Понятие корреляционной зависимости

Многие задачи требуют установить и оценить зависимость между двумя или несколькими случайными величинами.

- Определение. Зависимость случайных величин называют *статистической*, если изменение одной величины влечет изменение распределения другой величины.
- Определение. Статистическая зависимость называется *корреляционной*, если при изменении одной величины изменяется среднее значение другой.

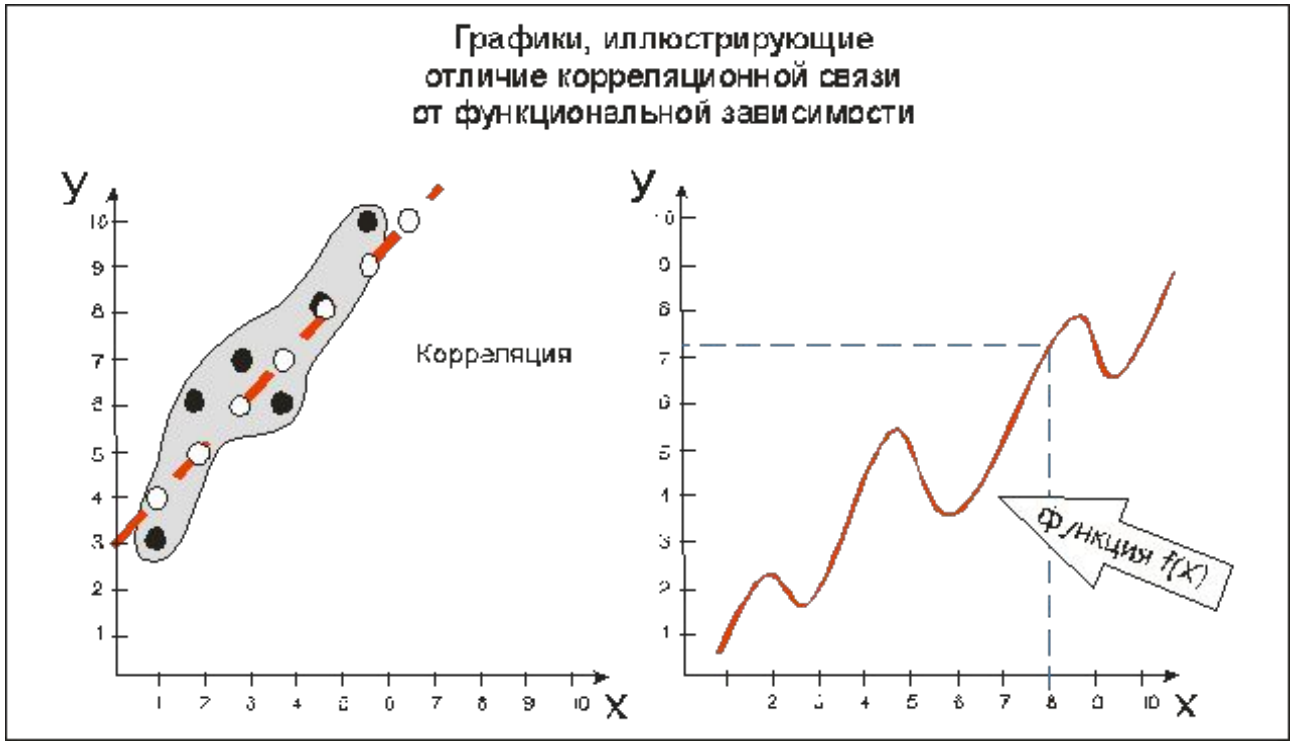
  Если случайная величина представляет некоторый признак (например, статистические наблюдения некой экономической величины), то под **корреляцией** понимают – меру согласованности одного признака с другим, или с несколькими, либо взаимную согласованность группы признаков.

