УДК 123; JEL Classification: A10, B40

Использование коэффициентов корреляции и конкордации

Шамсувалеева А.М.[1], Орлов А.И.[2]

[1]студент 4-го курса, кафедра «Экономика и организация производства» МГТУ им. Н.Э. Баумана, г. Москва, *Alina.Shamsuvaleeva@yandex.ru*;

[2]профессор, д.э.н., д.т.н., к.ф.-м.н., профессор кафедры «Экономика и организация производства», МГТУ имени Н.Э. Баумана, г. Москва, *prof-orlov@mail.ru.*

Аннотация: *В работе проведен анализ частоты использования коэффициентов корреляции и конкордации в различных тематиках. Всего рассмотрено 28 тематик, по которым производился поиск результатов в научной электронной библиотеке eLIBRARY.RU по ключевым словам.*
Ключевые слова: *коэффициент корреляции, коэффициент корреляции Спирмена, коэффициент корреляции Кендалла, коэффициент корреляции Пирсона, коэффициент конкордации.*

Use of correlation and concordance coefficients

Alina Shamsuvaleeva[1], Alexander Orlov[2]

[1]student of department «Economics and organization of production», Bauman Moscow State Technical University, Moscow;

[2]professor of department «Economics and organization of production», doctor of econ. sc., doctor of techn.sc., cand. of math., professor, Bauman Moscow State Technical University, Moscow.

Abstract: *The paper analyzes the frequency of using correlation and concordance coefficients in different subjects. In total, 28 subjects were considered, for which the results were searched in the scientific electronic library eLIBRARY.RU using keywords.*
Keywords: *correlation coefficient, the Spearman correlation coefficient, the Kendall correlation coefficient, the Pearson correlation coefficient, concordance coefficient.*

Введение

В современных исследованиях при анализе данных часто возникает необходимость в оценке взаимосвязей между различными показателями. Одними из наиболее распространенных и эффективных инструментов для решения этих задач являются коэффициенты корреляции и конкордации. Коэффициенты корреляции и конкордации являются статистическими показателями, которые позволяют обнаружить наличие взаимосвязи между двумя или несколькими переменными и оценить ее степень. Данные коэффициенты применяются в социальных, экономических, биологических и других системах.

Статья разделена на несколько частей. В первой части рассмотрены теоретические основы применения коэффициентов корреляции и конкордации и приведены формулы для расчета этих показателей. Во второй части описана методика поиска данных для проведения исследования. В третьей части проведен анализ полученных данных и выявлены наиболее часто используемые коэффициентов в конкретных тематиках.

Базовые понятия

Термин «корреляция» означает «связь». Применительно к анализу данных этот термин обычно используется в сочетании «коэффициент корреляции» [1].

Пусть исходными данными является набор случайных векторов $(x_i,\ y_i) = (x_i(\omega),\ y_i(\omega)),\ i = 1,\ 2,\ ...,\ n.$ Выборочным коэффициентом корреляции, более подробно, выборочным **линейным парным коэффициентом корреляции К. Пирсона**, как известно, называется число

$$r_n = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Если $r_n = 1,$ то $y_i = ax_i + b$ при некоторых a и b, причем *a > 0*. Если же причем *a < 0*. $r_n = -1,$ то $y_i = ax_i + b$ Таким образом, близость коэффициента корреляции к 1 (по абсолютной величине) говорит о достаточно тесной линейной связи.

Если $(x_i,\ y_i) = (x_i(\omega),\ y_i(\omega)),\ i = 1,\ 2,\ ...,\ n,$ случайные вектора независимы и одинаково распределены, то выборочный коэффициент корреляции сходится к теоретическому при безграничном возрастании объема выборки:

$$r_n \to \rho = \frac{M(x_1 - M(x_1))M(y_1 - M(y_1))}{\sqrt{D(x_1)}\sqrt{D(y_1)}}$$

(сходимость по вероятности в предположении, что существуют дисперсии координат случайного вектора).

