

Об ошибках обращения с коэффициентом корреляции Пирсона

Максимова О.В.^{1,2)}

¹⁾ФГБУ «Институт глобального климата и экологии имени академика Ю.А. Израэля»,
Россия, 107258, Москва, ул. Глебовская, 20Б

²⁾Университет МИСИС,
Россия, 119049, Москва, Ленинский пр-кт, 4

*Адрес для переписки: o-maximova@yandex.ru

Реферат. В статье рассматриваются опорные вопросы, важные для понимания при работе с коэффициентом корреляции Пирсона. Всегда ли правомерна и надежна интерпретация полученных результатов по рассчитанному коэффициенту корреляции Пирсона и что нужно при этом учитывать? Ответы на этот и многие другие вопросы даны в статье. Корреляционный анализ ничего не доказывает и не устанавливает причинно-следственных связей. Коэффициент корреляции не может сам по себе без знания природы данных использоваться для демонстрации того, истинна или ложна теория. Приведены частные примеры и контрпримеры, отражающие ключевые свойства рассматриваемого коэффициента корреляции, которые позволяют достигать понимания правильного применения его на практике. Даны общие практические рекомендации с целью получения достоверных выводов.

Ключевые слова. Коэффициент корреляции Пирсона, выборка, выбросы, ложная корреляция, скалярное произведение, линейная регрессия.

On the errors of handling the Pearson correlation coefficient

O.V. Maksimova^{1,2)}

¹⁾Yu.A. Izrael Institute of Global Climate and Ecology,
20B, Glebovskaya str., 107058, Moscow, Russian Federation

²⁾ University of Science and Technology,
4, Leninsky pr., Moscow, 119049, Russian Federation

*Correspondence address: o-maximova@yandex.ru

Abstract. The article discusses the basic questions that are important to understand when working with the Pearson correlation coefficient. Is it always

valid and reliable to interpret the results obtained using the calculated Pearson correlation coefficient and what should be taken into account? Answers to this and many other questions are given in the article. Correlation analysis does not prove anything and does not establish cause-and-effect relationships. The correlation coefficient alone can not, without knowledge of the nature of the data, be used to demonstrate whether a theory is true or false. Specific examples and counter-examples are given that reflect the key properties of the correlation coefficient in question, enabling an understanding of its correct application in practice. General practical advice is given in order to draw valid conclusions.

Keywords. Pearson correlation coefficient, sampling, outliers, false correlation, scalar product, linear regression.

*П.Л. Чебышев и его последователи остаются постоянно
на реальной почве, руководствуясь взглядом,
что только те изыскания имеют цену,
которые вызываются приложениями (научными или
практическими), и только те теории действительно полезны,
которые вытекают из рассмотрения частных случаев.*

А.М. Ляпунов

Вводная часть

Коэффициентом корреляции Пирсона оперируют во многих исследованиях: от биологических до политических. Затрагиваемые здесь вопросы не новы и общеизвестны. Однако, как отмечал в своих работах выдающийся физик Ричард Фейнман, о многих научных истинах имеет смысл задумываться заново и повторять их (Фейнман, 2023).

Обратимся к историческим причинам и задачам, приведшим ученых того времени к необходимости ввести корреляцию. Термин «корреляция» уходит корнями в исследования в области ботаники и антропологии. Ввел этот термин французский палеонтолог Жорж Кювье в XVIII в., чтобы описать «связь как синхронность». Его глубокие познания в анатомии животных позволили восстанавливать облик вымерших существ по их сохранившимся костям. В статистике же слово «корреляция» первым стал использовать выдающийся английский географ, метеоролог и статистик Фрэнсис Гальтон¹⁾ в конце XIX в. Все началось с горошка... (рис.1). Как и Мендель, Гальтон экспериментировал с выращиванием душистого горошка, он проводил анализ размеров его потомства, которое, как оказалось, регрессировало от распределения его родителей (работа Менделя в то время была неизвестна, она была заново открыта в 1900 г.).

¹⁾<http://galton.org/> (дата обращения 11.11.2024)

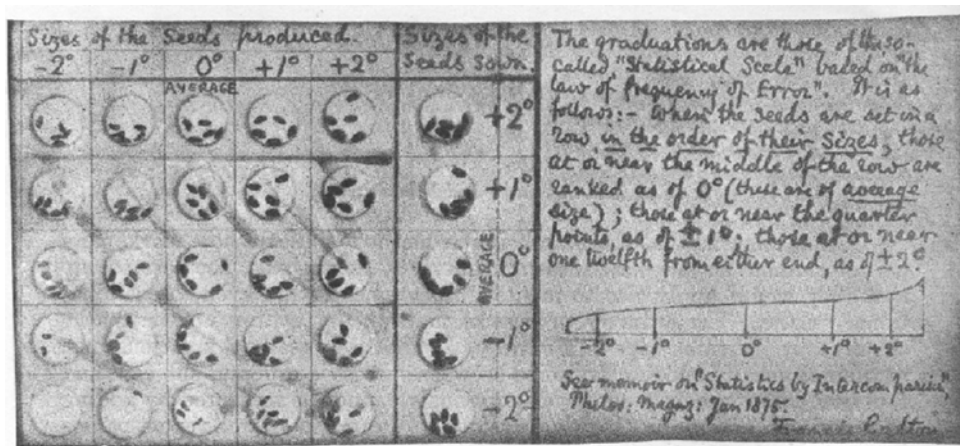


Рисунок 1. Результаты экспериментов с потомством семян душистого горошка, проводимых Френсисом Гальтоном в 1894 г.

