УДК 519.233.5

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

© 2008 г. **А. М. Гржибовский**
Национальный институт общественного здоровья, г. Осло, Норвегия

В статье рассматривается применение корреляционного анализа с расчетом коэффициентов корреляции Пирсона, Спирмена и Кендалла с использованием пакета статистических программ SPSS. Изложенный материал дает общие сведения об оценке степени тесноты взаимосвязи между переменными и призван вызвать интерес читателей журнала к прочтению специализированной литературы перед началом работы над будущими публикациями.
Ключевые слова: корреляционный анализ, коэффициенты корреляции, SPSS.

По мнению некоторых исследователей, в медицинской научной печати слишком часто используется корреляционный анализ с представлением тех или иных коэффициентов корреляции без достаточного разъяснения, что они означают [21]. В некоторых российских биомедицинских изданиях корреляционный анализ занимает второе место по частоте встречаемости после критерия Стьюдента [4], однако аргументация применения этого вида анализа и интерпретация результатов, за исключением констатации факта установления сильной или слабой корреляционной связи, встречается крайне редко [3]. В данной статье будет кратко представлен корреляционный анализ для двух переменных с расчетом коэффициентов корреляции Пирсона (Pearson), Спирмена (Spearman) и Кендалла (Kendall) с использованием пакета статистических программ SPSS. Для демонстрации расчетов коэффициентов корреляции в SPSS будет использоваться фрагмент данных Северодвинского когортного исследования [14]. Для примера отобраны только дети первородящих женщин, рожденные в срок, от одноплодных беременностей. Это ограничение пригодится при решении вопроса о распределении. Файл «Human_Ecology_2008_9.sav» можно скачать с сайта журнала «Экология человека»: www.nsmu.ru/nauka_sgmu/rio/eco_human. Переменные «id», «vozrast», «srok», «pol», «dlina», «ves» обозначают идентификационный номер участниц исследования, возраст (полных лет), гестационный возраст, пол ребенка, длину и массу тела ребенка при рождении соответственно.

Термин «корреляция» был впервые применен Ж. Кювье в 1806 году. Математическое обоснование метода предложено О. Браве в 1846 году, а применимо к биомедицинским исследованиям (речь идет только о коэффициенте корреляции Пирсона) – Ф. Гальтоном в 1886-м [6]. Коэффициент корреляции Пирсона обозначается как r, Спирмена – как ρ (греческая строчная буква «ро») или r_s, а Кендалла – как τ (греческая строчная буква тау). Различные коэффициенты оценивают силу статистической взаимосвязи между признаками по-разному, следовательно, интерпретировать их следует тоже по-разному. Так, например, коэффициенты корреляции Пирсона, Спирмена и Кендалла, все равные, скажем, 0,5, означают вовсе не одно и то же.

Корреляционный анализ встречается в отечественной биомедицинской литературе чаще, чем в зарубежной, вероятно, из-за его кажущейся простоты, однако целесообразность его применения во многих случаях остается сомнительной. Представление результатов далеко не всегда является корректным, а интерпретация их довольно часто ошибочно включает сообщения о причинно-следственных связях

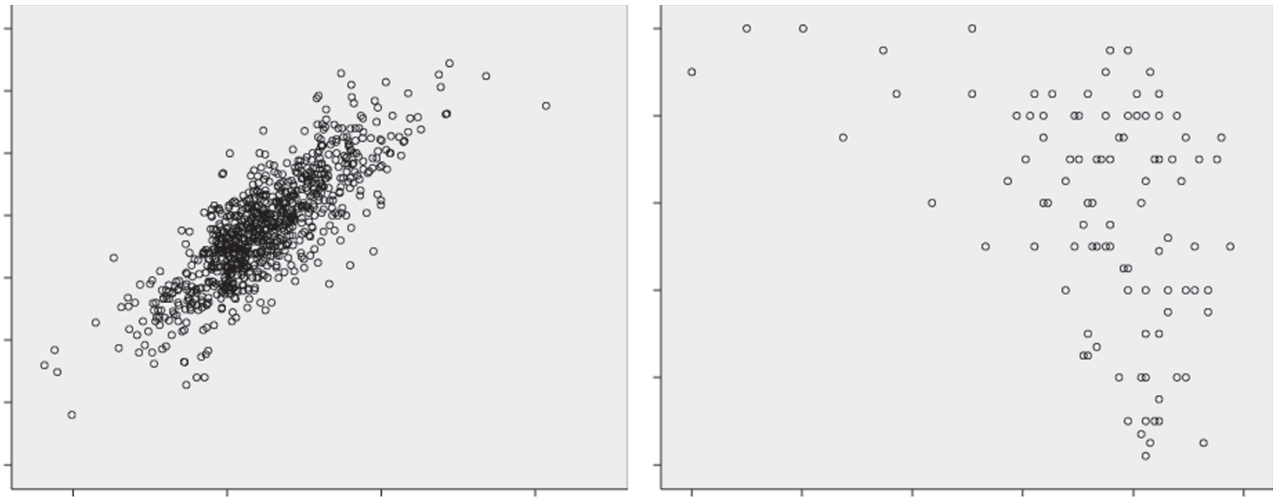

Рис. 1. Вид скаттерограммы при соблюдении (слева) и несоблюдении (справа) условия гомоскедастичности

и даже обнаружении «достоверных различий между группами».

На самом деле корреляционный анализ позволяет определить только силу и направление взаимосвязи между переменными.

Коэффициент корреляции Пирсона используется наиболее часто, хотя его следует применять только при соблюдении следующих условий:

• Обе переменные являются количественными и непрерывными

• Как минимум один из признаков (а лучше оба) имеет нормальное распределение (поэтому расчет этого коэффициента является параметрическим методом оценки взаимосвязи признаков)

• Зависимость между переменными носит линейный характер

• Гомоскедастичность (вариабельность одной переменной не зависит от значений другой переменной)

• Независимость участников исследования друг от друга (признаки X и Y у одного участника исследования независимы от признаков X и Y у другого)

• Парность наблюдений (признак X и признак Y изучаются у одних и тех же участников исследования)

• Достаточно большой объем выборки, как минимум 25 наблюдений [12]

• Для адекватной проекции расчетов на генеральную совокупность выборка должна быть репрезентативной.

Таким образом, перед принятием решения о применении коэффициента корреляции Пирсона исследователям необходимо знать тип данных; распределение изучаемых признаков в генеральной совокупности (популяции), а если это неизвестно, то проверить распределение обеих переменных в выборке; построить скаттерограммы (графики разброса) для того, чтобы убедиться в том, что связь между переменными носит линейный характер, а также чтобы проверить условие гомоскедастичности (рис. 1). При соблюдении этого условия разброс данных переменной Y будет приблизительно одинаковым для всех значений переменной X. Если вариабельность переменной Y меняется в зависимости от значений переменной X (скаттерограмма имеет вид треугольника, трапеции и т. п.), то тогда коэффициент корреляции Пирсона не будет должным образом отражать взаимосвязи между переменными. В правой скаттерограмме на рис. 1 видно, что разброс значений переменной, отложенной по оси ординат, увеличивается по мере увеличения значений переменной, отложенной по оси абсцисс. Два последних необходимых условия применения коэффициента (достаточный объем и репрезентативность выборки) должны приниматься во внимание еще на этапе планирования исследования.

