УДК 519.24

Исследование статистических свойств оценок и критериев, используемых при анализе корреляционных связей[1]

Б.Ю. Лемешко, А.Д. Дамдинжапова
Новосибирский государственный технический университет

Исследуются статистические свойства оценок ранговых коэффициентов корреляции Спирмена и Кендалла и парного коэффициента корреляции Пирсона. Сравнивается мощность критериев, используемых для проверки некоррелированности.

Ключевые слова: коэффициент корреляции Спирмена, коэффициент корреляции Кендалла, парный коэффициент корреляции Пирсона, проверка гипотез

1. Введение

При анализе данных статистические методы играют особую роль. Они позволяют дать обоснованные рекомендации в тех случаях, когда опора на детерминированные знания из соответствующей прикладной области не даёт возможности сделать однозначные выводы. В таких ситуациях приходится анализировать события случайного характера, применяя различные методы математической статистики. Методы корреляционного анализа представляют собой один из важнейших разделов статистического анализа.

В процессе решения задач корреляционного анализа выявляется наличие и характер взаимосвязи величин, взаимозависимости величин при устранении влияния совокупности других или зависимости одной случайной величины от группы величин в исследуемом наборе данных. Наиболее распространенным способом выявление наличия связи (меры тесноты линейной связи) между некоторыми наборами данных оказывается вычисление оценок коэффициента линейной корреляции Пирсона, который может быть использован в качестве меры взаимозависимости.

С другой стороны, возможна ситуация, когда связь между исследуемыми данными не является строго линейной. В таком случае вместе с коэффициентами парной корреляции анализируются корреляционные отношения.

Аппарат классического корреляционного анализа опирается на предположение о принадлежности анализируемых данных многомерному нормальному закону. Такое предположение в условиях реальных приложений зачастую выполняется очень приблизительно. А значит свойства оценок и статистик критериев, используемых в корреляционном анализе, в условиях нарушения стандартного предположения о нормальности в той или иной мере отличаются от имеющих место в предположении о нормальности.

Стремление сделать выводы более устойчивыми к нарушению стандартного предположения привело к появлению ранговых коэффициентов корреляции, которые свободны от предположения о многомерной нормальности анализируемых величин.

[1] Работа выполнена при поддержке Министерства образования и науки РФ в рамках проектной части государственного задания (№ 2.541.2014/К).

Большинство результатов в математической статистике имеет асимптотический характер. На практике же всегда имеют дело с ограниченными объемами наблюдений. И свойства используемых статистик в таких случаях порой существенно отличаются от асимптотических. Не являются исключением и асимптотические свойства оценок и статистик корреляционного анализа, имеющие место при $n \to \infty$ в предположении о многомерной нормальности [1, 2, 3]. На практике же исследователю важно знать, начиная с каких объемов выборок n, он может пользоваться классическими результатами (соответствующими предельными законами). Заметим, что асимптотические свойства оценок коэффициентов парной корреляции зависят от истинного значения этого коэффициента.

Цель данной работы заключалась: в исследовании распределений оценок парных коэффициентов корреляции в зависимости от истинных значений этих коэффициентов при различных объёмах выборок; в исследовании связи между парным коэффициентом корреляции и ранговыми коэффициентами корреляции; в исследовании того, каким образом эта связь меняется в зависимости от истинного значения парного коэффициента корреляции r_{ij} и от объёма выборок n.

Взаимозависимость двух компонент $X^{(i)}$ и $X^{(j)}$ случайного вектора характеризуется парным коэффициентом корреляции r_{ij}. Если известна оценка ковариационной матрицы $\hat{\Sigma}$, то оценка максимального правдоподобия (ОМП) парного коэффициента корреляции может быть найдена в соответствии с выражением

$$\hat{r}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}, \tag{1}$$

Относительно коэффициента парной корреляции может проверяться два вида гипотез: о значимости коэффициента корреляции $H_0 : r_{ij} = 0$ и о равенстве его номинальному значению $H_0 : r_{ij} = r_0$.

