Корреляционный анализ

Введение

До этого момента мы рассматривали только отдельные переменные и их характеристики, однако в практике мы редко работаем только с одной переменной. Как правило, у нас есть многомерное пространство признаков, и нас интересуют взаимосвязи между ними.

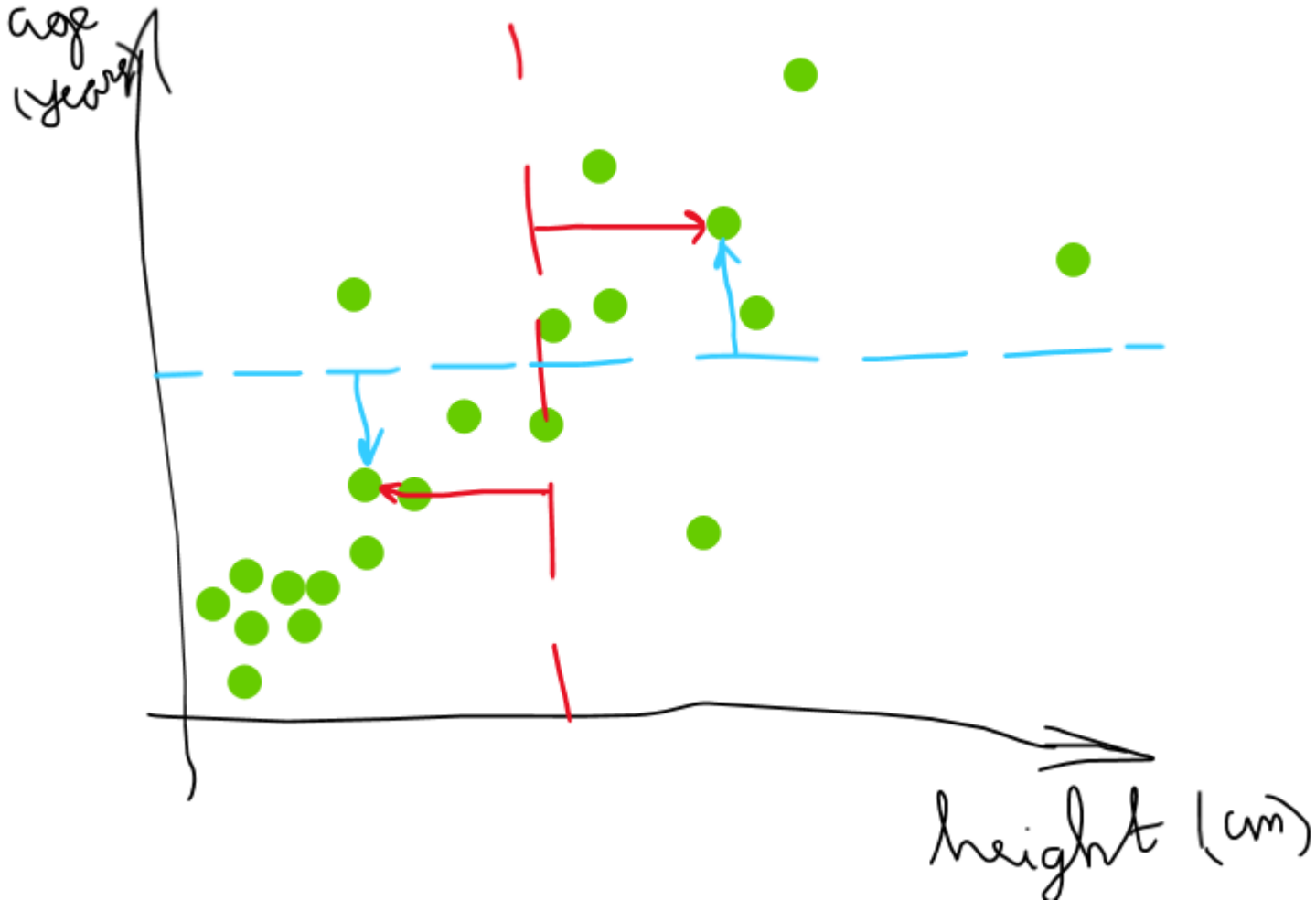
Ковариация

Мы хотим описать закономерности как можно более простым способом и используя уже имеющиеся у нас знания.

