3. Элементы статистического анализа (продолжение)

3.3 Поиск статистических взаимосвязей между признаками

3.3.1 Основные понятия

Взаимосвязи между признаками

Взаимосвязь между признаками выражается в некоторой закономерности встречаемости значений этих признаков.

В самом общем виде взаимосвязи делят на

- ***функциональные***,
- ***корреляционные***.

Функциональная зависимость

Если каждому возможному значению СВ ***X*** соответствует одно определенное возможное значение СВ ***Y***, то ***Y*** называется функцией случайного аргумента ***X***:

$$Y = \varphi(X).$$

Пример.

X	-2	2	3
p	0,2	0,5	0,3

Y	4	9
p	0,7	0,3

На практике строгая функциональная зависимость реализуется редко.

Пример: СВ *X* - объем (в руб.) сделок, заключенных менеджером в течение календарного периода,
СВ *Y* - доход менеджера с учетом % от заключенных сделок

Корреляционная зависимость

Статистической называется зависимость, при которой изменение одной из СВ влечет за собой изменение закона распределения другой СВ.

В частности, статистическая зависимость может проявляться в том, что при изменении одной СВ изменяется среднее значение другой СВ.

В этом случае статистическая зависимость называется ***корреляционной***.

Корреляционная зависимость

Важно:

- Корреляционная связь описывает тенденцию;

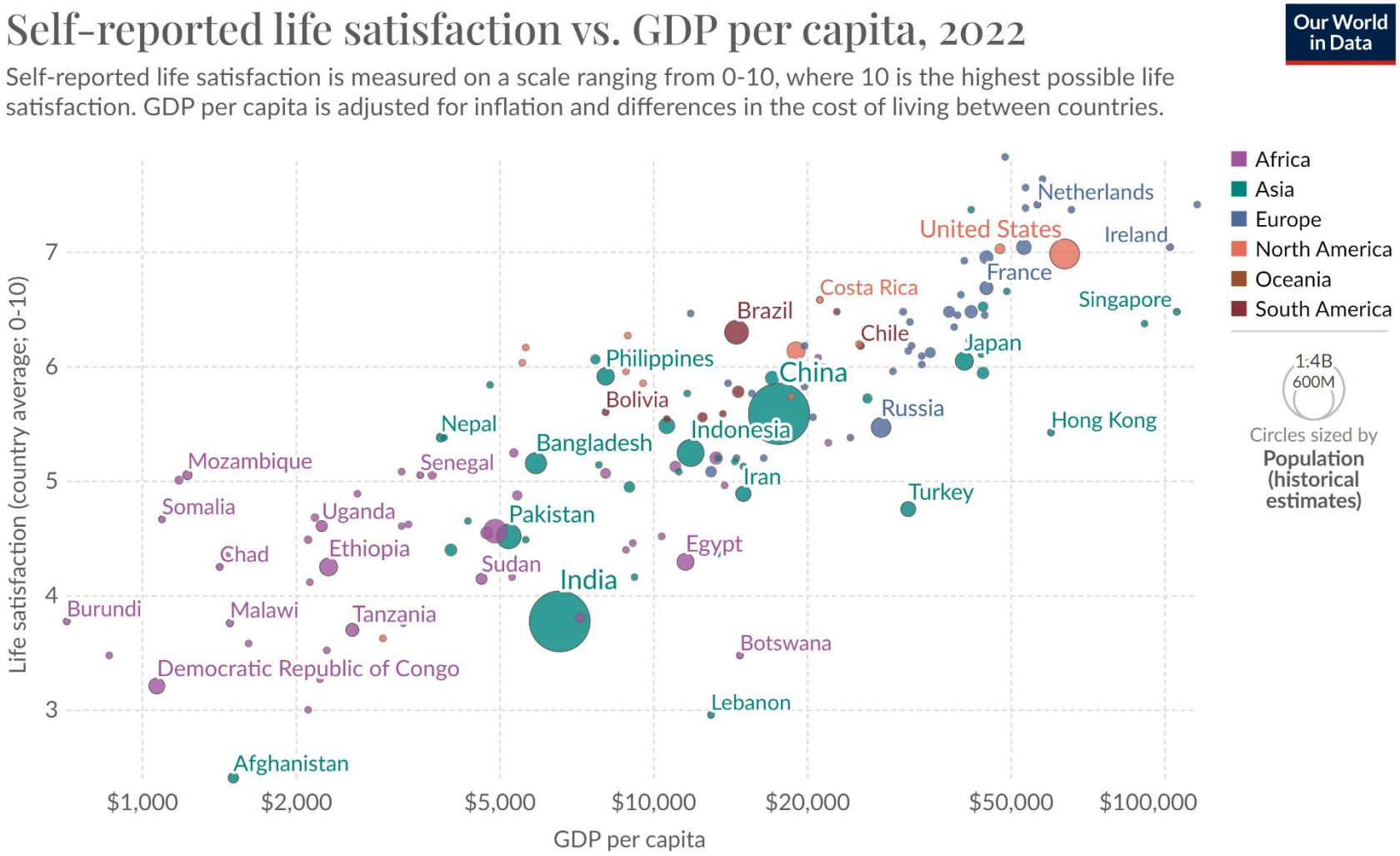
  при этом изменение одной величины может не иметь никаких последствий в отдельно взятом наблюдении другой величины.