Figure 1. Results of experiments with the progeny of sweet pea seeds conducted by Francis Galton in 1894

Позднее Гальтон загорелся идеей измерения характеристик человека, проявляя все больший интерес к наследственности. Однако ему стало ясно, что статистическое исследование данных надежно только при работе с большими выборками, ввиду чего он создал «Антропометрическую лабораторию», которая в 1885 г. на Международной выставке здравоохранения, проходившей в Лондоне, тестировала посетителей этой выставки с помощью инструментов, которые Гальтон часто проектировал сам и изготавливал по своим чертежам (рис. 2). Примером подобной взаимосвязи, найденной Гальтоном, является корреляция длины ноги у пропорционально сложенного человека с длиной руки. Эта большая работа, и более поздние результаты привели к введению теоретических понятий «регрессия» и «корреляция». Но сам Гальтон не мог полноценно математически обработать такой объем данных. Статистический подход Гальтона заинтриговал выдающегося математика того времени, впоследствии и ставшего его другом и учеником, Карла Пирсона, который основал биометрическую школу и обеспечил решающее формирование теоретических концепций, заложенных Гальтоном.

Так начался длинный путь применения до сих пор востребованного коэффициента корреляции Пирсона, который активно применяется и сегодня, в частности, для обоснования надежности результатов наблюдений (Rummel, 1976). Но всегда ли правомерна и надежна интерпретация полученных результатов по рассчитанному коэффициенту корреляции и что нужно при этом учитывать? Ответы на этот и многие другие вопросы будут даны далее.



Рисунок 2. «Антропометрическая лаборатория», которая была включена в Международную выставку здравоохранения, проходившую в Лондоне в 1885 г.

Picture 2. 'Anthropometric Laboratory' which was included in the International Health Exhibition held in London in 1885

Что особенного в формуле коэффициента корреляции Пирсона?

Часто недопонимание основ построения коэффициента корреляции приводит к неверным расчетам, и как следствие, ошибочным прогнозам и сомнительным выводам (Максимова, 2023). Вспомним формулу *выборочного коэффициента корреляции Пирсона* величин X и Y :

$$\hat{r}_{XY} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\hat{\sigma}_X \cdot \hat{\sigma}_Y}, \quad (1)$$

где

$$\hat{\sigma}_X = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}, \quad \hat{\sigma}_Y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}}.$$

В первую очередь отметим, что в коэффициенте корреляции присутствует случайность, так как его находят по данным, в которых также заложена случайность. Поэтому к нему нельзя относиться как к постоянной величине.

Во-вторых, у тех, кто впервые знакомится с формулой (1), возникает вопрос, почему в знаменателе этой формулы стоят именно оценки стандарт-

ных отклонений $\hat{\sigma}_X$ и $\hat{\sigma}_Y$, а не исправленные оценки $s_X = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$ и $s_Y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}$, которые часто используются при анализе выборочных

данных в качестве оценки стандартных отклонений. На самом деле, этот вопрос очень непрост. Отчасти ответ кроется в лучших статистических свойствах выборочного коэффициента корреляции, рассчитанного с применением $\hat{\sigma}_X$ и $\hat{\sigma}_Y$. Выборочный коэффициент корреляции $\hat{r}_{XY}$, рассчитанный по формуле (1) с использованием $\hat{\sigma}_X$ и $\hat{\sigma}_Y$, является состоятельной оценкой коэффициента корреляции, но смещенной. Такое смещение может быть существенным для малых выборок, но величина смещения убывает с увеличением объема выборки и при $n > 30$ уже становится практически пренебрежимой (Кобзарь, 2012). Но, в целом, понятно, что использование s_X и s_Y вместо $\hat{\sigma}_X$ и $\hat{\sigma}_Y$ в формуле (1) может привести к совершенно иным результатам и иногда к дальнейшим ошибочным расчетам корреляционно-регрессионного анализа, так как выборочный коэффициент корреляции и угловой коэффициент прямой линии регрессии связаны между собой формульно. Приведем пример.

