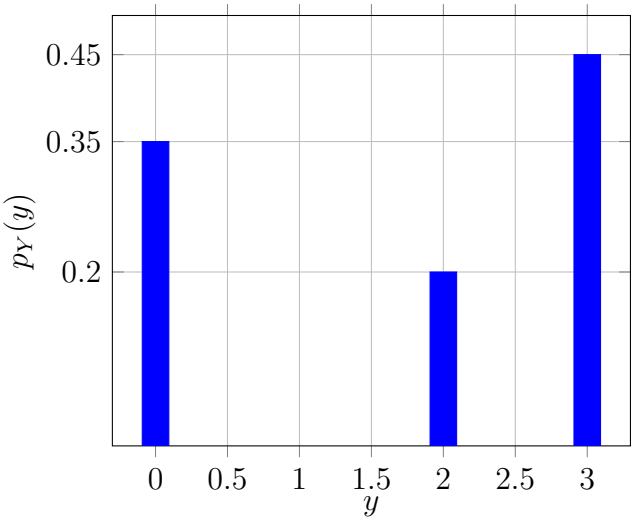
Лекция 7. Преобразование распределений. Типы случайных величин. Ковариация и корреляция.

24 марта 2022 г.

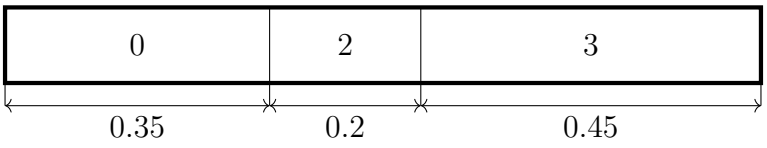
1 Эмитация одного распределения другим

Практически любое современное вычислительное устройство умеет генерировать (псевдо)случайные числа, но чаще всего только равномерное распределение. Этого на самом деле достаточно, чтобы получить случайную величину, следующую любому другому распределению. В данной части мы предполагаем, что мы умеем получать с.в. $X \sim U(0,1)$ и хотим получить с.в. Y, для которой знаем функцию распределения $F_Y(y)$.

Рассмотрим простой случай. Y – дискретная с.в. с конечным набором значений. Пусть ее функция вероятности выглядит так:

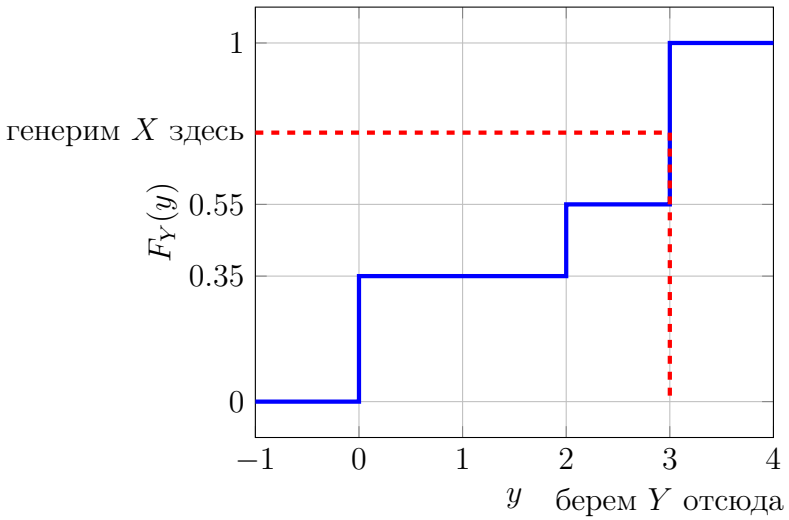


Как мы можем поступить. Возьмем отрезок единичной длины, разобъем его на отрезки длиной, соответствующей вероятности каждого значения с.в., сгенерим X и вернем значение, соответствующее тому значению, в которое попал X .