  Например: установлена корреляционная взаимосвязь между ВВП на душу населения и «уровнем счастья» (по субъективной самооценке) в разных странах; но это не означает, что при увеличении ВВП на 1 у. е. каждый конкретный человек станет на 1 пункт счастливее.

- Выявление корреляционной взаимосвязи между двумя величинами не означает установление каких-либо причинно-следственных связей.

Корреляционная зависимость

Источник данных.



Корреляционная зависимость

Корреляционная связь может быть

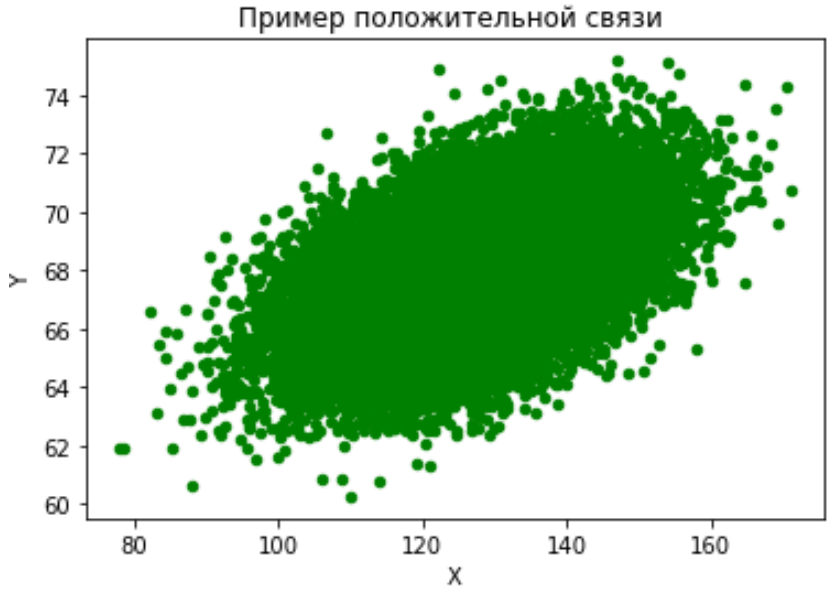
- *положительной* (*прямой*),
- *отрицательной* (*обратной*).

