

Лекция 11

Корреляционный анализ, часть I:
коэффициенты корреляции Пирсона и Спирмена.

О чём речь

Выборка, в которой каждое наблюдение характеризуется двумя признаками:

$$(X_1, Y_1), \dots, (X_n, Y_n)$$

Вообще-то, не обязательно двумя...

Задачи корреляционного анализа:

- определить, есть ли связь между признаками
(проверка гипотезы о независимости);
- измерить тесноту связи.

Для начала предположим, что (X, Y) - количественные признаки.

График рассеяния

(scatter plot)

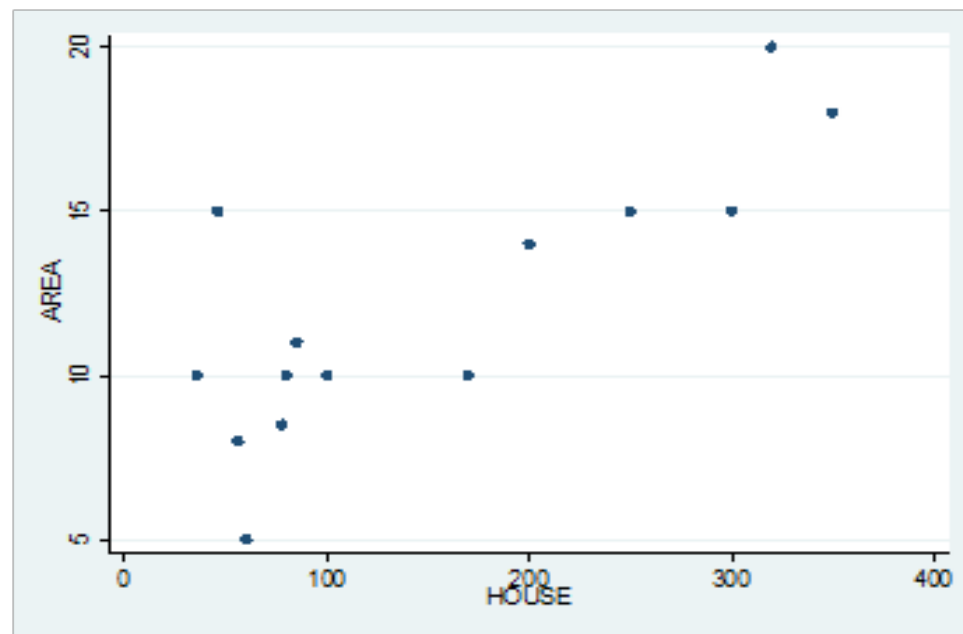
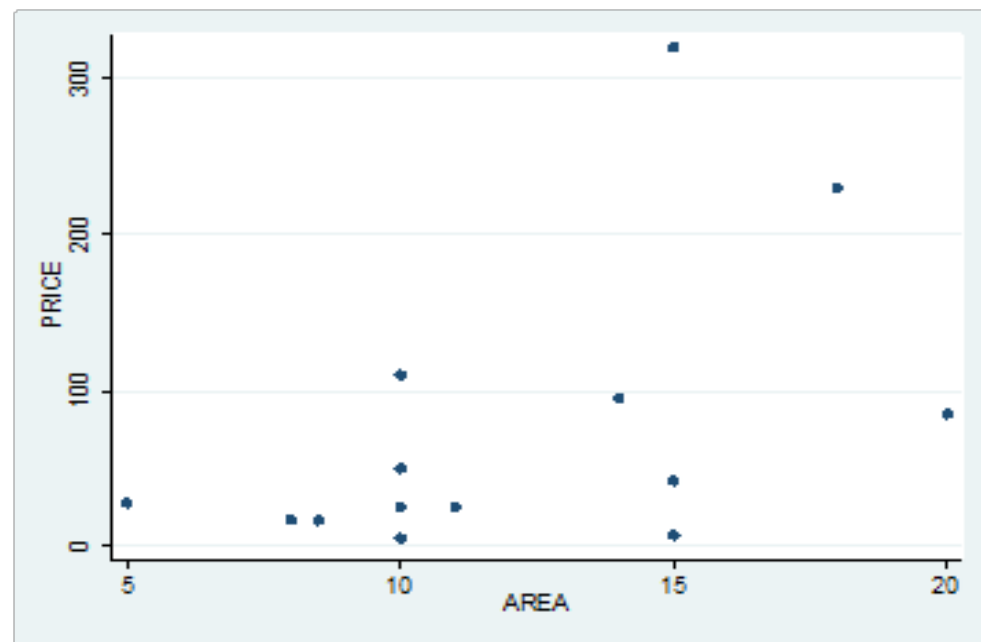
Наблюдение i — точка с координатами (X_i, Y_i) .

Пример: данные о характеристиках 14 коттеджных участков.

price — цена участка, тыс. долл.

area — площадь участка, сотки.

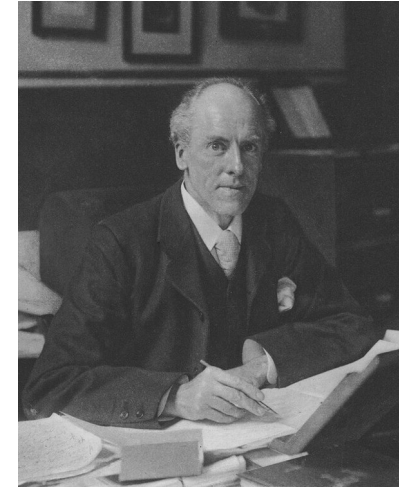
house — площадь дома, м².



о чём эти графики говорят?

как количественно описать связь?

Теоретический коэффициент корреляции Пирсона



Карл Пирсон

Ковариация:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))].$$

что это значит и зачем это?