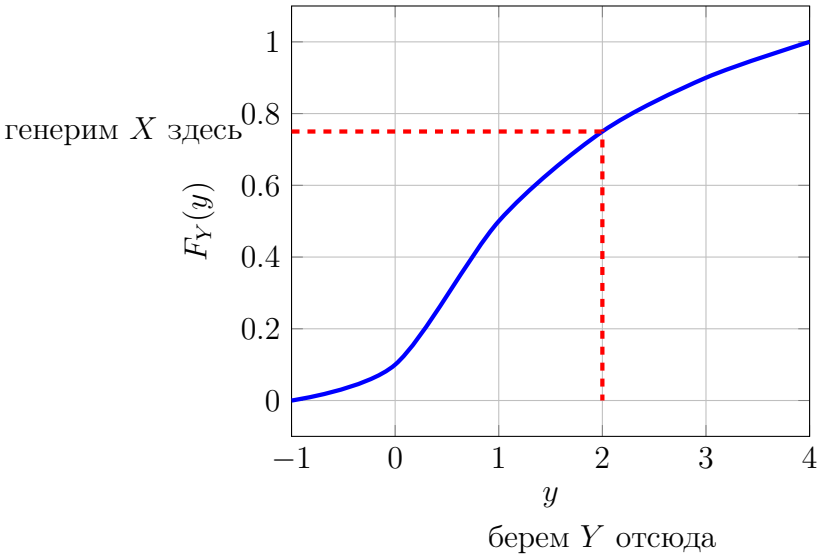


Таким методом можно сэмитировать любую с.в. с конечным множеством значений размера n за время запроса к $X \sim U(0,1)$ плюс время $\Theta(\log(n))$, необходимое для нахождения отрезка, в который мы попали, двоичным поиском.

Интерпретация наших действий на основе функции распределения:



Давайте используем эту интерпретацию для непрерывной Y в случае с монотонной $F_Y(y)$.



То есть после выбора X мы берем $Y = F_Y^{-1}(X)$. Почему это работает:

$$\Pr(Y \le y) = \Pr(F_Y^{-1}(X) \le y) = \Pr(X \le F_Y(y)) = F_Y(y)$$

2 Уточнение про типы с.в.

Какие бывают с.в.:

- Сингулярные — сконцентрированные на множестве меры ноль (имеется в виду мера Лебега на $\mathbb{R}$)
- Абсолютно непрерывные — те, у которых есть плотность вероятности
- Смешанные — те, функция распределения которых является взвешенной суммой функций распределений какой-то сингулярной и какой-то непрерывной с.в. То есть для любой с.в. X сущестуют такие с.в. Y и Z (Y — сингулярная, Z — абсолютно непрерывная) и число $p \in [0, 1]$, что

  $$F_X = pF_Y + (1 - p)F_Z,$$

  что является равенством функций, то есть их значения на всех $x \in \mathbb{R}$ совпадают. То есть с вероятностью p с.в. X принимает значение сингулярной с.в., а с вероятностью $(1 - p)$ — абсолютно непрерывной

Почему мы раньше говорили о дискретных а теперь их не упоминаем, но говорим о каких-то сингулярных? Потому что дискретные с.в. (те, которые имеют функцию вероятностей) есть подмножество сингулярных, но не все сингулярные имеют функцию вероятностей. Пример сингулярного недискретного распределения — то, у которого функция распределения есть лестница Кантора — функция, неубывающая и непрерывная на отрезке $[0, 1]$, которая возрастает на нем от 0 до 1, но при этом почти всюду (то есть на всех точках, за исключением множества меры ноль) имеет нулевую производную.

Поэтому уместнее разбить сингулярные с.в. на два типа: дискретные (именно в том понимании, в котором мы имели с ними делао) и сингулярно-непрерывные (те, у которых функция распределения возрастает на множестве меры ноль, но не имеют точек разрыва). Таким образом, для любой с.в. X существуют такие с.в. Y, Z и S (Y — дискретная, Z — абсолютно непрерывная и S — сингулярно-непрерывная) и числа $p \in [0, 1], q \in [0, 1 - p]$, что

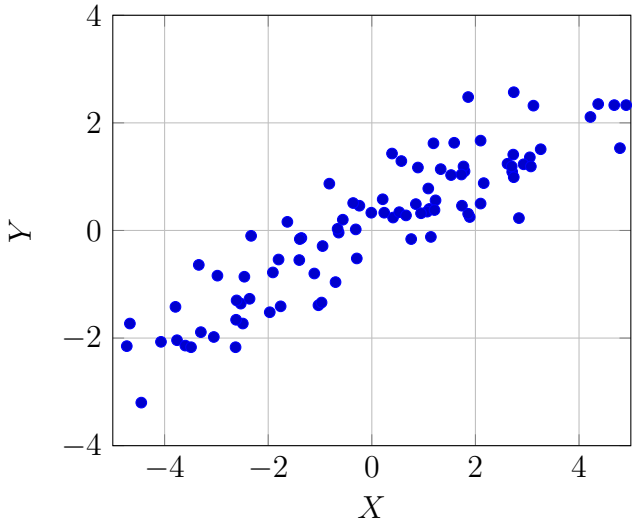
$$F_X = pF_Y + qF_Z + (1 - p - q)F_S.$$

Это доказывается довольно сложно и выходит за рамки нашего курса, но если интересно можете почитать про теорему Лебега о разложении меры (Lebesgue's decomposition).