Для построения скаттерограммы в SPSS следует выбрать в выпадающем меню «Graph» меню «Interactive», в нем выбрать «Scatterplot», как показано на рис. 2. В результате появится окно «Create Scatterplot», в котором предлагается переменные переместить из левого поля в поля, располагающиеся около системы координат в правой части окна. Для нашего примера будем на оси абсцисс откладывать значения длины новорожденных, а по оси ординат значения их массы тела (рис. 3).

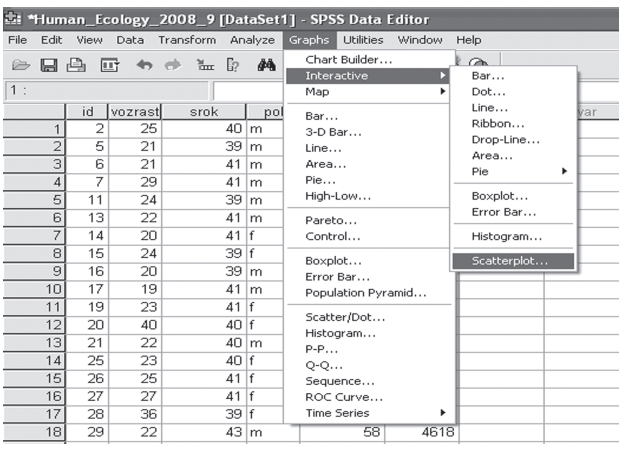

Рис. 2. Окно «SPSS Data Editor» и выбор меню для построения скаттерограмм

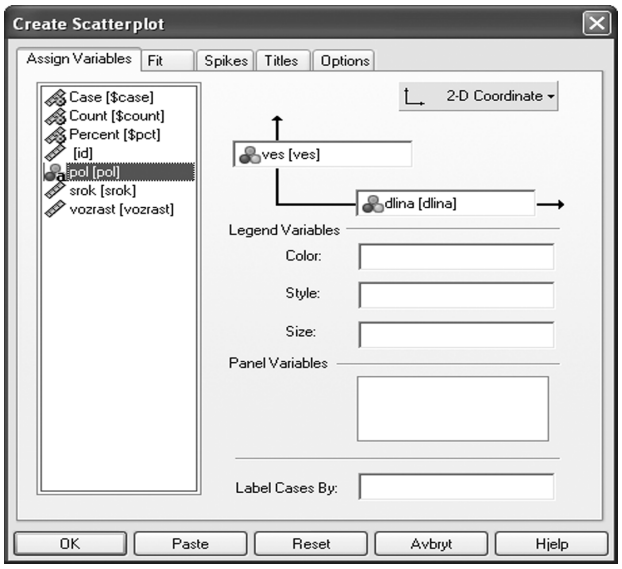

Рис. 3. Диалоговое окно «Create Scatterplot»

SPSS также дает возможность построить несколько скаттерограмм одновременно. Например, если бы мы хотели посмотреть взаимосвязь между длиной и весом новорожденных раздельно для мальчиков и девочек, то можно было бы перенести переменную «pol» в поле «Panel Variables», в результате чего SPSS создала бы две скаттерограммы – одну для мальчиков, другую для девочек. Если же необходимо представить обе скаттерограммы в одной системе координат, то группировочную переменную (pol) следует поместить в одно из полей в области «Legend Variables».

При помещении группировочной переменной в поле «Color» скаттерограммы для мальчиков и девочек будут построены разными цветами; при помещении переменной «pol» в поле «Style» условные обозначения для обоих полов будут различными (по умолчанию кружки и треугольники). Помещение переменной «pol» в поле «Size» приведет к тому, что обозначения для мальчиков и девочек будут разных размеров, что, однако, не очень удобно при визуальной оценке результатов. Помимо меню «Interactive» скаттерограммы можно строить, используя меню «Scatter/Dot» (рис. 4), которое можно найти в выпадающем меню «Graphs».

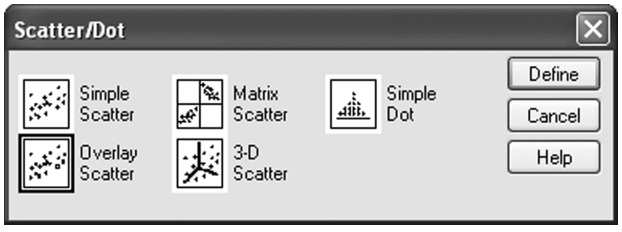

Рис. 4. Окно «Scatter/Dot»

Для построения простой скаттерограммы следует выбрать «Simple Scatter», после чего появится окно «Simple Scatterplot» (рис. 5), в котором также можно переместить интересующие нас переменные из левого поля в одно из правых в зависимости от поставленной задачи. На рис. 5 показано, как выбрать переменные для построения скаттерограммы с длиной новорожденных, отложенной на оси абсцисс, и массой тела – на оси ординат, причем в пределах одной системы координат разными цветами будут показаны скаттерограммы для мальчиков и девочек.

О построении других типов скаттерограмм в SPSS можно прочитать в специальных пособиях по использованию SPSS [1, 7, 8].

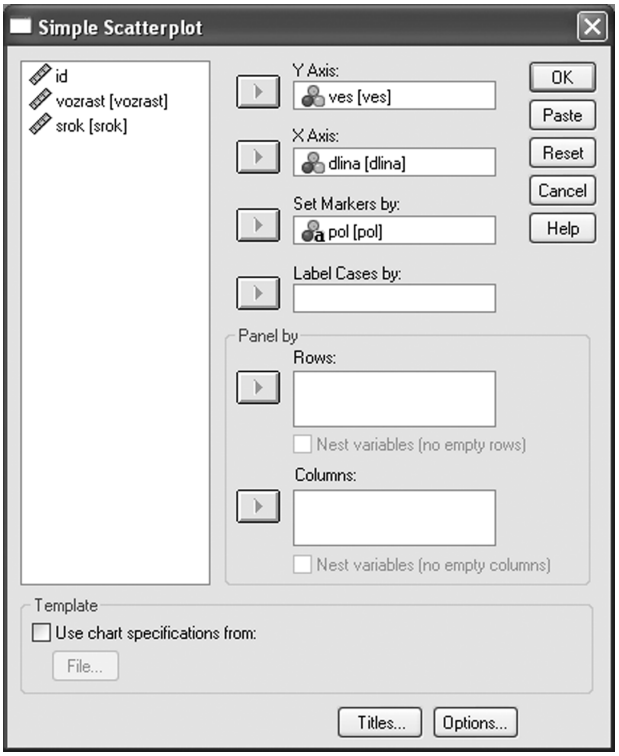

Рис. 5. Окно «Simple Scatterplot»

Для нашего примера при построении скаттерограммы четко видно, что зависимость носит линейный характер (дети обоих полов анализировались вместе), причем условие гомоскедастичности соблюдается, так как вариабельность массы тела новорожденных приблизительно одинакова для всех значений длины (рис. 6). Известно, что и длина, и масса тела новорожденных, родившихся в срок от одноплодных беременностей, в генеральной совокупности имеют нормальное распределение. В данное исследование каждая женщина была включена только один раз, то есть наблюдения можно с достаточной долей уверенности считать независимыми. Объем выборки составляет 869 человек; выборка является достаточно репрезентативной, так как исследование имело сплошной характер, то есть в него включались практически все беременные г. Северодвинска, вставшие на учет в женские консультации в 1999 году [14]. Таким образом, условия для применения коэффициента корреляции Пирсона соблюдены.

