Кафедра математической статистики ИВМиИТ КФУ 25 декабря 2020 г.

Научно-популярная статья к выпуску подготовил Казанцев А.В.

Автор: Мирослав Меньшов, магистрант 2 курса кафедры математической статистики

Коэффициент корреляции Пирсона

Автор статьи – Мирослав Меньшов

Мирослав Меньшов – бакалавр математики (КФУ, 2019). Выпускная квалификационная работа: Меньшов М.А. «Мощностные свойства инвариантного критерия, различающего распределения гамма и обобщенного показательного». Научный руководитель – профессор И.Н. Володин

1. Введение

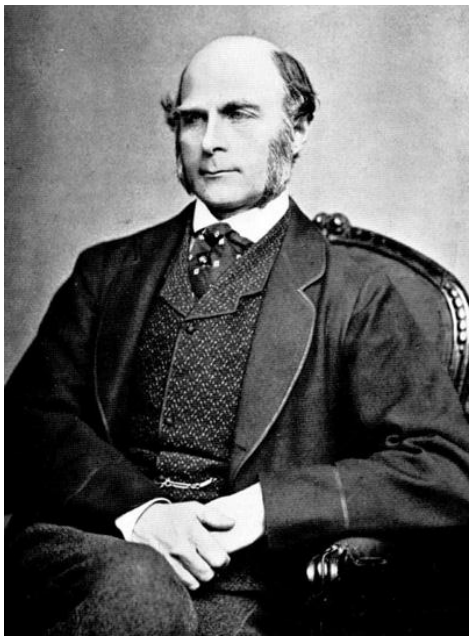
В математической статистике, коэффициент корреляции Пирсона, известный также как коэффициент парной корреляции или коэффициент корреляции произведения моментов Пирсона, представляет собой статистику, которая измеряет величину линейной связи (корреляцию) между двумя переменными. Он принимает значения от -1 до +1. Значение коэффициента +1 означает наличие полной положительной линейной связи, а значение -1 – наличие полной отрицательной линейной связи.

Огюст Браве (1811-1863)

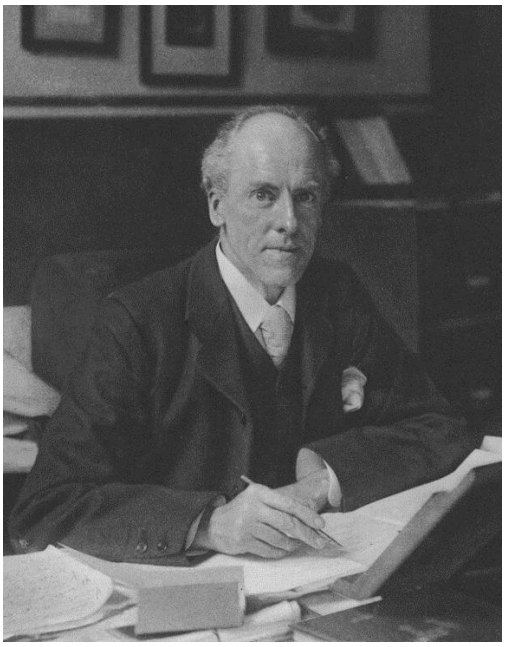
Родоначальниками коэффициента корреляции являются французский физик Огюст Браве и английский ученый Фрэнсис Гальтон. В 1844 году Огюст Браве опубликовал статью о статистической концепции корреляции и пришел к определению коэффициента корреляции.

А Фрэнсис Гальтон независимо заново открыл концепцию корреляции в 1888 году и продемонстрировал ее применение в изучении наследственности, антропологии и психологии.

Пирсон, основываясь на родственных идеях предшественников, разработал коэффициент в терминах «произведения моментов», корреляция Пирсона – это ковариация двух переменных, деленная на произведение их стандартных отклонений, ковариация же представляется как среднее значение произведения средних скорректированных случайных величин, в этом и заключается принцип «произведения моментов». Таким образом, название коэффициента является примером закона Стиглера, который гласит, что ни одно научное открытие не названо в честь его первооткрывателя.