Пример 1. На рис. 3 представлена диаграмма рассеяния для смоделированных случайно значений двух признаков X и Y . Коэффициент корреляции, рассчитанный по формуле (1), равен $\hat{r}_{XY} = 0.86$. Однако, если вместо $\hat{\sigma}_X$ и $\hat{\sigma}_Y$ в формулу (1) подставить s_X и s_Y , то получим значение $\hat{r}_{XY} = 0.69$. Данная разница в полученных значениях весьма ощутима для прикладных исследований, особенно в условиях наличия выборок малого объема. Но с ростом n эта разница визуально становится не так заметна. Однако оценивать величину расхождения, как мы увидим в следующем пункте, имеет смысл только совместно с объемом выборки.

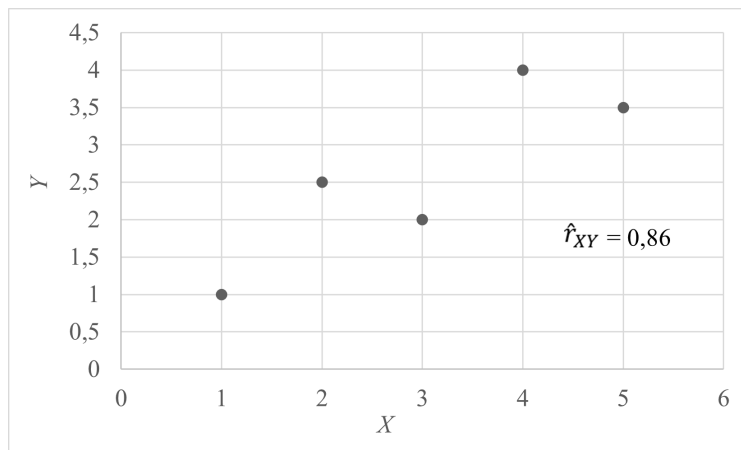


Рисунок 3. Диаграмма рассеяния для смоделированных значений двух признаков X и Y

Figure 3. Scatter diagram for modeled values of two attributes X and Y

Общим выводом служит факт, что потеря полного диапазона значений одной из переменных или недостаточность объема данных могут давать промежуточные, не всегда правдоподобные, результаты, и по мере накопления данных необходимо проверять выдвигаемую изначально гипотезу и соответствующие выводы, увеличивая тем самым их надежность.

Коэффициент корреляции и шкала Чеддока, так ли все однозначно?

Если по выборке значение коэффициента корреляции $\hat{r}_{XY} \neq 0$, то возникает вопрос о том, считать его большим или нет для дальнейшей работы. Во многих прикладных исследованиях для ответа на этот вопрос пользуются шкалой Чеддока (табл. 1). Однако не все так однозначно.

Таблица 1. Шкала Чеддока для оценки тесноты корреляционной связи

Table 1. Cheddock scale for assessing the closeness of the correlation

Количественная мера тесноты связи $ \hat{r}_{XY} $	Качественная характеристика силы связи
0.1-0.3	Слабая
0.3-0.5	Умеренная
0.5-0.7	Заметная
0.7-0.9	Сильная
0.9-1.0	Очень сильная

Дело в том, что коэффициент корреляции чувствителен как к самим данным (а именно, к выделяющимся наблюдениям – выбросам), так и к их количеству (Орлов, 2018; Гржибовский и др., 2020). А критические значения критерия Стьюдента проверки значимости коэффициента корреляции зависят от объема выборки, поэтому использование шкалы Чеддока для интерпретации результата в целом некорректно, и однозначно сразу выводы о его величине делать нельзя. Для признания значимыми небольших коэффициентов корреляции требуются гораздо более объемные выборки, чем для больших. Это хорошо иллюстрирует следующий пример.

Пример 2. На рис. 4 приведены две диаграммы рассеяния модельных данных: а) объема выборки $n = 50$ и $\hat{r}_{XY} = 0.4$, б) объема выборки $n = 8$ и $\hat{r}_{XY} = 0.6$. Несложно проверить, что в первом случае следует признать выборочный коэффициент корреляции $\hat{r}_{XY} = 0.4$ значимым на уровне 5% ($|t_{\text{расч}}| = 3.023 > t_{\text{кр}} = 2.011$), а во втором $\hat{r}_{XY} = 0.6$ – незначимым на уровне 5 % ($|t_{\text{расч}}| = 1.837 < t_{\text{кр}} = 2.447$).

Пример 2 демонстрирует, что ошибочно принимать решение о том, велик ли коэффициент корреляции или нет, без анализа объема выборки и, соответственно, не применив соответствующий статистический критерий проверки его значимости.

