Лекция №8 Корреляционный анализ

Цель лекции:

- ✓ Познакомиться с понятием корреляции
- ✓ Научиться рассчитывать и интерпретировать коэффициент корреляции Пирсона.
- ✓ Изучить ковариацию
- ✓ Рассчитать и интерпретируем коэффициент корреляции Спирмена.
- ✓ Рассмотреть реальный пример применения коэффициента корреляции Спирмена.

Материал прошлого урока:

На прошлых занятиях мы рассматривали тестирования гипотез и построение доверительных интервалов. На этом уроке и следующем уроках мы познакомимся с корреляционным и регрессионным анализами, которые позволяют оценить тесноту линейной связи и показать, как изменяется зависимая переменная при изменении независимой переменной.

План урока:

1. Корреляция
2. Интерпретация коэффициента корреляции
3. Слабые стороны корреляционного анализа
4. Ковариация
5. Коэффициент корреляции Спирмена

Корреляция

В реальной жизни перед нами часто встает задача, где надо понять, а есть ли взаимосвязь между двумя и более случайными величинами (СВ). И здесь на помощь приходит корреляционный и регрессионный анализы. Начнем изучение с корреляционного анализа. Так что же такое корреляция?

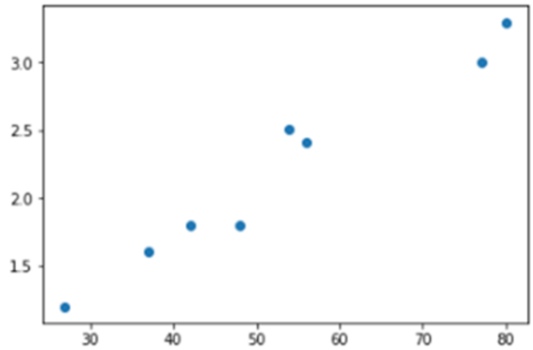
Корреляция – это математический показатель, по которому можно судить о наличии статистической взаимосвязи между двумя и более случайными величинами.

Но чтобы нам оценить в цифрах, насколько тесна линейная взаимосвязь, мы используем для расчета коэффициент корреляции. Иными словами, коэффициент корреляции – это коэффициент, который показывает, насколько велика линейная зависимость между случайными величинами.

Давайте взглянем на таблицу ниже:

Площадь	Цена
27	1.2
37	1.6
42	1.8
48	1.8
57	2.5
56	2.6
77	3
80	3.3

Здесь видим две переменные, площадь и цена квартиры. Мы расположили площадь по возрастанию и видим, что с ростом этой СВ в целом растет и цена. Лучше всего оценивать с помощью графика, который позволяет взглянуть на СВ целиком.



По графику также видим, что расположение данных напоминает прямую, что свидетельствует о наличии линейной зависимости. Но как же понять, насколько велика эта линейная взаимосвязь. И вот здесь приходит на помощь коэффициент корреляции.

С помощью функции corrcoef () из пакета numpy рассчитаем коэффициент корреляции между ценой(p) и площадью (s).

```
s=np.array([27, 37, 42, 48, 57, 56, 77, 80])
s
array([27, 37, 42, 48, 57, 56, 77, 80])

p = np.array([1.2, 1.6, 1.8, 1.8, 2.5, 2.6, 3, 3.3])
p
array([1.2, 1.6, 1.8, 1.8, 2.5, 2.6, 3. , 3.3])

np.corrcoef (p,s)
array([[1.        , 0.97857682],
       [0.97857682, 1.        ]])
```

Коэффициент корреляции 0.978. Единицы в этом массиве показывают корреляцию величины с самой собой.

Интерпретация коэффициента корреляции

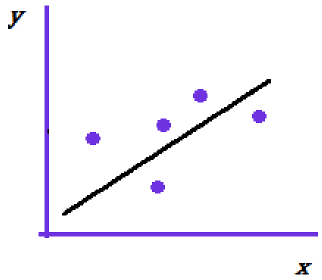
Коэффициент корреляции обозначается r или R и принимает значения [-1, 1]. Теснота линейной взаимосвязи определяется по модулю, чем ближе по модулю к 1, тем сильнее линейная взаимосвязь. Знак показывает прямая или обратная взаимосвязь между случайными величинами.