Фрэнсис Гальтон (1822-1911)



Карл Пирсон (1857-1936)

2. Формы представления

1) Коэффициент корреляции Пирсона, применяемый к совокупности, обозначается **ρ** и называется **коэффициентом корреляции Пирсона совокупности**. Для случайных величин X и Y формула вычисления коэффициента **ρ** представляется в следующем виде:

$$\boldsymbol{\rho_{X,Y}} = \frac{\boldsymbol{cov(X,Y)}}{\boldsymbol{\sigma_X \sigma_Y}},$$

где σ_X, σ_Y – стандартные отклонения, соответствующие случайным величинам X и Y, а cov – коэффициент ковариации.

Коэффициент ковариации, стоящий в числителе, может быть выражен в терминах математического ожидания и средних значений, тогда получится следующее преобразование исходной формулы:

$$\boldsymbol{\rho_{X,Y}} = \frac{\boldsymbol{E[(X-\mu_X)(Y-\mu_Y)]}}{\boldsymbol{\sigma_X \sigma_Y}},$$

где μ_X – среднее значение, принимаемое случайной величиной X (аналогично для μ_Y).

Эту формулу можно выразить через нецентрированные моменты, имеем:

$$\boldsymbol{\mu_X = E[X],\ \mu_Y = E[Y],}$$

$$\boldsymbol{\sigma_X^2 = E[(X - E[X])^2] = E[X^2] - (E[X])^2,}$$

$$\boldsymbol{\sigma_Y^2 = E[(Y - E[Y])^2] = E[Y^2] - (E[Y])^2,}$$

$$\boldsymbol{E[(X - \mu_X)(Y - \mu_Y)] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y].}$$

Подставим полученные выражения в преобразованную формулу:

$$\boldsymbol{\rho_{X,Y} = \frac{E[XY] - E[X]E[Y]}{\sqrt{E[X^2] - (E[X])^2}\sqrt{E[Y^2] - (E[Y])^2}}.}$$

2) В случае, когда коэффициент корреляции Пирсона применяется к выборке, он обозначается **r** и называется **коэффициентом корреляции Пирсона выборки**. Формулу для него можно получить, подставив выборочные оценки ковариации и дисперсий в выражение для коэффициента корреляции Пирсона совокупности. Необходимо учитывать парность данных. Имеем:

$$\boldsymbol{r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},}$$

где x_i, y_i – элементы выборки, n – размер выборки, а $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – выборочное среднее значение параметра *x* (аналогично для $\bar{y}$).

Используем формулы выборочных средних и раскрытие сумм, чтобы преобразовать выше представленную формулу:

$$\boldsymbol{r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.}$$

Полученная формула предлагает удобный алгоритм для численного расчета выборочных корреляций в один проход. Но иногда данный алгоритм, при некоторых значениях выборок, может быть вычислительно неустойчивым, возможно накопление ошибок.

Сократим числитель и знаменатель на n, и запишем полученное выражение в более компактном и удобном виде, используя соответствующие формулы выборочных средних значений, получим:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum y_i^2 - n\bar{y}^2}}.$$

Для коэффициента корреляции Пирсона выборки также существует эквивалентная формула, представляющая собой несмещенное среднее значение произведения стандартных оценок рассматриваемых параметров.

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_x}\right)\left(\frac{y_i - \bar{y}}{S_y}\right),$$

где $S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ – стандартное отклонение (несмещенное) выборки для параметра *x* (аналогично S_y).

3. Математические свойства

Рассмотрим математические свойства коэффициента корреляции Пирсона.

1) Диапазон абсолютных значений коэффициентов корреляции Пирсона выборки и совокупности есть интервал от 0 до +1.

2) Коэффициенты, равные +1 и -1, в случае выборочной корреляции, соответствуют точкам двумерных данных, лежащих точно на прямой линии, в случае же корреляции совокупности, они соответствуют двумерной функции распределения, полностью расположенной на прямой.

3) Коэффициент корреляции Пирсона обладает свойством симметрии, т.е. corr(x,y) = corr(y,x).

4) Коэффициент корреляции Пирсона обладает свойством инвариантности при отдельных изменениях значений параметров сдвига и масштабных параметров в двух переменных. Это означает, что при выполнении преобразований вида: $X \to a + bX,\ Y \to c + dY$, где a,b,c и d – константы (b,d >0), коэффициент корреляции не изменит своего значения. Более же общие линейные преобразования данных изменяют значения коэффициентов корреляции.

