УДК 004.78

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ ЧИСЛОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Хоменок К.Г.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

Научный руководитель: Ролич О.Ч. – канд. техн. наук, доцент, доцент кафедры ПИКС

Аннотация. В данной работе исследуется корреляционный анализ числовых последовательностей, который включает в себя степень линейной связи между двумя числовыми последовательностями.

Ключевые слова: корреляционный анализ, коэффициент корреляции, линейная связь

Введение. В информационных системах корреляционный анализ числовых последовательностей является одним из наиболее распространенных методов анализа данных, который используется для определения степени линейной зависимости между двумя числовыми последовательностями. Данный метод находит применение в различных областях науки и практики, таких как экономика, медицина, социология и т.д.

Основная часть. Одним из основных методов, используемых для анализа связи между числовыми переменными, является корреляционный анализ. Этот метод позволяет определить, насколько сильно две или более переменных связаны между собой [1].

Для проведения корреляционного анализа необходимо иметь две или более переменные, которые измеряются в количественных единицах. Примерами таких переменных могут быть число проданных товаров и объем рекламных затрат, время, затраченное на выполнение задания и качество выполнения, уровень дохода и степень образования и т.д.

Для расчета корреляции используются коэффициенты корреляции, такие как Пирсонов коэффициент корреляции и ранговый коэффициент Спирмена.

Коэффициент корреляции Пирсона используется, в том случае, когда данные имеют нормальное распределение и относятся к метрической шкале измерений. Данный коэффициент измеряет линейную связь между двумя переменными и принимает значения от -1 до 1. Значение 1 означает, что между переменными существует положительная линейная связь, значение -1 означает, что между переменными существует отрицательная линейная связь, а значение 0 означает отсутствие линейной связи между переменными. Коэффициент корреляции Пирсона может быть вычислен по формуле, которую я упоминал в предыдущем ответе [2].

Иначе, коэффициент корреляции Спирмена используется, только тогда, когда данные не имеют нормального распределения или не относятся к метрической шкале измерений. Этот коэффициент измеряет монотонную связь между двумя переменными и принимает значения от -1 до 1, где 1 означает сильную монотонную связь, -1 означает сильную обратную монотонную связь, а 0 означает отсутствие монотонной связи между переменными. Коэффициент корреляции Спирмена может быть вычислен на основе ранговых позиций значений переменных.

Отсюда стоит сделать вывод, что для корреляционного анализа числовых последовательностей лучшего всего использовать коэффициент корреляции Пирсона, поскольку числовые последовательности обычно имеют нормальное распределение и измеряются в метрических единицах. Однако, если данные не соответствуют этим условиям, то может быть целесообразным использовать коэффициент корреляции Спирмена.

Таким образом, коэффициенты корреляции Пирсона и Спирмена – два распространенных метода для измерения степени связи между двумя переменными.

В этом контексте, необходимо рассмотреть оба коэффициента корреляции: Спирмена и Пирсона, который является непараметрической мерой связи между двумя переменными. Коэффициент корреляции Спирмена измеряет монотонную связь, тогда как коэффициент корреляции Пирсона – линейную.

Коэффициент корреляции Спирмена вычисляется по формуле *1*:

$$p = 1 - \frac{6}{n(n-1)(n+1)} \sum_{i=1}^{n} (R_i - S_i)^2 \quad (1)$$

В данном случае n – количество наблюдений в выборке; R_i – ранг наблюдения x_i в ряду x; S_i – ранг наблюдения y_i в ряду Y; коэффициент p принимает значения из отрезка [-1; 1].

В формуле *1* переменные R и S – это ранги значений каждой из двух переменных. Ранг – это порядковый номер значения в упорядоченной последовательности. Например, если имеется последовательность [4, 7, 2, 5], то ранги будут [2, 4, 1, 3].

Дробь в данном выражении является нормализующим коэффициентом, который делит сумму квадратов разностей рангов на определенной значение, чтобы коэффициент корреляции мог принимать допустимые значения: от -1 до 1. Сумма квадратов разностей рангов между двумя переменными означает сумму квадратов отклонений от среднего ранга.

Далее, необходимо рассмотреть коэффициент корреляции Пирсона. Пирсонов коэффициент корреляции используется для измерения линейной связи между двумя непрерывными переменными.

Пусть даны две выборки $x^m = (x_1, \dots, x_m), y^m = (y_1, \dots, y_m)$; коэффициент корреляции Пирсона рассчитывается по формуле *2* [3]*:*

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i + \underline{x})(y_i - \underline{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \underline{x})^2 \sum_{i=1}^{m}(y_i - \underline{y})^2}} = \frac{cov(x,y)}{\sqrt{s_x^2 s_y^2}} \quad (2)$$

где $\underline{x}, \underline{y}$ – выборочные средние x^m и y^m; s_x^2, s_y^2 – выборочные дисперсии; $r_{xy} \in [-1; 1]$.

