УДК 681.3.01. 616-07:311

В. А. Каширин[1], В. П. Леонов[2], А. В. Томашевский[3]
[1]ГУ «Запорожская медицинская академия последипломного образования МЗ Украины», Запорожье, Украина
[2] БИОМЕТРИКА, Томск, Россия
[3]Запорожский национальный технический университет, Запорожье, Украина

V. A. Kashirin[1], V. P. Leonov[2], A. V. Tomashevskiy[3]
[1]SI «Zaporizhia medical academy of post-graduate education Ministry of health of Ukraine», Zaporizhia , Ukraine
[2]BIOMETRICA, Tomsk, Russia
[3]Zaporizhia National Technical University, Zaporizhia, Ukraine

КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ СТАТИСТИЧЕСКОГО АНАЛИЗА БИОМЕДИЦИНСКОЙ ИНФОРМАЦИИ

(Часть третья – корреляционный анализ)

Computer technologies of statistical analysis of biomedical information

(Part Three – Correlation Analysis)

Резюме

Представлены практические рекомендации по проведению в программе STATISTICA парной линейной, ранговой и частной корреляции.

Ключевые слова: *корреляционный анализ.*

Abstract

The practical recommendations for conducting in STATISTICA program pair linear, rank and partial correlation are given.

Keywords: *correlation analysis.*

Корреляция – несколько видов статистической взаимосвязи двух или нескольких случайных величин (показателей). Статистически значимый один из видов корреляции между случайными величинами, является свидетельством существования некоторой их взаимозависимости в данной выборке, что, однако, не обязательно должно наблюдаться с аналогичными случайными величинами в другой выборке. В то же время, отсутствие одного вида корреляции между величинами ещё не значит, что между ними вообще нет никакой взаимосвязи.

Для определения и характеристики статистической взаимосвязи используется корреляционный анализ, который решает при этом следующие задачи:

1. Существует ли связь между изменениями значений исследуемых показателей.
2. Степень (силу) выраженности этой связи (коэффициент корреляции).

ИСПОЛЬЗУЕМЫЕ ТЕРМИНЫ

Прямая корреляция – однонаправленное изменение показателей (их повышение ↑–↑ или понижение ↓–↓).

Обратная корреляция – разнонаправленное изменение показателей (↑–↓ или ↓–↑).

Коэффициент корреляции – имеет значения от –1 до +1. При прямой корреляции имеет положительное (+) значение. При обратной корреляции имеет отрицательное (–) значение. Нулевое значение показывает, что признаки независимы.

При оценке величины линейной корреляции между количественными показателями используется коэффициент корреляции Пирсона.

ОПРЕДЕЛЕНИЕ ПАРНОГО КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ ПИРСОНА

Для оценки силы корреляционных связей между двумя или несколькими количественными показателями (в данном примере между некоторыми значениями показателей Т-клеточного иммунитета – CD3+, CD4+, CD8+, CD16+) необходимо выполнить следующую последовательность действий:

1. Импортировать данные из листа программы MS Excel в рабочую книгу программы STATISTICA, где активировать Statistics, затем Basic Statistics/Tables, что приведет к открытию соответствующего окна, в котором выбрать и активировать Correlation matrices и нажать OK (рис. 1).