4. Интерпретация

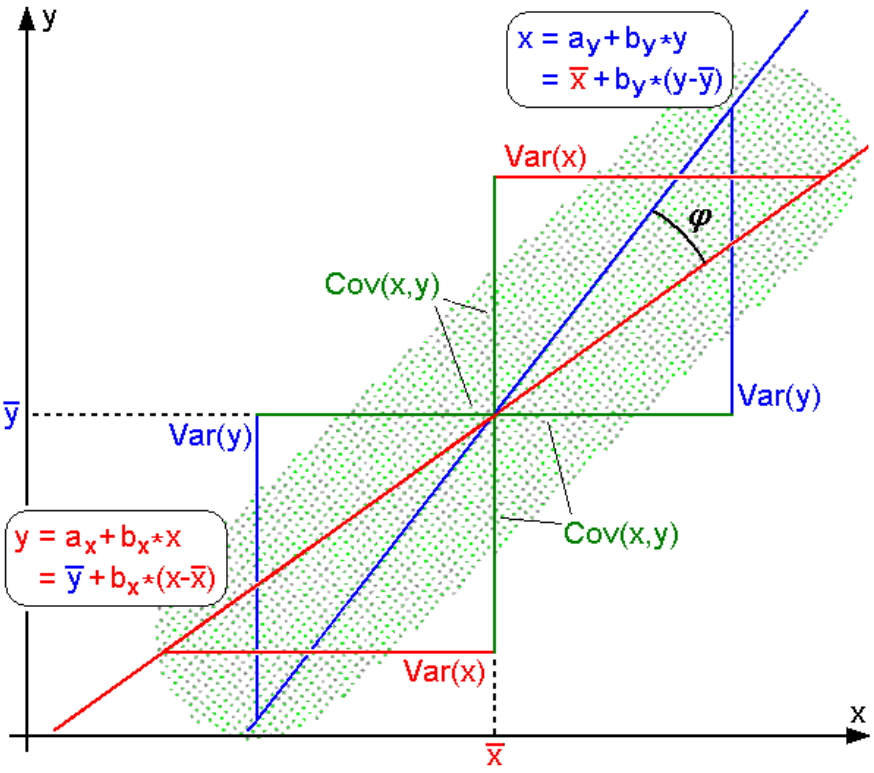
Величина коэффициента корреляции равная -1 или +1 означает, что линейное уравнение вида Y=aX+b идеально описывает связь между исследуемыми параметрами. Значение 0 означает отсутствие корреляции и, следовательно, отсутствие линейной связи между переменными. Величина коэффициента равная +1 показывает, что значения параметров лежат на прямой такой, что параметр Y увеличивается, с увеличением значений X. При величине коэффициента равной -1 наоборот, данные распределены на прямой линии, для которой Y уменьшается, с увеличением значений X.

Знак коэффициента корреляции совпадает со знаком коэффициента ковариации. Обратим внимание на выражение $(x_i - \bar{x})(y_i - \bar{y})$. Оно положительно тогда и только тогда, когда и x_i и y_i лежат по одну сторону от своих средних значений. Это означает, что для того чтобы коэффициент корреляции был положительным, x_i и y_i должны иметь тенденцию быть одновременно больше или одновременно меньше, чем их средние значения, тогда результирующая сумма в формуле ковариации будет положительна. Если же x_i и y_i имеют тенденцию лежать на противоположных позициях от своих средних значений, то коэффициент корреляции будет принимать отрицательное значение. Чем сильнее будут появляться эти тенденции, тем ближе окажется абсолютное значение коэффициента корреляции к единице.

Рассмотрим случай нецентрированных данных. Пусть $y = a_x + b_x x$ и $x = a_y + b_y y$ – построенные линейные регрессии y по x и x по y соответственно. Если стандартные отклонения данных S_x и S_y равны, то формула коэффициента имеет следующий вид:

$$\boldsymbol{r = \sec\varphi - \tan\varphi},$$

где φ - угол между линиями регрессий $y = a_x + b_x x$ и $x = a_y + b_y y$.



Если коэффициент корреляции положительный, то φ измеряется против часовой стрелки в первом квадранте, образованном вокруг точки пересечения линий регрессий. Если же коэффициент корреляции отрицательный, то измерение угла производится против часовой стрелки от четвертого во второй квадрант.