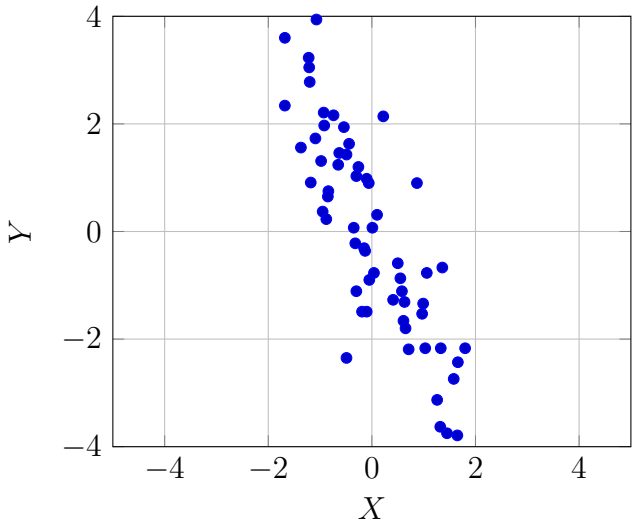
В рамках нашего курса мы не будем сталкиваться с сингулярно-непрерывными распределениями, поэтому мы будем считать, что любая с.в. может быть представлена в виде комбинации дискретной и непрерывной.

3 Ковариация

До сих пор мы про две с.в. могли сказать только то, зависимы ли они, или нет. Но иногда хочется определить степень этой зависимости, то есть понять, сколько информации одна с.в. дает про другую с.в. Рассмотрим пример. Пусть две с.в. X и Y имеют нулевое матожидание. Если они независимы, то $E[XY] = E[X]E[Y] = 0$. Но рассмотрим такие случаи: пусть каждая точка на графике равновероятна.



В таком случае велика вероятность, что либо X и Y оба меньше нуля, либо оба больше нуля, то есть $E[XY] > 0$. В следующем случае будет наоборот: X и Y с большой вероятностью имеют разные знаки, значит, $E[XY] < 0$.



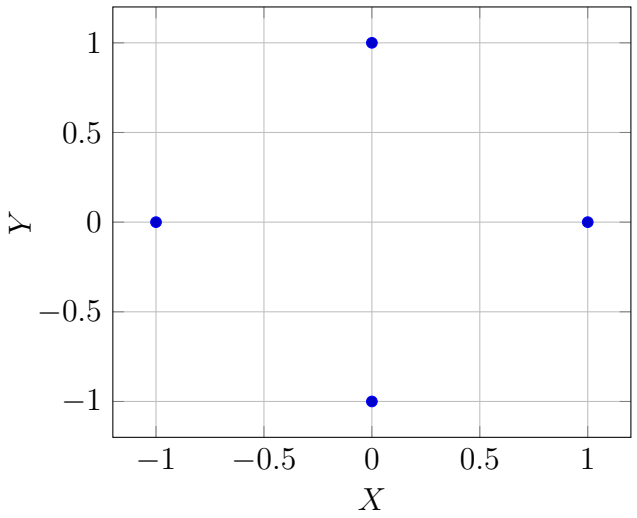
Чтобы измерить степень зависимости была введена величина, называемая *ковариацией* случайных величин X и Y.

$$\mathrm{Cov}(X,Y) = E\Big[(X - E[X])(Y - E[Y])\Big]$$

Если две с.в. независимы, то для них будет верно

$$\mathrm{Cov}(X,Y) = E(X - E(X))E(Y - E(Y)) = 0 \cdot 0 = 0$$

Правда обратное неверно: с.в. с нулевой ковариацией могут быть зависимы. Рассмотрим пример. Пусть четыре пары (X, Y) равновероятны: $(1, 0), (0, 1), (-1, 0), (0, -1)$



Заметим, что матожидания X и Y равны нулю, то есть $\mathrm{Cov}(X,Y) = E[XY]$, но XY всегда равен нулю, поэтому ковариация нулевая. При этом X и Y явно зависимы: если $X = 1$, то Y однозначно равен нулю.

4 Свойства ковариации

Полезная формула.

