СТАТИСТИЧЕСКИЙ АНАЛИЗ РЕЗУЛЬТАТОВ КОНТРОЛЬНЫХ РАБОТ СТУДЕНТОВ ПО ТЕОРИИ ВЕРОЯТНОСТЕЙ

Соболева Е.С.
науч. рук. к.т.н. Ю. Я. Кацман
Томский политехнический университет, Институт кибернетики
ess18@tpu.ru

Введение

В настоящий момент методы сбора и обработки числовых данных, которые являются сутью статистического исследования, нужны для повседневной жизни в современном цивилизованном обществе. Аппарат математической статистики является наиболее мощным инструментом для исследования закономерностей и отсеивания случайностей.

Целью данной работы является исследование статистических взаимосвязей между оценками (переменными), полученными в результате выполнения контрольных работ. Задачей данной работы является измерение тесноты связи двух признаков (переменных) между собой (корреляция Пирсона и ранговые корреляции) для построения в дальнейшем регрессионной модели (линейной и/или нелинейной). Все исследования проведены с использованием пакета Statistica [1].

Коэффициент корреляции Пирсона

Корреляционный анализ позволяет установить степень взаимосвязи двух и более случайных величин.

Наиболее простым способом нахождения тесноты связи переменных является корреляция Пирсона. Корреляция Пирсона позволяет определить силу линейной зависимости между величинами. В исходных данных были представлены баллы (оценки) по 4 контрольным. Результаты анализа представлены в табл.1.

Таблица 1. Корреляционная матрица

Pair of Variables	Valid N	Pearson P	p-level
Score1 & Score1			
Score1 & Score2	160	0,2753	0,000
Score1 & Score3	160	0,4238	0,000
Score1 & Score4	160	0,2683	0,001
Score2 & Score1	160	0,2753	0,000
Score2 & Score2			
Score2 & Score3	160	0,2978	0,000
Score2 & Score4	160	0,3102	0,000
Score3 & Score1	160	0,4238	0,000
Score3 & Score2	160	0,2978	0,000
Score3 & Score3			
Score3 & Score4	160	0,3866	0,000
Score4 & Score1	160	0,2683	0,001
Score4 & Score2	160	0,3102	0,000
Score4 & Score3	160	0,3866	0,000
Score4 & Score4			

Коэффициент корреляции Пирсона (P) приведенный в третьем столбце значимо отличается от ноля для всех пар переменных. То есть при уровне значимости α= 0.05 нулевую гипотезу можно принять с вероятностью менее 0.001. Так как коэффициент корреляции отличается от ±1, это свидетельствует о нелинейной зависимости переменных.

Корреляция Пирсона корректна только для переменных, распределенных по нормальному закону в непрерывных шкалах. Учитывая, что исходные данные (оценки) даны в рангах, а сами переменные отличаются от гауссовых, в работе были получены коэффициенты ранговых корреляций.

Коэффициент ранговой корреляции Спирмена

Коэффициент ранговой корреляции Спирмена находится по формуле:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(r_i - s_i)^2}{n^3 - n},$$

где r_i и s_i – ранги i-го объекта по двум переменным, n– число пар наблюдений [2].

Данный коэффициент показывает степень статистической зависимости между двумя переменными.

При полной линейной связи коэффициент Спирмена равен единице. При обратной связи он равен -1. В любом другом случае коэффициент Спирмена по модулю будет меньше единицы.

При проверке значимости p использовался тот факт, что в случае справедливости нулевой гипотезы об отсутствии корреляционной связи между переменными при $n > 10$статистика

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}},$$

где имеет t-распределение Стьюдента с k = n – 2 степенями свободы. Поэтому ρ значим на уровне α, если фактически наблюдаемое значение t по абсолютной величине будет больше критического, приведенного в таблице распределений Стьюдента при данном количестве степеней свободы на данном уровне значимости [2]. Результаты исследований представлены в табл. 2.