Например: используя дисперсию (variance) можно исследовать *совместную изменчивость нескольких переменных*, a.k.a. ***co-variance*** 🤩

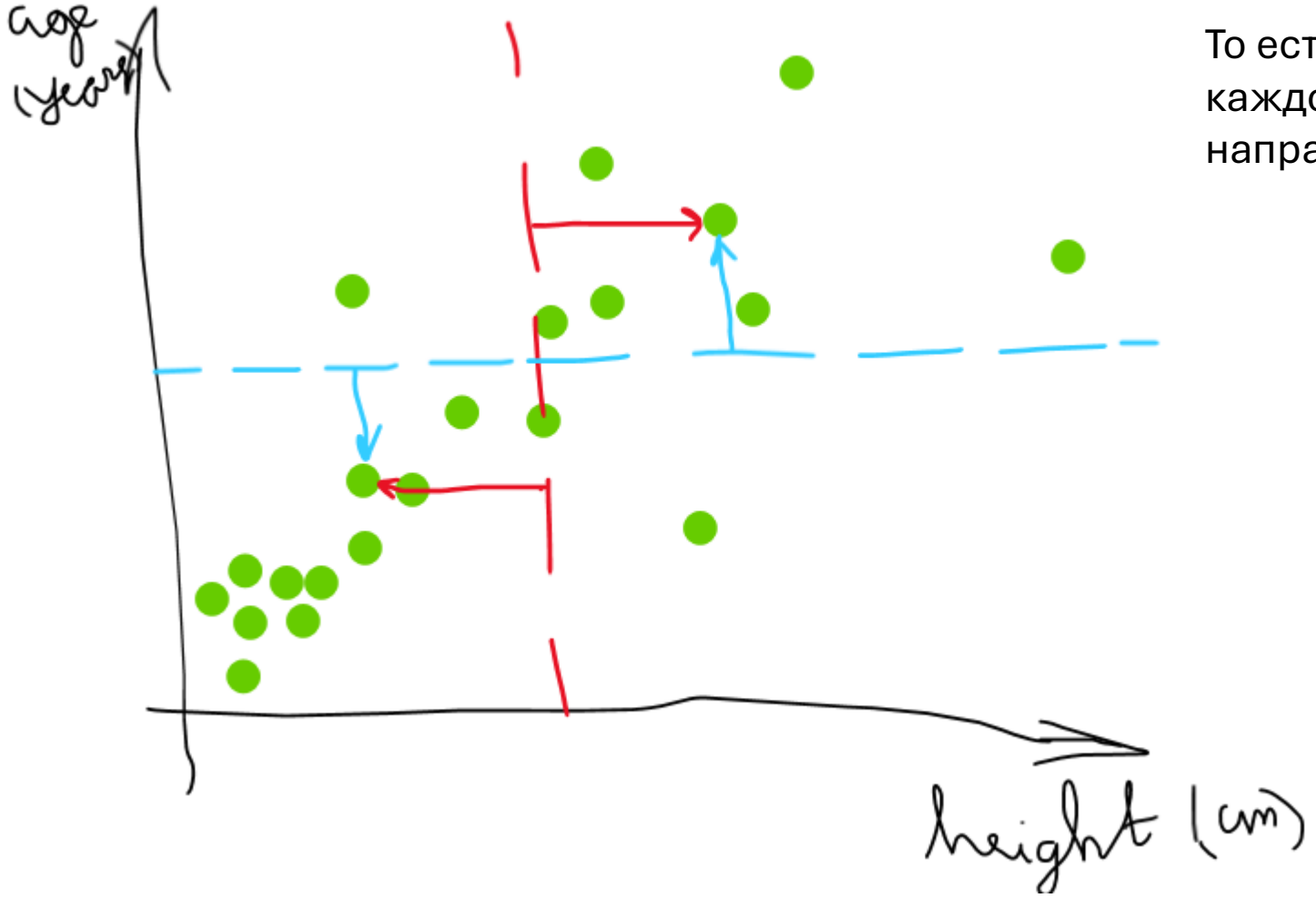
Ковариация

Звучит логически. А как измерить ковариацию? Подумаем графически:



Ковариация

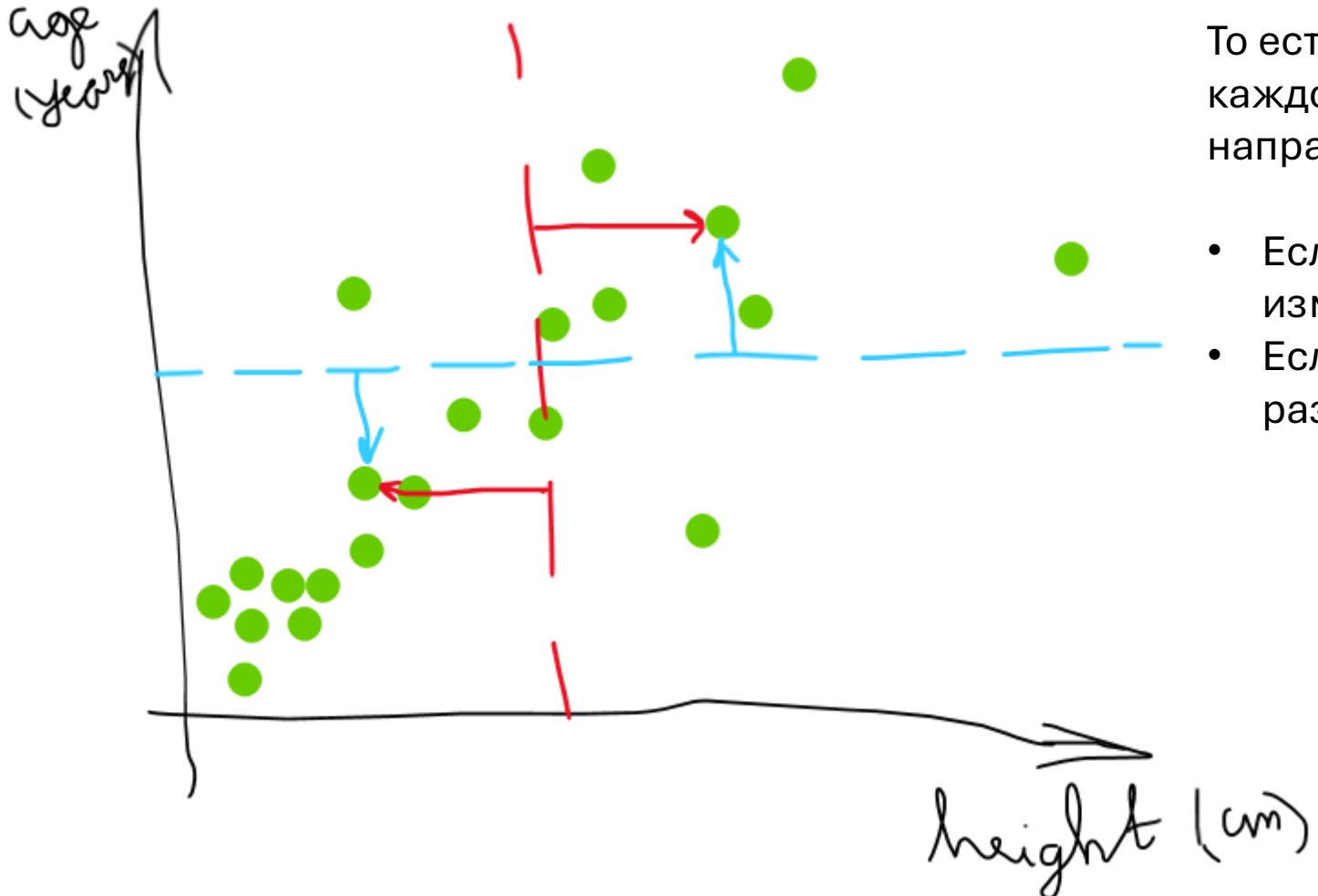
Звучит логически. А как измерить ковариацию? Подумаем графически:



То есть по со-направленности отклонений по каждой переменной мы можем судить о направлении связи.