Ложная корреляция

- **Корреляционная зависимость** указывает на причинно-следственную связь изменений двух признаков. Однако, корреляционные методы не выявляют этой причинности, а лишь указывают на наличие некоторого соответствия. Признаки могут находиться не только во взаимной зависимости друг от друга, но и оба зависеть от какого-либо третьего воздействия, не включенного в область рассмотрения. Например, между двумя временными рядами (переменные, состоящие из наблюдений отстоящих на равные промежутки времени друг от друга) может быть сильная корреляционная зависимость, однако эта зависимость будет ложной, так как переменные сами зависят от времени.
- Таким образом, более корректно употреблять понятие корреляционная связь.

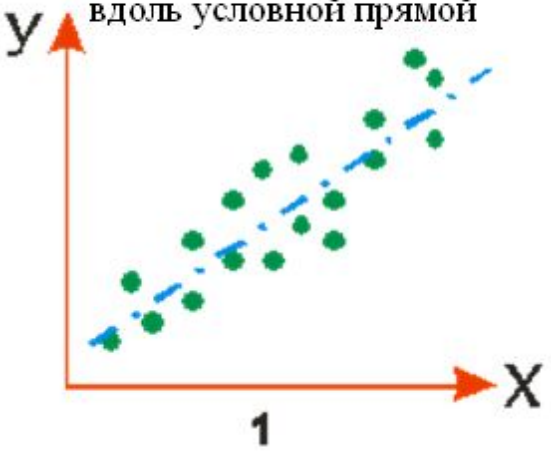
Отличие корреляционной от функциональной зависимости

Функциональная зависимость предполагает взаимно однозначное соответствие аргумента ***x*** и функции ***y=f(x)***, вероятностная же зависимость допускает некий условный диапазон, в который предположительно (с такой-то долей вероятности) попадает значение признака y_i при значении x_i признака **x**.

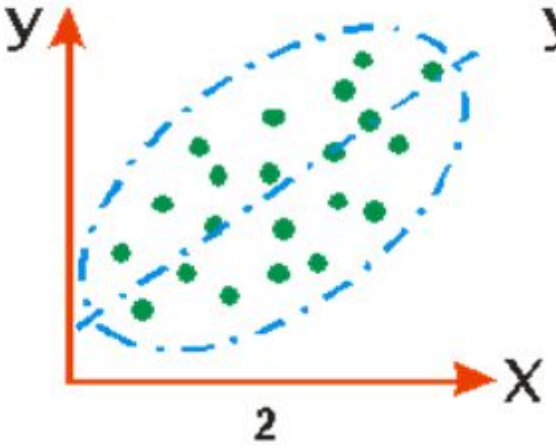


Примеры корреляционной зависимости

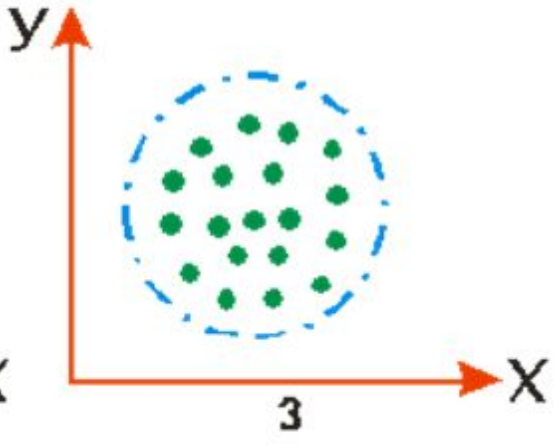
имеется значимая положительная корреляция: r>0,8
точки расположены примерно вдоль условной прямой



имеется некоторая корреляция, точки еще расположены вдоль прямой, но уже хаотично, вписываются в эллипс 0,5<r<0,6