При проверке гипотезы вида $H_0 : r_{ij} = 0$ опираются на статистику

$$t = \frac{\sqrt{n-2}\,\hat{r}_{ij}}{\sqrt{1-\hat{r}_{ij}^2}}, \tag{2}$$

где $\hat{r}_{ij}$ – ОМП парного коэффициента корреляции между компонентами $X^{(i)}$ и $X^{(j)}$, которая при справедливости гипотезы H_0 и выполнении стандартного предположения о принадлежности двумерной случайной величины нормальному закону подчиняется t_{n-2}-распределению Стьюдента с числом степеней свободы $n-2$. При нарушении стандартного предположения о нормальности распределение этой статистики также хорошо описывается t_{n-2}-распределением Стьюдента, если многомерный закон симметричен и с несильно “тяжёлыми хвостами” [4, 5].

При проверке гипотезы вида $H_0 : r_{ij} = r_0$ и её справедливости распределение статистики

$$z_0 = \sqrt{n-3}\left(\frac{1}{2}\ln\left(\frac{1+\hat{r}_{ij}}{1-\hat{r}_{ij}}\right) - \frac{1}{2}\ln\left(\frac{1+r_0}{1-r_0}\right) - \left(\frac{r_0}{2(n-1)}\right)\right), \tag{3}$$

подчиняется стандартному нормальному закону [4, 5]. При нарушении предположения о нормальности распределения статистики (3) (при тех же ограничениях на многомерный закон), не будет сильно отличаться от стандартного нормального закона при $r_0 \le 0.15$ [4, 5].

2. Вычисление оценки коэффициента ранговой корреляции Спирмена

Ряд объектов, упорядоченных в соответствии со степенью проявления некоторого свойства, называются ранжированными. Каждому числу такого ряда присваивается ранг. Ранги обозначают порядковыми числительными $1, 2, .., n$, где n – количество объектов. Меры взаимосвязи между парой признаков, каждый из которых ранжирует изучаемую совокупность объектов, называют коэффициентами ранговой корреляции.

На практике нередки случаи, когда несколько значений исходной выборки одинаковы, тогда им нужно приписывать одинаковые ранги. Несколько подряд идущих одинаковых значений образуют связку, такие элементы называются *связанными*. Каждый из связанных элементов получает ранг, равный среднему арифметическому рангов, которые имели бы элементы связки, если бы они были различны.

Коэффициенты ранговой корреляции могут использоваться не только для анализа взаимосвязи ранговых признаков, но для определения силы связи между ранговыми и количественными признаками, а также между двумя количественными признаками. В таких случаях значения количественных признаков упорядочиваются, и им приписываются соответствующие ранги.

В настоящее время коэффициенты ранговой корреляции получили широкое распространение, так как в отличие от традиционного коэффициента линейной корреляции Пирсона они позволяют успешно обнаруживать нелинейные зависимости и не зависят от вида распределения.

Статистика критерия строится на основе выборочного коэффициента ранговой корреляции Спирмена, который может быть вычислен по следующей формуле [3]:

$$\tau_{ij}^{(S)} = 1 - \frac{6\sum_{k=1}^{n}\left(r_k^{(i)} - r_k^{(j)}\right)^2}{n(n^2-1)}, \qquad (4)$$

где $r_k^{(j)}$, $r_k^{(j)}$ – ранги k-х объектов в наборах данных $X^{(i)}$ и $X^{(j)}$.

Наиболее распространенная статистика при проверке гипотез относительно коэффициента Спирмена имеет вид:

$$\tau_1^{(S)} = \tau_{ij}^{(S)}\sqrt{n-1}. \qquad (5)$$

где n – объем исследуемой выборки. При справедливости проверяемой гипотезы H_0 о некоррелированности статистика (5) должна подчиняться стандартному нормальному закону.

Используется также статистика вида

$$\tau_2^{(S)} = \tau_{ij}^{(S)}\sqrt{(n-2)\left(1-\tau_{ij}^{(S)}\right)}, \qquad (6)$$

которая при справедливости H_0 подчиняется распределению Стьюдента с $(n-2)$ степенями свободы, где n – объем исследуемой выборки.