В ситуациях, когда все условия для применения коэффициента корреляции Пирсона соблюдаются, этот метод является наиболее подходящим для определе-

ния корреляционной зависимости между изучаемыми признаками. Однако если условия не соблюдаются, коэффициент корреляции Пирсона может дать искаженные результаты, а потому в таких ситуациях следует применять непараметрические коэффициенты корреляции (Спирмена или Кендалла).

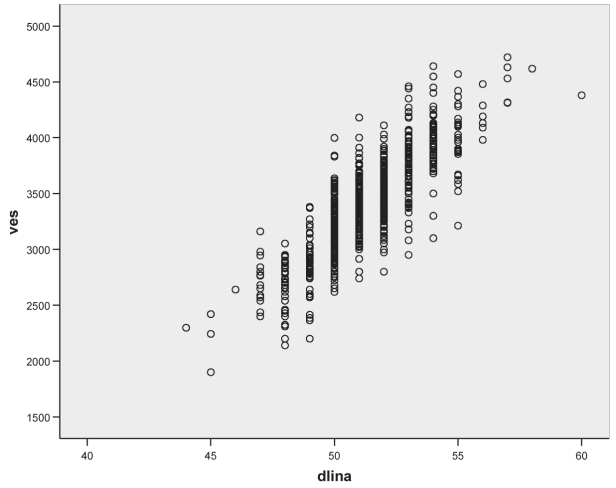

Рис 6. График корреляционной взаимосвязи между длиной и массой тела новорожденных в г. Северодвинске

Технические моменты расчета коэффициента корреляции Пирсона подробно описаны в отечественной литературе, например в [6], поэтому остановимся только на его применении с помощью SPSS, на интерпретации полученных значений, а также некоторых наиболее часто встречающихся ошибках, связанных с применением данного коэффициента.

Для проведения корреляционного анализа нужно в меню «Analyze» выбрать меню «Correlate», в котором, в свою очередь, выбрать «Bivariate», как показано на рис. 7.

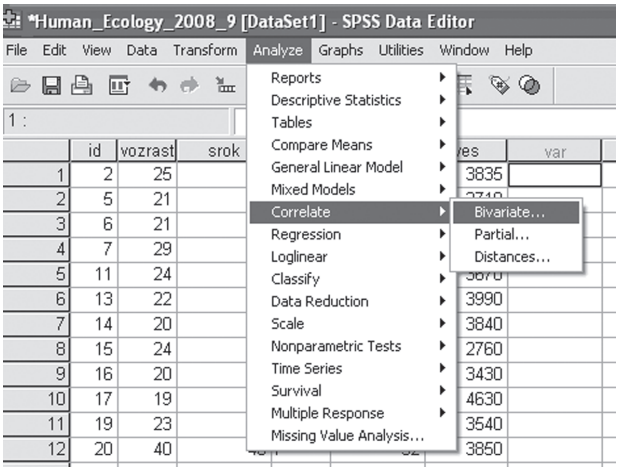

Рис. 7. «SPSS Data Editor» и выбор меню для проведения корреляционного анализа

В открывшемся диалоговом окне «Bivariate Correlations» (рис. 8) в левом поле следует выбрать две изучаемые переменные и переместить их в правое поле «Variables». Под полями для переменных следует отметить, какой из коэффициентов корреляции нужно рассчитать. По умолчанию рассчитывается только коэффициент корреляции Пирсона. Для нашего примера с целью экономии места отмечены все три рассматриваемые в статье коэффициента. По умолчанию уровень значимости рассчитывается для двухстороннего теста (отмечено Two-tailed в области Test of Significance), а статистически значимые отличия рассчитанных коэффициентов корреляции от нуля отмечаются для наглядности звездочками (отмечено Flag significant correlations).

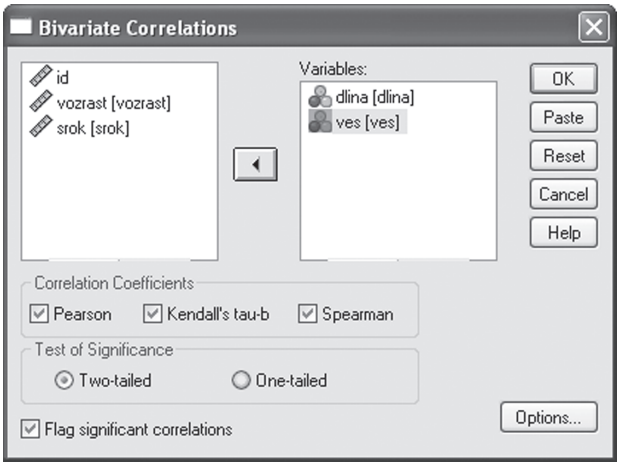

Рис. 8. Диалоговое окно «Bivariate Correlations»

Запуск расчетов осуществляется нажатием на «OK». Таблицы результатов представлены на рис. 9 и 10. Поскольку коэффициент корреляции Пирсона является параметрическим, а два других коэффициента – непараметрическими, то и представляются они SPSS в разных таблицах. Коэффициент корреляции Пирсона равен для нашего примера 0,83, причем он согласно данным второй строки таблицы (Sig. 2-tailed) статистически значимо отличается от нуля ($p < 0{,}001$). Поскольку для коэффициента корреляции не имеет значения, какая переменная является зависимой, а какая независимой (корреляция симметрична), то значения коэффициента для пары переменных «длина – масса тела» и пары «масса тела – длина» одинаковы.

Correlations

		dlina	ves
dlina	Pearson Correlation	1	,833**
	Sig. (2-tailed)		,000
	N	869	869
ves	Pearson Correlation	,833**	1
	Sig. (2-tailed)	,000	
	N	869	869

**. Correlation is significant at the 0.01 level

Рис. 9. Результаты расчетов коэффициента корреляции Пирсона в SPSS

Представляя результаты корреляционного анализа, рекомендуется показывать абсолютное значение

коэффициента корреляции, достигнутый уровень значимости и количество наблюдений, на основании которых был получен данный коэффициент. Для нашего примера: r = 0,83, p < 0,001, n = 869. Если задачей исследования ставится определение генерального параметра (коэффициента корреляции для генеральной совокупности), то необходимо представить доверительный интервал для полученного коэффицента.