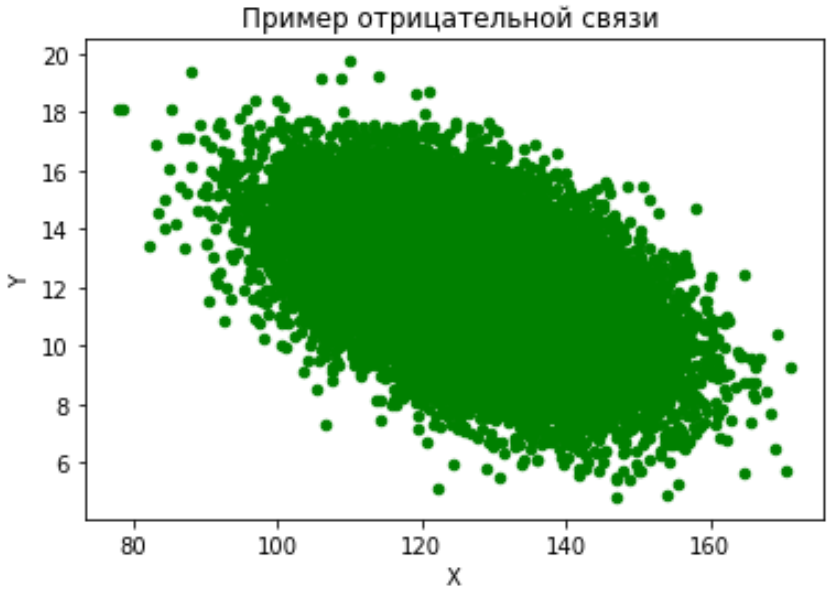
Положительная связь характеризуется тем, что рост значений одного признака сопровождается ростом значений другого признака (пример выше).

В случае ***отрицательной связи*** рост значений одного признака сопровождается уменьшением значений другого признака.

Положительные и отрицательные связи

Пример.





3.3.2 Выявление линейной взаимосвязи двух признаков (количественные данные)

Коэффициент корреляции

Для оценки силы взаимосвязи между количественными или порядковыми признаками используется мера, называемая ***коэффициентом корреляции***.

Эта мера вычисляется по-разному в зависимости от вида данных (количественные с распределением, близким к нормальному, либо не являющиеся таковыми).

Мера взаимосвязи количественных признаков

Мера линейной взаимосвязи количественных признаков ***X*** и ***Y*** – ***коэффициент корреляции Пирсона***:

$$r_{в} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}} ,$$

где n – число наблюдений (объем выборки);

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i , \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

Свойства коэффициента корреляции Пирсона

- Для любых СВ X и Y $-1 \le r_{в} \le 1$.

  При этом:

  - знак $r_{в}$ совпадает с характером связи (положительная или отрицательная);
  - чем ближе значение $|r_{в}|$ к 1, тем сильнее линейная связь между величинами X и Y ;
  - значение $r_{в}$, близкое нулю, означает отсутствие линейной связи (но возможна нелинейная связь).

- Величина $r_{в}$ чувствительна к выбросам.

Как и все метрики, основанные на вычислении средних

Свойства коэффициента корреляции Пирсона

Итог:

коэффициент корреляции Пирсона – хорошая мера линейной связи количественных признаков, если данные не содержат аномалий (в идеале – распределение, близкое к нормальному, без выбросов).

Если какие-то из указанных условий не выполнены, то следует использовать другие меры связи (будут рассмотрены далее).

Значимость коэффициента корреляции Пирсона

Важно:

вся генеральная совокупность обычно не доступна для анализа, и коэффициент корреляции (как и другие статистические показатели) вычисляется по выборке.

Таким образом, мы имеем ***выборочный коэффициент корреляции***, который является оценкой коэффициента корреляции $r_г$ генеральной совокупности.

Значимость коэффициента корреляции Пирсона

Предположим:

выборочный коэффициент корреляции $r_в$ оказался отличным от нуля.

Вопрос:

является ли это отличие значимым (означает ли оно, что и коэффициент корреляции $r_г$ генеральной совокупности также отличен от нуля, и можно предполагать наличие линейной связи между признаками в генеральной совокупности)?

Значимость коэффициента корреляции Пирсона

Для ответа на поставленный вопрос – проверка статистической гипотезы о значимости выборочного коэффициента корреляции.

H_0: $r_г = 0$ (корреляция отсутствует);

H_1: $r_г \neq 0$ (корреляция существует).

Применяется общая схема проверки статистических гипотез.