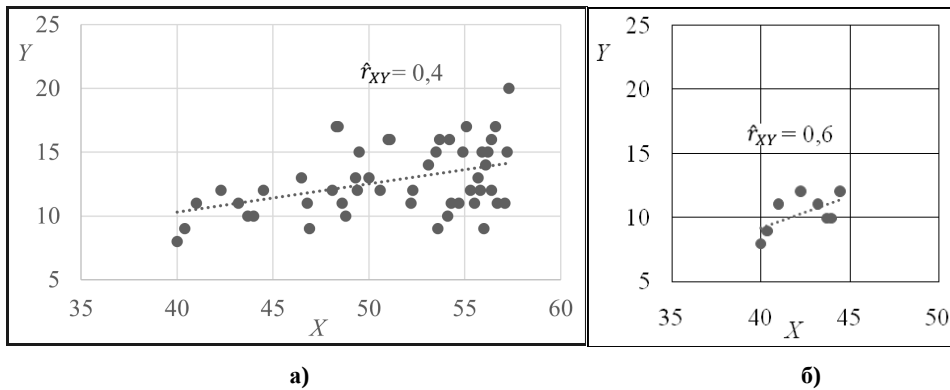


Рисунок 4. Диаграммы рассеяний для: а) объема выборки $n = 50$, б) объема выборки $n = 8$

Picture 4. Scatter plots for: a) sample size $n = 50$, b) sample size $n = 8$

Коэффициент корреляции и данные, распределенные не по нормальному закону

Вообще говоря, коэффициент корреляции Пирсона можно рассчитать между любыми количественными и интересующими исследователя признаками. Это связано с тем, что по своей структуре коэффициент корреляции (1) представляет собой скалярное произведение векторов наблюдений после незначительных преобразований (Максимова, 2023). Однако для ответа на вопрос проверки его значимости необходим критерий Стьюдента, который накладывает ряд ограничений.

Первое ограничение – это объем данных. Установлено, что при объеме выборки $n > 10$ распределение случайной величины $t = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2}$ удовлетворительно аппроксимируется распределением Стьюдента с $(n-2)$ степенями свободы (Кобзарь, 2012).

Второе ограничение – нормальный закон распределения данных. Группа ученых провела исследования распределений статистики Стьюдента корреляционного анализа для законов, отличающихся от нормального (более островершинных и плосковершинных, но симметричных), в которых было показано, что значимого изменения предельных распределений статистик не происходит. Это существенно расширяет сферу корректного применения методов классического корреляционного анализа в приложениях (Лемешко, Помадин, 2002). Но это вовсе не значит, что мы можем пренебрегать законом распределения данных, мы лишь допускаем возможность работы с коэффициентом корреляции и его интерпретации при нарушении этого условия. Поэтому, как отмечалось ранее, важно накапливать данные и проводить повторные исследования для принятия/отвержения выдвинутой гипотезы. Как отмечал великий математик Рональд Фишер: «Нулевая гипотеза никогда не доказывается и не подтверждается, но, возможно, опровергается в ходе

экспериментов. Можно сказать, что каждый эксперимент существует только для того, чтобы факты могли опровергнуть нулевую гипотезу».

Коэффициент корреляции и причинно-следственные связи

Всем известны случаи, когда неверная интерпретация высокого коэффициента корреляции приводит к ложным умозаключениям. И некоторые из таких умозаключений весьма увлекательны, как, например, высокая связь между числом свивших гнезда в южных районах Швеции аистов и рождаемостью в эти же годы (Matthews, 2000), между числом утонувших в бассейне и числом фильмов, выпущенных с Николасом Кейджем (Tyler Vigen, 2015), между количеством экономических преступлений и численностью студентов в регионе (Благовещенский, 2009) и др. Вся причина в том, что расчетная формула (1) дает результат, который показывает сопряженность изменения признаков, но никак не позволяет устанавливать причинно-следственные связи. В случае, когда такая причинно-следственная связь не имеет смысла, говорят о *ложной корреляции* (Благовещенский, 2009). Причины такого феномена могут быть разные:

а) действие третьего фактора, вызывающее одновременный рост двух изучаемых признаков,

б) наличие целых групп (кластеров) или единичных выделяющихся наблюдений, внутри которых корреляции нет, а при этом все они, объединенные в одну выборку, группируются достаточно хорошо около одной прямой и дают высокое значение коэффициента корреляции,

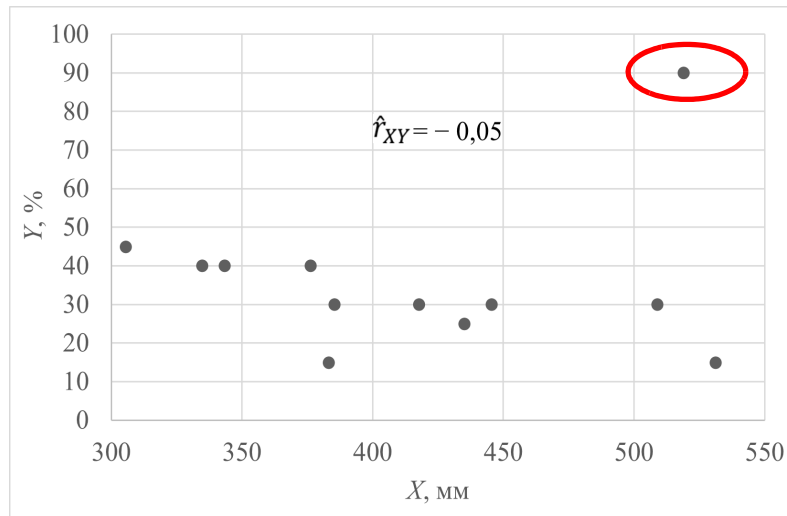
в) случайные совпадения, обусловленные конфигурацией/особенностями выборки.