Correlations

			dlina	ves
Kendall's tau_b	dlina	Correlation Coefficient	1,000	,681**
		Sig. (2-tailed)	.	,000
		N	869	869
	ves	Correlation Coefficient	,681**	1,000
		Sig. (2-tailed)	,000	.
		N	869	869
Spearman's rho	dlina	Correlation Coefficient	1,000	,821**
		Sig. (2-tailed)	.	,000
		N	869	869
	ves	Correlation Coefficient	,821**	1,000
		Sig. (2-tailed)	,000	.
		N	869	869

**. Correlation is significant at the 0.01 level (2-tailed).

Рис. 10. Результаты расчетов коэффицентов корреляции Кендалла и Спирмена в SPSS

SPSS не рассчитывает доверительные интервалы для коэффициентов корреляции, однако это не должно считаться поводом для их игнорирования, так как интервальная оценка любого генерального параметра всегда более информативна, чем точечная. Доверительные интервалы для коэффициента корреляции Пирсона можно рассчитать, используя онлайн-калькулятор на веб-странице http://faculty.vassar.edu/lowry/rho.html (рис. 11). В верхнее поле следует ввести рассчитанный коэффициент корреляции Пирсона (r), а в нижнее — объем выборки (n). В качестве разделительного знака используется точка, а не запятая.

Рис. 11. Внешний вид онлайн-калькулятора (http://faculty.vassar.edu/lowry/rho.html) для расчета доверительных интервалов для коэффициента корреляции Пирсона

В основе расчетов лежит z-преобразование Фишера. Нижняя (Z_L) и верхняя (Z_U) границы преобразованного 95 % доверительного интервала для коэффициента корреляции Пирсона будут равны:

$$Z_L = 0{,}5 \cdot \ln\left(\frac{1+r}{1-r}\right) - \frac{1{,}96}{\sqrt{n-3}} \quad \text{и}$$

$$Z_U = 0{,}5 \cdot \ln\left(\frac{1+r}{1-r}\right) + \frac{1{,}96}{\sqrt{n-3}},$$

где ln обозначает натуральный логарифм, а n — объем выборки. Само же значение коэффициента корреляции для генеральной совокупности, рассчитанное по данным выборки, будет в 95 % случаев находиться в интервале

$$\text{от } \frac{e^2 \cdot (Z_L) - 1}{e^2 \cdot (Z_r) + 1} \quad \text{до} \quad \frac{e^2 \cdot (Z_U) - 1}{e^2 \cdot (Z_r) + 1},$$

где e — число Эйлера (e ≈ 2,7). В рассматриваемом примера коэффициент корреляции Пирсона для взаимосвязи между длиной и массой тела новорожденных в г. Северодвинске был равен 0,83 и статистически значимо отличался от 0 (p < 0,001). Применив преобразование Фишера, рассчитаем сначала Z_L и Z_U:

$$Z_L = 0{,}5 \cdot \ln\left(\frac{1+0{,}83}{1-0{,}83}\right) - \frac{1{,}96}{\sqrt{869-3}} = 1{,}12;$$

$$Z_U = 0{,}5 \cdot \ln\left(\frac{1+0{,}83}{1-0{,}83}\right) + \frac{1{,}96}{\sqrt{869-3}} = 1{,}25,$$

что соответствует следующим нижней и верхней границам 95 % доверительного интервала:

$$\frac{e^2 \cdot (1{,}12) - 1}{e^2 \cdot (1{,}12) + 1} = 0{,}81 \quad \text{и} \quad \frac{e^2 \cdot (1{,}25) - 1}{e^2 \cdot (1{,}25) + 1} = 0{,}85.$$

Использование онлайн-калькулятора на странице http://faculty.vassar.edu/lowry/rho.html дает аналогичный результат (рис. 12), причем автоматически рассчитывается не только 95 %, но и 99 % доверительный интервал для коэффициента корреляции. Для «ручных» вычислений 99 % доверительного интервала 1,96 в формуле следует заменить на 2,58.

0.95 and 0.99 Confidence Intervals of rho

	Lower Limit	Upper Limit
0.95	0.812	0.852
0.99	0.805	0.857

Рис. 12. Рассчитанные с помощью онлайн-калькулятора 95 % и 99 % доверительные интервалы для коэффициента корреляции Пирсона

Поскольку доверительные интервалы используются для оценки коэффициента для генеральной совокупности, они обозначены не как r, а как rho (не путать с ρ, с помощью которого обозначается коэффициент корреляции Спирмена).

Помимо доверительных интервалов с помощью преобразования Фишера и онлайн-калькуляторов можно рассчитать, отличается ли полученный коэффициент корреляции от известного или предполагаемого

популяционного значения коэффициента корреляции (ρ). В основе расчетов лежит формула

$$z = \frac{0{,}5 \cdot \ln\left(\frac{1+r}{1-r}\right) - 0{,}5 \cdot \ln\left(\frac{1+\rho}{1-\rho}\right)}{1/\sqrt{n-3}},$$

в которой r — значение коэффициента корреляции, рассчитанное по данным выборочной совокупности, а ρ — популяционное значение, с которым проводится сравнение. Рассчитанное значение z сравнивается с табличными значениями. Для статистически значимых различий на уровне доверительной вероятности 95 % z = 1,96. Вышеприведенная формула используется в онлайн-калькуляторе на странице http://faculty.vassar.edu/lowry/VassarStats.html. На рис. 13 представлен пример ввода данных для сравнения коэффициента корреляции из данного примера с фиксированным значением 0,8. Расчет осуществляется путем нажатия на кнопку «Calculate».

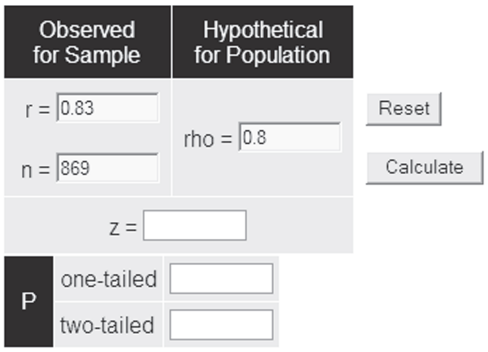

Рис. 13. Внешний вид онлайн-калькулятора (http://faculty.vassar.edu/lowry/VassarStats.html) для сравнения коэффициента корреляции Пирсона с фиксированным значением

Результаты расчетов представлены на рис. 14, из них видно, что выборочный коэффициент корреляции статистически значимо отличается от 0,8 (p = 0,009 для двустороннего теста), что неудивительно, так как рассчитанный ранее 95 % доверительный интервал (0,81–0,85) не включал в себя значение 0,8.

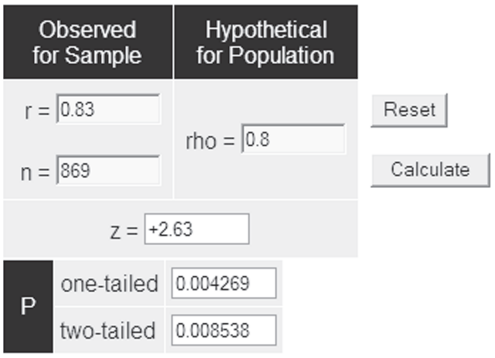

Рис. 14. Результаты сравнения выборочного коэффициента корреляции Пирсона с фиксированным значением с помощью онлайн-калькулятора (http://faculty.vassar.edu/lowry/VassarStats.html)

Описание методов и примеров сравнения коэффициентов корреляции Пирсона для двух независимых выборок (на примере оценки зависимости между индексом массы тела (ИМТ) и чувствительностью к инсулину для групп с наличием и отсутствием гипертиреоза) и для ситуаций, когда нужно сравнить степень тесноты взаимосвязи одной и той же переменной с двумя другими, представлены в [13].