В качестве статистического критерия используется статистика

$$T = r_в \cdot \frac{\sqrt{n-2}}{\sqrt{1-r_в^2}},$$

(вычисляется по данным выборки).

Алгоритм проверки нулевой гипотезы

1. По данным выборки найти значение выборочного коэффициента корреляции $r_в$ и $t_{набл}$ – наблюдаемое значение СВ *T*.

2. Зная, что СВ ***T*** имеет *распределение Стьюдента* с ***n*** – 2 степенями свободы, найти ***p-value*** – вероятность того, что при условии справедливости нулевой гипотезы (признаки не коррелированы) по данным выборки будет получен выборочный коэффициент корреляции, не меньший $r_в$:

$$P(T \geq t_{набл}).$$

3. Сравнить значение ***p-value*** с заранее выбранным уровнем значимости ***a***:

   в случае ***p-value*** > ***a*** нулевая гипотеза принимается;

   в противном случае – отвергается.

Вычисление коэффициента корреляции Пирсона и проверка его значимости

Вычисления с помощью библиотеки Scipy.

Функция **pearsonr()** (модуль **stats** библиотеки **Scipy**) позволяет вычислить коэффициент корреляции Пирсона и выполнить оценку его значимости.

Возвращает значение выборочного коэффициента корреляции $r_в$ и величину ***p-value***.

Документация функции.

3.3.3 Выявление взаимосвязи двух признаков (порядковые данные)

Исследование порядковых данных

Для случаев, когда

- признаки измерены в порядковой шкале,
- либо данные содержат выбросы,
- либо распределение существенно отличается от нормального,

существуют другие меры взаимосвязи – ***коэффициенты ранговой корреляции Спирмена и (тау) Кендалла.***

Исследование порядковых данных

Эти показатели имеют общие черты с коэффициентом корреляции Пирсона:

- характеризуют наличие взаимосвязи между признаками, ее силу и направленность;
- могут принимать значения в диапазоне [-1, 1].

Отличие от коэффициента корреляции Пирсона:

основаны не на абсолютных значениях признаков, а на рангах.

Коэффициент корреляции Спирмена

1. Объекты выборки следует расположить в порядке «убывания» значений признака X;

   каждому из объектов приписать ранг $x_i = i$, равный порядковому номеру объекта.

2. Объекты выборки расположить в порядке «убывания» значений признака Y;

   каждому из объектов приписать ранг y_i, равный порядковому номеру объекта (для удобства сравнения индекс i по-прежнему равен порядковому номеру объекта по признаку X).

Итог – две последовательности рангов; причем в общем случае $x_i \neq y_i$.

Коэффициент корреляции Спирмена

3. Обозначим $d_i = x_i - y_i$.
4. Коэффициент ранговой корреляции Спирмена равен

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

где n – объем выборки.

Если несколько объектов имеют одно и то же значение по какому-то из признаков, то каждому из этих объектов приписывается ранг, равный среднему арифметическому порядковых номеров этих объектов

Коэффициент корреляции Кендалла

1. Объекты выборки следует расположить в порядке «убывания» значений признака X ;

   каждому из объектов приписать ранг $x_i = i$, равный порядковому номеру объекта.

2. Объекты выборки расположить в порядке «убывания» значений признака Y ;

   каждому из объектов приписать ранг y_i , равный порядковому номеру объекта (для удобства сравнения индекс i по-прежнему равен порядковому номеру объекта по признаку X).

Те же действия, что при вычислении коэффициента ранговой корреляции Спирмена

Коэффициент корреляции Кендалла

3. Обозначим $R = \sum_{i=1}^{n-1} R_i$,

   где R_i - число рангов, больших y_i ,

4. Коэффициент ранговой корреляции Кендалла равен

$$\tau = \frac{4R}{n(n-1)} - 1 ,$$

   где n – объем выборки.

Значимость коэффициентов ранговой корреляции

Вопросы о значимости коэффициентов ранговой корреляции Спирмена и Кендалла ставятся по аналогии с вопросом о значимости коэффициента корреляции Пирсона.

