УНИВЕРСИТЕТ ИТМО

Основы анализа данных. Корреляционный анализ.

Лекция 5

КМАИ

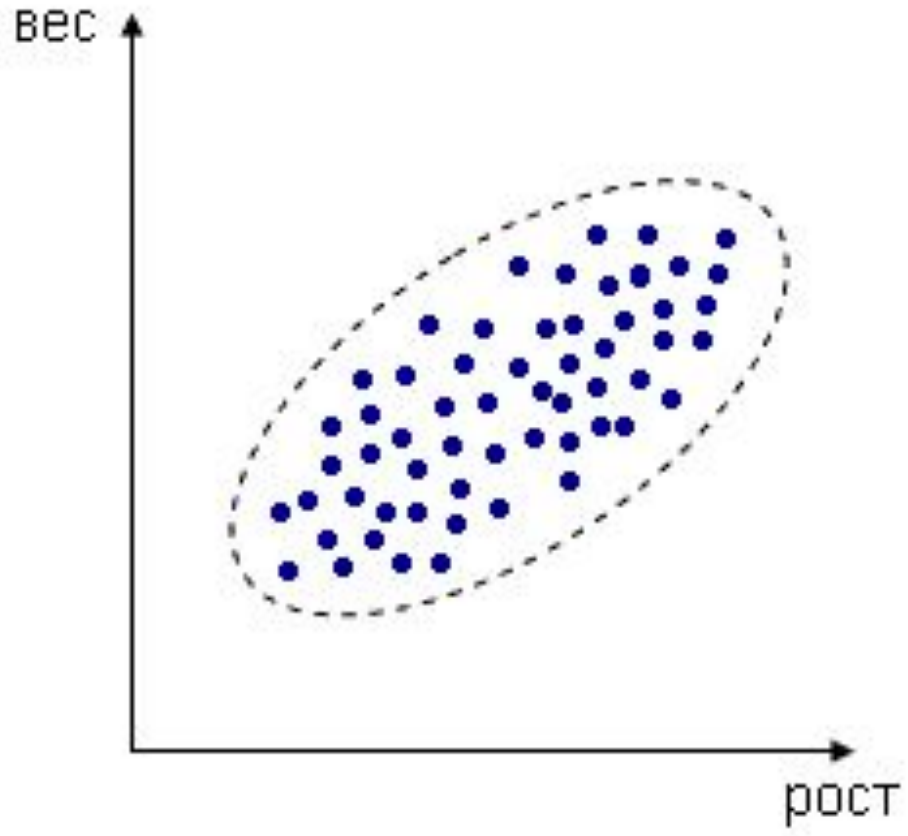
Определение.
Виды зависимостей

Показатели корреляции

Проверка значимости корреляции

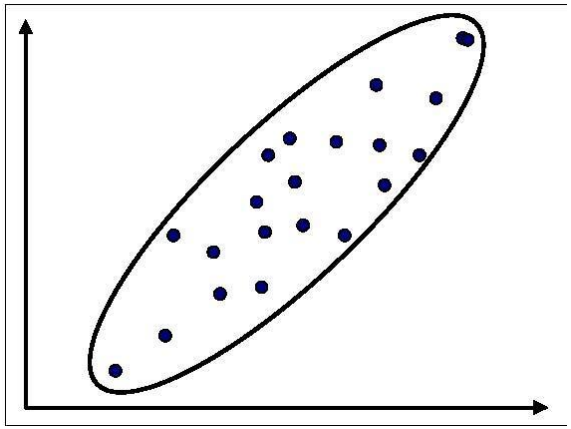
Определение

Корреляционный анализ - метод, позволяющий обнаружить зависимость между несколькими случайными величинами.