Коэффициенты корреляции типа r_n используются во многих алгоритмах многомерного статистического анализа.

В теоретических рассмотрениях часто считают, что случайные вектора $(x_i,\ y_i) = (x_i(\omega),\ y_i(\omega)),\ i = 1,\ 2,\ ...,\ n,$ имеют двумерное нормальное распределение. Распределения реальных данных, как правило, отличны от нормальных [2, 3].

Например, равенство 0 теоретического коэффициента корреляции эквивалентно независимости случайных величин. Поэтому проверка независимости сводится к проверке статистической гипотезы о равенстве 0 теоретического коэффициента корреляции. Эта гипотеза принимается, если $|r_n| < C(n,\ \alpha)$, где $C(n,\ \alpha)$ — некоторое граничное значение, зависящее от объема выборки *n* и уровня значимости α .

Для расчета непараметрического коэффициента ранговой корреляции Спирмена необходимо сделать следующее. Для каждого x_i рассчитать его ранг r_i в вариационном ряду, построенном по выборке $x_1,\ x_2,\ ...,\ x_n.$ Для каждого y_i

рассчитать его ранг q_i в вариационном ряду, построенном по выборке $y_1,\ y_2,\ ...,\ y_n$. Для набора из *n* пар $(r_i,\ q_i),\ i = 1,\ 2,\ ...,\ n,$ вычислить линейный коэффициент корреляции. Он называется коэффициентом ранговой корреляции, поскольку определяется через ранги.

Коэффициент ранговой корреляции Спирмена равен

$$\rho_n = 1 - \frac{6\sum_{i=1}^{n}(r_i - q_i)^2}{n^3 - n}.$$

Коэффициент ранговой корреляции Спирмена остается постоянным при любом строго возрастающем преобразовании шкалы измерения результатов наблюдений.

Широко используется также коэффициент ранговой корреляции τ Кендалла, коэффициент ранговой конкордации Кендалла и Б. Смита и др.

Коэффициент ранговой корреляции τ Кендалла определяется так [4]. Пусть N — количество тех упорядоченных пар индексов *(i, j), i < j,* для которых эксперты одинаково упорядочивают объекты, т. е. для которых либо одновременно $r_i < r_j$, $q_i < q_j$, либо одновременно $r_i > r_j$, $q_i > q_j$. Тогда

$$\tau = \frac{4N}{n(n-1)} - 1.$$

Если экспертные упорядочения совпадают, то коэффициент ранговой корреляции Кендалла принимает максимальное значение τ = 1. Если экспертные упорядочения совпадают, то коэффициент ранговой корреляции Кендалла принимает максимальное значение τ = 1. Если эксперты дают прямо противоположные упорядочения, их мнения противоречат друг другу для любой пары объектов, то коэффициент ранговой корреляции Кендалла минимален, τ = –1.

Если экспертов $m > 2$, то данные ими m упорядочений можно записать в виде матрицы, *i*-я строка которой содержит ранжировку, полученную от *i*-го эксперта, а столбцы соответствуют *n* объектам экспертизы, рассматриваемым в данном исследовании:

$$\begin{Vmatrix} r_{1,1} & r_{1,2} & \text{K} & r_{1,n} \\ r_{2,1} & r_{2,2} & \text{K} & r_{2,n} \\ \text{K} & \text{K} & \text{K} & \text{K} \\ r_{m,1} & r_{m,2} & \text{L} & r_{m,n} \end{Vmatrix}.$$

В качестве единой выборочной меры связи *m* признаков Кендалл и Бэбингтон Смит предложили коэффициент согласованности *W*, называемый также **коэффициентом конкордации** (от лат. *concordare* — привести в соответствие, упорядочить):

$$W = \frac{12S_W}{m^2(n^3 - n)},$$

где

$$S_W = \sum_{i=1}^{n}\left[\sum_{j=1}^{m} r_{i,j} - \frac{m(n+1)}{2}\right]^2.$$

Можно показать, что среднее арифметическое коэффициентов ранговой корреляции Спирмена ρ для $m(m-1)/2$ пар признаков равно $(mW-1)/(m-1)$. В частности, если $m = 2$, то $\rho = -2W - 1$.