Корреляция:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)D(Y)}} = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{E(X - E(X))^2 E(Y - E(Y))^2}}.$$

а это зачем?

Свойства:

1°. $-1 \leq \text{Corr}(X, Y) \leq +1;$

2°. $\text{Corr}(X, Y) = \pm 1 \Leftrightarrow Y = aX + b, \quad a \neq 0;$

3°. $\text{Corr}(aX + b, cY + d) = \text{sgn}(a) \text{sgn}(c) \text{Corr}(X, Y), \quad a \neq 0, \quad c \neq 0;$

4°. X, Y независимы $\Rightarrow \text{Corr}(X, Y) = 0.$

Выборочный коэффициент корреляции Пирсона

$$r_{X,Y} = \hat{\text{Corr}}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \times \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

$r_{X,Y}$ — состоятельная оценка для $\text{Corr}(X, Y)$.

Другие свойства?

$$-1 \leq r_{X,Y} \leq 1$$

строго линейная
обратная связь

$$Y_i = aX_i + b, \quad a < 0$$

строго линейная
прямая связь

$$Y_i = aX_i + b, \quad a > 0$$

Коэффициент корреляции Пирсона отражает тесноту *линейной* связи между признаками и её направление.

а что если $r_{X,Y} = 0$?

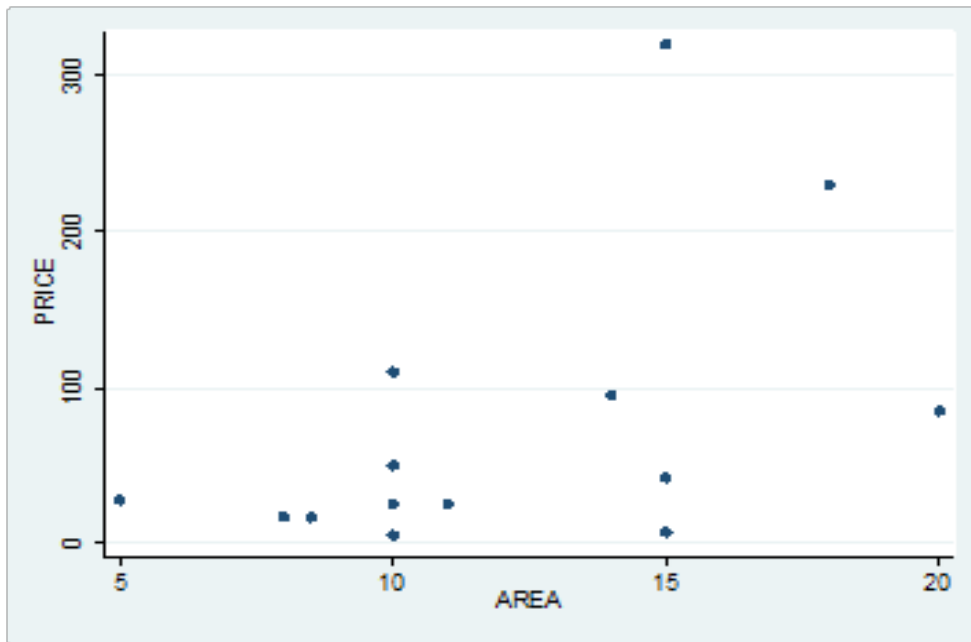
Коттеджи и корреляции

Данные о характеристиках 14 коттеджных участков.

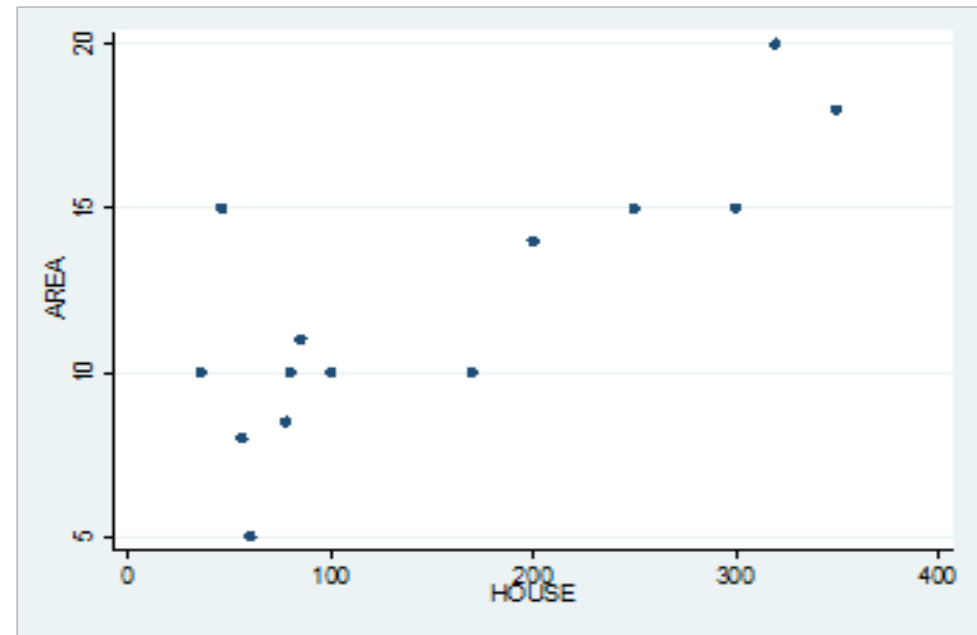
price — цена участка, тыс. долл.

area — площадь участка, сотки.

house — площадь дома, м².



