определяется значением p , которое должно быть меньше выбранного уровня (для медико-биологических исследований $\alpha = 0,05$). При большой численности выборки даже слабые связи могут стать статистически значимыми.

Например, при различных численностях выборки ($n_1 < 90, n_2 > 100$) для одного и того же значения $R_{xy} = 0,200$ возможны две различные ситуации:

1. при $n_1 < 90, p > 0,05$ – корреляция статистически не значима;
2. при $n_2 > 100$ значение $p < 0,05$ – связь статистически достоверна.

Величина корреляции не всегда отражает силу связи. Соответственно, p -уровень значимости не всегда отражает надежность связи. Наиболее распространенные причины – «выбросы», «ложные» корреляции, нелинейные связи.

Если коэффициент корреляции возвести в квадрат, то полученное значение R^2 представляет собой часть общей вариации двух переменных. Можно при значимой корреляции величину R^2 трактовать как часть согласованных изменений наблюдаемых переменных.

Однако выявленная корреляция между переменными x и y не может интерпретироваться как безусловная причинно-следственная связь. Возможных вариантов при наличии статистически значимой корреляции – 3.

1. Переменная x влияет на y .
2. Переменная y влияет на x .
3. Переменные x и y находятся под опосредованным влиянием некоторой третьей переменной.

Эти три ситуации следует очень тщательно анализировать при трактовке наблюдаемых переменных в зависимости от изучаемого явления.

5. ДОСТОВЕРНОСТЬ КОРРЕЛЯЦИЙ [1, 2, 3, 6]

Какие по численности выборки можно корректно изучать с применением корреляционного анализа?

Как выше показали, что значимость коэффициента корреляции меняется в зависимости от размера выборки. Кроме того существенным условием для заключения о значимости корреляции является предположение, что распределение остатков (отклонения от линии регрессии) для зависимой переменной y следует нормальному распределению, а также, что изменчивость остатков является одинаковой для всех значений независимой переменной x . Тем не менее, исследования выполненные по методу Монте-Карло, показывают, что эти предположения не всегда являются обязательными для выборок достаточно большого размера и, когда отклонения от нормальности не очень большие. Невозможно сформулировать универсальные жесткие рекомендации о численности выборок. Однако результаты, полученные методами статистического моделирования Монте-Карло, выполненные независимо многими исследователями, позволили сформулировать следующее правило.

Если размер выборки составляет 94 или более единиц наблюдений, то небольшими отклонениями от нормальности можно пренебречь. Но надо помнить, что численность в 94 единицы (как граничного

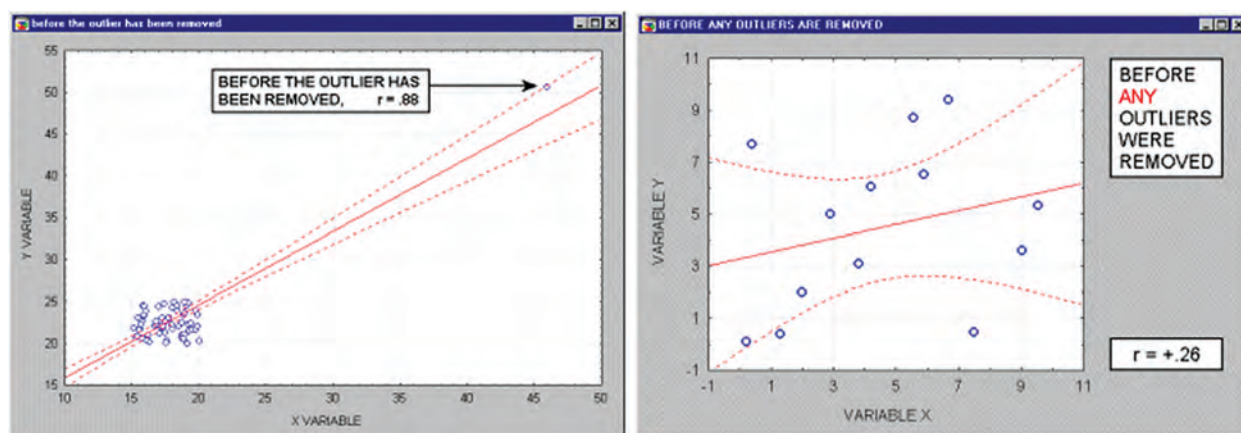


Рис. 2. Пример «выброса».