Все три коэффициента |ρ|, |τ| и W принимают значения из отрезка [0; 1] и используются для проверки нулевой гипотезы H_0 о независимости признаков. Признаки называются независимыми, если для наугад выбранного столбца матрицы ранги (порядковые номера) $r_{1,j}$, $r_{2,j}$, ..., $r_{m,j}$ являются взаимно независимыми случайными величинами. В терминах теории экспертных оценок гипотеза H_0 — это гипотеза о том, что случайные ранжировки независимы и равномерно распределены на множестве всех ранжировок (без связей).

Если рассматриваемый коэффициент (|ρ|, |τ| и W) не превосходит заданного граничного значения, то гипотеза H_0 принимается, если превосходит — отклоняется в пользу альтернативной гипотезы общего вида, т. е. гипотезы о том, что совместное распределение ранжировок отличается от совместного распределения независимых одинаково распределенных ранжировок. При этом остается неизвестным, нарушается ли предположение независимости, или предположение равномерности распределения, или и то и другое вместе. Например, нулевая гипотеза отклоняется, если все эксперты повторяют ответ первого из них, но сам этот ответ равномерно распределен. Или тогда, когда половина экспертов выбирает одну определенную ранжировку или похожие на нее, а вторая половина экспертов — другую определенную ранжировку (или похожую на нее). В этом случае нет равномерности распределения, и нулевая гипотеза отклоняется, хотя говорить о согласованности экспертов не приходится. Если же нулевая гипотеза принимается, то ни о какой согласованности мнений экспертов говорить нельзя.

Если гипотеза H_0 верна, то

$$M(\rho) = 0,\ M(\tau) = 0,\ M(W) = \frac{1}{m},$$

$$D(\rho) = \frac{1}{n-1},\ D(\tau) = \frac{2(2n+5)}{9n(n-1)},\ D(W) = \frac{2(m-1)}{m^3(n-1)}.$$

Результаты поиска статей в поисковых системах и Российском индексе научного цитирования (РИНЦ)

Поиск статей производился в научной электронной библиотеке eLIBRARY.RU по ключевым словам: «Корреляция», «Корреляция Спирмена», «Корреляция Кендалла», «Корреляция Пирсона», «Конкордация» в целом и конкретно по 28 тематикам.

Тематики:

1. Общественные науки в целом.
2. Философия.
3. История. Исторические науки.
4. Социология.
5. Демография.
6. Экономика. Экономические науки.
7. Государство и право. Юридические науки.
8. Политика. Политические науки.
9. Науковедение.
10. Культура. Культурология.
11. Народное образование. Педагогика.
12. Психология.

13. Искусство. Искусствоведение.
14. Массовая коммуникация. Журналистика. Средства массовой информации.
15. Математика.
16. Кибернетика.
17. Физика.
18. Химия.
19. Биология.
20. Легкая промышленность.
21. Пищевая промышленность.
22. Сельское и лесное хозяйство.
23. Медицина и здравоохранение.
24. Физическая культура и спорт.
25. Военное дело.
26. Организация и управление.
27. Статистика.
28. Патентное дело. Изобретательство. Рационализаторство.

С теоретической точки зрения не выявлены иные способы расчетов коэффициентов корреляции и конкордации. Поэтому имеет смысл рассмотреть данную тему с точки зрения применения на практике.

Анализ полученных результатов

Тема «Коэффициенты корреляции и конкордации» представляет интерес с точки зрения распространённости в той или иной области знаний и частоте использования конкретных методов расчета коэффициента корреляции.

Уточним, что сумма запросов по отдельным методам не равна общему количеству статей по запросу «Корреляция», так как не все авторы указывают в своих работах конкретную методику расчета и представляют формулы без названия. Так же часто одна статья может относиться к нескольким тематикам. Особенно часто данную ситуацию можно заметить при поиске по тематикам «Философия» и «История. Исторические науки».