Коэффициент корреляции Пирсона называют также теснотой линейной связи: $|r_{xy}| = 1 \Rightarrow x, y$ – линейно зависимы; $r_{xy} = 0 \Rightarrow x, y$ – линейно независимы.

Формула *2* коэффициента корреляции Пирсона между двумя выборками x и y. В числителе стоит ковариация между x и y, а в знаменателе – произведение среднеквадратических отклонений (стандартных отклонений) каждой из выборок.

Таким образом, оба коэффициента корреляции, Пирсона и Спирмена, являются мерами степени линейной связи между двумя переменными. Однако, они имеют различные применения в зависимости от типа данных и целей анализа.

Одним из этапов анализа является расчет коэффициента корреляции, который может быть посчитан с помощью различных методов. В данной работе мы рассмотрим пример использования языка программирования Python для проведения корреляционного анализа и расчета коэффициента корреляции (см. рисунок *1*).

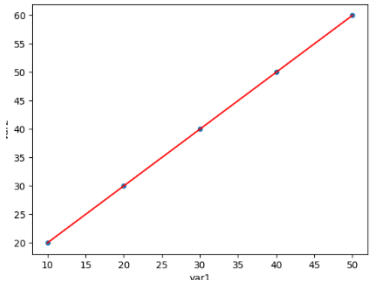

Рисунок *1* – Полученный график корреляционного анализа чисел

Также коэффициент корреляции Пирсона составил 1.0 для следующего DataFrame: var1 = [10, 20, 30, 40, 50]; var2 = [20, 30, 40, 50, 60] (см. рисунок *2)*.

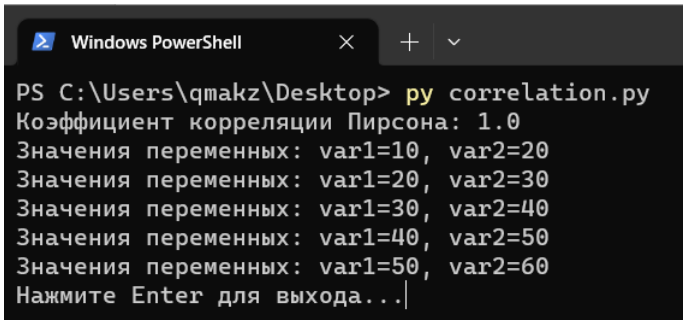

Рисунок *2* – Результат выполнения программного кода

Для корреляционного анализа числовых последовательностей была использована библиотека *Pandas* для работы с данными в виде таблицы, а также библиотека *Matplotlib* для визуализации данных. Сначала был создан *DataFrame*, содержащий две переменные, затем был рассчитан коэффициент корреляции Пирсона с помощью метода *corr*(). Для визуализации данных был использован метод *plot*() с аргументом *kind='scatter'*, который создает диаграмму рассеяния для двух переменных. Наконец, значение коэффициента корреляции было выведено на экран с помощью функции *print*().

Заключение. В данной научной работе был проведен анализ корреляционных связей между числовыми последовательностями. Было показано, что корреляционный анализ является важным инструментом для исследования зависимостей между переменными и может быть использован в различных областях науки. Для проведения корреляционного анализа были использованы статистические методы, такие как коэффициент корреляции Пирсона, который позволяет определить степень линейной зависимости между двумя переменными. Также были использованы инструменты программирования на языке *Python*, такие как библиотека *Pandas* для работы с данными и *Matplotlib* для визуализации результатов.

Список литературы

1. Фаерман, В. А. Корреляционный анализ в методах цифровой обработки сигналов / В. А. Фаерман, В. С. Аврамчук.– 2020. – С. 76 – 78. – Режим доступа: https://www.lib.tpu.ru/fulltext/c/2012/C04/033.pdf – Дата доступа : 15.03.2023.

2. Machinelearning [Электронный ресурс]. : Коэффициент корреляции Спирмена [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Коэффициент корреляции Спирмена . – Дата доступа : 15.03.2023

3. Machinelearning [Электронный ресурс] : Коэффициент корреляции Пирсона [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Коэффициент корреляции Пирсона . – Дата доступа : 15.03.2023

UDC 004.78

CORRELATION ANALYSIS OF NUMERICAL SEQUENCES

Khomenok K.G.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Rolich O.Ch. – PhD, associate professor, associate professor of the Department of ICSD

Annotation. This paper examines the correlation analysis of numerical sequences, which involves the degree of linear relationship between two numerical sequences.

Keywords: correlation analysis, correlation coefficient, linear relationship