В дальнейших исследованиях ограничимся рассмотрением статистику (5).

Исследования методами статистического моделирования свойств статистики (5) показали, что при объёмах выборок порядка $n \approx 30 \sim 100$ распределение статистики достаточно удовлетворительно описывается стандартным нормальным законом. Отмечена зависимость скорости сходимости к предельному закону от распределения, которому принадлежат наблюдаемые выборки.

3. Вычисление оценки коэффициента ранговой корреляции Кендалла

Оценка рангового коэффициента корреляции Кендалла определяется соотношением [3]:

$$\tau_{ij}^{(K)} = 1 - \frac{4 \cdot I(X^{(i)}, X^{(j)})}{n(n-1)}, \tag{7}$$

где $I(X^{(i)}, X^{(j)})$ – количество инверсий в наборе данных $X^{(i)}$ по отношению к $X^{(j)}$. Количество инверсий подсчитывается в соответствии с соотношением

$$I\left(X^{(i)}, X^{(j)}\right) = \sum_{q=1}^{n-l} \sum_{l=q+1}^{n} v_{ql}^{(i,j)}, \tag{8}$$

где $v_{ql}^{(i,j)} = \begin{cases} 1, & \text{если } x_q^{(i)} > x_l^{j}; \\ 0, & \text{иначе.} \end{cases}$ Статистика (7) принимает значения на отрезке [–1, 1].

В [2, 4] предложена следующая статистика

$$\tau_*^{(K)} = \tau^{(K)} \sqrt{\frac{9n(n-1)}{2(2n+5)}}, \tag{9}$$

которая при справедливости H_0 должна подчиняться стандартному нормальному закону.

Проведенные нами исследования показали, что статистика (9) при $n \geq 40$ удовлетворительно описывается стандартным нормальным законом. При этом какой-либо зависимости от вида закона, которому принадлежат наблюдаемые выборки, не отмечено.

4. Связь между парным коэффициентом корреляции Пирсона и ранговыми

Взаимозависимость двух компонент случайного вектора характеризуется коэффициентом корреляции Пирсона r_{ij}. Он представляет собой меру тесноты линейной связи. Коэффициент корреляции можно использовать в качестве некоторой меры взаимозависимости для нормального закона. При независимости двух случайных величин, коэффициент корреляции равен нулю, но при этом обратное утверждение в общем случае не верно. Это представляет трудность интерпретации r_{ij} как коэффициента взаимозависимости в общем случае. Однако, оно справедливо в случае нормальности случайных величин.

В случае выборок из нормального распределения коэффициент корреляции Кендалла может быть использован для расчета оценки парного коэффициента корреляции Пирсона в соответствии с соотношением:

$$r_{ij}^{k} = \sin\frac{\pi\tau_{ij}^{k}}{2}. \tag{11}$$

Аналогично, к коэффициенту парной корреляции Пирсона можно перейти от рангового коэффициента корреляции Спирмена в соответствии с выражением:

$$r_{ij}^{S} = 2\sin\frac{\pi\tau_{ij}^{S}}{6}. \tag{12}$$

5. Исследование статистических свойств различных оценок коэффициентов корреляции.

В связи с наличием соотношений (11) и (12) вызывает интерес, насколько отличаются статистические свойства оценок (11) и (12) от свойств ОМП $\hat{r}_{ij}$ [6]. В частности, насколько отличаются их распределения при конкретных объёмах выборок.

На рис. 1 представлены функции распределения ОМП $\hat{r}_{ij}$ (1) парного коэффициента корреляции Пирсона и распределения оценок (11) и (12) этого коэффициента, полученных из ранговых коэффициентов Спирмена и Кендалла в случае справедливости гипотезы $H_0 : r_{ij} = 0$. Количество экспериментов было принято равным 10000.

Как видим, функции распределений приведенных оценок показывают приемлемую близость, но при этом оценка дисперсии парного коэффициента корреляции Пирсона имеет меньшее значение, чем дисперсии оценок (11) и (12).