Рассмотрим случай центрированных данных, это данные, для которых был совершен сдвиг соответствующих переменных их выборочными средними значениями, чтобы каждая переменная имела нулевое среднее. В этом случае коэффициент корреляции можно интерпретировать как косинус угла между двумя наблюдаемыми векторами в N-мерном пространстве, для каждой переменной производилось N наблюдений.

Существуют различные критерии для интерпретации величины коэффициента корреляции. Но они по своей сути являются довольно субъективными. Интерпретация величины коэффициента корреляции непосредственно зависит от целей и области исследования. Например, корреляция 0.8 может быть обозначена как низкая, если проверяется физический закон с помощью высокоточных приборов, но в социальных науках может считаться очень высокой, так как там возможно влияние множества усложняющих факторов.

5. Статистическое применение

Математическая статистика, в плане использования и применения коэффициента корреляции Пирсона, в основном фокусируется на одной из следующих целей:

I. Проверка нулевой гипотезы о равенстве нулю истинного коэффициента корреляции **ρ**, на основе значения выборочного коэффициента корреляции **r.**

II. Построение доверительного интервала, который имеет заданную вероятность содержания **ρ**, при повторной выборке.

Рассмотрим некоторые примеры:

1) **Перестановочные тесты** обеспечивают прямой подход к выполнению проверки гипотез и построению доверительных интервалов. Перестановочный тест для коэффициента корреляции Пирсона включает в себя два этапа:

1. На основе исходных парных данных (x_i, y_i), путем перераспределения пар, создаются новые наборы данных вида $(x_i, y_{i'})$, где i' – перестановка множества 1,…,n.

2. На основе новых парных данных строится коэффициент корреляции **r.**

Шаг (1) и шаг (2) необходимо выполнить большое число раз. Величина p-значения для теста перестановок определяется как доля полученных значений **r,** которые больше, чем коэффициент корреляции Пирсона, рассчитанный на исходных данных.

2) **Бутстрэп** – практический программный метод исследования распределения статистик вероятностных распределений, основанный на многократной генерации выборок методом статистического моделирования на базе имеющейся выборки. Метод бутстрэпа может быть применен для построения доверительных интервалов для значений коэффициента корреляции Пирсона. В непараметрическом бутстрэпе, подобно тесту перестановки, n пар данных (x_i, y_i) многократно пересчитываются, с вычислением соответствующих значений коэффициентов **r** для них. Далее происходит построение эмпирического распределения пересчитанных значений **r,** и с его помощью аппроксимируется выборочное распределение статистики.

3) Часто на практике проверка гипотез и построение доверительных интервалов для коэффициента корреляции **ρ**, осуществляются с использованием **преобразования Фишера**:

$$\boldsymbol{F(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \operatorname{arctanh}(r).}$$

F(r) приближенно следует нормальному распределению со средним значением $F(\rho) = \operatorname{arctanh}(\rho)$ и стандартной ошибкой $= \frac{1}{\sqrt{n-3}}$, n – размер выборки. При справедливости нулевой гипотезы (простая гипотеза $H_0: \rho = \rho_o$), и предполагая, что выборки независимы и одинаково распределены и следуют двумерному нормальному распределению, стандартная оценка аппроксимации будет равна:

$$\boldsymbol{z = \frac{x - \mu_x}{S_x} = [F(r) - F(\rho_0)]\sqrt{n-3}.}$$

Следовательно, приблизительное значение величины p-value может быть получено из известной таблицы вероятностей для критерия Фишера. Чтобы получить доверительный интервал для **ρ,** необходимо сначала вычислить доверительный интервал для **F(ρ)**, а затем применить обратное преобразование Фишера.

6. Статистические свойства

Как указывалось ранее, коэффициент корреляции Пирсона совокупности определяется в терминах моментов, следовательно, он **существует** для тех двумерных распределений вероятностей, для которых определены ковариации совокупности и неравные нулю предельные дисперсии совокупности.

Рассмотрим свойства, связанные с размером выборки:

1) Для среднего или большого размера выборки и нормальной совокупности и в случае двумерного нормального распределения, коэффициент корреляции выборки представляет собой оценку максимального правдоподобия для коэффициента корреляции совокупности. Эта оценка является асимптотически несмещенной и эффективной.