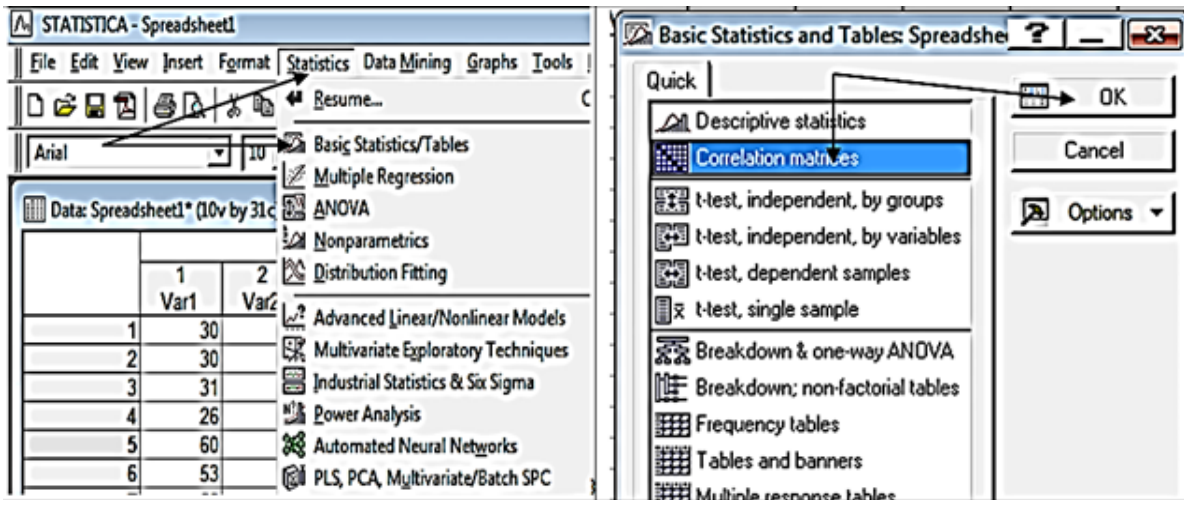

Рис. 1. Таблица пакета STATISTICA и окно «Basic Statistics/Tables».
Указать Statistics → Basic Statistics/Tables → Correlation matrices → OK

2. В открывшемся окне «Product-Moment and Partial Correlations» указать Quick, активировать кнопку One variable list, а затем, в окне «Select the variables for the analysis» выделить исследуемые показатели и нажать кнопку ОК (рис. 2).

3. Во вновь открывшемся окне «Product-Moment and Partial Correlations» активировать Summary Correlation matrix, что позволит получить результаты проведенного корреляционного анализа, где в матрице красным цветом будут выделены значимые коэффициенты корреляции с доверительной вероятностью 95% (рис. 3), которые и следует указать в текстовой части, при обсуждении результатов проведенного анализа: CD3+ и CD4+ (r = 0,43); CD4+ и CD8+ (r = 0,59).

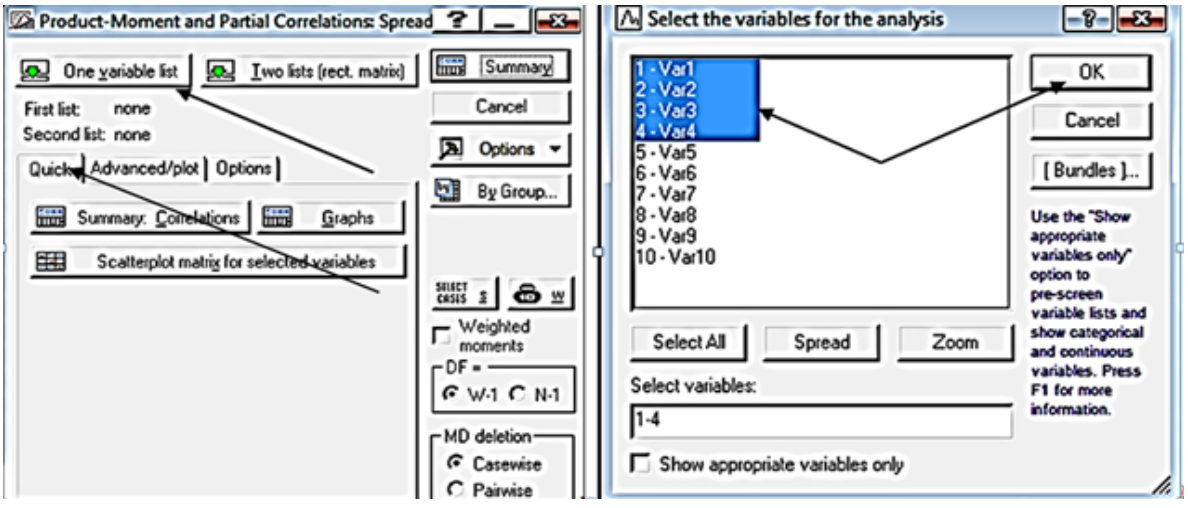

Рис. 2. Окна «Product-Moment and Partial Correlations» и «Select the variables for the analysis».
Указать: One variable list → выделить показатели → ОК