Приведем далее примеры по каждому случаю.

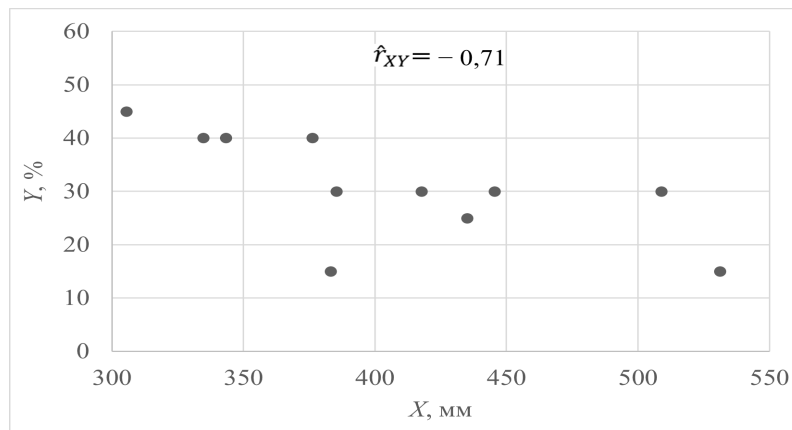
Пример 3. Примером ложной корреляции для случая а) может служить связь между числом преступлений в городе и числом обучающихся студентов (Благовещенский, 2009). Но и тот, и другой признак связаны с численностью населения города, которая напрямую влияет как на численность студентов этого города, так и на количество преступлений. Тем самым, в больших городах, как правило, будет и большее число студентов, и число преступлений, а в малых – наоборот.

Пример 4. Для случая б) приведем модельный пример, основанный на реальных данных. На рис. 5 представлены диаграммы рассеяния связи смоделированных данных суммарных годовых осадков с величиной депигментации хвои сосны обыкновенной с выделяющимся наблюдением (а) и без него (б). Причины такого выделяющегося наблюдения могут быть разные: повреждение дерева бобрами, ошибка в записи и в измерении, прошедший полосой низовой пожар и т.п. После удаления выброса коэффициент корреляции изменится со значения $\hat{r}_{XY} = -0.05$ до значения $\hat{r}_{XY} = -0.71$.

Важно иметь в виду, что в таком случае, если мы собираемся просто выбросить некоторое наблюдение или целый кластер наблюдений без их анализа, то мы теряем заметную часть информации, которая в них спрятана.



а)



б)

Рисунок 5. Диаграммы рассеяний связи моделированных данных суммарных годовых осадков с величиной депигментации: а) с выбросом, б) без выброса

Figure 5. Scatter plots of the relationship of modeled total annual precipitation data with depigmentation value: а) with outlier, б) without outlier

Пример 5. Случай в) встречается редко. Пример такой ситуации приведен в работе (Максимова, Кухта, Коротков, 2023): при анализе воздействия климатических факторов на линейные приросты сосны обыкновенной в заповеднике «Кивач» выявлено наличие значимой связи между осадками и температурами на заданном временном промежутке. Однако известно, что такая связь не носит причинный характер, и ее можно считать случайной.

Итак, корреляционный анализ ничего не доказывает. Коэффициент корреляции не может сам по себе без знания природы данных использоваться для демонстрации того, истинна или ложна теория.

Корреляция во времени и пространстве

Если данные представлены временными рядами, то простой расчет коэффициента корреляции может давать некорректные результаты. Рассмотрим пример, приведенный в работе (Максимова, Кухта, 2024).

Пример 6. Изучается связь двух временных рядов модельных значений. На рис. 6 изображены значения смоделированного примера: слева два ряда во времени (ряд 1 и ряд 2), а справа – диаграмма рассеяния в пространстве для смоделированных значений. Показатели на рис. 6 слева явно демонстрируют как близость данных во времени, так и синхронность по монотонности, которая нарушается лишь для двух временных периодов ($N = 5, N = 15$).