Ковариация с.в. с самой собой равна ее дисперсии, причем у нас была очень милая формула:

$$\mathrm{Cov}(X,X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2.$$

Есть ли что-то подобное для ковариации? Конечно, есть.

$$\begin{aligned}\mathrm{Cov}(X,Y) = E\Big[(X - E[X])(Y - E[Y])\Big] &= E\Big[XY - YE[X] - XE[Y] + E[X]E[Y]\Big] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] = E[XY] - E[X]E[Y].\end{aligned}$$

$$\operatorname{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

Линейные преобразования

Линейное преобразование одной с.в.:

$$\begin{aligned}\operatorname{Cov}(aX + b, Y) &= E[(aX + b)Y] - E[aX + b]E[Y] \\ &= aE[XY] + bE[Y] - aE[X]E[Y] - bE[Y] \\ &= a(E[XY] - E[X]E[Y]) = a\operatorname{Cov}(X, Y)\end{aligned}$$

Заметим, что прибавление константы к одной с.в. ничего не меняет, так как не меняется распределение $X - E[X]$, а умножение на константу как раз увеличивает этот множитель в ту же константу.

Одна с.в. – сумма двух других:

$$\begin{aligned}\operatorname{Cov}(X + Y, Z) &= E\Big[(X + Y - E[X + Y])(Z - E[Z])\Big] \\ &= E\Big[(X - E[X])(Z - E[Z])\Big] + E\Big[(Y - E[Y])(Z - E[Z])\Big] \\ &= \operatorname{Cov}(X, Y) + \operatorname{Cov}(Y, Z)\end{aligned}$$

Ковариация и дисперсия суммы с.в.

Распишем дисперсию суммы двух с.в. по формуле

$$\begin{aligned}\operatorname{Var}(X + Y) &= E\Big[(X + Y - E[X + Y])^2\Big] \\ &= E\Big[((X - E[X]) + (Y - E[Y]))^2\Big] \\ &= E[(X - E[X])^2] + E[(Y - E[Y])^2] + 2E\Big[(X - E[X])(Y - E[Y])\Big] \\ &= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X, Y)\end{aligned}$$

Что насчет более двух с.в.? Предположим, что у нас есть набор с.в. $X_1, \ldots, X_n$, и матожидание каждой с.в. равно нулю. Тогда имеем

$$\begin{aligned}\operatorname{Var}(X_1 + \cdots + X_n) &= E((X_1 + \cdots + X_n)^2) \\ &= E\left(\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j\right) \\ &= \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum_{i \neq j} \operatorname{Cov}(X_i, X_j)\end{aligned}$$

Если хотим показать то же самое для с.в. с ненулевым ожидаением, то заметим, что добавление константы к с.в. не меняет ни ее дисперсию, ни ковариацию с другой с.в., поэтому

$$\begin{aligned}\mathrm{Var}(X_1 + \cdots + X_n) &= \mathrm{Var}((X_1 - E[X_1]) + \cdots + (X_n - E[X_n])) \\ &= \sum_{i=1}^{n} \mathrm{Var}(X_i - E[X_i]) + \sum_{i \neq j} \mathrm{Cov}((X_i - E[X_i]), (X_j - E[X_j])) \\ &= \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)\end{aligned}$$

5 Коэффициент корреляции

Ковариация – очень странная величина. Ее размер по сути зависит от разброса двух с.в., которые являются ее аргументом. То есть по фразе "ковариация X и Y равна единице" очень сложно сказать, много это, или мало, так как мы не знаем масштаба X и Y. Например, когда мы пытаемся посчитать ковариацию температуры воздуха и давления, то в зависимости от используемых величин измерения (цельсий или фаренгейты, паскали или атмосферы) будет разная ковариация. Причем у самой ковариации всегда есть единица измерения, равная единице измерения X, умноженная на Y (например, градус цельсия * паскаль), что еще больше усложняет осознание степени зависимости пары с.в. Поэтому была введена мера зависимости с.в., называемая коэффициентом корреляции $\rho(X, Y)$.