Correlations (Spreadsheet1)
Marked correlations are significant at p < ,05000
N=31 (Casewise deletion of missing data)

Variable	Means	Std.Dev.	Var1	Var2	Var3	Var4
Var1	46,38710	13,65449	1,000000	0,428944	0,136147	0,044273
Var2	31,93548	6,35576	0,428944	1,000000	0,595698	-0,183905
Var3	35,67742	3,72726	0,136147	0,595698	1,000000	0,209481
Var4	29,80645	10,16340	0,044273	-0,183905	0,209481	1,000000

Рис. 3. Результат корреляционного анализа Пирсона

ЗАМЕЧАНИЯ

В анализируемом материале отдельные признаки могут иметь в массиве данных разные количества измеренных значений. В таком случае, вместо кнопки Quick можно использовать кнопку Options. При использовании Options отображаются коэффициенты корреляции, достигнутые уровни статистической значимости каждого коэффициента корреляции, а также то количество наблюдений, в которых все анализируемые признаки не имели пропущенных значений.

При наличии разного количества пропусков в разных анализируемых признаках, общее количество анализируемых наблюдений существенно снижается. Поэтому целесообразно при оценке коэффициентов корреляций делать это не для большого количества признаков, а раздельно по всем интересующим парам признаков. Тогда количество анализируемых наблюдений не будет столь сильно уменьшаться.

Следует помнить, что полученные коэффициенты могут являться проявлениями «ложной корреляции».

ЛОЖНАЯ КОРРЕЛЯЦИЯ

Кажущаяся простота оценки данных, полученных при проведении корреляционного анализа, может подтолкнуть исследователя к ошибочным выводам о наличии причинно-следственных отношений между парами признаков, в то время как коэффициенты корреляции устанавливают лишь статистические взаимосвязи и, более того, могут оказаться проявлениями ложной корреляции.

Так, рассматривая сложные, многочасовые хирургические вмешательства можно выявить прямую корреляцию между объемом перелитой пациенту крови и количеством хирургов, участвовавших в операции. Из этого, однако, не следует вывод – «большее количество хирургов обуславливает большую кровопотерю» и, тем более, не имеет смысла попытка минимизировать кровопотерю путем уменьшения количества участвующих в операции врачей.

Ложная корреляция вызывается «общей причиной», называемой «агентом ложной корреляции», устранить искажающее влияние, которого можно если:

– логически исключить из анализа явно абсурдные переменные. К примеру, заболеваемость детей респираторными заболеваниями скорее будет коррелировать с наличием инфекции в ближайшем окружении, чем с количеством докторов наук по специальности «педиатрия»;

– рассматривать не пару, а множество «потенциально важных» значений, используя при этом коэффициент частной корреляции, с помощью которого можно оценить степень тесноты линейной связи между показателями, «очищенную» от опосредованного влияния других факторов.

ОПРЕДЕЛЕНИЕ ЧАСТНОГО КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

1. Подготовка исходных данных выполнить, как и при определении парного коэффициента корреляции (импортировать исходные данные в рабочую книгу программы STATISTICA). Активировать Statistics, а затем Basic Statistics/Tables, что приведет к открытию окна – «Basic Statistics/Tables», в котором выбрать и активировать Correlation matrices, а затем нажать кнопку OK.

2. В открывшемся окне «Product-Moment and Partial Correlations» последовательно активировать кнопки Advanced/Plot и Partial Correlations.

3. В окне «Select two-variable lists» выделить исследуемые показатели (удерживая клавишу Ctrl) в первом подокне (first variable list) для вычисления частного коэффициента корреляции и во втором подокне (second var. list) указать «фиксируемый» показатель и нажать кнопку OK (рис. 4).

3. В появившемся окне «Partial Correlations» (рис. 5) показан частный коэффициент корреляции для CD3+ и CD4+. Анализ подтвердил наличие корреляционной связи, причем значение коэффициента возросло: $r = 0{,}42$ и $r = 0{,}48$.