Ковариация

Звучит логически. А как измерить ковариацию? Подумаем графически:

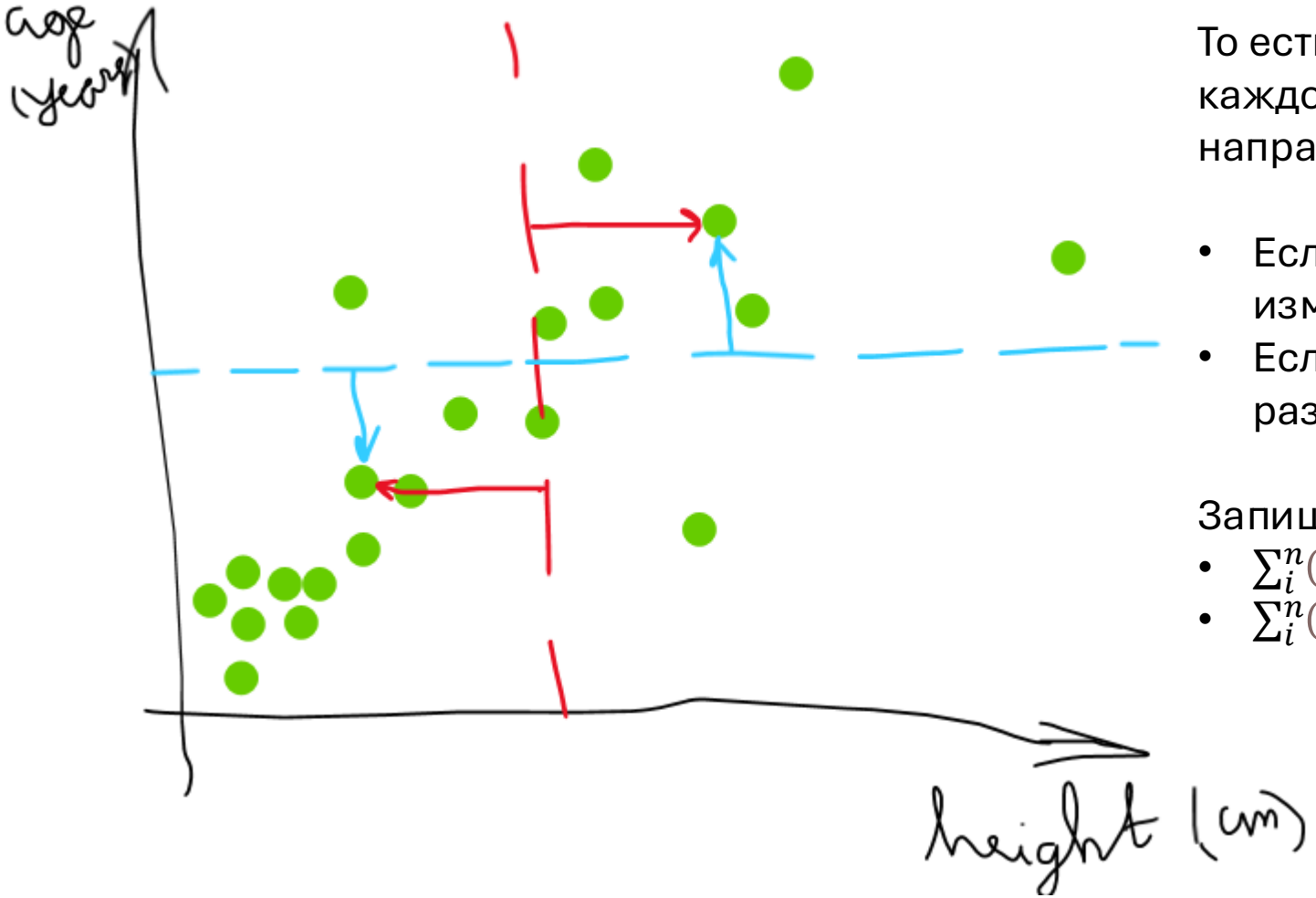


То есть по со-направленности отклонений по каждой переменной мы можем судить о направлении связи.