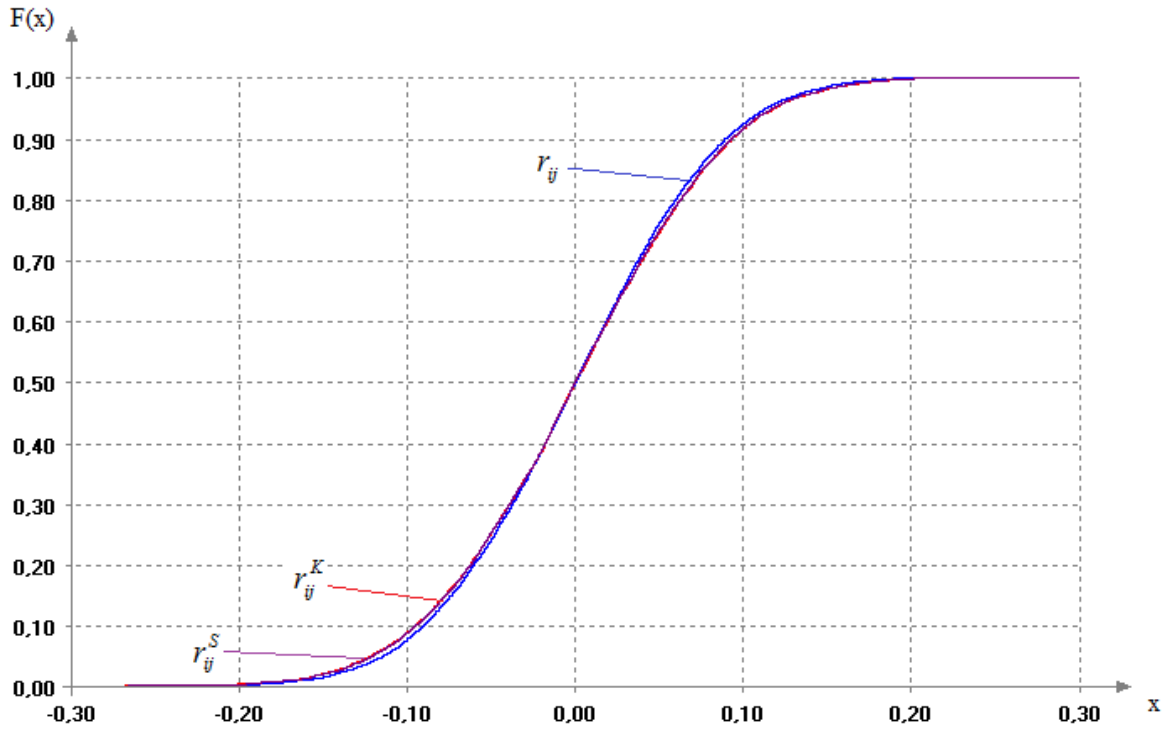

Рис. 1. Функции распределений ОМП $\hat{r}_{ij}$ парного коэффициента корреляции Пирсона и оценок (11) и (12) при истинном значении корреляции $r = 0.0$ и объеме выборки $n = 200$

На рис. 2 демонстрируется зависимость распределений рассматриваемых оценок от объемов выборки. Как показано на рисунке, при малых объемах выборок ($n = 10$) у распределений оценок (11) и (12), вычисляемых на основании ранговых коэффициентов корреляции Спирмена и Кендалла, ярко выражена дискретность, в связи с чем они существенно отличаются от распределений ОМП $\hat{r}_{ij}$ коэффициента Пирсона. С увеличением объема выборок ступенчатость функций распределений оценок (11) и (12), вызванная использованием рангов в этих оценках, исчезает, а сами распределения оказываются ближе к распределениям ОМП $\hat{r}_{ij}$.

В качестве положительного факта можно отметить несмещенность оценок (11), (12) при $H_0 : r_{ij} = 0$.