Следует помнить, что критерий корреляции Пирсона – параметрический и условием его применения является нормальное распределение сопоставляемых количественных показателей.

При необходимости проведения корреляционного анализа показателей, распределение которых отличается от нормального или измеренных в порядковой шкале, следует использовать ранговую корреляцию.

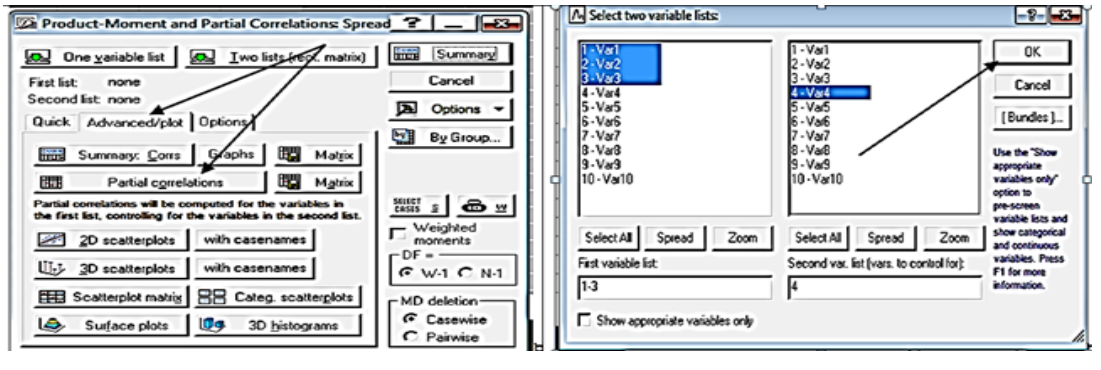

Рис. 4. Окно «Product-Moment and Partial Correlations». Указать Advanced/Plot → Partial Correlations. Окно «Select two-variable lists»: выделить показатели → OK

Partial Correlations (Spreadsheet1)
Marked correlations are significant at p < ,05000
N=31 (Casewise deletion of missing data)

Variable	Means	Std.Dev.	Var1	Var2
Var1	46,38710	13,65449	1,000000	0,482411
Var2	31,93548	6,35576	0,482411	1,000000

Рис. 5. Окно «Partial Correlations» с результатом анализа

РАНГОВЫЙ КОРРЕЛЯЦИОННЫЙ АНАЛИЗ СПИРМЕНА

Ранговая корреляция применяется для выявления взаимосвязи между количественными или качественными ранжированными показателями (упорядоченными по возрастанию или по убыванию их значения).

Достоинством коэффициентов ранговой корреляции является возможность их использования независимо от характера распределения коррелирующих признаков.

Рассмотрим методику проведения ранговой корреляции на примере взаимозависимости показателя пятилетней безрецидивной выживаемости больных раком гортани со следующими показателями:

– S (Survival): 1, 2, 3, 4, 5 – годы;

– I (Index): 1, 2, 3, 4 – критерии адаптационных реакций по Л. Х. Гаркави;

– T (Tumor): 1, 2, 3, 4 – распространенность новообразования;

– N (Nodes): 0, 1, 2, 3 – характеристика лимфатических узлов (регионарные метастазы);

– O (Operation): 1, 2, 3, 4 – объем операции;

– RT (Radiation therapy): 0, 1, 2 – (1, 2: 40 или 60 Гр);

– G (Histology): 1, 2, 3 – морфологическая характеристика опухоли;

которые внесем в рабочий лист программы MS Excel, а далее необходимо выполнить следующие действия:

1. Импортировать из MS Excel исходные данные в рабочую книгу программы STATISTICA. В верхней строчке окна рабочей книги последовательно активировать Statistics Nonparametrics (рис. 6).