Значение r	Интерпретация линейной зависимости

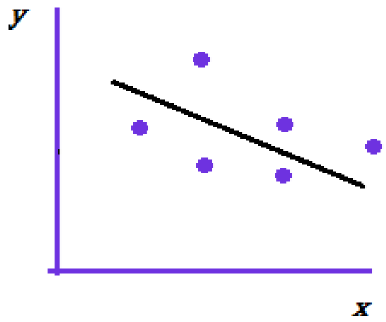
0 - 0.1	нет линейной зависимости
0.1 - 0.3	очень слабая
0.3 - 0.5	слабая
0.5 - 0.7	средняя (заметная)
0.7 - 0.9	сильная
0.9 – 1	очень сильная

Т.е. коэффициент корреляции -1 и 1 показывают одинаково сильную линейную зависимость. Только одна из них будет обратная (-1), а другая прямая зависимость (1).

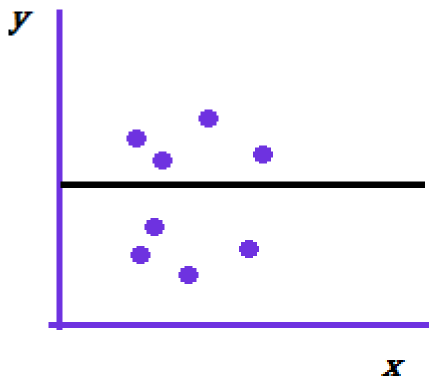
Прямая зависимость означает, что рост одной случайной величины сопровождается ростом другой случайной величины. Например, с увеличением расстояния возрастает стоимость билета.



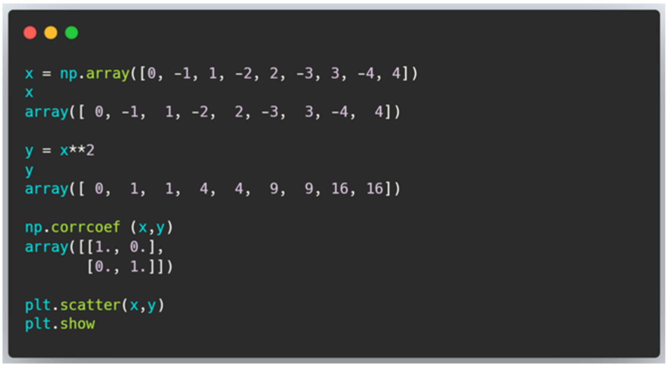
Обратная зависимость означает, что с увеличением одной случайной величины уменьшается другая случайная величина. Например, выше температура, меньше времени занимает растопить лед.



И если коэффициент корреляции равен или близок к нулю, то это говорит лишь о том, что между СВ нет линейной зависимости, но возможна какая - то другая зависимость, поэтому в таком случае рекомендуется построить график.



Одним из самых распространенных примеров из книг по статистики, который иллюстрирует справедливость вышеупомянутого факта, является квадратичная зависимость $y = x^2$. На графике четко видна параболла, но коэффициент корреляции показывает ноль.

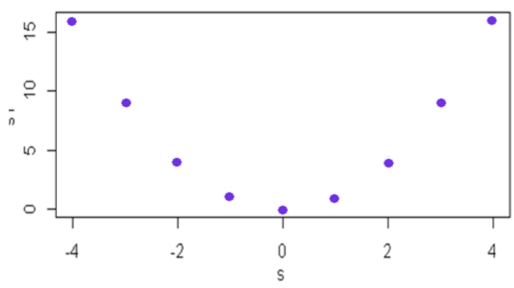


```
x = np.array([0, -1, 1, -2, 2, -3, 3, -4, 4])
x
array([ 0, -1,  1, -2,  2, -3,  3, -4,  4])

y = x**2
y
array([ 0,  1,  1,  4,  4,  9,  9, 16, 16])

np.corrcoef (x,y)
array([[1., 0.],
       [0., 1.]])

plt.scatter(x,y)
plt.show
```



Слабые стороны корреляционного анализа

1. То, что коэффициент корреляции показывает ноль при наличии, например, квадратичной зависимости можно уже отнести к недостаткам корреляционного анализа.
2. Еще одним недостатком может служить то, что случайные величины могут коррелировать по случайности. Проиллюстрируем это.

Мы возьмем одинаковой длины массив a и массив b и будем случайным образом изменять только массив b. Посмотрим, что показывает коэффициент корреляции.