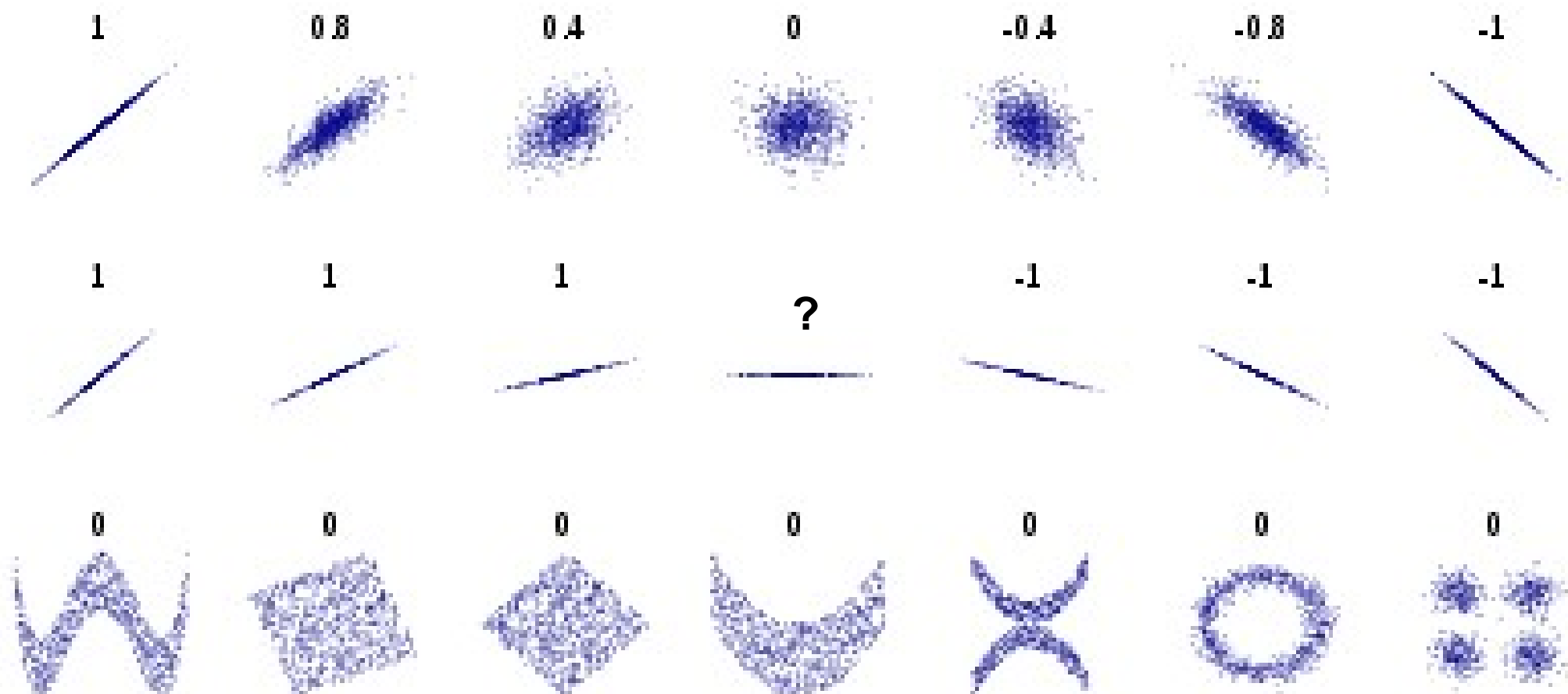
$$r_{price, area} = 0.504$$



$$r_{area, house} = 0.803$$

Тут можно поболтать о статистической и причинно-следственной связи.

Корреляции и картинка из Википедии



Что **не** измеряет коэффициент корреляции Пирсона?

Проверка гипотезы о независимости

с помощью коэффициента корреляции Пирсона

Выборка $(X_1, Y_1), \dots, (X_n, Y_n)$ из двумерного нормального распределения:

$$(X_i, Y_i) \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho \sigma_X \sigma_Y \\ \rho \sigma_X \sigma_Y & \sigma_Y^2 \end{pmatrix}\right);$$

$(X_1, Y_1), \dots, (X_n, Y_n)$ независимы.

$$H_0: \rho = 0 \quad (X, Y \text{ независимы})$$

$$H_A: \rho \neq 0 \quad (X, Y \text{ зависимы})$$

Помните: в общем случае независимость \neq некоррелированность.

Проверка гипотезы о независимости

с помощью коэффициента корреляции Пирсона

Выборка $(X_1, Y_1), \dots, (X_n, Y_n)$ из двумерного нормального распределения:

$$(X_i, Y_i) \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho \sigma_X \sigma_Y \\ \rho \sigma_X \sigma_Y & \sigma_Y^2 \end{pmatrix}\right);$$

$(X_1, Y_1), \dots, (X_n, Y_n)$ независимы.

$$H_0: \rho = 0 \quad (X, Y \text{ независимы})$$

$$H_A: \rho \neq 0 \quad (X, Y \text{ зависимы})$$

Помните: в общем случае независимость \neq некоррелированность.

Статистика:

$$t = \frac{r_{X,Y} \sqrt{n-2}}{\sqrt{1-r_{X,Y}^2}} \stackrel{H_0}{\sim} t_{n-2}$$

Сравните:

$$F = \frac{R^2 / (k-1)}{(1-R^2) / (n-k)} \stackrel{H_0}{\sim} F_{k-1, n-k}$$

Проверка гипотезы о независимости

с помощью коэффициента корреляции Пирсона

Выборка $(X_1, Y_1), \dots, (X_n, Y_n)$ из двумерного нормального распределения:

$$(X_i, Y_i) \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho \sigma_X \sigma_Y \\ \rho \sigma_X \sigma_Y & \sigma_Y^2 \end{pmatrix}\right);$$

$(X_1, Y_1), \dots, (X_n, Y_n)$ независимы.

$$H_0: \rho = 0 \quad (X, Y \text{ независимы})$$

$$H_A: \rho \neq 0 \quad (X, Y \text{ зависимы})$$

Помните: в общем случае независимость \neq некоррелированность.