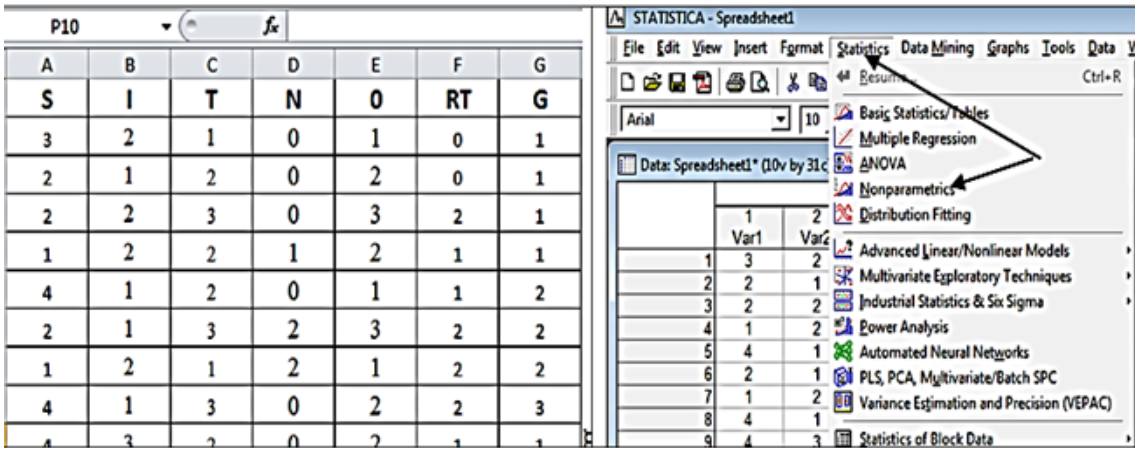

Рис. 6. Таблица в программе MS Excel с внесенными данными.
Рабочий лист программы STATISTICA: указать Statistics → Nonparametrics

2. В окне «Nonparametrics Statistics» указать Correlations (Spearmen, Kendall tau. gamma) и нажать ОК, а затем, в открывшемся окне «Spreadsheeat1» указать Detailed report и активировать Variables (рис. 7).

3. В открывшемся окне «Select two variable lists» обозначить анализируемые показатели и нажать ОК, что приведет к повторному открытию окна «Spreadsheeat1», где следует указать Spearman R (рис. 8).

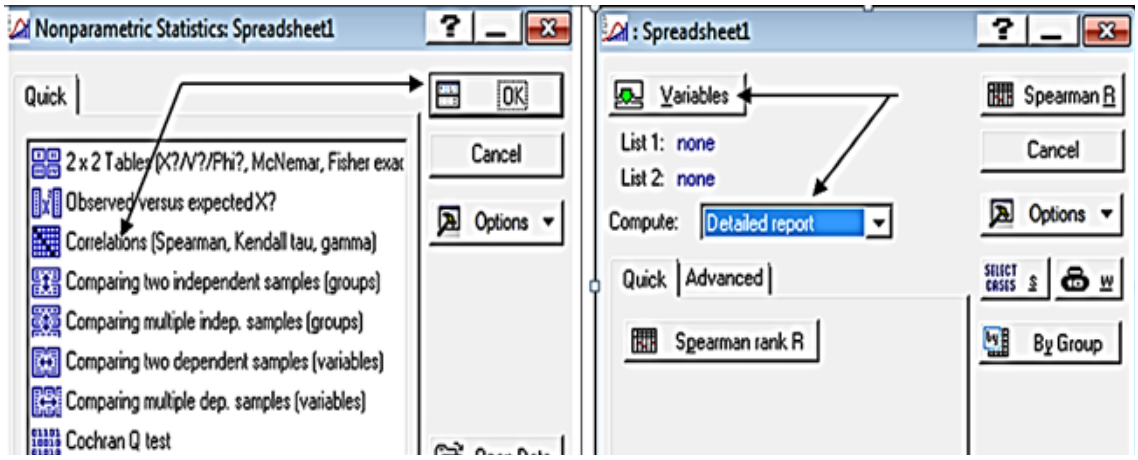

Рис. 7. Окно «Nonparametrics Statistics», указать Correlations (Spearmen, Kendall tau. gamma).
Окно «Spreadsheeat1», указать Detailed report → Variables