В первом случае коэффициент -0.416 – слабая обратная зависимость, во втором случае – 0.68 – заметная обратная зависимость. А последний вариант массива b был набран неслучайным образом. Случайная величина a росла и случайную величину b я набрала таким образом, что почти все значения тоже растут. И получили коэффициент корреляции 0.9 – сильная прямая линейная взаимосвязь. Но тем не менее и в предыдущих СВ линейная зависимость прослеживалась, хотя это были абсолютно случайные СВ.

```
a = np.array([1, 2, 3, 4, 5])
a
array([1, 2, 3, 4, 5])

b = np.array ([7, 4, 6, 9, 0])
b
array([7, 4, 6, 9, 0])

np.corrcoef( a,b)
array([[ 1.        , -0.41602515],
       [-0.41602515,  1.        ]])


b = np.array ([11, 12, 0.8, 9, 0.4])
b
array([11. , 12. ,  0.8,  9. ,  0.4])

np.corrcoef( a,b)
array([[ 1.        , -0.68080746],
       [-0.68080746,  1.        ]])


b = np.array ([0.5, 0.7, 0.9, 0.8, 1])
b
array([0.5, 0.7, 0.9, 0.8, 1. ])

np.corrcoef( a,b)
array([[1.        , 0.90419443],
       [0.90419443, 1.        ]])
```

3. Высокая корреляции двух величин может свидетельствовать о том, что есть третья скрытая переменная. Например, с увеличением, числа кафе в городе, растет и число больниц. На самом же деле между СВ нет никакой зависимости, но есть третья скрытая переменная, плотность населения. Чем больше город, тем больше кафе и больниц.
4. И к последнему недостатку можно отнести то, что можно перепутать причинно-следственную связи, т.е. что является причиной, а что следствием. Т.к. мы не всегда работаем с такими очевидными переменными, как, к примеру, температура и скорость таяния льда, то подобный недостаток тоже нужно держать в памяти.

Ковариация

Ковариация – это величина, определяющая зависимость двух случайных величин.

Найти ее можно по формуле:

$$cov_{xy} = M(XY) - M\,(X) * M(Y)$$

где M - математическое ожидание

Масштаб ковариации зависит от дисперсии, поэтому по ковариации нельзя судить о силе взаимосвязи СВ, но ее можно нормировать, поместив значения в [-1; 1]. Таким образом, мы получим коэффициент корреляции Пирсона, который мы уже сегодня рассчитывали с помощью функции corrcoef().

$$r = \frac{cov_{xy}}{\sigma_x * \sigma_y}$$

Давайте рассчитаем ковариацию для цены и площади – случайных величин, с которыми мы сегодня уже работали.

```
p
array([1.2, 1.6, 1.8, 1.8, 2.5, 2.6, 3. , 3.3])
s
array([27, 37, 42, 48, 57, 56, 77, 88])

cov = np.mean(p*s) - np.mean(p) * np.mean (s)

cov
11.662500000000023

np.cov (p,s)
array([[  0.53928571,  13.32857143],
       [ 13.32857143, 344.        ]])
```

Ковариация, рассчитанная функцией отличается от значения ковариации, рассчитанной по формуле (11.66 и 13.28). Дело в том, что ковариация может быть как смещенная, так и несмещенная.

Давайте рассчитаем коэффициент корреляции Пирсона через смещенную и несмещенную ковариацию. Мы должны получить коэффициент корреляции Пирсона 0.978

```
np.corrcoef (p,s)
array([[1.        , 0.97857682],
       [0.97857682, 1.        ]])
```

Согласно формуле для расчета коэффициента корреляции мы должны ковариацию разделить на произведение стандартных отклонений. Т.е. если мы рассчитаем несмещенную ковариацию, то и делить мы должны на произведение несмещенных стандартных отклонений.

```
np.cov( p,s, ddof = 1)

array([[  0.53928571,  13.32857143],
       [ 13.32857143, 344.        ]])

np.std(p, ddof = 1)
0.7343607521414215

np.std(s, ddof = 1)
18.547236990991408

13.32857143/ (0.7343607521414215 * 18.547236990991408 )

0.9785768206878758
```

Если же мы используем смещенную ковариацию, то и делить ее будем на произведение смещенных стандартных отклонений.

```
np.cov (p,s, ddof = 0 )
array([[  0.471875,  11.6625  ],
       [ 11.6625  , 301.      ]])

np.std(p, ddof = 0)
0.6869315832017042

np.std(s, ddof = 0)
17.349351572897472

11.6625 / (0.6869315832017042 * 17.349351572897472)

0.9785768205829909
```

Мы видим, что значения коэффициента корреляции совпадают между собой и равны тому значению, которое мы получили через функцию corrcoef.