Как интерпретировать коэффициент корреляции Пирсона и что он означает? Во многих учебных пособиях, например в [5], сообщается, что $r \geq 0{,}7$ говорит о наличии сильной связи между признаками, $0{,}3 < r < 0{,}7$ — о связи средней силы, $0 < r < 0{,}3$ — о слабой связи, 0 — об отсутствии линейной связи между переменными, а 1 — о наличии полной или функциональной связи между признаками. Все перечисленное выше относилось к положительной зависимости. При отрицательной корреляционной зависимости коэффициент корреляции имеет отрицательные значения, величина которых интерпретируется так же, как и для положительной зависимости. Данная классификация весьма условна, так как если для медицинских или социологических исследований коэффициент корреляции 0,8 может считаться высоким, то для некоторых исследований в области, скажем, физики такой коэффициент будет считаться очень низким. Весь спектр возможных значений находится между −1 и 1. Коэффициент корреляции Пирсона является безразмерной величиной и не зависит от единиц измерения переменных. Для понимания степени тесноты взаимосвязи между признаками лучше пользоваться коэффициентом детерминации (r^2), который, как следует из его обозначения, рассчитывается путем возведения коэффициента корреляции Пирсона во вторую степень. Коэффициент детерминации показывает, какую долю вариабельности одного из изучаемых признаков способен объяснить другой признак. Таким образом, видно, что приведенная выше классификация подразумевает под сильной связью ситуацию, когда одна из переменных способна объяснить от 49 % вариабельности другой переменной. Естественно, возникают сомнения в наличии сильных связей, если одна переменная способна объяснить лишь половину вариабельности другой. Еще один пример: коэффициент корреляции между ИМТ и систолическим артериальным давлением (САД) в некоторых странах Африки и Юго-Восточной Азии составляет в среднем 0,25 при уровне значимости $p < 0{,}01$ [23]. Из этого следует, что только 6,25 % вариабельности САД в изучаемых странах можно объяснить, зная ИМТ. Значит, на долю прочих факторов приходится 93,75 %. Кроме того, линейность зависимости между ИМТ и САД также вызывает определенные сомнения, а коэффициент корреляции Пирсона предназначен только для анализа линейных зависимостей. На рис. 15 представлены различные скаттерограммы с указанием коэффициента корреляции Пирсона, причем бросается в глаза, что на

скаттерограммах, расположенных в нижнем ряду, также имеются определенные связи между переменными, которые невозможно оценить с помощью коэффициента корреляции Пирсона.

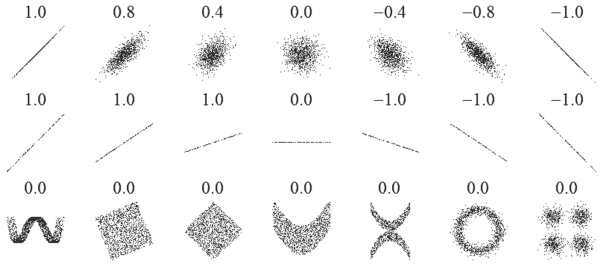

Рис. 15. Значения коэффициента корреляции Пирсона для определения взаимосвязи между двумя признаками, изображенные в виде скаттерограмм (Источник: http://upload.wikimedia.org/wikipedia/ru/3/3f/Corr-example2.png)

Еще одним стимулом для использования скаттерограмм для графического представления данных перед принятием решения о применении коэффициента корреляции Пирсона является высокая чувствительность этого коэффициента к наличию выскакивающих величин (выбросов). Так, на рис. 16 слева изображена скаттерограмма взаимосвязи двух признаков для выборки объемом 25 человек. Рассчитанный коэффициент корреляции Пирсона составил 0,9. После включения всего лишь одного «нетипичного» случая (в левом верхнем углу правой скаттерограммы) r уменьшился до 0,5. Более существенные выбросы могут полностью «уничтожить» зависимость, однако всегда следует разбираться, является ли выброс следствием ошибки регистрации данных, или же это истинные значения переменных.

Из ошибок и неточностей, встречающихся при применении коэффициента корреляции Пирсона в отечественной медицинской периодике, можно упомянуть следующие:

• Применение метода при несоблюдении необходимых условий

• Интерпретация корреляционной связи как причинно-следственной

• Расчет коэффициентов корреляции для всех пар переменных по принципу «сравним все со всем, авось что и найдем»

• Неполное представление результатов корреляционного анализа (в некоторых работах авторы сообщают в разделе «Методы» о применении корреляционного анализа, однако не удается найти даже коэффициентов корреляции)

• Представление только точечных оценок (игнорирование доверительных интервалов)

• Использование шаблонной фразы об использовании «корреляционно-регрессионного анализа» в случаях, когда использовался только корреляционный анализ

• Отождествление статистически значимых коэффициентов корреляции с клинически важными

• Отсутствие обсуждения, почему были получены те или иные коэффициенты корреляции (истинная зависимость? ложная зависимость? наличие других переменных, тесно коррелирующих с обеими изучаемыми переменными?)

• Заключения о полном отсутствии взаимосвязи между признаками при r близком к 0 при наличии нелинейной взаимосвязи

• Редкое применение скаттерограмм.

Зарубежные исследователи, проводившие анализ применения корреляционного анализа в наиболее престижных медицинских журналах США и Великобритании, наиболее часто встречающимися проблемами называли игнорирование доверительных интервалов, неполное представление результатов, а именно отсутствие данных об объеме выборки, а также придание слишком большого значения статистической значимости при оценке важности коэффициентов [19, 22].

Всегда следует помнить, что в результате корреляционного анализа невозможно установить причинно-следственные связи, поэтому выводов о том, что один

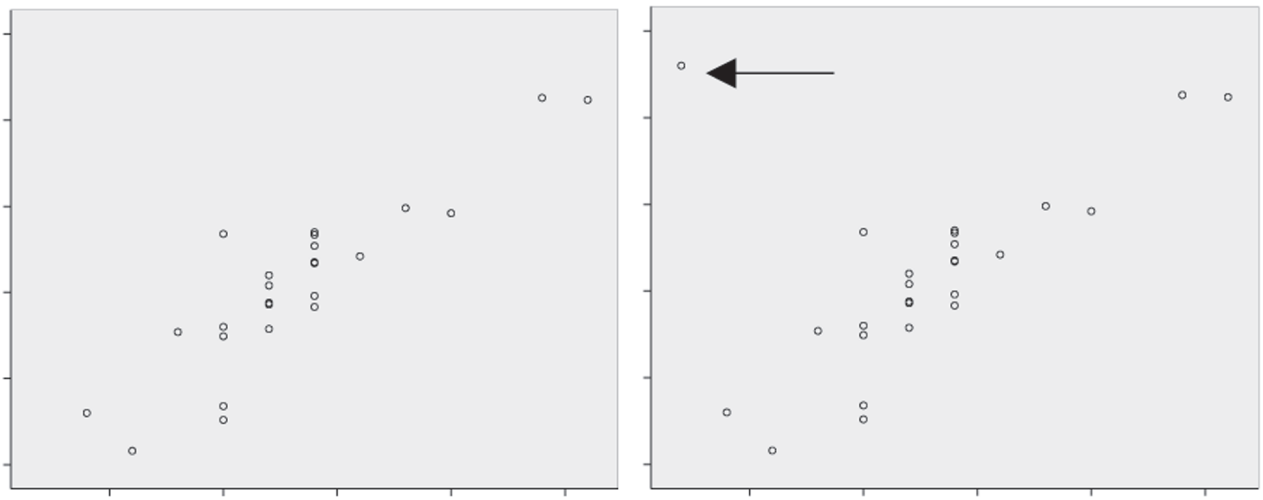

Рис. 16. Скаттерограммы для r = 0,9 (слева) и r = 0,5 (справа). Различия вызваны единственным выбросом, обозначенным стрелкой.