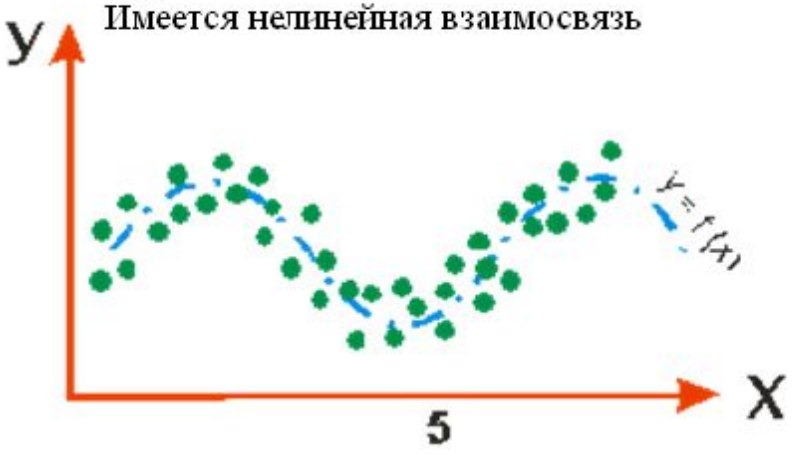


корреляция отсутствует: точки расположены хаотично (вписываются в окружность) r=0



Имеется значимая положительная корреляция r=+1, точки расположены вдоль прямой
Иначе: функциальная зависимость

Диаграмма рассеяния показывает однозначное соответствие: точки расположены вдоль линии y=cos(x)
Однако r=0!
Имеется нелинейная взаимосвязь



Коэффициент корреляции Пирсона

Коэффициент корреляции Пирсона характеризует наличие линейной связи между признаками,

$$r_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

де x_i — значения, принимаемые в выборке X,

$$\overline{X} = \frac{1}{n} \sum_{t=1}^{n} X_t$$

y_i — значения, принимаемые в выборке Y;

$\overline{x}$ — средняя по X, $\overline{y}$ — средняя по Y.

$$\overline{Y} = \frac{1}{n} \sum_{t=1}^{n} Y_t$$

Ведем обозначения: ковариация признаков X иY

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Средние квадратичные отклонения $\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}}$ и $\sigma_X = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$

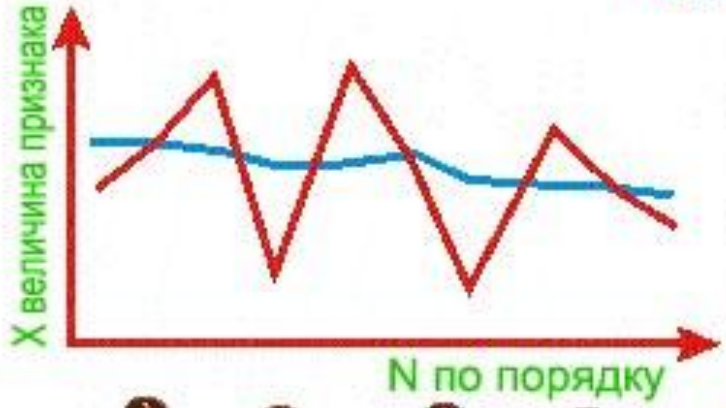
Тогда:

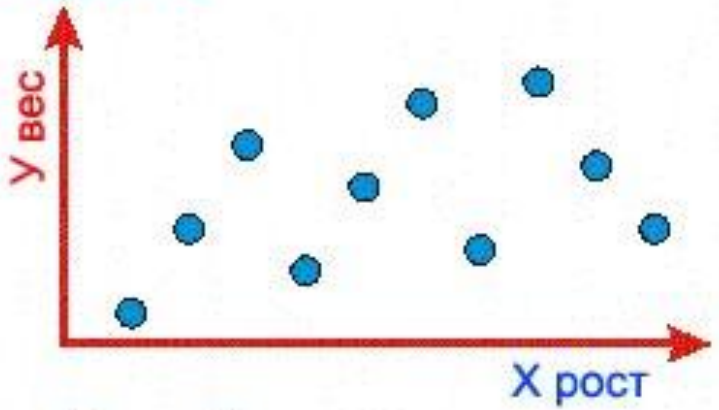
$$r_{xy} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \cdot \sigma_y}$$