- Если произведение отклонений положительно – изменения сонаправлены.
- Если произведение отклонений отрицательно – разнонаправлены.

Ковариация

Звучит логически. А как измерить ковариацию? Подумаем графически:



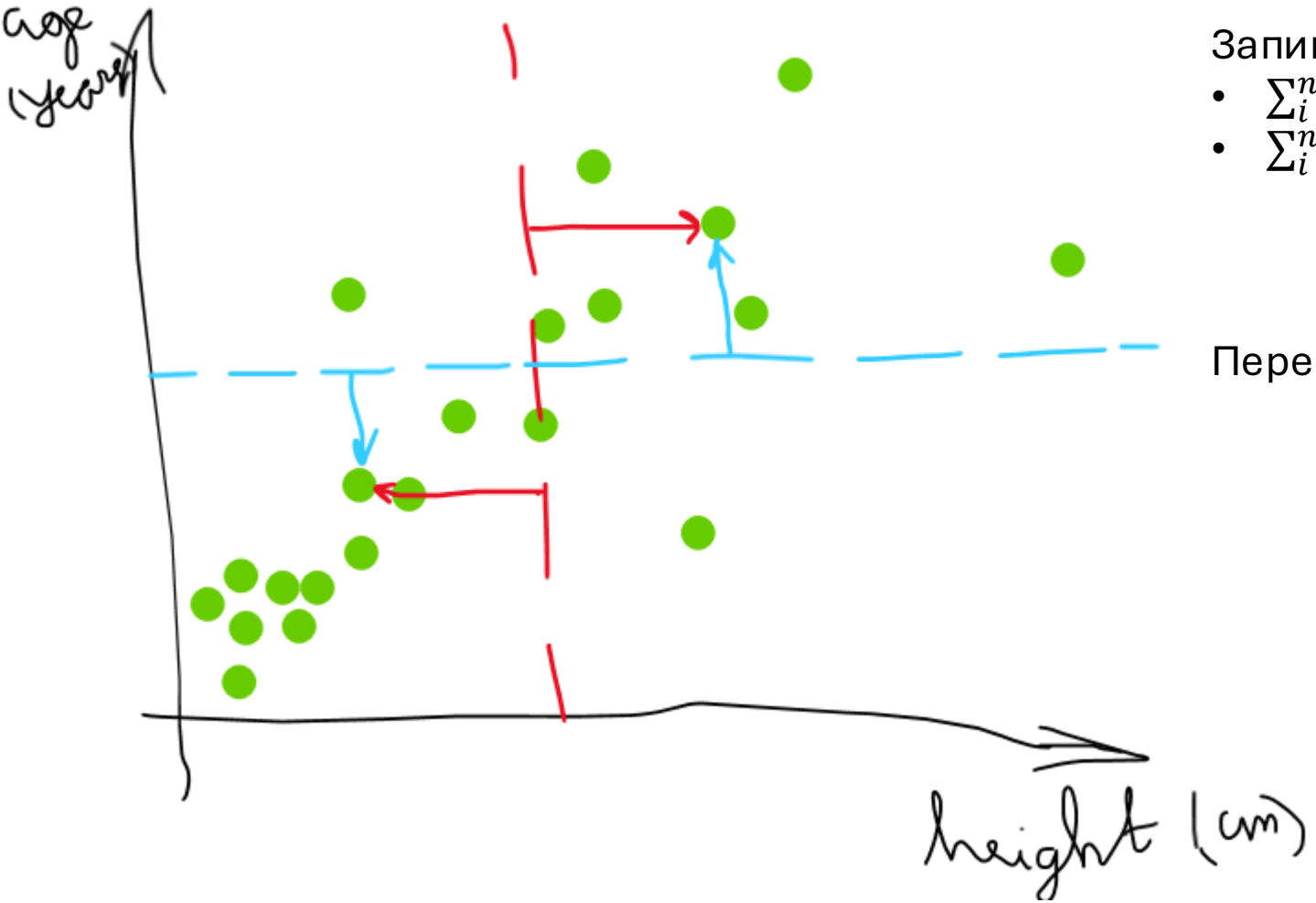
То есть по со-направленности отклонений по каждой переменной мы можем судить о направлении связи.

