Корреляция: простая, частная и условная

Борис Демешев*

5 августа 2015 г.

Аннотация

Корреляция — это способ описать силу линейной зависимости между двумя случайными величинами одним числом. Каков геометрический смысл корреляции? Что такое частная корреляция? Как связаны частная и условная корреляция?

Ключевые слова: корреляция, частная корреляция, условная корреляция, косинус, проекция.

1 Сколько вешать в граммах?

Почему мы измеряем температуру тела с помощью градусника?

1. Измерить температуру очень удобно
2. Это измерение несёт в себе информацию о здоровье

Сложное описание здоровья сводится измерением температуры к одной цифре. Естественно, куча информации теряется в этой цифре и бессмысленно лечить человека, руководствуясь только температурой его тела. Температура 39° говорит, что не так, но что — непонятно. А температура 36.6° ещё не говорит о том, что у человека идеальное здоровье. Однако процедура очень проста и в некоторых ситуациях (например, при обыкновенной простуде) её достаточно для принятия решения о приёме жаропонижающего.

Если бы для измерения температуры нужно было специальное устройство размером с половину комнаты, никто бы дома её не мерял. Простота измерения очень важна!

Подобном образом дела обстоят и с описанием зависимости между случайными величинами. Зависимость между случайными величинами полностью описывается их совместной функцией распределения, $F(x, y) = ¶(X \leq x, Y \leq y)$. Вместо сложной функции распределения мы хотим получить одно число. Некую «силу зависимости». Назовём это мифическое число $Dep(X, Y)$.

Что мы требуем от этого числа?

1. Посчитать это число очень удобно
2. Это число несёт в себе информацию о зависимости

*НИУ ВШЭ, Москва.

В каком смысле «удобно» считать?

Очень часто возникают суммы случайных величин, поэтому было бы здорово, чтобы у суммы легко считалась наша характеристика $Dep(X, Y)$. Проще всего было бы, если бы:

$$Dep(X + Z, Y) = Dep(X, Y) + Dep(Z, Y)$$

И, конечно, мы ждём, что у независимых случайных величин нулевая сила зависимости, $Dep(X, Y) = 0$, а ненулевая сила зависимости, $Dep(X, Y) \neq 0$, была бы возможна только у зависимых случайных величин.

Этим двум требуемым свойствам (простота подсчёта и информация о зависимости) отвечает ковариация.

Определение 1. *Ковариация величин X и Y измеряет ...*

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

здесь рассказать про прямоугольники и площадь с плюсом/минусом?

2 Корреляция по-русски

Обычно в учебниках даётся такое определение корреляции

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}. \tag{2.1}$$

Естественно, возникает вопрос: «С какого перепугу? Почему это мы делим ковариацию на что-то там?»

Мы дадим определение корреляции словами:

Определение 2. *Корреляция между случайными величинами X и Y показывает на сколько своих стандартных отклонений в среднем растёт случайная величина Y при росте случайной величины X на одно своё стандартное отклонение.*

А теперь из этого словесного определения мы получим формулу 1.1. Разложим величину Y на два слагаемых. Первое слагаемое вбирает в себя всю ту часть Y, которая линейно зависит от X, а второе — всё оставшееся:

$$\frac{Y}{\sigma_Y} = \rho \cdot \frac{X}{\sigma_X} + \varepsilon$$

В этой формуле видно, что с ростом X на одно стандартное отклонений σ_X правая часть изменится в среднем на ρ, и, следовательно, величина Y в среднем изменится на $\rho \cdot \sigma_Y$.

Естественно, мы хотим, чтобы с ростом X величина ε в среднем не менялась, то есть хотим нулевую «силу зависимости» между ними, $\mathrm{Cov}(X, \varepsilon) = 0$.

$$\mathrm{Cov}\left(X, \frac{Y}{\sigma_Y} - \rho \cdot \frac{X}{\sigma_X}\right) = 0$$

По свойствам ковариации получаем

$$\mathrm{Cov}(X,Y)/\sigma_Y - \rho\,\mathrm{Cov}(X,X)/\sigma_X = 0$$

И, тадам, выражаем корреляцию, ρ:

$$\rho = \frac{\mathrm{Cov}(X,Y)/\sigma_Y}{\mathrm{Cov}(X,X)/\sigma_X} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

Несмотря на асимметричность исходного разложения (эпсилон прибавляется в правой части уравнения к величине X), результирующая формула для корреляции получается симметричной. Из этого следует, что ровно такой же результат получится, если начать с разложения:

$$\frac{X}{\sigma_X} = \rho \cdot \frac{Y}{\sigma_Y} + \varepsilon$$

Из определения неочевидно, что корреляция лежит в пределах от -1 до 1. ...