Статистика:

$$t = \frac{r_{X,Y} \sqrt{n-2}}{\sqrt{1-r_{X,Y}^2}} \stackrel{H_0}{\sim} t_{n-2}$$

Решающее правило:

$$|t| > t_{n-2, \frac{\alpha}{2}} \Rightarrow H_0 \text{ отвергается, выявлена связь между } X \text{ и } Y.$$

$$|t| < t_{n-2, \frac{\alpha}{2}} \Rightarrow \text{нет оснований отвергнуть } H_0, \text{ связь не выявлена.}$$

Речь

«проверка гипотезы о равенстве коэффициента корреляции нулю» =

= «проверка значимости коэффициента корреляции»

«коэффициент корреляции значимо отличается от нуля» =

= «коэффициент корреляции значим» =

= «гипотеза о равенстве коэффициента корреляции нулю отвергается»

«коэффициент корреляции незначим» =

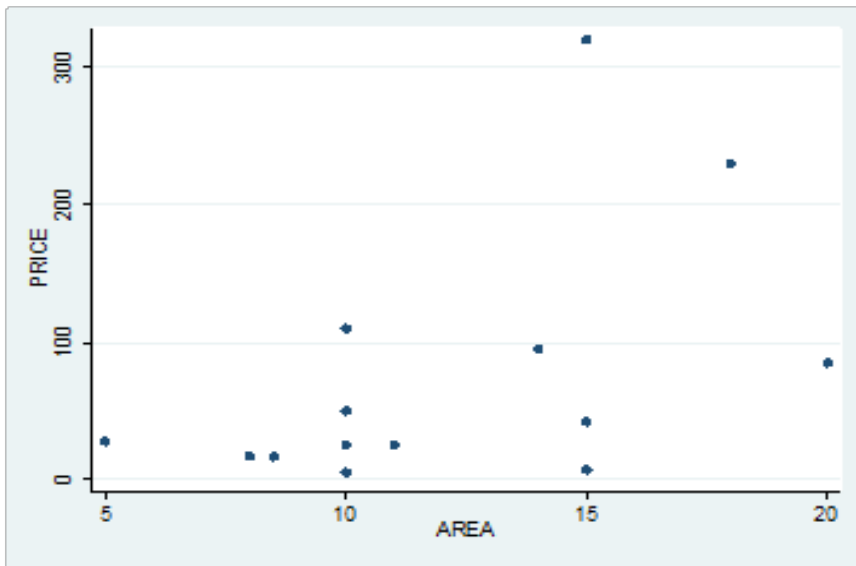
= «нет оснований отвергнуть гипотезу о равенстве коэффициента
корреляции нулю»

Проверка значимости связи между ценой и площадью участка

$$H_0: \rho_{price, area} = 0$$

$$H_A: \rho_{price, area} \neq 0$$

Уровень значимости $\alpha = 0.05$



Выборочная корреляция: $r_{price, area} = 0.504$

$$\begin{aligned} \text{Статистика: } t &= \frac{r_{price, area} \sqrt{n-2}}{\sqrt{1-r_{price, area}^2}} = \frac{0.504 \sqrt{14-2}}{\sqrt{1-0.504^2}} = \\ &= 2.021. \end{aligned}$$

Критическое значение: $t_{n-2, \frac{\alpha}{2}} = t_{12, \frac{0.05}{2}} = 2.179.$

Вывод: $|t| = 2.021 < 2.179 \Rightarrow$ нет оснований отвергнуть основную гипотезу.

Связь между ценой участка и его площадью не выявлена.

Матрица выборочных корреляций

для характеристик коттеджных участков

dist — расстояние до МКАД, км;

house — площадь дома, м²;

price — цена, тыс. долл.;

area — площадь участка, сотки.

	<u>dist</u>	house	price	area
<u>dist</u>	1	−0.578**	−0.675***	−0.225
house	−0.578**	1	0.782***	0.803***
price	−0.675***	0.782***	1	0.504*
area	−0.225	0.803***	0.504*	1

* — связь значима на уровне 10%,

** — на уровне 5%,

*** — на уровне 1%.

Коэффициент ранговой корреляции Спирмена

(Spearman rank correlation coefficient)

Это тот же коэффициент Пирсона, но применённый к рангам:

$$r_{X,Y}^S = r_{\text{rank}(X), \text{rank}(Y)}$$



Чарльз Эдвард
Спирмен

№	<u>price</u>	<u>price,</u> ранг	<u>area</u>	<u>area,</u> ранг
1	25	5.5	11	8
2	28	7	5	1
3	17	4	8	2
4	50	9	10	5.5
5	110	12	10	5.5
6	7	2	15	11
7	42	8	15	11
8	85	10	20	14
9	5	1	10	5.5
10	230	13	18	13
11	95	11	14	9
12	320	14	15	11
13	25	5.5	10	5.5
14	16.5	3	8.5	3