значения) может рассматриваться только в случае отсутствия «выбросов» и отсутствия нескольких «горбов».

Очень серьезной причиной, приводящей к нарушению достоверности корреляционных связей, являются «выбросы» (или «выскакивающие значения»). Выбросами являются нетипичные для данной задачи исследования, редкие наблюдения, далеко отстоящие от других наблюдаемых значений в группе. В силу того, что близость расположения точки определяется не простым расстоянием, а квадратом его удаленности (чтобы элиминировать знак удаленности: больше или меньше), выбросы оказывают большое влияние на наклон линии регрессии и, следовательно, на значение коэффициента корреляции. Один выброс способен значительно изменить наклон линии регрессии и, следовательно, значение корреляции. Очень наглядным является пример, приведенный в документации к пакету Statistica, (рис. 2). Еще раз отметим, что даже один выброс может быть полностью ответственным за высокие значения корреляции, которая без выброса будет близка к нулю.

Заметим, что влияние «выбросов» особенно ощутимо в выборках малой численности наблюдений. В приведенном примере с исключением одного значения «выброса» видно, что корреляция изменилась со значения 0,88 до значения 0,26, что является очень значимым изменением.

Принято считать, что выбросы представляют собой случайные ошибки, которые исследователь не в состоянии контролировать. Одним из общепринятых методов удаления выбросов является изучения рассеяния единиц наблюдения при построении корреляции. Заметим, что выбросы могут не только искусственно увеличивать значение коэффициента корреляции, но и также снижать уровень «реальных» корреляций.

Таким образом, вышеприведенное иллюстрирует классическое правило [1, 2, 3, 6], которое рекомендуется запомнить и ему следовать: никогда не следует формулировать важные выводы, основываясь только на значении коэффициента корреляции.

К сожалению, определение выброса несет в себе некоторую степень субъективности, и решение принимается на основе опыта и знаний исследователя. Следует также отметить, что в случаях, когда выбросов выявляется достаточно много, это может свидетельствовать о наличии некоторого другого явления и следует выделить выбросы в отдельную подгруппу для другого анализа, не входящего в анализ первоначальной группы. Наличие многих выскакивающих значений может свидетельствовать о наличии неоднородных групп, т.е. наличии некоторого фактора, смещающего корреляционную зависимость. Для медицинских исследований типовым может быть пример, когда в одну группу свели больных с различными патогенетическими механизмами, имеющими различные зависимости между переменными. Так на рис. 3 приведены взаимосвязи пролактина с показателем ФСГ (фолликулостимулирующим гормоном) в подгруппах женщин с первичным и вторичным бесплодием. На рисунке видно, что взаимосвязи

переменных в каждой из подгрупп образуют свои подмножества рассеяний («облаков»), которые, по-видимому, следует изучать отдельно.

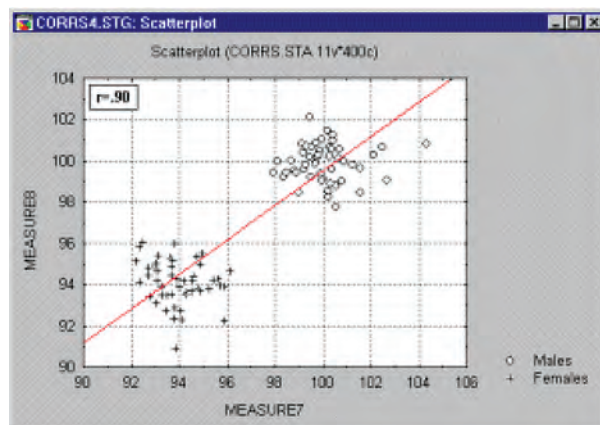


Рис. 3. Взаимосвязи пролактина с показателем ФСГ (фолликулостимулирующим гормоном) в подгруппах женщин с первичным и вторичным бесплодием.