Стоит обратить внимание на немного контр-интуитивный факт. Если бы зависимость между X и Y была бы жесткой детерминистической, и с ростом X на единицу величина Y росла бы на Δ, то с ростом Y на единицу величина X росла бы на $1/\delta$. Для случайных величин обращения не происходит. Если с ростом X на одно своё стандартное отклонение величина Y в среднем растёт на ρ своих стандартных отклонений, то и с ростом Y на одно своё стандартное отклонение величина X в среднем растёт на ρ своих стандартных отклонений.

? парадокс возвращения к среднему ?

3 Геометрический смысл корреляции

Давайте рисовать случаные величины векторами-стрелочками! Не в том смысле, что у стрелочки случайное направление или длина, а в том смысле, что направление и длина стрелочки описывают характеристики этой случайной величины.

Любую геометрию можно задать, задав скалярное произведение. Действительно, если мы умеем считать скалярное произведение двух любых векторов, $<\vec{a},\vec{b}>$, то длина вектора считается ровно как в 9-м классе:

$$|\vec{a}| = \sqrt{<\vec{a},\vec{a}>}$$

И также любой девятикласник помнит, что косинус угла между векторами считается как

$$\cos(\vec{a},\vec{b}) = \frac{<\vec{a},\vec{b}>}{|\vec{a}||\vec{b}|}$$

Мы определим скалярное произведение двух случайных величин как их ковариацию:

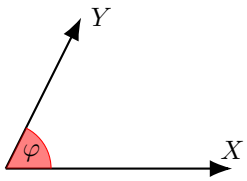
$$<X,Y> = \mathrm{Cov}(X,Y)$$

При таком подходе длиной случайной величины окажется стандартное отклонение:

$$\sqrt{\mathrm{Cov}(X,X)} = \sqrt{\mathrm{Var}(X)} = \sigma_X$$

А корреляция окажется косинусом угла между случайными величинами:

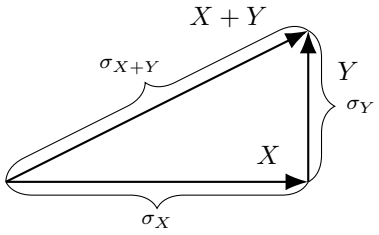
$$\cos\varphi = \cos(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \mathrm{Corr}(X,Y)$$



Значит в нашей геометрии длина стрелочки — стандартное отклонение случайной величины, а косинус угла между двумя стрелочками — это корреляция двух случайных величин. Дисперсия, следовательно, это квадрат длины случайной величины. Перпендикулярными случайными величинами будут те, косинус угла между которыми равен нулю, то есть некоррелированные.

Например, сформулируем в данной геометрии теорему Пифагора. Если случайные величины X и Y перпендикулярны (корреляция или ковариация равны нулю), то дисперсия их суммы (квадрат длины гипотенузы) равен сумме их дисперсий (сумму квадратов длин катетов):

$$\mathrm{Var}(X+Y) = \sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y = \mathrm{Var}(X) + \mathrm{Var}(Y)$$



Введение геометрии позволяет говорить о проекции. Например, можно спроецировать случайную величину Y на множество случайных величин пропорциональных величине X, $\{cX | c \in \mathbb{R}\}$. Если на обычной плоскости спроецировать вектор $\vec{a}$ на прямую, порожденную вектором $\vec{b}$, то получится $\cos(\vec{a},\vec{b}) \cdot \vec{b}$. По аналогии, если спроецировать случайную величину Y на множество $\{cX | c \in \mathbb{R}\}$, то получится $\mathrm{Corr}(X,Y) \cdot X$. Другими словами, среди случайных величин пропорциональных X величина $\hat{Y} = \mathrm{Corr}(X,Y) \cdot X$ — самая похожая на величину Y.