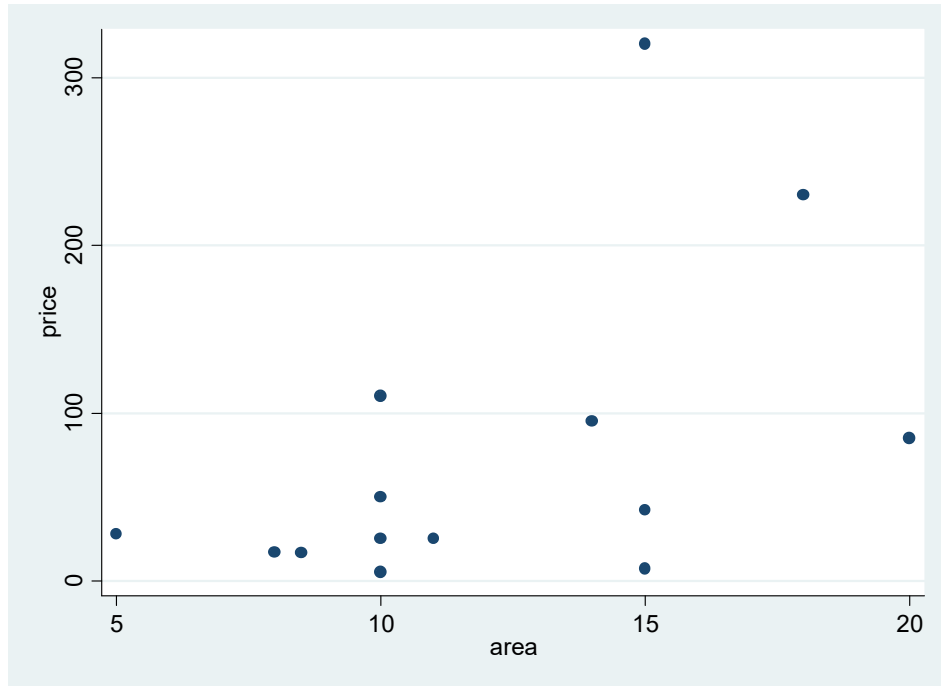
$$r_{\text{price}, \text{area}} = 0.504;$$

$$r_{\text{price}, \text{area}}^S = r_{\text{rank}(\text{price}), \text{rank}(\text{area})} = 0.464.$$

В чём выгода ранжировать?

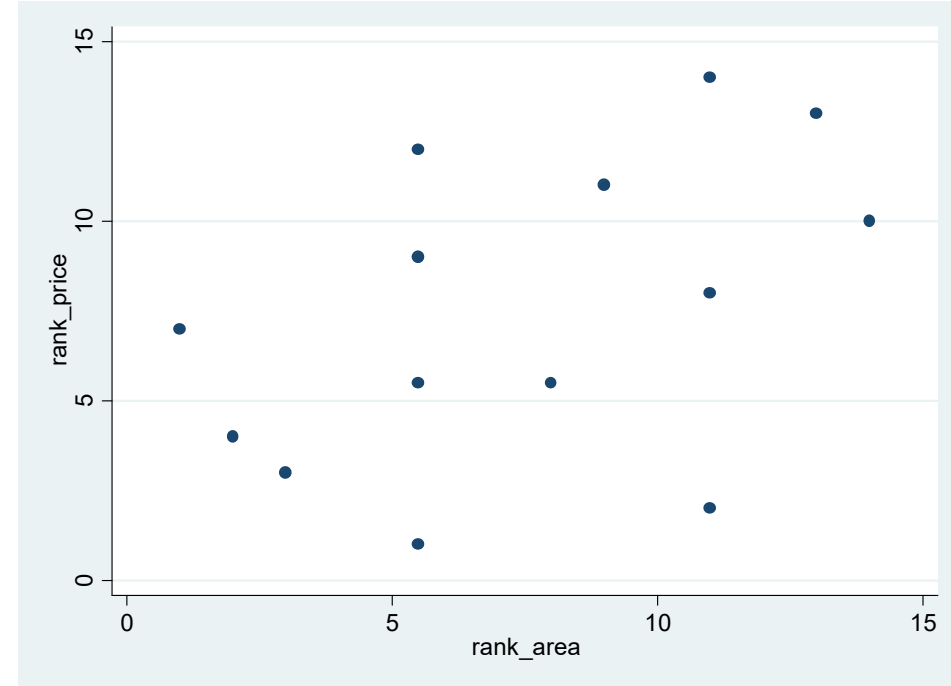
Тысячи долларов, сотки и ранги

исходные данные



$$r_{price, area} = 0.504$$

ранги



$$r_{price, area}^S = 0.464$$

ну и что?

Что означает коэффициент Спирмена?

$$-1 \leq r_{X,Y}^S \leq +1$$

Когда достигаются крайние значения?

Что означает коэффициент Спирмена?

$$-1 \leq r_{X,Y}^S \leq +1$$

Когда достигаются крайние значения?

rank(X)	rank(Y)
1	1
2	2
3	3
4	4
5	5

$$r_{X,Y}^S = 1$$

rank(X)	rank(Y)
1	5
2	4
3	3
4	2
5	1

$$r_{X,Y}^S = -1$$


Что означает полная согласованность рангов? А полное расхождение?

Что означает коэффициент Спирмена?

Коэффициент Спирмена измеряет тесноту *монотонной* связи между признаками и её направление.