С увеличением модуля истинного коэффициента корреляции (при $r_{ij} = r$) различие между распределениями ОМП $\hat{r}_{ij}$ и оценок (11) и (12) становятся более явными. На рис. 3

приведены плотности распределений оценок при истинном значении парного коэффициента корреляции $r_{ij} = 0.5$, а на рис. 4 – при $r_{ij} = 0.99$.

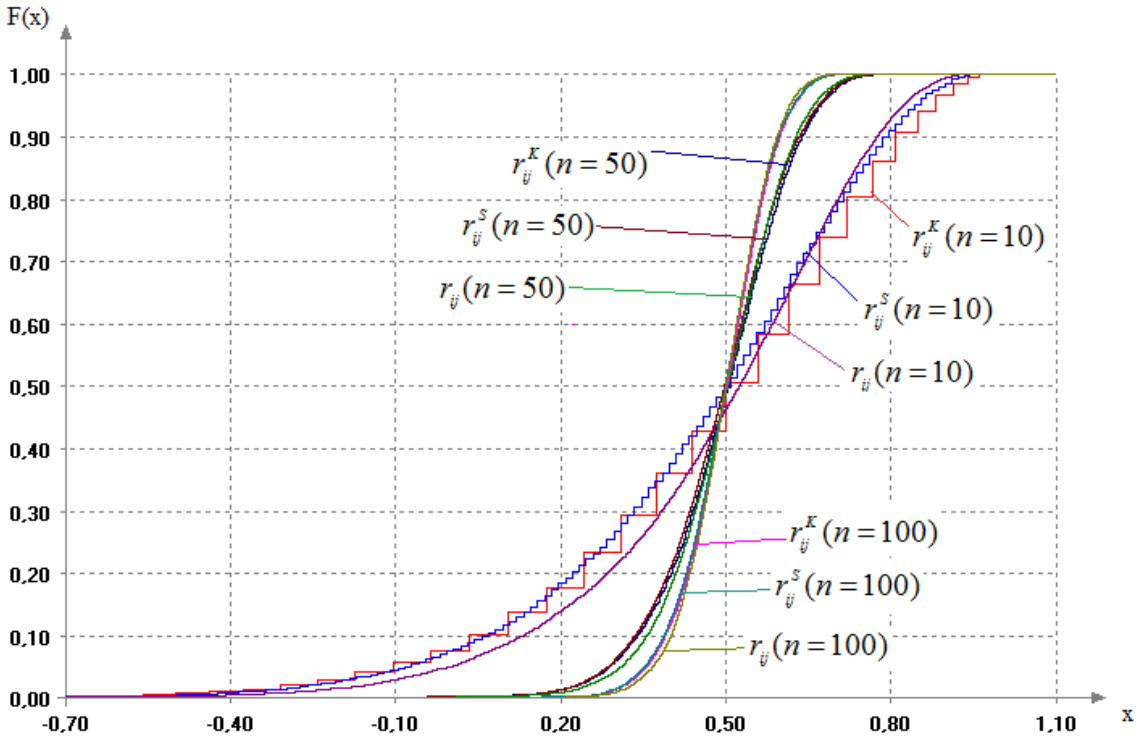

Рис. 2. Функции распределения ОМП $\hat{r}_{ij}$ парного коэффициента корреляции Пирсона и распределения оценок (11) и (12) при $r = 0.5$, $n = 10, 50, 100$