из изучаемых признаков вызывает другой лишь на основании корреляционного анализа, делать нельзя. Установленные корреляционные связи являются лишь статистическими, хотя некоторые из них могут быть и функциональными. В одном из часто используемых в качестве примера исследований была установлена сильная положительная корреляционная взаимосвязь между количеством гнезд аистов и количеством новорожденных в Копенгагене в ранние послевоенные годы, однако этот результат сложно считать доказательством того, что детей приносят аисты [цит. по 9]. Данная взаимосвязь является лишь статистической. Статистические взаимосвязи могут быть вызваны наличием третьей переменной, которая тесно связана с обеими изучаемыми в ходе корреляционного анализа переменными. Так, например, вероятность рождения ребенка с синдромом Дауна тесно коррелирует с количеством родов у матери до настоящей беременности. Эта взаимосвязь, как нетрудно догадаться, обусловлена тем, что возраст матери тесно связан с обеими переменными, что и приводит к обнаружению корреляционной, но никак не причинно-следственной связи между переменными.

При скошенных распределениях, а также при наличии истинных выбросов (если исследователи решают их оставить для анализа) лучше использовать непараметрические коэффициенты корреляции Спирмена или Кендалла, первый из которых в зарубежной литературе применяется значительно чаще [21]. В российской биомедицинской литературе коэффициент Кендалла применяется настолько редко, что складывается впечатление, будто отечественные исследователи с ним просто незнакомы.

Для расчета обоих непараметрических коэффициентов характерно использование не исходных значений признаков, а их рангов, что позволяет применять их для распределений, отличающихся от нормального. Использование рангов также позволяет применять непараметрические коэффициенты корреляции не только для количественных, но и для порядковых (ранговых, ординальных) данных. Технические подробности расчета коэффициента корреляции Спирмена в статье не приводятся, так как они описываются практически во всех пособиях по основам статистики.

Коэффициент корреляции Спирмена также является безразмерной величиной, принимающей значения от -1 до 1. Значение 1 говорит о наличии полного совпадения между рангами изучаемых переменных, -1 – о том, что ранги полностью противоположны. При полном отсутствии взаимосвязи между рангами переменных коэффициент корреляции Спирмена будет равен 0. Возведеный в квадрат, он также называется коэффициентом детерминации, который можно обозначить как ρ^2. Его можно интерпретировать как долю вариабельности рангов одной переменной, которую можно объяснить с помощью рангов другой переменной. Данная интерпретация достаточно громоздка и не совсем понятна с практической точки зрения, поэтому, несмотря на большую популярность коэффициента Спирмена, многие авторы склонны считать его менее практичным, чем коэффициент Кендалла [10, 11, 20, 21]. Для нашего примера с длиной и массой тела новорожденных $\rho^2 = 0{,}82^2 = 0{,}67$. Поскольку распределение обеих переменных в данном примере близко к нормальному, различия между коэффициентами корреляции Пирсона и Спирмена, а также их коэффициентами детерминации незначительны. В подобных ситуациях всегда лучше применять коэффициент корреляции Пирсона, так как он обладает большей статистической мощностью и его значительно проще интерпретировать.

Как и для r, SPSS не рассчитывает доверительные интервалы для ρ, но рассчитывает уровень значимости для проверки нулевой гипотезы о равенстве коэффициента нулю. Рассчитать доверительные интервалы для ρ несложно, используя уже известное преобразование [11]. Дисперсия коэффициента корреляции Спирмена не равна таковой для коэффициента корреляции Пирсона, поэтому, несмотря на общее сходство принципа расчета доверительных интервалов для обоих коэффициентов корреляции, формулы для расчета Z_L и Z_U для 95 % доверительного интервала для коэффициента корреляции Спирмена будут отличаться:

$$Z_L = 0{,}5 \cdot \ln\left(\frac{1+\rho}{1-\rho}\right) - \frac{1{,}96 \cdot \sqrt{1+0{,}5\cdot\rho^2}}{\sqrt{n-3}};$$

$$Z_U = 0{,}5 \cdot \ln\left(\frac{1+\rho}{1-\rho}\right) + \frac{1{,}96 \cdot \sqrt{1+0{,}5\cdot\rho^2}}{\sqrt{n-3}},$$

после чего полученные значения следует подставить в уже известную формулу для расчета верхней и нижней границ 95 % доверительного интервала:

$$\text{от} \quad \frac{e^2\cdot(Z_L)-1}{e^2\cdot(Z_L)+1} \qquad \text{до} \qquad \frac{e^2\cdot(Z_U)-1}{e^2\cdot(Z_U)+1}.$$

Приведенная здесь формула не единственная, но, по мнению D. Bonnett & T. Wright [11], она является наиболее адекватной для расчета доверительного интервала для коэффициена корреляции Спирмена. Для нашего примера коэффициент корреляции Спирмена равен 0,82 (см. рис 10). Использование вышеприведеных формул дает следующие значения для вспомогательных величин Z_L и Z_U:

$$Z_L = 0{,}5 \cdot \ln\left(\frac{1+0{,}82}{1-0{,}82}\right) - \frac{1{,}96 \cdot \sqrt{1+0{,}5\cdot 0{,}82^2}}{\sqrt{869-3}} = 1{,}08,$$

$$Z_U = 0{,}5 \cdot \ln\left(\frac{1+0{,}82}{1-0{,}82}\right) + \frac{1{,}96 \cdot \sqrt{1+0{,}5\cdot 0{,}82^2}}{\sqrt{869-3}} = 1{,}23,$$

а значение коэффициента корреляции для генеральной совокупности с 95 % надежностью будет располагаться в пределах

$$\text{от} \quad \frac{e^2 \cdot (1{,}08) - 1}{e^2 \cdot (1{,}08) + 1} = 0{,}79 \quad \text{до} \quad \frac{e^2 \cdot (1{,}23) - 1}{e^2 \cdot (1{,}23) + 1} = 0{,}84 .$$

Многие авторы считают, что из непараметрических коэффициентов корреляции наиболее просто интерпретировать коэффициент корреляции Кендалла [10, 11, 20, 21]. Учитывая, что этот коэффициент реже всего представлен в отечественной биомедицинской литературе, остановимся на нем несколько подробнее.