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

NB: он определен только для с.в. с ненулевой дисперсией.
Его основные свойсва:

1. $\rho(X, Y) \in [-1, 1]$
2. Если X и Y независимы, то $\rho(X, Y) = 0$, как и ковариация (обратное не всегда верно, см. пример для нулевой ковариации)
3. Если $|\rho(X, Y)| = 1$, то X и Y линейно зависимы, то есть $(X - E[X]) = c(Y - E[Y])$. Как следствие: $\rho(X, X) = \frac{\mathrm{Cov}(X,X)}{\sigma_X^2} = 1$.
4. Линейные преобразования любой из двух с.в. не меняют модуль коэффициента корреляции: $\rho(aX + b, Y) = \frac{a\,\mathrm{Cov}(X,Y)}{|a|\sigma_X \sigma_Y} = \mathrm{sign}(a)\rho(X, Y)$

Для доказательства первого и третьего свойств рассмотрим с.в. X и Y с нулевыми матожиданиями и единичной дисперсией (для общего случая легко доказать все, перейдя к с.в. $X' = \frac{X-E[X]}{\sigma_X}$ и $Y' = \frac{Y-E[Y]}{\sigma_Y}$, но мы опустим это для упрощения вычислений). Рассмотрим с.в. $Z = (X - \rho(X,Y)Y)^2$. Она неотрицательна, значит, и ее матожидание неотрицательно.

$$\begin{aligned} E[Z] &= E[(X-\rho Y)^2] = E[X^2] - 2\rho E[XY] + \rho^2 E[Y^2] \\ &= \operatorname{Var}(X) - 2\rho\cdot\rho + \rho^2 \operatorname{Var}(Y) = 1 - \rho^2 \geq 0 \end{aligned}$$

Отсюда следует, что во-первых, $\rho^2 \leq 1$, а во-вторых, если $|\rho| = 1$, то $E[Z] = 0$, что возможно только если Z всегда равен нулю, что в свою очередь подразумевает, что X всегда равен $\pm Y$ (в зависимости от знака ρ).

6 Интерпретация корреляции

Если две случайных величины зависимы, то чаще всего это не значит, что одна определяет другую и наоборот. Например, проведя исследование оценок в школе, можно обнаружить, что у учеников с хорошей оценкой по математике также хорошие оценки по литературе, однако это не значит, что знания по математике помогают в изучении литературы и наоборот. Это скорее означает, что у двух этих случайных величин есть какой-то общий фактор, который влияет на них обеих. В случае с оценками это может быть, например, степень усердия, которое тот или иной ученик прилагает к учебе.

Чуть более формально, давайте представим, что есть три случайных величины Z, U и V с нулевыми матожиданиями и единичными дисперсиями, которые все независимы друг от друга. И рассмотрим две других с.в. $X = Z + U$ и $Y = Z + V$. Посчитаем $\rho(X,Y)$

$$\begin{aligned} \operatorname{Var}(X) = \operatorname{Var}(Y) &= \operatorname{Var}(Z) + \operatorname{Var}(U) = 2 \\ \sigma_X = \sigma_Y &= \sqrt{(2)} \\ E[XY] &= E[Z^2 + UZ + VZ + UV] = \operatorname{Var}(Z) = 1 \\ \rho(X,Y) &= \frac{E[XY]}{\sigma_X \sigma_Y} = \frac{1}{2} \end{aligned}$$

То есть коэффициент корреляции определяет вклад какой-то причины, которая влияет на обе с.в., в каждую с.в.

7 Пример важности корреляции

Допустим, вы решили инвестировать. И у вас есть 10\$, и вы вкладываете по доллару в какие-то 10 компаний. Средняя выручка с каждой компании равна 1\$ и среднеквадратичное отклонение выручки с каждой компании есть 1.3\$. Очевидно, ваша суммарная

выручка равна вашему вкладу, но каково среднеквадратичное отклонение? Если выручки от разных компаний независимы, то

$$\mathrm{Var}(X_1 + \cdots + X_{10}) = \sum_{i=1}^{10} \mathrm{Var}(X_i) = 10 \cdot 1.3^2 = 16.9$$
$$\sigma_{X_1+\cdots+X_10} = \sqrt{16.9} \approx 4.1$$

То есть сумма довольно хорошо сконцентрирована, больше, чем при вкладе всего лишь в одну компанию. На большую прибыль надеяться не стоит, но и много проиграть шансов не так много. Теперь допустим, что у всех выручек есть какой-то общий фактор, который под ними лежит, и коэффициент корелляции между любыми X_i и X_j равен 0.9. Тогда

$$\mathrm{Var}(X_1 + \cdots + X_{10}) = \sum_{i=1}^{10} \mathrm{Var}(X_i) - \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) = 10 \cdot 1.3^2 + 90 \cdot 0.9 \cdot (1.3)^2 \approx 154$$
$$\sigma_{X_1+\cdots+X_{10}} = \sqrt{16.9} \approx 12.4$$

То есть если падают акции хотя бы одной компании, то с большой вероятностью падают и акции других, что приводит к большим потерям. Примерно это и случилось в 2008 году, когда рухнул весь рынок сразу.