Понятие проекции позволяет интепретировать квадрат корреляции. Квадрат косинуса равен отношению квадрата длины прилежащего катета $\mathrm{Var}(\hat{Y})$ к квадрату гипотенузы $\mathrm{Var}(Y)$.

(картинка)

Следовательно, $\mathrm{Corr}(X,Y)^2 = \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)}$, то есть квадрат корреляции показывает долю дисперсии Y, которую можно объяснить с помощью величин пропорциональных X.

4 Корреляция и независимость

Теорема 1. *Случайные величины X и Y независимы тогда и только тогда, когда некоррелированы любые функции $f(X)$ и $g(Y)$.*

Другими словами для независимости X и Y необходима некоррелированность пар X и Y, X^2 и $\cos(Y)$, $\exp(X)$ и $1/Y$, и так далее. Из этого следует, что некоррелированность X и Y является необходимым, но недостаточным условием для независимости.

Можно выделить три «степени» независимости случайных величин X и Y:

Некоррелированность Y и X	$\mathrm{Corr}(X,Y)=0$
$\mathbb{E}(Y\|X)=\mathbb{E}(Y)$	$\mathrm{Corr}(f(X),Y)=0$ для всех $f()$
Независимость Y и X	$\mathrm{Corr}(f(X),g(Y))=0$ для всех $f()$ и $g()$

Многие ошибочно считают, что если величина X имеет нормальное распределение $N(\mu_X,\sigma_X^2)$ и величина Y имеет нормальное распределение $N(\mu_Y,\sigma_Y^2)$, и X и Y некоррелированы, то они независимы. Это неверно.

Контрпример. Случайная величина X имеет стандартное нормальное распределение $N(0;1)$, случайная величина Z независима от X и равновероятно принимает значения -1 и $+1$. Определим величину Y как их произведение, $Y=XZ$.

В этом примере величины X и Y зависимы, так как $|X|=|Y|$. Однако Y распределена нормально стандартно и $\mathrm{Corr}(X,Y)=0$.

Правильная теорема звучит так:

Теорема 2. *Если некоррелированные случайные величины X и Y имеют совместное нормальное распределение, то X и Y независимы.*

Попутно упомянем ещё одно неожиданное свойство предъявленного контрпримера. Если случайные величины нормальны по отдельности, то вполне возможно, что их сумма ненормальна. Для пары величин, имеющих совместное нормальное распределение, это невозможно.

5 Частная корреляция

Определение 3. *Частная корреляция между величинами X и Y при фиксированной величине Z показывает на сколько своих стандартных отклонений σ_Y в среднем вырастет Y при росте величины X на одно своё стандартное отклонение σ_X и постоянном значении величины Z.*

Для нахождения частной корреляции используется разложение

$$\frac{Y}{\sigma_Y}=\rho_{XY|Z}\cdot\frac{X}{\sigma_X}+\rho_{YZ|X}\frac{Z}{\sigma_Z}+\varepsilon$$

Альтернативный подход к подсчёту частной корреляции следующий:

1. Спроецируем X на множество величин, некоррелированных с Z. Получим $\tilde{X}$.
2. Спроецируем Y на множество величин, некоррелированных с Z. Получим $\tilde{Y}$.
3. Частная корреляция между X и Y при фиксированной Z — это обычная корреляция между $\tilde{X}$ и $\tilde{Y}$.

(картинка ...)

Два подхода к определению частной корреляции эквивалентны в силу теоремы Фриша-Ву-Ловелла (Frisch–Waugh–Lovell). Обычно эта теорема формулируется применительно к регрессии, а здесь мы приведём её вариат для случайных величин.

Теорема 3. *Если имеют место разложения:*

$$Y = a_1 Z_1 + a_2 Z_2 + \ldots + a_n Z_n + \tilde{Y},\ \text{где}\ \ \tilde{Y} \perp Z_1, Z_2, \ldots, Z_n$$

и

$$X = b_1 Z_1 + b_2 Z_2 + \ldots + b_n Z_n + \tilde{X},\ \text{где}\ \ \tilde{X} \perp Z_1, Z_2, \ldots, Z_n$$

То в разложениях

$$\tilde{Y} = d\tilde{X} + \varepsilon,\ \text{где}\ \ \varepsilon \perp \tilde{X}$$

и

$$Y = c_1 Z_1 + c_2 Z_2 + \ldots + c_n Z_n + dX + u,\ \text{где}\ \ u \perp Z_1, Z_2, \ldots, Z_n, X$$

коэффициенты при $\tilde{X}$ и X совпадают.