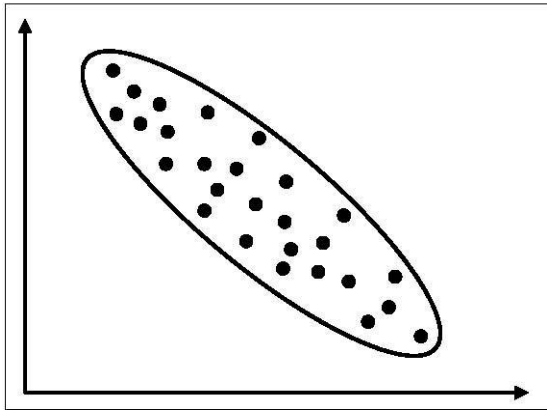


Виды зависимостей

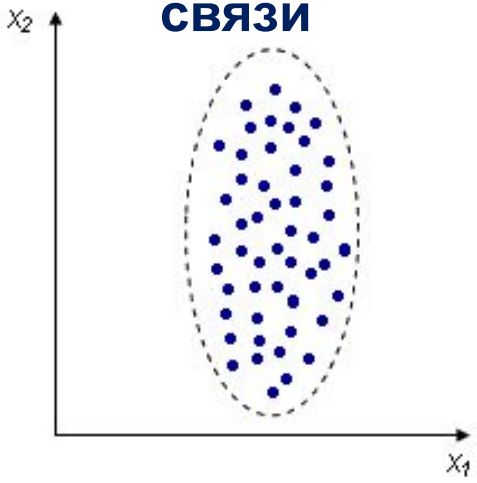
Положительная статистическая связь



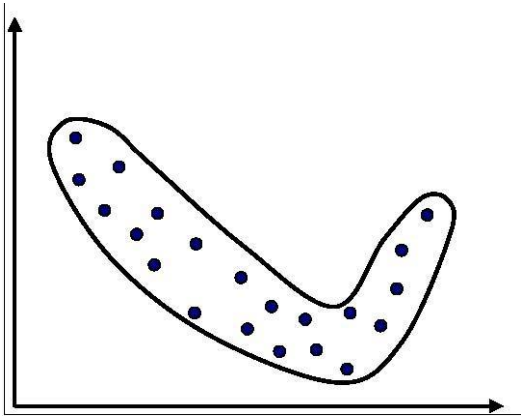
Отрицательная статистическая связь



Отсутствие статистической связи



Нелинейная статистическая связь



Ограничения корреляционного анализа

Ограничения корреляционного анализа:

1. Применение возможно при наличии достаточного количества наблюдений для изучения.
2. Необходимо, чтобы совокупность значений всех факторных и результативного признаков подчинялась многомерному нормальному распределению.
3. Исходная совокупность значений должна быть качественно однородной.
4. Сам по себе факт корреляционной зависимости не даёт основания утверждать, что одна из переменных предшествует или является причиной изменений, или то, что переменные вообще причинно связаны между собой, а не наблюдается действие третьего фактор.