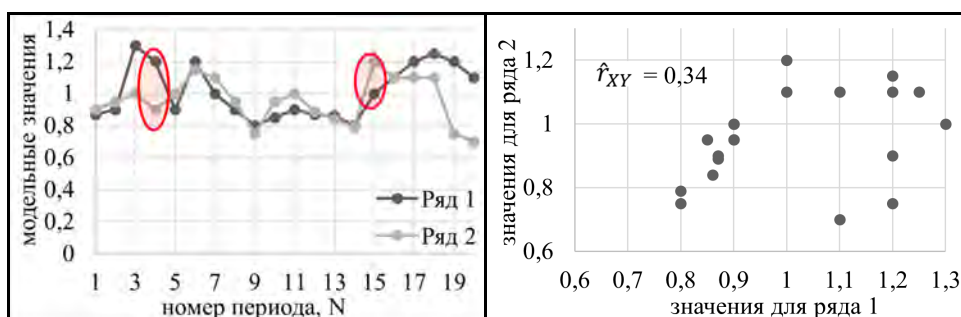


Рисунок 6. Смоделированные значения двух динамических рядов и их диаграмма рассеяния

Picture 6. Simulation values of two dynamic series and their scatter plot

Казалось бы, возникает гипотеза о тесной связи. Однако восприятие результатов меняется, если построить диаграмму рассеяния для рядов 1 и 2, убрав компоненту времени. На рис. 6 справа при удалении временной компоненты диаграмма рассеяния демонстрирует слабую связь: коэффициент корреляции составляет лишь $\hat{r}_{XY} = 0.34$. Тем самым, продемонстрировано, что коэффициент корреляции Пирсона лишь отчасти учитывает сопряженность динамических рядов по характеру монотонности на каждом временном промежутке. В такой постановке задачи нужны специальные методы анализа рядов динамики (Мхитарян и др., 2008).

Диаграмма рассеяния и первичные выводы

Известно, что визуализация данных, пройдя долгий исторический путь совершенствования, часто играет важную роль в разведывательном анализе (Friendly, 2007). На рис. 7 приведены диаграммы рассеяния, составленные для одних и тех же данных, но при разном выбранном масштабе для вертикальной оси. Заметно при этом, как меняется восприятие корреляции, которая составила лишь значение $\hat{r}_{XY} = 0.1$. Это, в свою очередь, показывает важность математических критериев для интерпретации результатов, которая обсуждалась выше.

Критерии не позволяют вводить себя в заблуждение графиками и кажущейся логичностью подобной аргументации. Но, помимо этого, на рис. 7 продемонстрирован и другой интересный факт, который можно встретить преимущественно в теории.

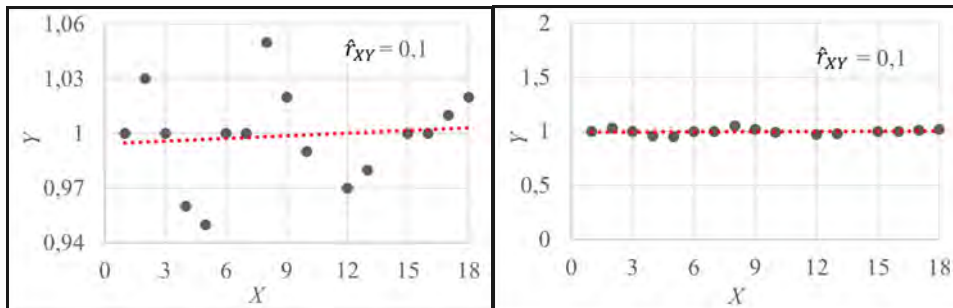


Рисунок 7. Диаграммы рассеяния для модельных данных в разных масштабах

Picture 7. Scatterplots for model data at different scales

Задумаясь, всегда ли «строгая» линейная связь между данными приводит к коэффициенту корреляции Пирсона $r_{XY} = \pm 1$? Невероятно, но – нет! Рассмотрим рис. 8, который можно воспринимать близким к рис. 7. Значения признака Y при изменении признака X на рис. 8 не меняются, однако точки ложатся на прямую. Чему равен коэффициент корреляции в этом случае? С одной стороны, следуя умозаключениям Кьюве, который вводил корреляцию как синхронность признаков, на рис. 8 $r_{XY} = 0$. Но вспоминая, что коэффициент корреляции показывает силу линейной связи, возникает догадка, что $r_{XY} = 1$ или $r_{XY} = -1$. Но ни тот, ни другой вариант неверен.

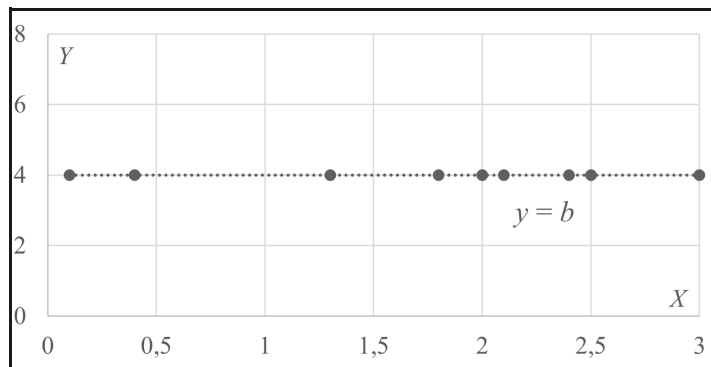


Рисунок 8. Пример линейной регрессии

Figure 8. Linear regression example

Дело в том, что в таком случае $\hat{\sigma}_Y = 0$, и коэффициент корреляции (1) не определен. На практике такая ситуация практически невозможна, за исключением случаев, когда округление значений признака столь грубо, что задает равные значения (и это можно исправить), либо когда действительно значения признака Y не изменяются с ростом X .