Построим таблицу 1 по количеству результатов поиска по пяти запросам по каждой из выбранных тематике.

Таблица 1. Количество результатов поиска

Тематика поиска	Корреляция	Корреляция Спирмена	Корреляция Кендалла	Корреляция Пирсона	Конкордация
Общий запрос	38 614	12 545	4 301	11 922	846
Общественные науки в целом	7 024	1 048	66	679	364
Философия	2 753	376	19	227	66
История. Исторические науки	2 315	167	16	148	44
Социология	2 008	485	30	295	115
Демография	483	77	1	65	23

Тематика поиска	Корреляция	Корреляция Спирмена	Корреляция Кендалла	Корреляция Пирсона	Конкордация
Экономика. Экономические науки	11 266	1 224	135	974	699
Государство и право. Юридические науки	2 802	398	45	321	258
Политика. Политические науки	946	82	9	97	51
Науковедение	865	93	15	67	57
Культура. Культурология	658	51	2	40	15
Народное образование. Педагогика	6 383	1 230	85	832	539
Психология	3 505	1 130	42	654	136
Искусство. Искусствоведение	365	27	1	12	6
Массовая коммуникация. Журналистика. Средства массовой информации	350	30	0	17	5
Математика	5 505	333	87	368	322
Кибернетика	2 517	158	56	236	192
Физика	7 464	274	41	322	68
Химия	9 005	821	45	585	80
Биология	11 660	1 339	73	947	129
Легкая промышленность	138	5	1	8	55
Пищевая промышленность	1 246	59	9	70	59
Сельское и лесное хозяйство	10 272	769	59	579	213
Медицина и здравоохранение	10 535	1 865	96	1 240	324
Физическая культура и спорт	1 572	401	14	303	66
Военное дело	273	18	2	13	31

Тематика поиска	Корреляция	Корреляция Спирмена	Корреляция Кендалла	Корреляция Пирсона	Конкордация
Организация и управление	978	116	21	95	134
Статистика	931	114	28	84	66
Патентное дело. Изобретательство. Рационализаторство	98	8	3	11	5

По данным, представленным в таблице 1, найдем максимальные и минимальные значения количества результатов и соответсвенные названия тематик.

По запросу «Корреляция» общее количество результатов равно 38 614. Максимальное количество результатов получено по тематике «Биология» и равно 11 660. Минимальное значение – «Патентное дело. Изобретательство. Рационализаторство», 98.

По запросу «Корреляция Спирмена» общее количество результатов равно 12 545. Максимальное количество результатов получено по тематике «Медицина и здравоохранение» и равно 1 865. Минимальное значение – «Легкая промышленность», 5.

По запросу «Корреляция Кендалла» общее количество результатов равно 4 301. Максимальное количество результатов получено по тематике «Экономика. Экономические науки» и равно 135. Минимальное значение – «Массовая коммуникация. Журналистика. Средства массовой информации», 0.

По запросу «Корреляция Пирсона» общее количество результатов равно 11 922. Максимальное количество результатов получено по тематике «Медицина и здравоохранение» и равно 1 240. Минимальное значение – «Легкая промышленность», 8.

По запросу «Конкордация» общее количество результатов равно 846. Максимальное количество результатов получено по тематике «Экономика. Экономические науки» и равно 699. Минимальное значение – «Массовая коммуникация. Журналистика. Средства массовой информации», 5.

Исходя из полученных данных, можно судить о том, что проверка согласованности мнений реже всех используется в тематике «Массовая коммуникация. Журналистика. Средства массовой информации» и «Легкая промышленность», чаще всего – «Медицина и здравоохранение» и «Экономика. Экономические науки».