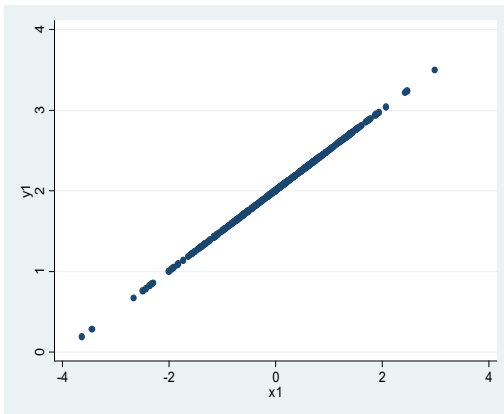
$$-1 \leq r_{X,Y}^S \leq +1$$

строго монотонная
обратная связь

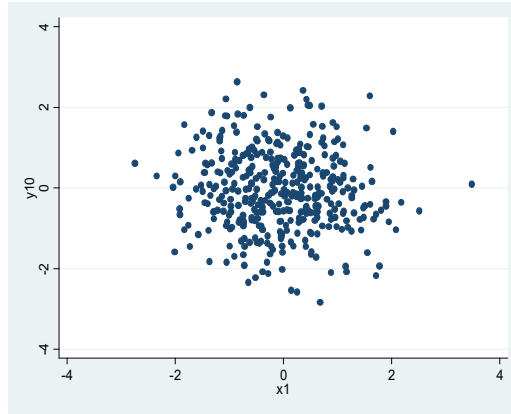


строго монотонная
прямая связь

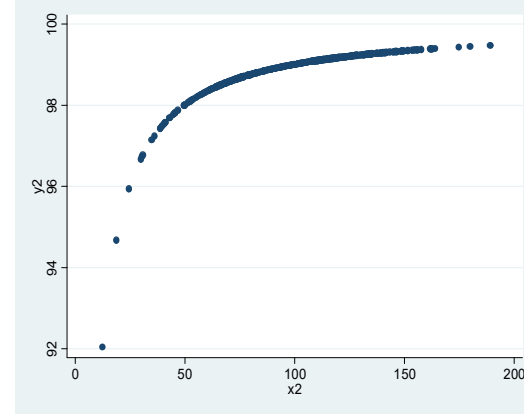
Картинки и корреляции



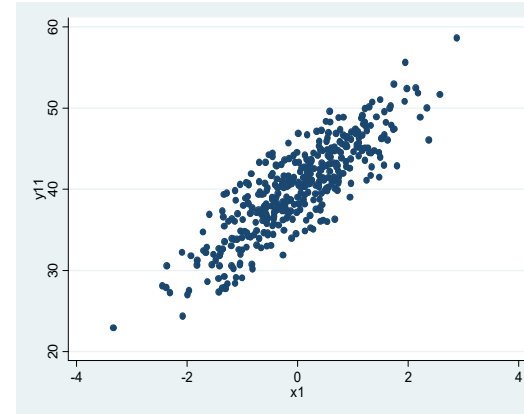
$$r = 1$$
$$r^S = 1$$



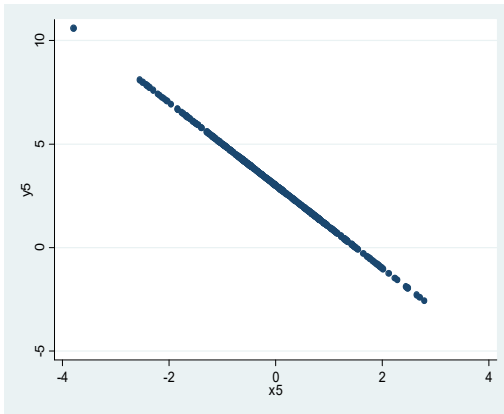
$$r = -0.08$$
$$r^S = -0.07$$



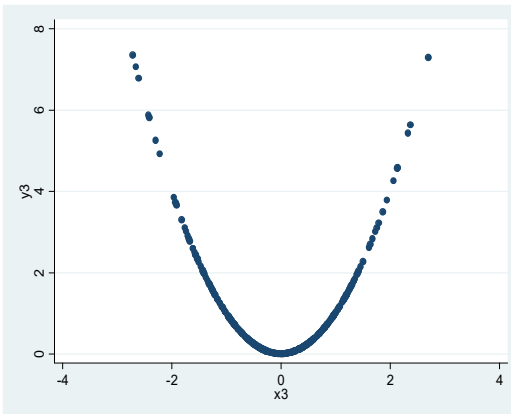
$$r = 0.78$$
$$r^S = 1$$



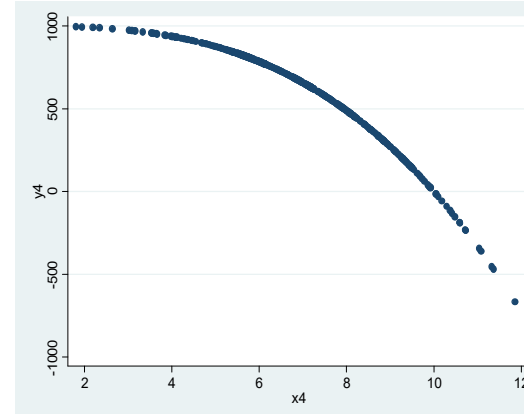
$$r = 0.85$$
$$r^S = 0.83$$



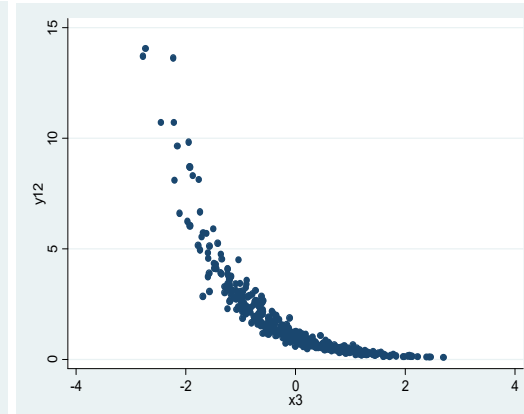
$$r = -1$$
$$r^S = -1$$



$$r = -0.14$$
$$r^S = -0.06$$



$$r = -0.95$$
$$r^S = -1$$



$$r = -0.80$$
$$r^S = -0.98$$

Проверка гипотезы о независимости

с помощью коэффициента ранговой корреляции Спирмена

Выборка из независимых и одинаково распределённых пар $(X_1, Y_1), \dots, (X_n, Y_n)$

H_0 : X_i, Y_i независимы

H_A : X_i, Y_i зависимы

Точный критерий

Приближение
(для $n > 12$?)

