РАСПРЕДЕЛЕНИЯ СТАТИСТИК, ПОСТРОЕННЫХ НА ОСНОВЕ КОЭФФИЦИЕНТА РАНГОВОЙ КОРРЕЛЯЦИИ СПИРМЕНА

Танасейчук А.В., Лемешко Б.Ю.
НГТУ, Новосибирск,
Email: awtan@yandex.ru

В настоящее время широкое распространение получили коэффициенты ранговой корреляции, так как в отличие от традиционного коэффициента линейной корреляции Пирсона они позволяют успешно обнаруживать нелинейные зависимости и более устойчивы к ошибкам измерения в реальных данных. Наиболее часто в литературе встречается ранговый коэффициент Спирмена и статистики, построенные на его основе [1-4]. Распределение самого коэффициента Спирмена и возможности применения к нему z-преобразования Фишера были исследованы в работе [5]. В данной статье исследованы распределения двух основных статистик, используемых для проверки гипотезы о значимости коэффициента ранговой корреляции Спирмена.

Для вычисления коэффициента ранговой корреляции Спирмена используют следующую формулу:

$$\rho_S^{kj} = 1 - \frac{6\sum_{i=1}^{n}(r_i^{(k)} - r_i^{(j)})^2}{n(n^2-1)},$$

где $r_i^{(j)}$ – ранг i-го объекта в j-ом наборе данных.

Наиболее распространенная статистика для проверки гипотез относительно коэффициента Спирмена имеет вид:

$$r_1 = \rho_S \cdot \sqrt{n-1}, \qquad (1)$$

где n – объем исследуемой выборки. Статистика (1) подчиняется стандартному нормальному распределению. Также используется статистика вида

$$r_2 = \frac{\rho_S \cdot \sqrt{n-2}}{\sqrt{1-\rho_S^2}}, \qquad (2)$$

которая подчиняется закону распределения Стьюдента с (n–2) степенями свободы, где n – объем исследуемой выборки.

Исследования проводились для малых объемов выборок (от 10 до 40 элементов) и для объема выборки в 100 наблюдений, моделирование проводилось для 5 различных начальных значений генератора случайных чисел, полученные результаты усреднялись. Для проверки согласия использовались три критерия: критерий отношения правдоподобия, критерий Хи-квадрат Пирсона (асимптотически оптимальное группирование для 7 интервалов) и критерий Колмогорова. Исследования проводились с использованием методики компьютерного моделирования статистических закономерностей [6-7].

В следующих таблицах мы приводим полученные результаты проверки согласия с теоретическими распределениями для статистик (1) и (2):

Таблица 1. Степень согласия распределения статистики (1) с теоретическим.

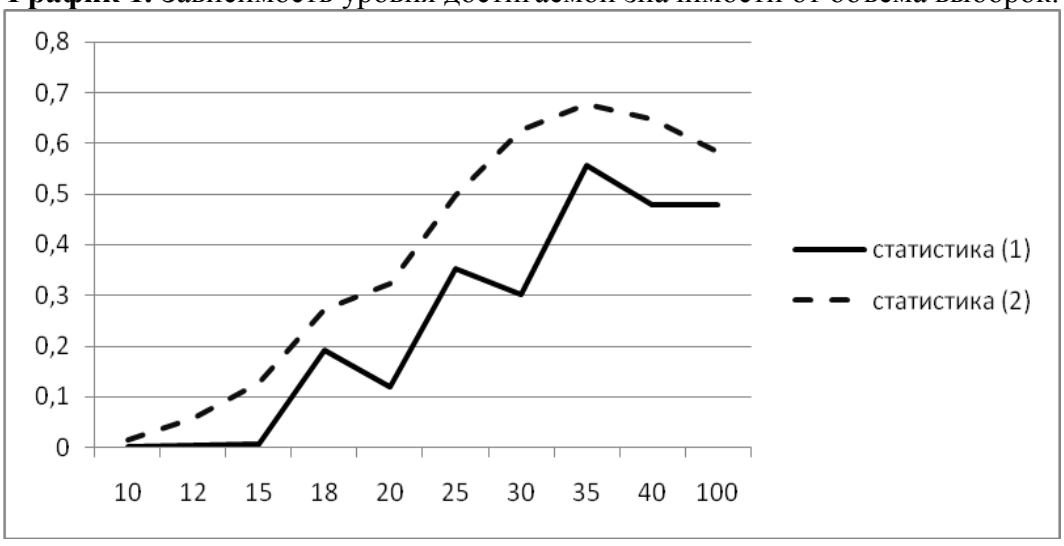
критерий / n	10	12	18	40	100
критерий отношения правдоподобия	1,28973e-10	4,74859e-05	0,04930221	0,3612818	0,4990026
критерий Хи-квадрат Пирсона (АОГ, 7 интервалов)	1,86972e-10	4,06633e-05	0,052514414	0,3580344	0,4985574
критерий Колмогорова	0,000100047	0,002443971	0,1918611	0,478708	0,47971

Таблица 2. Степень согласия распределения статистики (2) с теоретическим.

критерий / n	10	12	18	40	100
критерий отношения правдоподобия	0,005351703	0,0161783	0,51615044	0,436334	0,570618
критерий Хи-квадрат Пирсона (АОГ, 7 интервалов)	0,005198959	0,01656126	0,5160539	0,4366316	0,569246
критерий Колмогорова	0,01482522	0,0590765	0,27345	0,646706	0,58441

На следующем графике приведена зависимость уровня достигнутой значимости по критерию Колмогорова от объема выборки для обоих статистик:

График 1. Зависимость уровня достигаемой значимости от объема выборок.



Из приведенных данных видно, что для малых объемов выборок распределение статистики (2) значительно лучше аппроксимируется своим теоретическим распределением, чем распределение статистики (1). Так, если в качестве уровня значимости принять значение $\alpha = 0.1$, то гипотеза о согласии распределения статистики (2) с распределением Стьюдента не отвергается начиная с объемов $n = 10 \div 12$, тогда как в случае статистики (1) согласие со стандартным нормальным распределением достигается лишь начиная с объема $n = 18$.

Таким образом, можно заключить, что для получения наиболее корректных статистических выводов корреляционного анализа с использованием ранговой корреляции Спирмена целесообразно использовать статистику вида (2), имеющую распределение Стьюдента, особенно в случаях, когда исследователь имеет дело с малыми объемами выборок.

Литература

1. Iman R.L., Conover W.J. *A distribution-free approach to inducing rank correlation among input variables.* Communications in Statistics – Simulation and Computation, 1982, pp. 311-334.
2. Ramsey P. H. *Critical Values for Spearman's Rank Order Correlation.* Journal of Educational and Behavioral Statistics, 1989, pp. 245-253.
3. Zayed H., Quade D. *On resistance of rank correlation.* Journal of statistical computation and simulation, 1997, pp. 59-81.
4. Kotlyar M., Fuhrman S., Ableson A., Somogyi R. *Spearman Correlation Identifies Statistically Significant Gene Expression Clusters in Spinal Cord Development and Injury.* Neurochemical Research, 2004, pp. 1133-1140.
5. 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.
6. Лемешко Б.Ю. *Компьютерные методы исследования статистических закономерностей.* Информационные системы и технологии: ИСТ`2000: Сб. научн. ст. – Новосибирск. 2001. – С.26-41.
7. Лемешко Б.Ю., Постовалов С.Н. *Система статистического анализа наблюдений и исследования статистических закономерностей.* Сб. "Моделирование, автоматизация и оптимизация наукоемких технологий". - Новосибирск: изд-во НГТУ, 2000. - С. 44-46.