В таблице 2 представим процентные соотношения результатов запроса: в 1 столбце представлены тематики поиска, во 2 столбце – доли результатов запроса «Корреляция» по конкретной тематике к общему количеству результатов по запросу без тематики, в 3 столбце – доли результатов запроса «Конкордация» по конкретной тематике к общему количеству результатов по запросу без тематики, в столбцах 4, 5 и 6 – процентное отношение количества результатов запроса «Корреляция Спирмена» по каждой тематике к количеству результатов запроса «Корреляция» по каждой тематике и т.д.

Таблица 2. Процентные соотношения результатов запроса

Тематика поиска	Корреляция, %	Конкродация, %	Корреляция Спирмена/ Корреляция, %	Корреляция Кендалла/ Корреляция, %	Корреляция Пирсона/ Корреляция, %
Общий запрос	–	–	32,49	11,14	30,87
Общественные науки в целом	18,19	43,03	14,92	0,94	9,67
Философия	7,13	7,80	13,66	0,69	8,25
История. Исторические науки	6,00	5,20	7,21	0,69	6,39
Социология	5,20	13,59	24,15	1,49	14,69
Демография	1,25	2,72	15,94	0,21	13,46
Экономика. Экономические науки	29,18	82,62	10,86	1,20	8,65
Государство и право. Юридические науки	7,26	30,50	14,20	1,61	11,46
Политика. Политические науки	2,45	6,03	8,67	0,95	10,25
Науковедение	2,24	6,74	10,75	1,73	7,75
Культура. Культурология	1,70	1,77	7,75	0,30	6,08
Народное образование. Педагогика	16,53	63,71	19,27	1,33	13,03
Психология	9,08	16,08	32,24	1,20	18,66
Искусство. Искусствоведение	0,95	0,71	7,40	0,27	3,29
Массовая коммуникация. Журналистика. Средства массовой информации	0,91	0,59	8,57	0	4,86
Математика	14,26	38,06	6,05	1,58	6,68
Кибернетика	6,52	22,70	6,28	2,22	9,38
Физика	19,33	8,04	3,67	0,55	4,31
Химия	23,32	9,46	9,12	0,50	6,50
Биология	30,20	15,25	11,48	0,63	8,12
Легкая промышленность	0,36	6,50	3,62	0,72	5,80
Пищевая промышленность	3,23	6,97	4,74	0,72	5,62
Сельское и лесное хозяйство	26,60	25,18	7,49	0,57	5,64
Медицина и здравоохранение	27,28	38,30	17,70	0,91	11,77
Физическая культура и спорт	4,07	7,80	25,51	0,89	19,27

Тематика поиска	Корреляция, %	Конкродация, %	Корреляция Спирмена/ Корреляция, %	Корреляция Кендалла/ Корреляция, %	Корреляция Пирсона/ Корреляция, %
Военное дело	0,71	3,66	6,59	0,73	4,76
Организация и управление	2,53	15,84	11,86	2,15	9,71
Статистика	2,41	7,80	12,24	3,01	9,02
Патентное дело. Изобретательство. Рационализаторство	0,25	0,59	8,16	3,06	11,22

По данным, представленным в таблице 2, найдем максимальные и минимальные доли количества результатов и соответсвенные названия тематик.

Наибольшая доля результатов по запросу «Корреляция» – 30,20% по тематике «Биология». Минимальная – «Патентное дело. Изобретательство. Рационализаторство», 0,25%. Результаты соответственно совпадают с полученными максимальными и минимальными значениями.

Наибольшая доля результатов по запросу «Корреляция Спирмена» – 32,24% по тематике «Психология». Минимальная – «Легкая промышленность», 3,62%. Доля максимальных значений не совпала – коэффициент корреляции Спирмена чаще всего используется в психологии. Доля минимальных значений соответствует тематике с минимальным количеством результатов запроса – реже всего коэффициент используется в легкой промышленности.

Наибольшая доля результатов по запросу «Корреляция Кендалла» – 3,06% по тематике «Патентное дело. Изобретательство. Рационализаторство». Минимальная – «Массовая коммуникация. Журналистика. Средства массовой информации», 0,25%. Доля максимальных значений не совпала – коэффициент корреляции Спирмена чаще всего используется в патентном деле. Доля минимальных значений соответствует тематике с минимальным количеством результатов запроса – реже всего коэффициент используется в массовых коммуникациях.