Таблица 2. Ранговые корреляции Спирмена

Pair of Variables	Valid N	Spearman R	p-level
Score1 & Score1			
Score1 & Score2	160	0,270030	0,000554
Score1 & Score3	160	0,420093	0,000000
Score1 & Score4	160	0,274177	0,000451
Score2 & Score1	160	0,270030	0,000554
Score2 & Score2			
Score2 & Score3	160	0,285398	0,000254
Score2 & Score4	160	0,305214	0,000087
Score3 & Score1	160	0,420093	0,000000
Score3 & Score2	160	0,285398	0,000254
Score3 & Score3			
Score3 & Score4	160	0,376949	0,000001
Score4 & Score1	160	0,274177	0,000451
Score4 & Score2	160	0,305214	0,000087
Score4 & Score3	160	0,376949	0,000001
Score4 & Score4			

Сравнив полученные значения распределения Стьюдента с табличными данными, можно сделать вывод о том, что коэффициент ранговой корреляции на 5% -ном уровне значимо отличается от ноля (верна альтернативная гипотеза). Значит связь между результатами контрольных достаточно тесная.

Коэффициент ранговой корреляции Кенделла

Коэффициент ранговой корреляции Кенделла вычисляется по формуле:

$$\tau = 1 - \frac{4K}{n(n-1)},$$

где K – статистика Кенделла [2].

Для определения K необходимо ранжировать объекты по одной переменной в порядке возрастания рангов и определить соответствующие им ранги по другой переменной. Статистика K равна общему числу инверсий (нарушений порядка, когда большее число стоит слева от меньшего) в ранговой последовательности (ранжировке). При полном совпадении двух ранжировок получим $K = 0$ и $\tau = 1$; при полной противоположности можно показать, что $\tau = -1$. Во всех остальных случаях $|\tau| < 1$.

Таблица 3. Ранговые корреляции Кенделла

Pair of Variables	Valid N	Kendall Tau	p-level
Score1 & Score1			
Score1 & Score2	160	0,281929	0,000016
Score1 & Score3	160	0,424211	0,000000
Score1 & Score4	160	0,277636	0,000026
Score2 & Score1	160	0,281929	0,000016
Score2 & Score2			
Score2 & Score3	160	0,287798	0,000008
Score2 & Score4	160	0,320601	0,000001
Score3 & Score1	160	0,424211	0,000000
Score3 & Score2	160	0,287798	0,000008
Score3 & Score3			
Score3 & Score4	160	0,371252	0,000000
Score4 & Score1	160	0,277636	0,000026
Score4 & Score2	160	0,320601	0,000001
Score4 & Score3	160	0,371252	0,000000
Score4 & Score4			

При проверке значимости τ исходят из того, что в случае справедливости нулевой гипотезы об отсутствии корреляционной связи между переменными (при $n > 10$) τ имеет приближенно нормальный закон распределения с математическим ожиданием, равным нулю, и средним квадратическим отклонением

$$S_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$$ [2].

Поэтому τ значим на уровне α если значение статистики

$$t = \frac{\tau - 0}{S_\tau} = \tau\sqrt{\frac{9n(n-1)}{2(2n+5)}}$$ [2]

больше критического значения, приведенного в таблице.

Анализ полученных результатов

Сравнивая значения коэффициентов, приведенных в таблицах 1-3, можно сделать вывод о том, что коэффициенты, рассчитываемые по формулам Пирсона, Спирмена и Кенделла имеют различия во 2-3 знаке после запятой. Таким образом, как для корреляции Пирсона, так и для ранговых корреляций Кенделла и Спирмена на уровне значимости $\alpha = 0.05$ следует признать верной альтернативную гипотезу – коэффициент корреляции значимо отличается от ноля.

Заключение

С помощью программы Statistica был проведен статистический анализ реальных данных. Исходные данные – оценки студентов кафедры ВТ за контрольные работы по теории вероятностей. Полученные результаты говорят о том, что между результатами сдачи контрольных работ существует значимая корреляция. Следующим этапом исследований предполагается построение и анализ регрессионной модели. В связи с тем, что в полученных результатах коэффициент корреляции не более 0.5, исследования будут проводиться в рамках нелинейного регрессионного анализа.

Список использованной литературы

1. Боровиков В. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. 2-е изд. – СПб.: Питер, 2003. – 688 с.

2. Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит; 2006. – 626-628 с.