Значение коэффициента корреляции

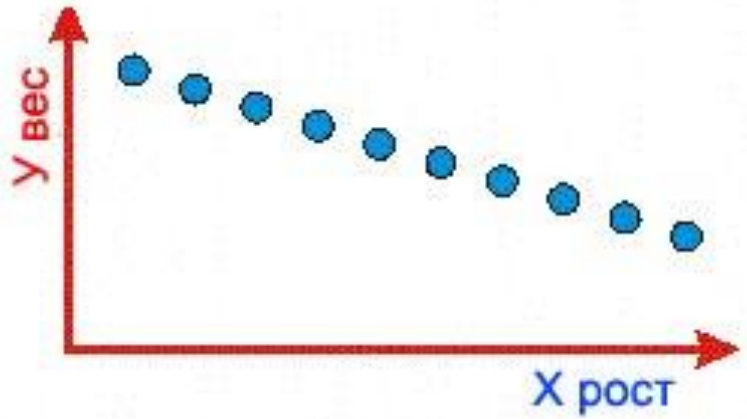
- **сильная**, или **тесная** при коэффициенте корреляции $r>0,70$;
- **средняя** при $0,50<r<0,69$;
- **умеренная** при $0,30<r<0,49$;
- **слабая** при $0,20<r<0,29$;
- **очень слабая** при $r<0,19$.
- Если коэффициент корреляции положительный, то связь между признаками прямая: увеличение одного признака приводит к увеличению другого
- Если коэффициент корреляции отрицательный, то связь между признаками обратная: увеличение одного признака приводит к уменьшению другого
- В случае, если $r=1, -1$, то связь между признаками функциональная!

Слабая линейная корреляция.
Почти при одинаковом росте все солдаты
то худые, то толстые.

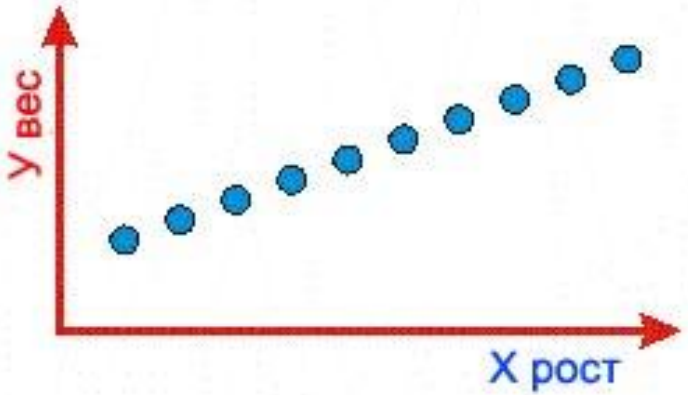




Сильная линейная корреляция.
Чем ниже солдат, тем он толще.
(Коэффициент корреляции отрицательный.)



Сильная линейная корреляция.
Чем выше солдат, тем он толще.



Непараметрические показатели корреляции

Определение. Под **качественным** подразумевается признак, который невозможно измерить точно, но он позволяет сравнить объекты между собой и расположить их в порядке убывания или возрастания качества.

Под ранжированием будем понимать упорядочивание объектов согласно убыванию качественного признака

Для оценки степени связи качественных признаков используют коэффициенты ранговой корреляции.

Коэффициент корреляции Спирмена — мера линейной связи между случайными величинами. Корреляция Спирмена является **ранговой**, то есть для оценки силы связи используются не численные значения, а соответствующие им ранги.

Коэффициент корреляции Кендалла — мера линейной связи между случайными величинами

Схема нахождения коэффициента Корреляции Спирмена

1. Определить, какие два признака или две иерархии признаков будут участвовать в сопоставлении как переменные *X* и *Y*.
2. Проранжировать значения переменной *X*, присваивая ранг 1 наименьшему значению, и т.д. Занести ранги в первый столбец таблицы по порядку номеров испытуемых или признаков.
3. Проранжировать значения переменной *У*, в соответствии с теми же правилами. Занести ранги во второй столбец таблицы по порядку номеров испытуемых или признаков.
4. Подсчитать разности *d* между рангами *X* и *Y* по каждой строке таблицы и занести в третий столбец таблицы.
5. Возвести каждую разность в квадрат: d^2. Эти значения занести в четвертый столбец таблицы.
6. Подсчитать сумму d^2.
7. При наличии одинаковых рангов рассчитать поправки:

$$T_a = \sum \left(a^3 - a\right)/12$$

$$T_b = \sum \left(b^3 - b\right)/12$$

где a - объем каждой группы одинаковых рангов в ранговом ряду *X*; b - объем каждой группы одинаковых рангов в ранговом ряду *Y*.