Коэффициент корреляции Спирмена

Коэффициент корреляции Спирмена называют ранговым коэффициентом корреляции. Он также показывает тесноту линейной связи, но в отличии от коэффициента корреляции Пирсона не требует нормальности распределений случайных величин и применяется для количественных и порядковых данных.

Рассчитаем коэффициент корреляции Спирмена в Python помощью функции spearmanr().

```
s
array([27, 37, 42, 48, 56, 57, 77, 80])

p
array([1.2, 1.6, 1.8, 1.8, 2.6, 2.5, 3. , 3.3])


stats.spearmanr(p, s)

SpearmanrResult(correlation=0.9700772721497398,
pvalue=6.548558831120599e-05)
```

Коэффициент корреляции Спирмена 0.97 Сильная корреляция.

Как рассчитывается коэффициент корреляции Спирмена?

Возьмем уже знакомые нам СВ площадь и цену , а затем присвоим им ранги в порядке возрастания. Т.е. самая маленькая площадь 27 – ранг 1, а самой большой площади 80 – ранг 8.

Как присваивать ранги, если значения повторяются, как, например, в массиве p, где два раза встречается цена 1.8? Расположив цены по возрастанию, величины 1.8 стоят на 3 и 4 местах. Тогда присваиваем им среднее арифметическое номеров элементов. Т.е.(3+4)/2 =3.5 Для каждого значения 1.8 будет ранг 3.5. И назовем эти СВ (сами значения рангов) s_2 и p_2. И уже к ним применим коэффициент корреляции Пирсона.

Условия применимости коэффициентов корреляции

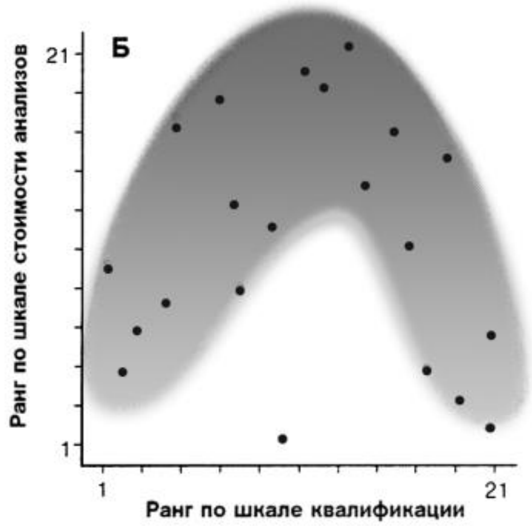
Коэффициент корреляции Пирсона	Коэффициент корреляции Спирмена
параметрический метод	непараметрический метод
нормальность	распределение может быть отличным от нормального
количественные данные	количественные и порядковые признаки
сделать проверку на U- образную кривую	сделать проверку на U- образную кривую

Рассмотрим интересный пример из книги Стентона Гланца «Медико- биологическая статистика».

В качестве примера автор книги берет реальное исследование*, в котором проводят корреляционный анализ между квалификацией врача и затратами на анализы, которые врач прописал при госпитализации пациента.

Врачи прошли аттестационную комиссию и получили оценки от 1 до 21 (ранги), где 21 – худшая квалификация. При анализе получился коэффициент корреляции Спирмена $r_s = -0.13$, что показывает очень слабую зависимости.

Но если мы посмотрим на график из этой книги, то увидим квадратичную зависимость. По графику видно, что меньше всего затрат на анализы у пациентов врачей с лучшей и худшей категорией, соответственно и количество назначаемых исследований этими врачами наименьшее.



Мы можем сделать выводы по графику, мы видим зависимость, но коэффициент корреляции Спирмена нам ничего не показал. Это говорит о том, что подобную U-образную зависимость никакой коэффициент корреляции не уловит.

---

* S. A. Schroeder, A. Schliftman, T. E. Piemine. Variation among physicians in use of laboratory tests: relation to quality of care. Med. Care, 12: 709–713, 1974

Есть еще один недостаток, который мы не обсудили. Посмотрите на схематичные рисунки.

На обоих графиках коэффициент корреляции равен 1, но на левом графике зависимая переменная y растет быстрее, чем на правом. Т.е. коэффициент корреляции не показывает, как быстро изменяется зависимая переменная y при изменении независимой переменной x. Ответить на этот вопрос сможет нам регрессионный анализ, которым мы займемся на следующем уроке.