Статистика:

$$r_{X,Y}^S$$

$$t = \frac{r_{X,Y}^S \sqrt{n-2}}{\sqrt{1-(r_{X,Y}^S)^2}} \stackrel{H_0}{\sim} t_{n-2}$$

Правило: отвергнуть H_0 , если

$$|r_{X,Y}^S| \geq r_{crit}^S\left(n, \frac{\alpha}{2}\right)$$

$$|t| > t_{n-2, \frac{\alpha}{2}}$$

односторонние альтернативы?

«проверка гипотезы о независимости с помощью коэффициента Спирмена» =
= «проверка значимости коэффициента Спирмена»

Upper Critical Values of Spearman's Rank Correlation Coefficient R_s

Note: In the table below, the critical values give significance levels as close as possible to but not exceeding the nominal α .

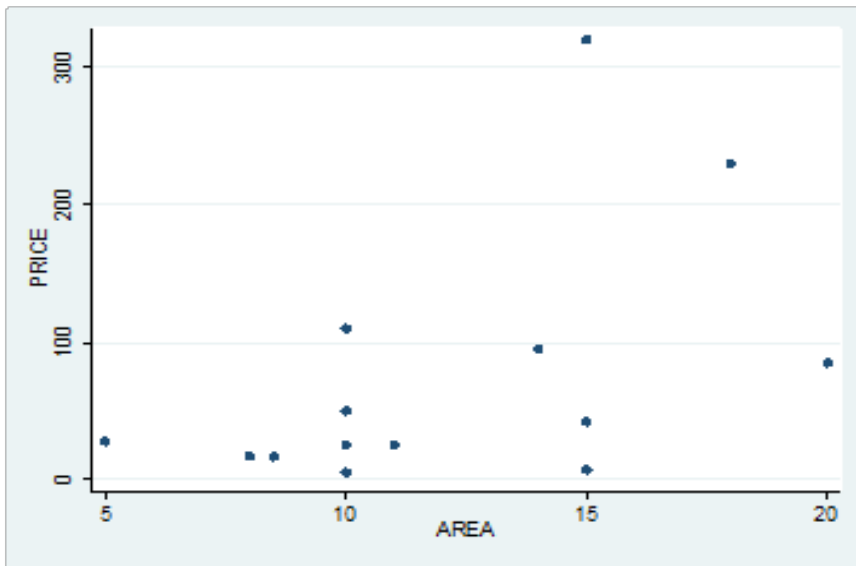
	Nominal α					
n	0.10	0.05	0.025	0.01	0.005	0.001
4	1.000	1.000	-	-	-	-
5	0.800	0.900	1.000	1.000	-	-
6	0.657	0.829	0.886	0.943	1.000	-
7	0.571	0.714	0.786	0.893	0.929	1.000
8	0.524	0.643	0.738	0.833	0.881	0.952
9	0.483	0.600	0.700	0.783	0.833	0.917
10	0.455	0.564	0.648	0.745	0.794	0.879
11	0.427	0.536	0.618	0.709	0.755	0.845
12	0.406	0.503	0.587	0.678	0.727	0.818
13	0.385	0.484	0.560	0.648	0.703	0.791
14	0.367	0.464	0.538	0.626	0.679	0.771
15	0.354	0.446	0.521	0.604	0.654	0.750
16	0.341	0.429	0.503	0.582	0.635	0.729
17	0.328	0.414	0.488	0.566	0.618	0.711
18	0.317	0.401	0.472	0.550	0.600	0.692
19	0.309	0.391	0.460	0.535	0.584	0.675
20	0.299	0.380	0.447	0.522	0.570	0.662
21	0.292	0.370	0.436	0.509	0.556	0.647
22	0.284	0.361	0.425	0.497	0.544	0.633
23	0.278	0.353	0.416	0.486	0.532	0.621
24	0.271	0.344	0.407	0.476	0.521	0.609
25	0.265	0.337	0.398	0.466	0.511	0.597
26	0.259	0.331	0.390	0.457	0.501	0.586
27	0.255	0.324	0.383	0.449	0.492	0.576
28	0.250	0.318	0.375	0.441	0.483	0.567
29	0.245	0.312	0.368	0.433	0.475	0.558

Проверка значимости связи между ценой и площадью участка

H_0 : *price* и *area* независимы

H_A : *price* и *area* зависимы

Уровень значимости $\alpha = 0.05$



Коэфф. Спирмена: $r_{price, area}^S = 0.464$.

Критическое значение: $r_{crit}^S\left(14, \frac{0.05}{2}\right) = 0.538$.

Вывод:

$|r_{price, area}^S| < 0.538 \Rightarrow$ нет оснований считать, что связь есть.

Или с помощью t -статистики: $t = \frac{r_{price, area}^S \sqrt{n-2}}{\sqrt{1 - (r_{price, area}^S)^2}} = \frac{0.464 \sqrt{14-2}}{\sqrt{1-0.464^2}} = 1.814$.

Критическое значение: $t_{n-2, \frac{\alpha}{2}} = t_{12, \frac{0.05}{2}} = 2.179$.

Вывод: $|t| = 1.814 < 2.179 \Rightarrow$ нет оснований отвергнуть основную гипотезу.

Связь между ценой участка и его площадью не выявлена.

Матрицы выборочных корреляций

для характеристик коттеджных участков

dist — расстояние до МКАД, км;

house — площадь дома, м²;

price — цена, тыс. долл.;

area — площадь участка, сотки.