Заключение

В работе даны ответы на часто задаваемые вопросы, такие как: почему знак коэффициента корреляции показывает направленность связи и почему так устроена его формула? Насколько правомерно пользоваться шкалой Чеддока без применения математических критериев в прикладных исследованиях? Всегда ли высокий выборочный коэффициент корреляции интерпретируем и значимо отличен от нуля? Насколько оправдан для интерпретации расчет корреляции между признаками во времени для оценки синхронности соответствующих рядов?

Результатом приведенных объяснений может служить общая практическая рекомендация с целью получения достоверных выводов, которая состоит в нескольких постулатах. При исследовании взаимосвязи между признаками необходимо:

- знать природу происхождения и получения данных для дальнейшего анализа; учитывать, проведены ли измерения в пространстве или во времени;
- проанализировать, насколько выявление корреляционной связи оправдано (если причинно-следственная связь не имеет смысла, то полученная связь может демонстрировать ложную корреляцию);
- анализировать первично данные на наличие выделяющихся наблюдений (выбросов) и целых кластеров выделяющихся наблюдений и выявлять их природу для понимания причины дальнейшей стратегии исследования;
- интерпретировать коэффициент корреляции осторожно, так как он носит случайный характер и зависит как от объема выборки, так и от строения самих данных, и пользоваться критериями, так как просто графическая визуализация может приводить к некорректным выводам.

Корреляционный анализ ничего не доказывает. Коэффициент корреляции не может сам по себе без знания природы данных использоваться для демонстрации того, истинна или ложна теория. Поэтому нельзя позволять вводить себя в заблуждение, например, графиками эмпирических наблюдений и кажущейся логичностью такой аргументации. Приведение частных примеров и контрпримеров, отражающих ключевые характерные свойства рассматриваемого коэффициента корреляции, позволяют достигать понимания правильного применения его на практике.

В заключение, вспомним высказывание: «Кто трижды смело предсказывал погоду и угадал, в глубине души немного верит в свой пророческий дар» (*Фридрих Ницше*). А вот повысить точность своего прогноза и принимаемых решений можно, применив иногда даже очень простые статистические инструменты.

Список литературы

Благовещенский, Ю.Н. (2009) *Тайны корреляционных связи в статистике, Серия «Библиотека Солев», М., «Научно-издательский центр ИНФРА-М», «Научная книга», 158 с.*

Гржибовский, А.М., Горбатова, М.А., Наркевич, А.Н., Виноградов, К.А. (2020) Объем выборки для корреляционного анализа, *Морская медицина*, т. 6, № 1, с.101-106, doi10.22328/2413-5747-2020-6-1-101-106.

Кобзарь, А.И. (2012) *Прикладная математическая статистика. Для инженеров и научных работников*, М., ФИЗМАТЛИТ, 816 с.

Лемешко, Б.Ю., Помадин, С.С. (2002) Корреляционный анализ наблюдений многомерных случайных величин при нарушении предположений о нормальности, *Сибирский журнал индустриальной математики*, т. 5, № 3(11), с. 115-130.

Максимова, О.В., Кухта, А.Е. (2024) Об особенностях сравнения динамических рядов параметров хода роста сосны обыкновенной в разных типах местообитаний, *Стандарты и качество*, № 13, с. 55-61.

Максимова, О.В., Кухта, А.Е., Коротков, С.А. (2023) Воздействие черного углерода и других климатических факторов на линейные приросты сосны обыкновенной (*Pinus Sylvestris* L.) на территории заповедника «Кивач», *Лесной вестник/Forestry Bulletin*, т. 27, № 5, с. 48-59, doi:10.18698/2542-1468-2023-5-48-59.

Максимова, О.В. (2023) *Математическая статистика и анализ данных. Учебное пособие*, М., Издательский Дом НИТУ МИСИС, 172 с.

Мхитарян, В.С., Архипова, М.Ю., Балаш, В.А. и др. (2008) *Эконометрика*, М., Изд-во Проспект, 384 с.

Орлов, А.И. (2018) Ошибки при использовании коэффициентов корреляции и детерминации, *Заводская лаборатория. Диагностика материалов*, № 84(3), с. 68-72, doi:10.26896/1028-6861-2018-84-3-68-72.

Фейнман Р. (2023) *Наука, не-наука и все-все-все*, М., АСТ, 192 с.

Friendly, M. (2007) A brief history of data visualization, *Handbook of Data Visualization. Springer Handbooks Comp. Statistics. Springer*, Berlin, Heidelberg, pp. 15-56, https://doi.org/10.1007/978-3-540-33037-0_2

Matthews, R. (2000) Storks Deliver Babies ($p = 0.008$), *Teaching Statistics*, 22(2), pp. 36-38.