2) Для большого размера выборки и ненормальной совокупности, коэффициент корреляции выборки сохраняет приблизительную несмещенность, но может быть не эффективным.

3) При большом размере выборки и согласованности средних значений выборки, дисперсий и ковариаций, коэффициент корреляции выборки является непротиворечивой оценкой коэффициента корреляции совокупности.

Выборочная статистика **r**, как и многие другие, часто используемые статистики, **не является надежной**, поэтому значение коэффициента корреляции может вводить в заблуждение, если в данных присутствуют выбросы. Проверка диаграммы разброса данных обычно помогает определить ситуации, когда отсутствие надежности может быть проблематично. В таких ситуациях может быть целесообразно воспользоваться надежной мерой ассоциации.

7. Иные вариации коэффициента корреляции

Рассмотрим случай, когда коррелируемые наблюдение имеют разную степень значимости, которую можно выразить с помощью весового вектора w. Для того чтобы вычислить **взвешенный коэффициент корреляции** между векторами x и y с вектором весов w, используются следующие формульные выражения:

Средневзвешенное значение:

$$\boldsymbol{m(x;w) = \frac{\sum x_i w_i}{\sum w_i}}.$$

Взвешенная ковариация:

$$\boldsymbol{cov(x,y;w) = \frac{\sum w_i (x_i - m(x;w))(y_i - m(y;w))}{\sum w_i}}.$$

Взвешенная корреляция:

$$\boldsymbol{corr(x,y;w) = \frac{cov(x,y;w)}{\sqrt{cov(x,x;w)cov(y,y;w)}}}.$$

В случае, когда данные не сосредоточены вокруг своих средних значений, используется специальный вариант коэффициента корреляции Пирсона – **корреляция отражения.** Выборочный коэффициент корреляции отражения равен:

$$\boldsymbol{corr_r(X,Y) = \frac{E[XY]}{\sqrt{E[X^2]E[Y^2]}} = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}.}$$

Во временных рядах, чтобы выявить корреляцию между быстрыми компонентами данных, используется специальный вариант коэффициента корреляции Пирсона – так называемая **масштабная корреляция**. Она определяется как средняя корреляция между короткими сегментами данных.

Пусть K – количество сегментов длины s, которые распределены на общей длине данных T; $K = round\left(\frac{T}{s}\right)$.

Масштабная корреляция по всем малым сегментам вычисляется по формуле:

$$\boldsymbol{\bar{r}_s = \frac{1}{K}\sum_{k=1}^{K} r_k},$$

где r_k – коэффициент корреляции Пирсона для k-ого сегмента. Данный коэффициент позволяет учитывать только вклад быстрых компонент, удаляя при этом вклад медленных компонент.

Список используемой литературы

[1] Rummel, R.J. (1976). "Understanding Correlation". ch. 5

[2] Pearson, Karl (20 June 1895). "Notes on regression and inheritance in the case of two parents". Proceedings of the Royal Society of London. 58: 240-242.

[3] Stigler, Stephen M. (1989). "Francis Galton's account of the invention of correlation". Statistical Science. 4 (2): 73-79.

[4] Wright, S. (1921). "Correlation and causation". Journal of Agricultural Research. 20 (7): 557-585.

[5] Moriya, N. (2008). "Noise-related multivariate optimal joint-analysis in longitudinal stochastic processes". Nova Science Publishers, Inc. pp. 223-260.

[6] Schmid, John, Jr. (December 1947). "The relationship between the coe-cient of correlation and the angle included between regression lines". The Journal of Educational Research. 41 (4): 311-313.

[7] Buda, Andrzej; Jarynowski, Andrzej (December 2010). Life Time of Correlations and its Applications. Wydawnictwo Niezalezne.

[8] Davey, Catherine E.; Grayden, David B.; Egan, Gary F.; Johnston, Leigh A. (January 2013). "Filtering induces correlation in fMRI resting state data". NeuroImage. 64: 728-740.

[9] Hotelling, Harold (1953). "New Light on the Correlation Coe-cient and its Transforms". Journal of the Royal Statistical Society. Series B (Methodological). 15 (2): 193-232.

[10] Devlin, Susan J.; Gnanadesikan, R.; Kettenring J.R. (1975). "Robust estimation and outlier detection with correlation coe-cients". Biometrika. 62 (3): 531-545.