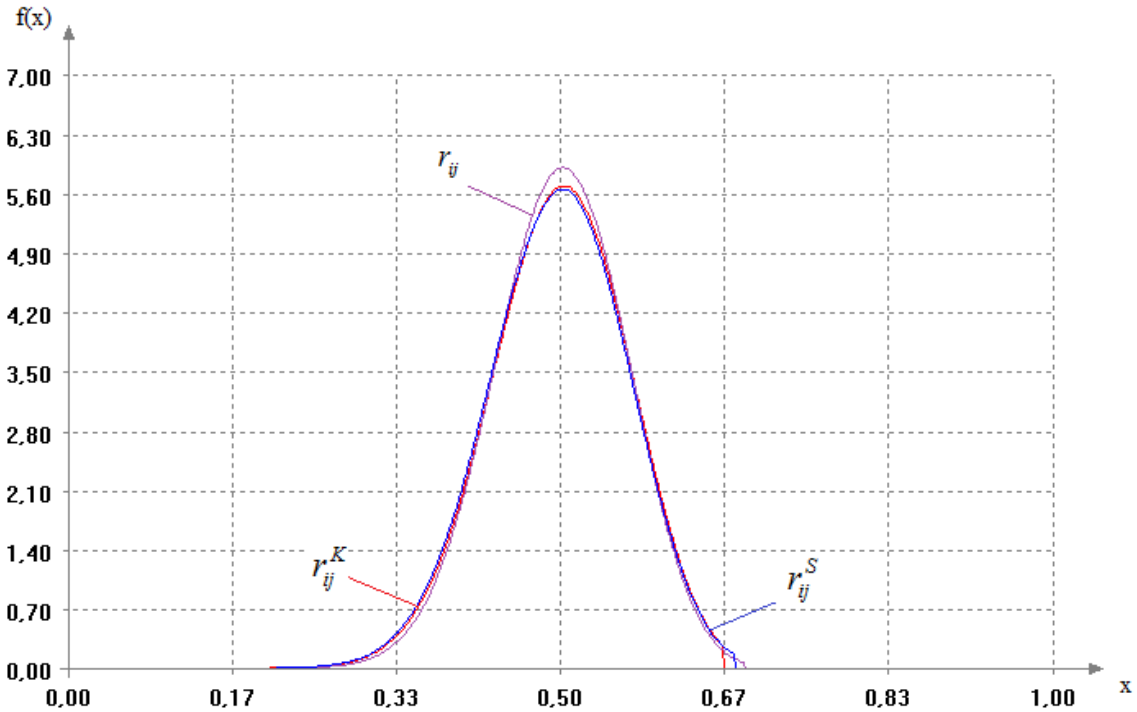

Рис 3. Плотности распределений ОМП $\hat{r}_{ij}$ парного коэффициента корреляции Пирсона и его оценок (11) и (12) при истинном значении $r = 0.5$, $n = 200$