Для получения ответа на эти вопросы – проверка соответствующих статистических гипотез.

Вычисление коэффициентов ранговой корреляции и проверка их значимости

Вычисления с помощью библиотеки Scipy.

Функция **spearmanr()** (модуль **stats** библиотеки **scipy**) позволяет вычислить коэффициент ранговой корреляции Спирмена и выполнить оценку его значимости.

Возвращает значение выборочного коэффициента корреляции Спирмена и величину ***p-value***.

Документация функции.

Вычисление коэффициентов ранговой корреляции и проверка их значимости

Вычисления с помощью библиотеки Scipy.

Функция **kendalltau()** (модуль **stats** библиотеки **scipy**) позволяет вычислить коэффициент ранговой корреляции Кендалла и выполнить оценку его значимости.

Возвращает значение выборочного коэффициента корреляции Кендалла и величину ***p-value***.

Документация функции.

3.3.4 Исследование взаимосвязи двух категориальных признаков

Для значений признаков, измеренных в категориальной (номинальной) шкале, недоступны не только числовые операции, но и упорядочивание.

В таких случаях инструментом анализа являются ***таблицы сопряженности признаков***.

Таблицы сопряженности

Предположим:

имеющаяся выборка содержит n наблюдений, причем

- для признака X были зарегистрированы значения $x_1, x_2, \dots, x_k$,
- а для признака Y - значения $y_1, y_2, \dots, y_l$;

при этом пары значений $X = x_i$, $Y = y_j$ наблюдались n_{ij} раз.

Таблицы сопряженности

Результаты наблюдений могут быть представлены в виде ***таблицы сопряженности признаков***:

	y_1	y_2	…	y_l	Всего
x_1	n_{11}	n_{12}	…	n_{1l}	$M_{1\cdot} = \sum_{j=1}^{l} n_{1j}$
x_2	n_{21}	n_{22}	…	n_{2l}	$M_{2\cdot} = \sum_{j=1}^{l} n_{2j}$
…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kl}	$M_{k\cdot} = \sum_{j=1}^{l} n_{kj}$
Всего	$M_{\cdot 1} = \sum_{i=1}^{k} n_{i1}$	$M_{\cdot 2} = \sum_{i=1}^{k} n_{i2}$	…	$M_{\cdot l} = \sum_{i=1}^{k} n_{il}$	n

Таблицы сопряженности

Величины $M_{i\bullet}$ и $M_{\bullet j}$

называются ***маргинальными частотами***.

Они определяют одномерные распределения признаков.

Постановка задачи

Пусть данные наблюдений СВ ***X*** и ***Y*** представлены таблицей сопряженности.

Требуется проверить нулевую гипотезу:

признаки ***X*** и ***Y*** не связаны между собой зависимостью

при альтернативной гипотезе:

признаки ***X*** и ***Y*** являются связанными.

Применение критерия χ^2

Статистический критерий χ^2 позволяет оценить меру расхождения эмпирических (наблюдаемых) и теоретических (вычисленных в предположении справедливости нулевой гипотезы) частот.

Обозначим:

$p_i = P(X = x_i)$, $q_j = P(Y = y_j)$ – теоретические вероятности;

$p_i^* = \frac{1}{n}\sum_{j=1}^{l} n_{ij} = \frac{1}{n} M_{i\,.}$, $q_j^* = \frac{1}{n}\sum_{i=1}^{k} n_{ij} = \frac{1}{n} M_{.\,j}$ –

статистические оценки вероятностей p_i и q_j (относительные частоты по данным таблицы сопряженности).