Наибольшая доля результатов по запросу «Корреляция Пирсона» – 19,27% по тематике «Физическая культура и спорт». Минимальная – «Искусство. Искусствоведение», 3,29%. Доли не совпали с максимальным и минимальным количеством полученных результатов. Можно сделать вывод о том, что коэффициент корреляции Пирсона чаще всего используется в физической культуре и спорте.

Наибольшая доля результатов по запросу «Конкордация» – 82,62% по тематике «Экономика. Экономические науки». Минимальная – «Массовая коммуникация. Журналистика. Средства массовой информации», 0,59%. Результаты соответственно совпадают с полученными максимальными и минимальными значениями. Чаще всего коэффициент Конкордации используется в экономике и реже всего в массовых коммуникациях.

На рисунке 1 представим распределение общего количества результатов по запросам «Корреляция Спирмена», «Корреляция Кендалла» и «Корреляция Пирсона».

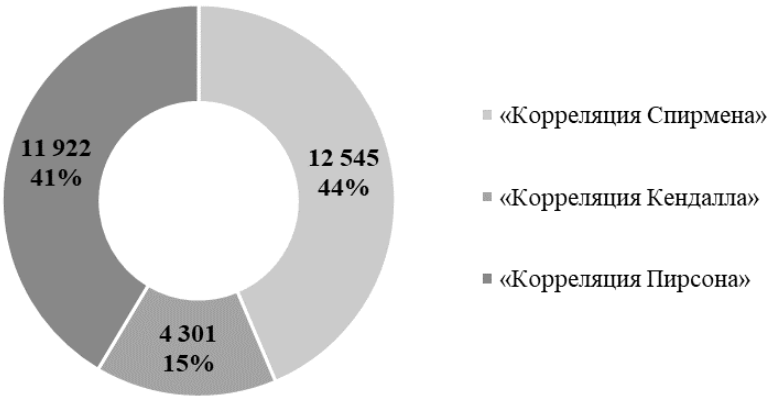

Рис. 1. Распределение общего количества результатов

Чаще всего авторы в своих публикациях используют коэффициент Корреляции Спирмена (44%) и реже всего коэффициент Корреляции Кендалла (15%). Такое распределение возможно из-за того, что коэффициент корреляции Пирсона оценивает только линейную связь и очень чувствителен к выбросам. Коэффициент корреляции Спирмена предпочтителен, когда данные имеют нелинейную связь, поэтому может более точно отразить связь между переменными. Коэффициент корреляции Кендалла используется в качестве альтернативного варианта коэффициента корреляции Спирмена, в связи с чем является наиболее редко используемым.

Выводы

Всего рассмотрено 28 тематик, по которым производился поиск результатов. Чаще всего коэффициенты корреляции и конкордации используются в тематиках: «Экономика. Экономические науки»., «Медицина и здравоохранение», «Биология»; реже всего – «Легкая промышленность», «Массовая коммуникация. Журналистика. Средства массовой информации» и «Патентное дело. Изобретательство. Рационализаторство». Наиболее часто в своих исследованиях авторы используют коэффициент корреляции Спирмена и реже всего коэффициент корреляции Кендалла. Коэффициент конкордации является наиболее редко используемым среди всех и наибольшую популярность имеет в тематике «Экономика. Экономические науки».

Литература

1. Эконометрика: учебник / З.С. Агаларов, А.И. Орлов. — М.: Издательско-торговая корпорация «Дашков и К°», 2021. — 380 с.
2. Орлов А.И. Прикладная статистика. Учебник. — М.: Экзамен, 2006. — 671 с.
3. Орлов А.И. Эконометрика: Учебник для вузов. — Изд. 3-е, перераб. и доп. — М.: Экзамен, 2004. — 576 с.
4. Кендэл М. Ранговые корреляции. — М.: Статистика, 1975. — 216 с.