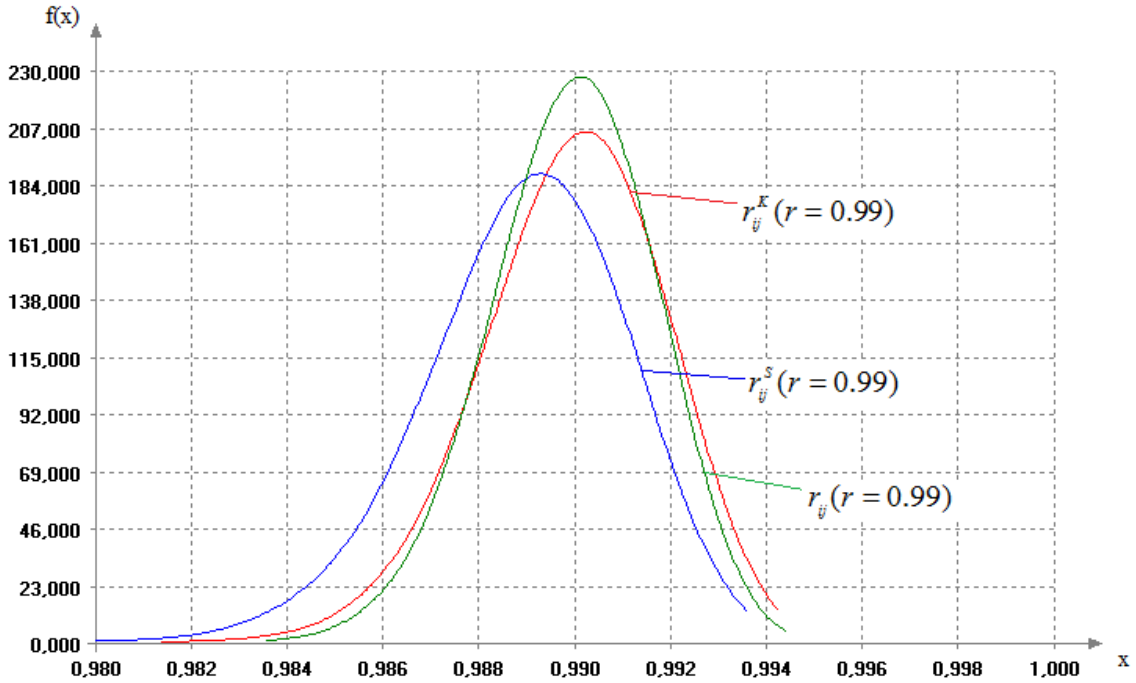

Рис. 4. Плотности оценки ОМП $\hat{r}_{ij}$ парного коэффициента корреляции и плотности оценок (11) и (12) при $r=0.99$, $n=200$

При истинном значении $r=0.99$ наблюдается явная асимметрия и сдвиг в распределениях. Плотность распределения оценки (11), полученной из коэффициента Спирмена, имеет сильный сдвиг влево относительно распределения ОМП коэффициента корреляции Пирсона, а распределение оценки (12), вычисляемой по коэффициенту Кендалла – сдвиг вправо. Можно обратить внимание и на то, что дисперсии оценок (11) и (12) больше дисперсии распределения ОМП $\hat{r}_{ij}$.

6. Сравнительный анализ мощности критериев проверки гипотез об отсутствии корреляции

Преимущества того или иного критерия при заданной вероятности ошибки 1-го рода α (отклонить верную гипотезу $H_0: r_{ij}=0$) можно судить по величине мощности $1-\beta$, где β – вероятность ошибки 2-го рода (не отклонить гипотезу $H_0: r_{ij}=0$ при справедливости конкурирующей гипотезы $H_1: r_{ij}=r$).

В приведенной ниже таблице 1 для уровней значимости $\alpha=0.1;\ 0.05;\ 0.01$ представлены полученные оценки мощности критериев, опирающихся на статистику (2) в случае использования коэффициента корреляции Пирсона, на статистику (5) в случае использования рангового коэффициента корреляции Спирмена, и на статистику (9) – в случае коэффициента корреляции Кендалла.

Полученные в результате исследования оценки мощности критериев, в статистиках которых используются оценки ранговых коэффициентов корреляции Спирмена и Кендалла, в большинстве случаев оказываются близкими к оценкам мощности классического критерия со статистикой (2), но всё-таки всегда уступают последнему.

В то же время можно констатировать, что в случае принадлежности выборок нормальному закону использование критериев со статистиками (5) и (9) вполне оправдано.

Таблица 1. Мощность критериев проверки гипотезы вида $H_0: r_{ij}=0$ относительно конкурирующей

гипотезы вида $H_1 : r_{ij} = r$

Истинное значение корреляции	Статистика с коэффициентом корреляции	Ошибка первого рода α	Объем выборки			
			10	50	100	200
0,1	Пирсона	0.1	0.159	0.279	0.396	0.559
		0.05	0.085	0.164	0.263	0.423
		0.01	0.020	0.055	0.084	0.194
	Спирмена	0.1	0.155	0.273	0.368	0.537
		0.05	0.088	0.164	0.240	0.400
		0.01	0.018	0.044	0.072	0.175
	Кендалла	0.1	0.164	0.274	0.368	0.538
		0.05	0.089	0.164	0.244	0.400
		0.01	0.025	0.047	0.076	0.169
0,3	Пирсона	0.1	0.344	0.805	0.964	0.999
		0.05	0.218	0.683	0.925	0.997
		0.01	0.069	0.431	0.766	0.983
	Спирмена	0.1	0.320	0.773	0.948	0.998
		0.05	0.200	0.654	0.899	0.994
		0.01	0.057	0.374	0.703	0.969
	Кендалла	0.1	0.329	0.768	0.948	0.998
		0.05	0.204	0.653	0.900	0.994
		0.01	0.074	0.382	0.710	0.967
0,5	Пирсона	0.1	0.614	0.994	1.000	1.000
		0.05	0.459	0.984	0.999	1.000
		0.01	0.208	0.934	0.996	1.000
	Спирмена	0.1	0.558	0.989	1.000	1.000
		0.05	0.408	0.974	0.999	1.000
		0.01	0.155	0.890	0.996	1.000
	Кендалла	0.1	0.564	0.988	1.000	1.000
		0.05	0.412	0.975	0.999	1.000
		0.01	0.192	0.895	0.996	1.000