Rummel, R.J. (1976) *Understanding Correlation*, ch. 5.

Tyler Vigen *Spurious Correlations*. Hachette Books (2015) 208

p.

References

Blagoveshchenskij, Yu.N. (2009) *Tajny korrelyacionnyh svyazi v statistike* [Secrets of correlation in statistics], Seriya «Biblioteka Soley», «Nauchno-izdatel'skij centr INFRA-M», «Nauchnaya kniga», Moscow, Russia, 158 p.

Grzhibovskij, A.M., Gorbatova, M.A., Narkevich, A.N., Vinogradov, K.A. (2020) Ob'em vyborki dlya korrelyacionnogo analiza [Required sample size for

correlation analysis], *Morskaya medicina*, vol. 6, no. 1, pp. 101-106, doi 10.22328/2413-5747-2020-6-1-101-106.

Kobzar', A.I. (2012) *Prikladnaya matematicheskaya statistika. Dlya inzhenerov I nauchnyh rabotnikov* [Applied mathematical statistics. For engineers and scientists], Fizmatlit, Moscow, Russia, 816 p.

Lemeshko, B.Yu., Pomadin, S.S. (2002) Korrelyacionnyj analiz nablyudenij mnogomernykh sluchajnykh velichin pri narushenii predpolozhenij o normal'nosti [Impact of black carbon and other climatic factors on linear growth of Scots pine (*Pinus Sylvestris* L.) in the territory of the Kivach nature reserve [Correlation analysis of observations of multidimensional random variables in violation of assumptions about normality], *Sibirskii zhurnal industrial'noi matematiki*, vol. 5, no. 3(11), pp. 115-130.

Maksimova, O.V., Kuhta, A.E. (2024) Ob osobennostyah sravneniya dinamicheskikh ryadov parametrov hoda rosta sosny obyknovennoj v raznyh tipah mestoobitanij, [On the Peculiarities of Comparison of Dynamic Series of Parameters of Scots Pine Growth Course in Different Types of Habitats], *Standarty i kachestvo-Standards and Quality*, no.13, pp. 55-61.

Maksimova, O.V., Kuhta, A.E., Korotkov, S.A. (2023) Vozdejstvie chernogo ugleroda I drugih klimaticheskikh faktorov na linejnye prirosty sosny obyknovennoj (*Pinus Sylvestris* L.) na territorii zapovednika «Kivach» [Impact of black carbon and other climatic factors on linear growth of Scots pine (*Pinus Sylvestris* L.) in the territory of the Kivach nature reserve], *Lesnoj vestnik*, vol. 27, no. 5. pp. 48-59, doi: 10.18698/2542-1468-2023-5-48-59.

Maksimova, O.V. (2023) *Matematicheskaya statistika I analiz dannyh: Uchebnoe posobie* [Mathematical statistics and data analysis], Izdatel'skij Dom NITU MISIS, Moscow, Russia, 172 p.

Mhitaryan, V.S., Arhipova, M.Yu., Balash, V.A. i dr. (2008) *Ekonometrika* [Econometrics], Izd-vo Prospekt, Moscow, Russia, 384 p.

Orlov, A.I. (2018) Oshibki pri ispol'zovanii koefficientov korrelyacii i determinacii [Errors in the use of correlation and determination coefficients], *Zavodskaya laboratoriya. Diagnostika materialov-Industrial laboratory. Materials diagnostics*, no. 84(3), pp. 68-72, doi: 10.26896/1028-6861-2018-84-3-68-72.

Feynman R. (2023) *Nauka, ne-nauka i vse-vse-vse* [Science, non-science and everything-everything-everything], Izd-vo AST, Moscow, Russia, 192 p.

Friendly, M. (2007) A brief history of data visualization, *Handbook of Data Visualization. Springer Handbooks Comp. Statistics. Springer*, Berlin, Heidelberg, pp. 15-56. https://doi.org/10.1007/978-3-540-33037-0_2.

Matthews, R. (2000) Storks Deliver Babies ($p=0.008$), *Teaching Statistics*, 22 (2), pp. 36-38.

Rummel, R.J. (1976) *Understanding Correlation*, ch. 5.

Tyler Vigen Spurious Correlations. Hachette Books (2015) 208 p.

Статья поступила в редакцию (Received): 15.11.2024.

Статья доработана после рецензирования (Revised): 25.11.2024.

Для цитирования / Forcitation

Максимова, О.В. (2024) Об ошибках обращения с коэффициентом корреляции Пирсона, *Экологический мониторинг и моделирование экосистем*, т. XXXV, № 3-4, с. 88-102, doi: 10.24412/2782-3237-2024-3-4-88-102.

Maksimova, O.V. (2024) On errors in handling Pearson's correlation coefficient, *Ecological monitoring and ecosystem modelling*, vol. XXXV, № 3-4, с. 88-102, doi: 10.24412/2782-3237-2024-3-4-88-102.