- Если произведение отклонений положительно – изменения сонаправлены.
- Если произведение отклонений отрицательно – разнонаправлены.

Запишем математически:

- $\sum_i^n (x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0$ – со-направлены
- $\sum_i^n (x_i - \bar{x}) \cdot (y_i - \bar{y}) < 0$ – разнонаправлены

Ковариация



Запишем математически:

- $\sum_i^n (x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0$ – со-направлены
- $\sum_i^n (x_i - \bar{x}) \cdot (y_i - \bar{y}) < 0$ – разнонаправлены

Перейдем к ковариации:

$$\boldsymbol{Cov} = \frac{\mathbf{1}}{\boldsymbol{n-1}} \sum_{\boldsymbol{i}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})$$

Ковариация

Важно! Ковариация позволяет исследовать **линейную** связь между переменными.

Ограничения ковариации:

1) Размерная величина – зависит от единиц измерения признаков
2) Зависит от дисперсий признаков, поэтому по её значению можно определить только направление связи (прямая или обратная), однако ничего нельзя сказать о силе связи

Корреляция

Раз ковариация зависит от дисперсии, то можно сделать некоторые математические преобразования, чтобы привести эмпирические распределения к какому-то одному виду — сделать так, чтобы они имели одинаковое среднее (математическое ожидание) и одинаковую дисперсию. Для этого используем стандартизацию:

$$x^* = \frac{x - \bar{x}}{\sigma}$$

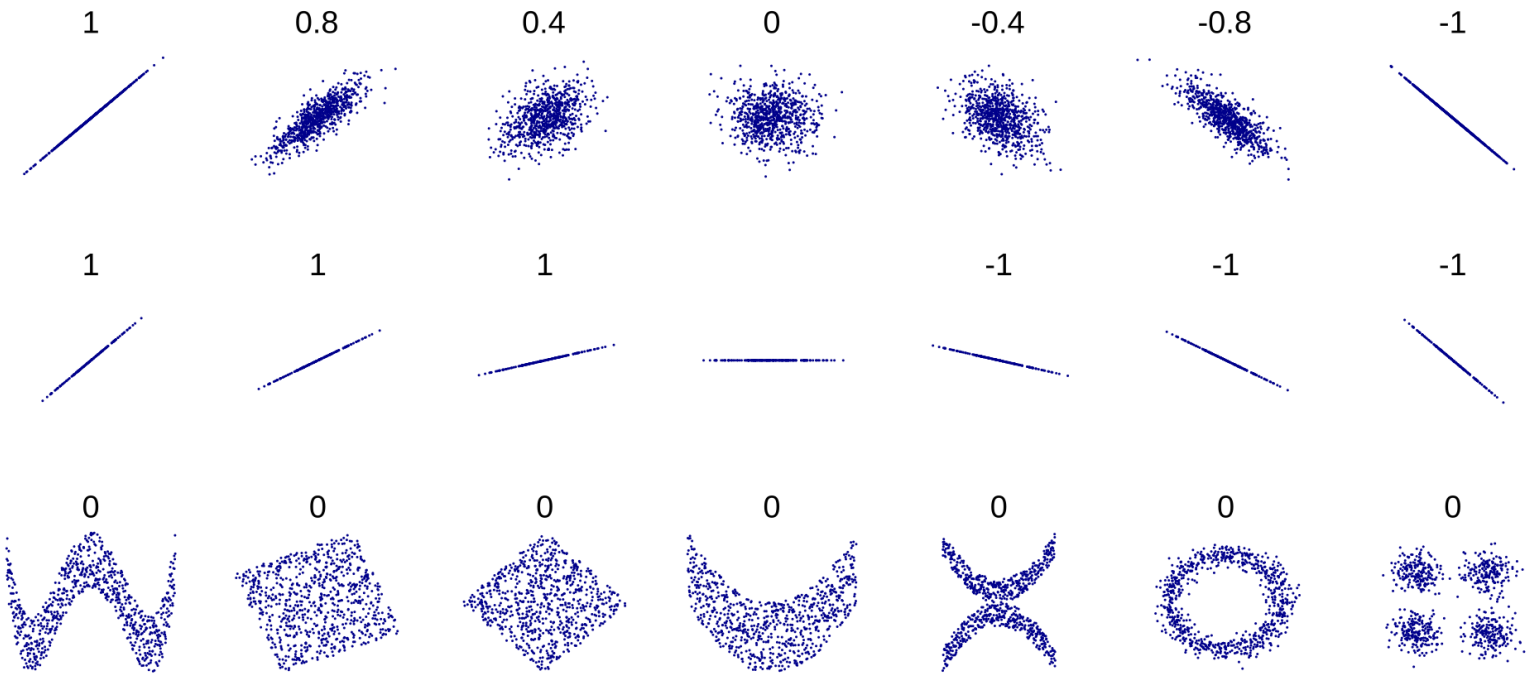
Чему будет равно среднее и стандартное отклонение переменной ***x****?