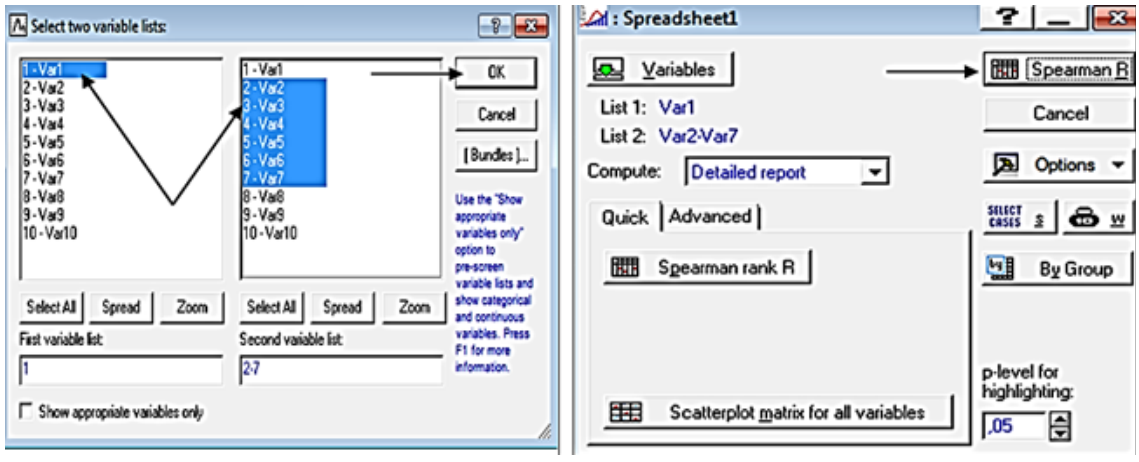

Рис. 8. Окно «Select two variable lists», обозначить показатели → ОК.
Окно «Spreadsheeat1», указать Spearman R

Рисунок 9 – результаты проведенного анализа, где в матрице красным цветом выделен статистически значимый коэффициент корреляции между Var 1 & Var 2 (S & I).

Spearman Rank Order Correlations (Spreadshee
MD pairwise deleted
Marked correlations are significant at p <,05000

Pair of Variables	Valid N	Spearman R	t(N-2)	p-level
Var1 & Var2	31	0,446780	2,689320	0,011746
Var1 & Var3	31	0,089334	0,483011	0,632715
Var1 & Var4	31	-0,049998	-0,269584	0,789390
Var1 & Var5	31	0,054103	0,291779	0,772532
Var1 & Var6	31	0,047664	0,256970	0,799017
Var1 & Var7	31	-0,101010	-0,546755	0,588730

Рис. 9. Результаты рангового корреляционного анализа.
Коэффициент корреляции для «выживаемость – адаптационная реакция» равен 0,45, при p = 0,01

Анализ полученных результатов позволяет сделать вывод, что показатель пятилетней безрецидивной выживаемости больных раком гортани имеет значимую прямую корреляцию с показателями адаптационных реакций, т.е. зависит от иммунобиологического состояния организма и не коррелирует с распространенностью новообразования, наличием или отсутствием метастазов в регионарных лимфатических узлах, морфологической характеристикой опухоли, объемом и компонентами проведенного специального лечения.

ЗАМЕЧАНИЯ

Следует четко различать понятия зависимости и корреляции. Зависимость величин обуславливает наличие корреляционной связи между ними, но не наоборот. Определение причинно-следственной связи между переменными – цель и возможность регрессионного анализа.

При проведении ранговой корреляции, для оценки данных необходима выборка от 5 до 40 наблюдений по каждой переменной.

РЕКОМЕНДОВАННАЯ ЛИТЕРАТУРА

1. Электронный учебник STATISTIKA (StatSoft).

http://statsoft.ru/home/textbook/modules/stbasic.html

2. Кендел М. Ранговые корреляции.

http://padaread.com/?book=35784&pg=10

3. Ланг Т. А., Сесик М. Как описывать статистику в медицине. http://kingmed.info/knigi/Meditsinskaya_informatika_i_biostatistika/book_3123/Kak_opisivat_statistiku_v_meditsine_Rukovodstvo_dlya_avtorov_redaktorov_i_retsenzentov-Lang_TA_Sesik_M-2011-djvu

Стаття надійшла до редакції 25.07.2017