Секции

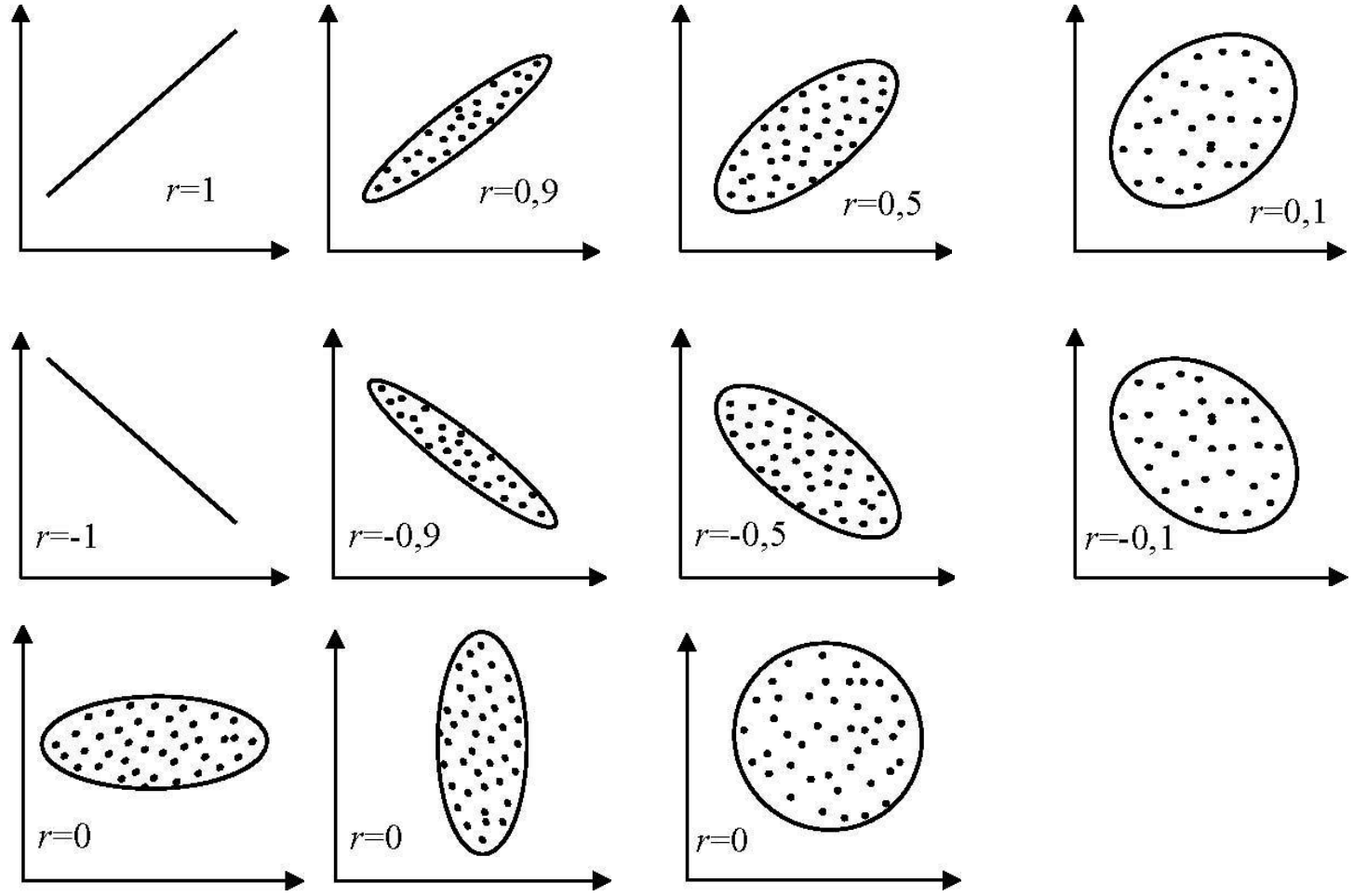
Определение.
Виды зависимостей

Показатели корреляции

Проверка значимости корреляции

Показатели корреляции

Коэффициент корреляции - количественная оценка тесноты взаимосвязи двух случайных величин.



Показатели корреляции

Линейный коэффициент корреляции (Пирсона) – параметрический показатель линейной зависимости двух величин, заданных на количественной или интервальной шкале.

$$r_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{(X - \bar{X})^2}\sqrt{(Y - \bar{Y})^2}} = \frac{cov_{XY}}{\sigma_X \sigma_Y}$$

$$cov_{XY} = M\left[\left(X - M(X)\right)\left(Y - M(Y)\right)\right] = M(XY) - M(X)M(Y)$$

Показатели корреляции

Коэффициент ранговой корреляции Кендалла – рименяется для выявления взаимосвязи между количественными или качественными показателями, если их можно ранжировать.

$$r_{XY} = \frac{2S}{n(n-1)}$$

Где

$$S = P - Q$$

P - суммарное число наблюдений, следующих за текущими наблюдениями с большим значением рангов Y

Q - суммарное число наблюдений, следующих за текущими наблюдениями с меньшим значением рангов Y. (равные ранги не учитываются!)

Коэффициент ранговой корреляции Спирмена.

X, Y - Признаки

$[(X_1, Y_1), \dots, (X_n, Y_n)]$ - наблюдения

R_1 R_n - ранги

d_i - расстояния между рангами

$$r_{XY} = 1 - \frac{6 \sum {d_i}^2}{n(n^2 - 1)}$$

Свойства коэффициента корреляции

1. Диапазон значений коэффициента корреляции

$$-1 \leq r_{XY} \leq 1$$

2. Граничные значения коэффициента корреляции

Коэффициент корреляции равен ± 1 тогда и только тогда, когда X и Y линейно зависимы

Интерпретация коэффициента корреляции

$\lvert r_{XY}\rvert = 1$	**Функциональная зависимость**
$0{,}7 \leq \lvert r_{XY}\rvert \leq 0{,}99$	**Сильная статистическая зависимость**
$0{,}5 \leq \lvert r_{XY}\rvert \leq 0{,}69$	**Средняя статистическая зависимость**
$0{,}2 \leq \lvert r_{XY}\rvert \leq 0{,}49$	**Слабая статистическая зависимость**
$0{,}09 \leq \lvert r_{XY}\rvert \leq 0{,}19$	**Очень слабая статистическая зависимость**
$\lvert r_{XY}\rvert < 0{,}09$	**Корреляции нет**

Определение.
Виды зависимостей

Показатели корреляции

Проверка значимости корреляции

Проверка значимости корреляции

r_{XY} - коэффициент корреляции, вычисленный по выборке объемом **n**

$H_0: r = 0$ - Гипотеза о равенстве нулю коэффициента корреляции генеральной совокупности (на уровне значимости α)

Варианты проверки гипотезы:

1. Сравнение расчетного значения коэффициента корреляции с табличным
2. Использование статистики, имеющей распределение Стьюдента (с последующим сравнением с табличным значением при заданном α)

$$t = r\sqrt{n-2}/\sqrt{1-r^2}$$

3. Вычисление уровня значимости α и принятие гипотезы, если меньше 5%.

Вопросы

1. Ковариация, коэффициент корреляции Пирсона.
2. Коэффициенты ранговой корреляции.
3. Определение корреляционного анализа. Виды зависимостей.
4. Ограничения корреляционного анализа. Шкала Чеддока.
5. Проверка гипотезы значимости корреляции.