Можно предполагать, что эффект непараметричности критериев со статистиками (5) и (9) играет некоторую положительную роль в случае принадлежности выборок законам, отличным от нормального. При проверке гипотез вида $H_0 : r_{ij} = 0$ это не имеет существенного значения, так как классический критерий со статистикой (2) устойчив к нарушению стандартного предположения о нормальности [7]. Однако при проверке гипотез вида $H_0 : r_{ij} = r$ критерии с ранговыми коэффициентами корреляции должны оказаться полезными.

6. Заключение

Таким образом, в данной работе исследованы статистические свойства оценок парных коэффициентов корреляции, ранговых коэффициентов корреляции, распределения этих оценок в зависимости от истинных значений корреляции и объёмов выборок.

Проведен сравнительный анализ мощности критериев, предназначенных для проверки гипотез вида $H_0 : r_{ij} = 0$ и опирающихся на различные коэффициенты корреляции.

Показано, что мощность критериев, использующих оценки ранговых коэффициентов корреляции Спирмена и Кендалла, не сильно уступает мощности классического критерия проверки гипотезы $H_0 : r_{ij} = 0$, опирающегося на ОМП коэффициента корреляции Пирсона.

Литература

1. *Андерсон Т.* Введение в многомерный статистический анализ. / Т. Андерсон. – М. : Физматгиз, 1963. – 500 с.
2. *Кендалл М.* Теория распределений. / М. Кендалл, А. Стьюарт. – М. : Наука, 1966. – 588 с.
3. *Кендалл М.* Статистические выводы и связи / М. Кендалл, А. Стьюарт. – М. : Наука, 1973. – 900 с.
4. *Помадин С.С.* Исследование распределений статистик многомерного анализа данных при нарушении предположений и нормальности. Диссертация на соискание ученой степени кандидата технических наук. – Новосибирск : НГТУ, 2004. – 136 с.
5. *Лемешко Б.Ю.* Статистический анализ данных, моделирование и исследование вероятностных закономерностей : Монография / Б.Ю. Лемешко, С.Н. Постовалов, Е.А. Чимитова. – Новосибирск : НГТУ, 2011. – 888 с.
6. *Лемешко Б.Ю., Танасейчук А.В.* Исследование распределения оценок коэффициента корреляции в зависимости от истинного значения корреляции // Материалы международной конференции «Актуальные проблемы электронного приборостроения» АПЭП-2006. Т.6, Новосибирск, 2006. – С. 91-94.

Лемешко Борис Юрьевич
Д.т.н., профессор, г.н.с. кафедры теоретической и прикладной информатики НГТУ (630073, Новосибирск, просп. Карла Маркса, 20), e-mail: Lemeshko@ami.nstu.ru

Дамдинжапова Арюна Дашиевна
Магистрант кафедры теоретической и прикладной информатики НГТУ (630073, Новосибирск, просп. Карла Маркса, 20), e-mail: t.aruyna_vik@mail.com.

The study of the statistical properties of estimates and the criteria used in the analysis of correlations

B. Yu. Lemeshko, A.D. Damdinzhapova

Novosibirsk State Technical University

We study the statistical properties of the estimates of rank correlation coefficient Spearman and Kendall and Pearson correlation coefficient pair. Compares power of tests used to test the uncorrelated.

Key words: Spearman's correlation coefficient, Kendall's correlation coefficient, paired Pearson's correlation coefficient, hypothesis testing