Можно представить, что речь идет о двух участниках исследования i и j, у которых в ходе исследования изучаются признаки X и Y. Изучаемыми признаками могут, например, быть рост и масса тела, индекс массы тела и артериальное давление, и т. п. Пару наблюдений можно обозначить как X_i, Y_i и X_j, Y_j. Если разности $X_j - X_i$ и $Y_j - Y_i$ будут одинаковы по знаку (либо $X_j > X_i$ и $Y_j > Y_i$, либо $X_j < X_i$ и $Y_j < Y_i$), то пару можно считать конкордантной. Количество конкордантных пар (проверсий) обозначается как C. Если разности $X_j - X_i$ и $Y_j - Y_i$ будут по знаку различаться (либо $X_j > X_i$ и $Y_j < Y_i$, либо $X_j < X_i$ и $Y_j > Y_i$), то такая пара называется дискордантной. Количество дискордантных пар (инверсий) обозначается как D. Если выборка состоит из n участников исследования, то возможно формирование $n(n-1)/2$ пар, для которых $1 \le i < j \le n$.

Коэффициент корреляции Кендалла рассчитывается по формуле [16, 18]:

$$\tau = \frac{2 \cdot (C - D)}{n \cdot (n-1)} .$$

Данный способ расчета коэффициента не учитывает одинаковых (связанных, равных) рангов (ties) и обозначается в литературе как tau-a (τ_a). Равные ранги возникают в тех случаях, когда у нескольких участников исследования изучаемый признак имеет одно и то же значение (например, одинаковый рост). Из формулы видно, что максимально возможное значение $\tau_a = 1$ достигается только в том случае, если все пары являются конкордантными. Аналогично, если все пары являются дискордантными, τ_a принимает минимально возможное значение -1. Если количество конкордантных и дискордантных пар равно, то $\tau_a = 0$, что говорит об отсутствии взаимосвязи между изучаемыми признаками.

Если C представляет собой количество конкордантных пар из возможных в выборочной совокупности $n(n-1)/2$ пар, то оценить вероятность того, что пара наблюдений будет конкордантной (π_c), можно с помощью формулы:

$$\pi_c = \frac{2 \cdot C}{n \cdot (n-1)} .$$

Аналогично вероятность того, что пара наблюдений будет дискордантной (π_d), можно оценить с помощью формулы:

$$\pi_d = \frac{2 \cdot D}{n \cdot (n-1)} .$$

Таким образом, для любой пары наблюдений, отобранных случайно, τ_a Кендалла может интерпретироваться как разность между вероятностью того, что пара будет конкордантной, и того, что она будет дискордантной, то есть

$$\tau_a = \pi_c - \pi_d .$$

Отрицательное значение τ будет говорить: вероятность того, что любая случайно отобранная пара наблюдений с характеристиками (X_i, Y_i и X_j, Y_j) будет скорее дискордантной, чем конкордантной, и наоборот.

Кроме того, в генеральной совокупности, для которой коэффициент корреляции Кендалла равен τ_a, вероятность того, что любая случайно отобранная пара наблюдений с характеристиками (X_i, Y_i и X_j, Y_j) окажется конкордантной, будет в $(1 + \tau_a)/(1 - \tau_a)$ раза выше, чем вероятность того, что эта пара будет дискордантной. Таким образом, если в исследовании с использованием случайно отобранной репрезентативной выборки был получен коэффициент корреляции Кендалла $\tau_a = 0{,}5$, это означает, что вероятность того, что любая случайно отобранная из генеральной совокупности пара будет конкордантной, в среднем в $(1 + 0{,}5)/(1 - 0{,}5) = 3$ раза выше, чем вероятность того, что эта пара будет дискордантной.

Оригинальный расчет коэффициента корреляции Кендалла с помощью графического изображения рангов был предложен D. Hill [15]. На рис. 17 схематично представлены ранги для двух переменных для 6 участников исследования. Сплошные прямые линии соединяют ранги для двух переменных для каждого из участников исследования. Так, например, участник исследования, для которого значения рангов обозначены квадратами, имеет ранг 2 для переменной X и ранг 3 для переменной Y.

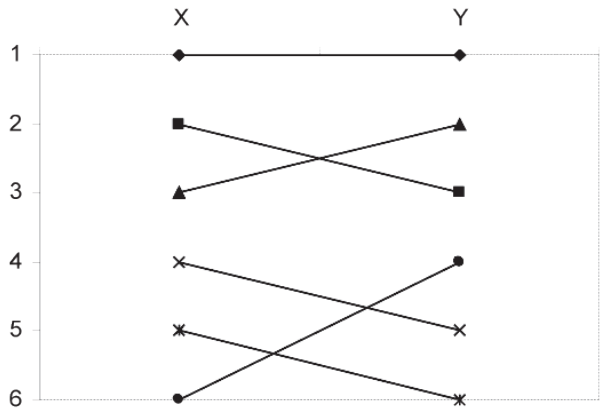

Рис. 17. Графическое представление рангов для двух изучаемых признаков (X и Y) для 6 участников исследования

Из рисунка видно, что некоторые линии, соединяющие ранги, пересекаются. Если посчитать количество пересечений и обозначить его как к, то коэффициент корреляции Кендалла можно рассчитать по формуле:

$$\tau_a = 1 - \frac{4 \cdot k}{n \cdot (n-1)},$$

где n – объем выборки. Для изображенного на рис. 17 примера количество пересечений – 3, а количество объем выборки – 6. После подставления этих значений в формулу получим $\tau_a = 0{,}6$. Более наглядное описание графического метода с примерами представлено в работе D. Wilkie [24].

Доверительный интервал для τ_a также можно рассчитать с помощью преобразования Фишера, которое дает достаточно адекватную интервальную оценку коэффициента корреляции Кендалла для генеральной совокупности при объеме выборки не менее 10 наблюдений и значении τ не более 0,8. Отличие будет заключаться в расчете вспомогательных значений Z_L и Z_U:

$$Z_L = 0{,}5 \cdot \ln\left(\frac{1+\tau}{1-\tau}\right) - \frac{1{,}96 \cdot \sqrt{0{,}437}}{\sqrt{n-4}};$$

$$Z_U = 0{,}5 \cdot \ln\left(\frac{1+\tau}{1-\tau}\right) + \frac{1{,}96 \cdot \sqrt{0{,}437}}{\sqrt{n-4}},$$

которые затем подставляют в формулу для расчета 95 % доверительного интервала:

$$\text{от} \quad \frac{e^2 \cdot (Z_L) - 1}{e^2 \cdot (Z_L) + 1} \quad \text{до} \quad \frac{e^2 \cdot (Z_U) - 1}{e^2 \cdot (Z_U) + 1}.$$

Интересно, что при соблюдении условия нормальности распределения имеется взаимосвязь между τ и r, которую можно выразить формулой $\tau = 0{,}5 \cdot \pi \cdot \sin^{-1}(r)$. Используя эту формулу для нашего примера, получим $\tau \approx 0{,}65$.

Следует помнить, что вышеописанный коэффициент корреляции Кендалла (τ_a) применяется для определения степени тесноты связи между переменными без учета равных (связанных) рангов. При наличии таковых, то есть когда два или более наблюдений по любой из переменных имеют одинаковые ранги, лучше применять другие разновидности коэффициента корреляции Кендалла, которые при расчете равные ранги учитывают. Такие ситуации неизбежны при изучении порядковых признаков, таких как, например, образование, степень тяжести заболевания и т. п.