Применение критерия χ^2

Если СВ X и Y независимы, то вероятность совместного наблюдения $X = x_i$ и $Y = y_j$ равна

$$P\left((X = x_i)\&\left(Y = y_j\right)\right) = p_i \cdot q_j,$$

а статистическая оценка вероятности этого события -

$$p_i^* \cdot q_j^* = \frac{1}{n^2} M_{i\,\cdot} \cdot M_{\cdot\,j}\,.$$

Применение критерия χ^2

Суммарное расхождение эмпирических и теоретических частот по всем наблюдениям оценивается величиной

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{\left(n_{ij} - n \cdot p_i^* \cdot q_j^*\right)^2}{n \cdot p_i^* \cdot q_j^*}$$

или

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{\left(n_{ij} - \frac{1}{n} M_{i\,.} \cdot M_{.\,j}\right)^2}{\frac{1}{n} M_{i\,.} \cdot M_{.\,j}} \ .$$

СВ χ^2 имеет так называемое распределение «хи-квадрат» с числом степеней свободы равным $(k-1)(l-1)$.

Алгоритм проверки нулевой гипотезы

1. По данным таблицы сопряженности найти наблюдаемое значение критерия $\chi^2_{\text{набл}}$ и число степеней свободы.
2. Найти ***p-value*** – вероятность того, что при условии справедливости нулевой гипотезы (признаки независимы) по данным выборки будет получено значение критерия, не меньшее $\chi^2_{\text{набл}}$:

$$P\left(\chi^2 \geq \chi^2_{\text{набл}}\right).$$

3. Сравнить значение ***p-value*** с заранее выбранным уровнем значимости ***a***:

   в случае ***p-value*** > ***a*** нулевая гипотеза принимается;

   в противном случае – отвергается.

Применение критерия χ^2

Замечание 1.

Критерий χ^2 считается надежным, только если в таблице сопряженности не слишком много клеток с небольшими частотами (количество клеток с частотами менее 5 не должно превышать 20%).

Замечание 2.

Критерий χ^2 позволяет только проверить наличие взаимосвязи между признаками, но не позволяет оценить силу и направленность связи.

Применение критерия χ^2

Вычисления с помощью Python-библиотек.

- Функция **crosstab()** (библиотека **Pandas**) позволяет сформировать таблицу сопряженности из двух массивов/списков/серий наблюдаемых значений признаков.

  Возвращает объект DataFrame, содержащий таблицу сопряженности.

  Документация функции.

Применение критерия χ^2

Вычисления с помощью Python-библиотек.

- Функция **chisquare()** (модуль **stats** библиотеки **scipy**) позволяет проверить гипотезу о независимости признаков.

Возвращает наблюдаемое значение χ^2 и величину ***p-value***.

Документация функции.

Применение критерия χ^2

Параметры:

f_obs – массив/список наблюдаемых частот;

f_exp – массив/список теоретических частот;

ddof – число наложенных связей: в данном случае ***k*** + ***l*** – 2; (по умолчанию ddof=0);

axis - устанавливает измерения, к которым применяется критерий;
при axis = None все значения f_obs обрабатываются как один набор;
по умолчанию axis=0.

Применение критерия χ^2

Вычисления с помощью Python-библиотек.

- Функция **chi2_contingency()** (модуль **stats** библиотеки **scipy**) позволяет проверить гипотезу о независимости признаков. В отличие от **chisquare()**, предварительное вычисление наблюдаемых и теоретических частот не требуется (уже реализовано в коде функции).

Возвращает наблюдаемое значение χ^2 и величину ***p-value***.

Документация функции.

Коэффициент Крамера

При исследовании зависимостей между признаками важно не только выявить наличие взаимосвязей, но и оценить силу связи (при ее наличии).
Например, в случае, когда нужно определить, какие факторы сильнее других влияют на некоторый целевой признак.

Для оценки силы взаимосвязи между категориальными признаками используется ***коэффициент Крамера***:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min\{\, k-1,\ l-1 \,\}}}$$

(с учетом принятых ранее обозначений).

Свойства коэффициента Крамера

- Для любых признаков ***X*** и ***Y*** $0 \le V \le 1$.

  При этом:

  - значение ***V***, близкое нулю, означает отсутствие связи между ***X*** и ***Y***;
  - значение ***V***, равное 1, означает полное совпадение.

- Значение ***V*** можно использовать для сравнения силы связи между различными парами признаков.
- Не позволяет судить о характере и направленности связи.