Итак, перед началом проведения корреляционного анализа необходимо проведение тщательного изучения диаграмм рассеяния изучаемых переменных. По крайней мере, этот шаг позволит выявить выбросы и наличие нелинейных отношений между переменными в различных диапазонах их изменения.

6. ВЫЯВЛЕНИЕ ЛОЖНЫХ КОРРЕЛЯЦИОННЫХ СВЯЗЕЙ

Особую проблему корреляционного анализа представляют ложные корреляции, обусловленные, в большей степени, наличием некоей третьей переменной – z , с которой выявлены корреляции изучаемых двух переменных [1, 2, 3, 6].

Учитывая, что методы корреляционного анализа не являются доказательством причинно-следственных связей, но если связь выявлена, то обосновать статистическую истинность или ложность этой связи необходимо. В ряду этого доказательства особое место принадлежит методу выявления ложных корреляций, возникающих при опосредованном влиянии третьей переменной.

Основная проблема с выявлением ложной корреляцией заключается в том, что не всегда известна эта третья переменная, из-за которой и существует ложная корреляция. Однако в тех случаях, когда известна эта третья переменная z , применяется метод частной корреляции, обосновывающий наличие взаимосвязей между x и y .

Итак, назначение частной корреляции заключается в обосновании факта, что связь между переменными x и y зависит от z , либо не зависит от z .

Нулевой гипотезой является равенство частной корреляции нулю в генеральной совокупности, альтернативная гипотеза заключается в утверждении наличия достоверной корреляционной связи между x и y , не зависящей от z .

Обозначим $R(x, y)$ – коэффициент корреляции между переменными x и y , $R(x, y, z)$ – коэффициент частной корреляции между переменными x, y и пере-

менной z . Если $R(x, y-z)$ близка к $R(x, y)$, то связь между x и y не зависит от z . Если $R(x, y-z)$ близка к 0, то связь между x и y обусловлена влиянием z .

Проверка значимости проводится с помощью t -критерия Стьюдента. Если величина $p < 0,05$, то связь между x и y значима; при $p > 0,05$ связь между x и y не выявлена.

ЛИТЕРАТУРА

1. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. Пер. с англ. – М.: Мир, 1982. – 488 с.

2. Боровиков В. Statistica: искусство анализа данных на компьютере. Для профессионалов. – СПб.: Питер, 2001. – 656 с.

3. Закс Л. Статистическое оценивание. Пер. с нем. В.Н. Варыгина / под ред. Ю.П. Адлера, В.Г. Горского. – М.: Статистика, 1976. – 598 с.

4. Ильин В.П. Методические особенности применения статистических непараметрических методов в анализе медико-биологических данных // Бюлл. ВСНЦ СО РАМН. – 2011. – № 5 (81). – С. 157–160.

5. Ильин В.П. Методические особенности применения t -критерия Стьюдента в медико-биологических исследованиях // Бюлл. ВСНЦ СО РАМН. – 2011. – № 5 (81). – С. 160–162.

6. Халафян А.А. Statistica 6. Статистический анализ данных. 3-е изд. Учебник. – М.: ООО «Бином-Пресс», 2008. – 512 с.

Сведения об авторах

Ильин Владимир Петрович – доктор биологических наук, профессор, ведущий научный сотрудник лаборатории физиологии и патологии эндокринной системы ФГБУ «НЦ проблем здоровья семьи и репродукции человека» СО РАМН (664003, г. Иркутск, ул. Тимирязева, 16; тел.: (3952) 20-76-36; e-mail: iphr@sbamsr.irk.ru)