SPSS не рассчитывает τ_a, а рассчитывает только τ_b в меню «Correlate». При отсутствии связанных рангов значения τ_a и τ_b будут равны, при небольшом их количестве – приблизительно равны, но в случаях, когда связанных рангов много, предпочтительнее использовать τ_b, так как он учитывает (связанные) ранги при расчетах. Кроме того, SPSS рассчитывает приблизительные значения стандартной ошибки для τ_b, что позволяет рассчитывать доверительные интервалы без применения громоздких формул. Приблизительная (асимптотическая) оценка стандартной ошибки τ_b не рассчитывается при использовании меню «Correlate», поэтому нужно использовать меню «Crosstabs». Краткое описание применения τ_b было описано в предыдущем номере журнала «Экология человека» [2]. Для нашего примера $\tau_b = 0{,}68$, а стандартная ошибка 0,014, значит, величина τ_b для генеральной совокупности будет с 95 % надежностью находиться в границах интервала от $0{,}68 - 1{,}96 \cdot 0{,}014$ до $0{,}68 + 1{,}96 \cdot 0{,}014$, то есть от 0,65 до 0,71. По мнению S. Arndt et al. [10], наличие связанных рангов не сильно усложняет интерпретацию значения коэффициента корреляции Кендалла, поэтому исходя из публикации S. Arndt et al. полученное значение 0,68 можно интерпретировать как вероятность того, что любая пара наблюдений будет конкордантна по изучаемым признакам, составит 0,68 или 68 %. На самом деле интерпретация τ_b более сложная, но для общего понимания принципа достаточно помнить, что все τ Кендалла показывают в том или ином виде вероятность того, что оба изучаемых признака изменяются одинаково (например, при увеличении роста увеличивается масса тела).

Помимо представленных в данной статье существует еще много других разновидностей корреляционного анализа как для количественных, так и для качественных переменных, как для двух, так и для нескольких переменных одновременно, с которыми можно ознакомиться в специализированной статистической литературе. Интересный обзор применения корреляционного анализа в медицинских исследований для различных типов переменных был недавно опубликован H. Kraemer [17]. В следующем выпуске будут кратко представлены основы линейного регрессионного анализа.

Список литературы

1. *Бююль А.* SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / А. Бююль, П. Цёфель. – Минск : ДиаСофт, 2005. – 608 с.

2. *Гржибовский А. М.* Анализ порядковых данных / А. М. Гржибовский // Экология человека. – 2008. – № 8. – С. 56–62.

3. *Гржибовский А. М.* Корреляционный анализ в медицинских исследованиях / А. М. Гржибовский // Бюллетень СГМУ. – 2000. – № 2. – С. 22–23.

4. *Гржибовский А.М.* Применение статистики в терапии: критический анализ публикаций / А. М. Гржибовский // Бюллетень СГМУ – 2000. – № 2. – С. 21–22.

5. *Зайцев В. М.* Прикладная медицинская статистика / В. М. Зайцев, В. Г. Лифляндский, В. И. Маринкин. – СПб. : Фолиант, 2003. – 428 с.

6. *Лакин Г. Ф.* Биометрия / Г. Ф. Лакин. – М. : Высшая школа, 1990. – 350 с.

7. *Наследов А. Д.* SPSS: Компьютерный анализ данных в психологии и социальных науках / А. Д. Наследов. – СПб. : Питер, 2005. – 416 с.

8. *Наследов А. Д.* SPSS 15: Профессиональный статистический анализ данных /А. Д. Наследов. – СПб. : Питер, 2007. – 416 с.

9. *Anderson M.* RSM simplified: optimizing processes using response surface methods for design of experiments / M. Anderson P., Whitcomb. – London : Taylor & Francis, 2005. – P. 39–42.

10. *Arndt S.* Correlating and predicting psychiatric symptom ratings: Spearman's r versus Kendall's tau correlation / S. Arndt, C. Turvey, N. Andreasen // Journal of Psychiatric Research. – 1999. – Vol. 33. – P. 97–104.

11. *Bonett D.* Sample size requirements for estimating Pearson, Kendall and Spearman correlations / D. Bonett, T. Wright // Psychometrica. – 2000. – Vol. 65. – P. 23–28.

12. *David* F. Tables of the ordinates and probability integral of the distribution of the correlation coefficient in small samples / F. David. – Cambridge : Cambridge University Press, 1938.

13. *Dawson B.* Basic and clinical biostatistics. Third edition / B. Dawson, R. Trapp. – Singapore : McGraw & Hill, 2001. – P. 188–189.

14. *Grjibovski A. M.* Social variations in fetal growth in Northwest Russia: an analysis of medical records / A. M. Grjibovski, L. O. Bygren, B. Svartbo, P. Magnus // Annals of Epidemiology. – 2003. – Vol. 13. – P. 599–605.

15. *Hill I.* Association football and statistical inference / I. Hill // Applied Statistics. – 1974. – Vol. 23. – P. 203–208.

16. *Kendall M.* A new method of rank correlation / M. Kendall // Biometrika. – 1938. – Vol. 30. – P. 91–93.

17. *Kraemer H.* Correlation coefficients in medical research: from product moment correlation to the odds ratio / H. Kraemer // Statistical Methods in Medical Research. – 2006. – Vol. 15. – P. 525–544.

18. *Kruskal W.* Ordinal measures of association / W. Kruskal // Journal of the American Statistical Association. – 1958. – Vol. 53. – P. 814–861.

19. *Kuo Y.* Extrapolation of correlation between 2 variables in 4 general medical journals / Y. Kuo // Journal of the American Medical Association. – 2002. – Vol. 287. – P. 2815–2817.

20. *Leach C.* Introduction to statistics: a nonparametric approach for the social sciences / C. Leach. – Chichester : Wiley, 1979. – 339 p.

21. *Noether G.* Why Kendall Tau? / G. Noether // Teaching Statistics. – 1981. – Vol. 3. – P. 41–43.

22. *Porter A.* Misuse of correlation and regression in three medical journals / A. Porter // Journal of the Royal Society of Medicine. – 1999. – Vol. 92. – P. 123–128.

23. *Tesfaye F.* Association between body mass index and blood pressure across three populations in Africa and Asia / F. Tesfaye, N. G. Nawi, H. Van Minh et al. // Journal of Human Hypertension. – 2007. – Vol. 21. – P. 28–37.

24. *Wilkie D.* Pictorial representation of Kendall's rank correlation coefficient / D. Wilkie // Teaching Statistics. – 1980. – Vol. 2. – P. 76–78.

CORRELATION ANALYSIS

A. M. Grjibovski

National Institute of Public Health, Oslo, Norway

The article gives a brief introduction about correlation analysis and calculations of Pearson, Spearman and Kendall correlation coefficients using SPSS software. The paper provides only general introduction about the analysis of the strength of statistical association between variables. The readers are encouraged to consult statistical literature prior to analysing own data and preparing manuscripts.

Key words: correlation analysis, correlation coefficients, SPSS.

Контактная информация:

Гржибовский Андрей Мечиславович – старший советник Национального института общественного здоровья, г. Осло, Норвегия

Адрес: Nasjonalt folkehelseinstitutt, Pb 4404 Nydalen, 0403 Oslo, Norway

Тел.: +47 21076392, +47 45268913; e-mail: angr@fhi.no

Статья поступила 20.08.2008 г.