Схема нахождения коэффициента Корреляции Спирмена

8. Расчитать коэффициент ранговой корреляции r_s по формуле:

при отсутствии одинаковых рангов

$$r_s = 1 - 6 \cdot \frac{\sum d^2}{N \cdot \left(N^2 - 1\right)}$$

при наличии одинаковых рангов

$$r_s = 1 - 6 \cdot \frac{\sum d^2 + T_a + T_b}{N \cdot \left(N^2 - 1\right)}$$

где $sum(d^2)$ - сумма квадратов разностей между рангами;

T_a и T_b - поправки на одинаковые ранги;

N - количество наблюдений признаков, участвовавших в ранжировании.

Проверка значимости коэффициента ранговой корреляции Спирмена

Нулевая и альтернативная гипотезы имеют вид:

H_0*: *коэффициент ранговой корреляции Спирмена r_s незначимый;

H_1*: *коэффициент ранговой корреляции Спирмена r_s значим.

- Расчитывается t-статистика по формуле:

$$t_{расч.} = \frac{r_s}{\sqrt{1 - r_s^2}} \sqrt{(n-2)}$$

- Определяется $t_{табл}$ по таблице Стьюдента со степенями свободы n-2 и уровнем значимости α
- *Если* $|t_{расч}| > t_{табл}$ *, то* $\boldsymbol{H_0}$ отклоняют на заданном уровне значимости, и считаем, что коэффициент ранговой корреляции Спирмена значимый.

Схема нахождения коэффициента Корреляции Кенделла

Пусть ранги объектов выборки объема:

по признаку A $\quad x_1, \quad x_2, \quad \ldots, \quad x_n$

по признаку B $\quad y_1, \quad y_2, \quad \ldots, \quad y_n$

Допустим, что правее y_1 имеется R_1 рангов, бóльших y_1, правее y_2 имеется R_2 рангов, бóльших y_2; ...; правее y_{n-1} имеется R_{n-1} рангов, бóльших y_{n-1}. Введем обозначение суммы рангов $R_i (i = 1, 2, \ldots, n-1)$:

$$R = R_1 + R_2 + \ldots + R_{n-1}$$

Выборочный коэффициент ранговой корреляции Кендалла определяется формулой

$$\tau_в = \frac{4R}{n(n-1)} - 1$$

где n – объем выборки, $R = \sum_{i=1}^{n-1} R_i$

В случае «полной прямой зависимости» признаков $\tau_в = 1$

В случае «противоположной зависимости» $\tau_в = -1$

Проверка значимости коэффициента ранговой корреляции Кендалла

Правило. Для того чтобы при уровне значимости α, проверить нулевую гипотезу о равенстве нулю генерального коэффициента ранговой корреляции τ Кендалла при конкурирующей гипотезе $H_1 : \tau_г \neq 0$, надо вычислить критическую точку:

$$T_{кр} = z_{кр}\sqrt{\frac{2(2n+5)}{9n(n-1)}}$$

где n – объем выборки; $z_{кр}$ – критическая точка двусторонней критической области, которую находят по таблице функции Лапласа по равенству $\Phi(z_{кр}) = \frac{1-\alpha}{2}$.

Если $|\tau_в| < T_{кр}$ – нет оснований отвергнуть нулевую гипотезу. Ранговая корреляционная связь между качественными признаками незначимая.

Если $|\tau_в| > T_{кр}$ – нулевую гипотезу отвергают. Между качественными признаками существует значимая ранговая корреляционная связь.