6 Условная корреляция

Определение 4. *Условная корреляция между величинами X и Y при известном значении величины Z показывает на сколько своих стандартных отклонений σ_Y в среднем вырастет Y при росте величины X на одно своё стандартное отклонение σ_X при заданном значении величины Z.*

Следует подчеркнуть одно существенное отличие условной корреляции от обычной и частной. Обычная и частная корреляция являеются константами. Условная корреляция $\mathrm{Corr}(X, Y|Z)$ является функцией от Z. Величина Z является случайной, поэтому и условная корреляция $\mathrm{Corr}(X, Y|Z)$ является случайной величиной.

Здесь регрессионное определение????

Чуть более формальное определение:

Определение 5.

$$\mathrm{Corr}(X, Y|Z) = \frac{\mathrm{Cov}(X, Y|Z)}{\sqrt{\mathrm{Var}(X|Z)\,\mathrm{Var}(Y|Z)}},$$

где $\mathrm{Cov}(X, Y|Z) = \mathbb{E}(XY|Z) - \mathbb{E}(X|Z)\mathbb{E}(Y|Z)$ и $\mathrm{Var}(X|Z) = \mathbb{E}(X^2|Z) - (\mathbb{E}(X|Z))^2$

Пример подсчета частной и условной корреляций.

Пример 1.

Закон распределения случайных величин X_1, X_2, X_3 задан двумя таблицами:

$X_3 = 0$

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	0.06	0.08
$X_1 = 1$	0.24	0.32

$X_3 = 1$

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	0.1	0
$X_1 = 1$	0	0.2

Найдите условную корреляцию $\mathrm{Corr}(X_1, X_2|X_3)$ и частную корреляцию $\mathrm{pCorr}_{X_3}(X_1, X_2)$.

Решение.

Эти две таблички на самом деле реализуют простую мысль: при $X_3 = 1$ величины X_1 и X_2 связаны детерминистически линейно, а при $X_3 = 0$ величины X_1 и X_2 независимы.

Считаем две вспомогательные условные корреляции, $\mathrm{Corr}(X_1, X_2|X_3 = 0) = 0$, $\mathrm{Corr}(X_1, X_2|X_3 = 1) = 1$.

Отсюда получаем, что $\mathrm{Corr}(X_1, X_2|X_3) = X_3$. Для дискретных случайных величин запись условного ожидания не однозначна, и, например, ответ $\mathrm{Corr}(X_1, X_2|X_3) = X_3^2$ также будет верным.

Пример 2.

Величины X_1, X_2, X_3 имеют совместное нормальное распредлеение с математическим ожиданием $\mathbb{E}(X) = (1, 2, -3)'$ и ковариационной матрицей

$$\mathrm{Var}(X) = \begin{pmatrix} 9 & 2 & -1 \\ 2 & 16 & 1 \\ -1 & 1 & 4 \end{pmatrix}$$

Найдите условную корреляцию $\mathrm{Corr}(X_1, X_2|X_3)$ и частную корреляцию $\mathrm{pCorr}_{X_3}(X_1, X_2)$.

Решение.

...

В данном примере частная и условная корреляция совпали. Это одно из приятных свойств многомерного нормального распределения:

Теорема 4. *Если величины X, Y и Z имеют совместное нормальное распределение, то частная и условная корреляции совпадают.*

Пример 3. AR(1) процесс

7 Выборочные характеристики

В теории обычную корреляцию и частную корреляцию можно посчитать, если известен закон распределения случайных величин. На практике закон распределения не известен, однако доступны наблюдения. Как по имеющимся наблюдениям оценить неизвестные корреляции?

Несколько способов оценки корреляции

Здесь пара картинок: википедийная с корреляциями и два ряда случайного блуждания/тренда

Несколько способов оценки частной корреляции