Корреляция

Ковариация двух стандартизованных величин называется **корреляцией** (**correlation**).

$$\boldsymbol{Cov}(\boldsymbol{x}^*, \boldsymbol{y}^*) = \boldsymbol{Corr}(\boldsymbol{x}, \boldsymbol{y})$$

Гипотезы:
$H_0: Corr(x, y) = 0$
$H_1: Corr(x, y) \neq 0$



Коэффициенты корреляции

1. Коэффициент корреляции Пирсона
2. Коэффициент корреляции Спирмена
3. Коэффициент корреляции Кендалла
4. Критерий хи-квадрат Пирсона

Коэффициент корреляции Пирсона

- Измеряет силу линейной связи между переменными
- Варьируется в диапазоне: [-1; 1]
- Для переменных, распределение которых соответствует нормальному
- Измеряет направление и силу связи

Абсолютное значение коэффициента	Интерпретация
0 – 0.2	Очень слабая
0.2 – 0.5	Слабая
0.5 – 0.7	Средняя
0.7 – 0.9	Сильная
0.9 – 1	Очень сильная

Коэффициент корреляции Спирмена

- Измеряет силу линейной связи между переменными
- Варьируется в диапазоне: [-1; 1]
- Для переменных, распределение которых отличается от нормального
- Измеряет направление и силу связи
- Используется, когда количество уникальных значений большое

Коэффициент корреляции Кендалла

- Измеряет силу линейной связи между переменными
- Варьируется в диапазоне: [-1; 1]
- Для переменных, распределение которых отличается от нормального
- Измеряет направление и силу связи
- Используется, когда количество уникальных значений небольшое

Коэффициент корреляции χ^2

- Проверяет взаимную независимость между категориальными переменными
- Варьируется в диапазоне: [0; +∞)
- НЕ определяет силу и направление связь
- Используется при наличии достаточного количества частот в ячейках таблицы сопряженности (>=95% ячеек должны содержать ожидаемую частоту >5)
- H0: две категориальные переменные независимы
- H1: две категориальные переменные зависимы

Коэффициенты корреляции

Коэффициент корреляции	Переменные	Допущения
Пирсона	Интервальные, отношений	Распределение нормальное
Спирмена	Ранговые, интервальные, отношений	Распределение ненормальное, много уникальных значений
Кендалла	Ранговые	Распределение ненормальное, мало уникальных значений
Хи-квадрат	Категориальные	>=95% ячеек содержат ожидаемую частоту >5