Спирмен:

	<u>dist</u>	house	price	area
<u>dist</u>	1	−0.690***	−0.813***	−0.262
house	−0.690***	1	0.889***	0.690***
price	−0.813***	0.889***	1	0.464*
area	−0.262	0.690***	0.464*	1

Пирсон:

	<u>dist</u>	house	price	area
<u>dist</u>	1	−0.578**	−0.675***	−0.225
house	−0.578**	1	0.782***	0.803***
price	−0.675***	0.782***	1	0.504*
area	−0.225	0.803***	0.504*	1

* — связь значима на уровне 10%,

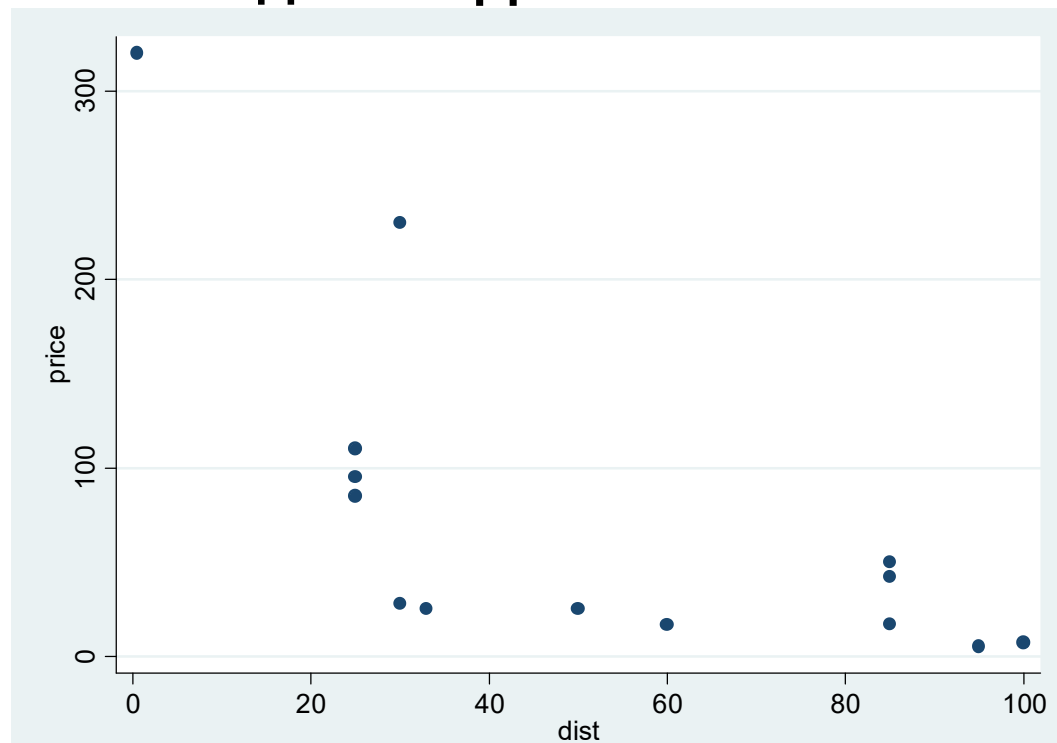
** — на уровне 5%,

*** — на уровне 1%.

Цена и расстояние до МКАД

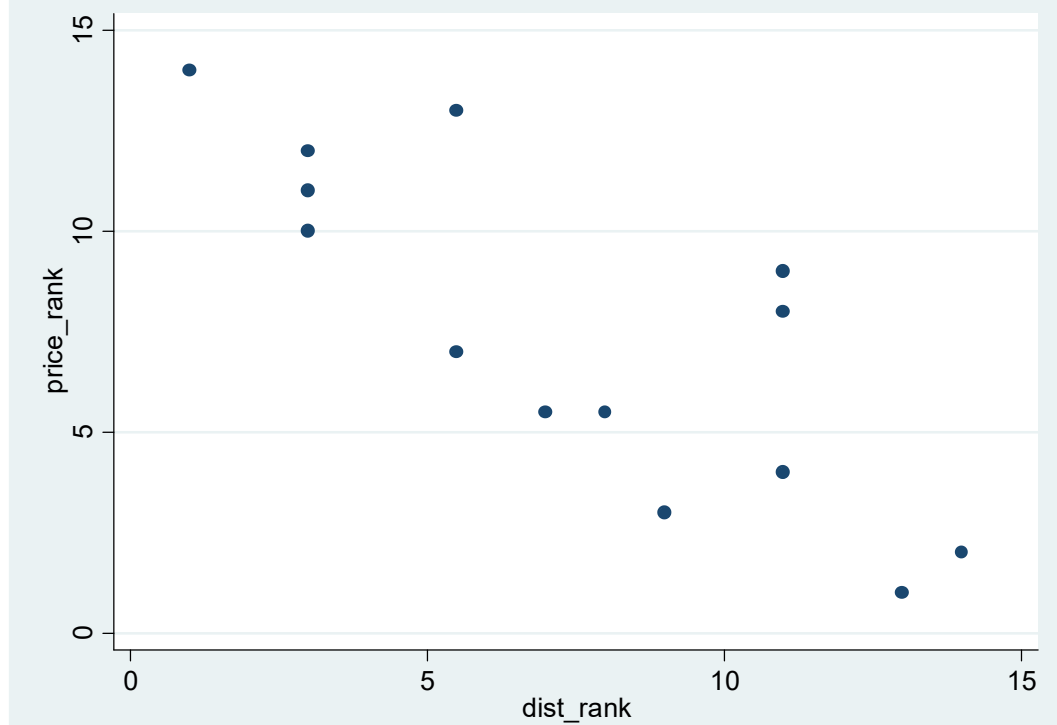
Исходные величины:

$$r_{price, dist} = -0.675$$

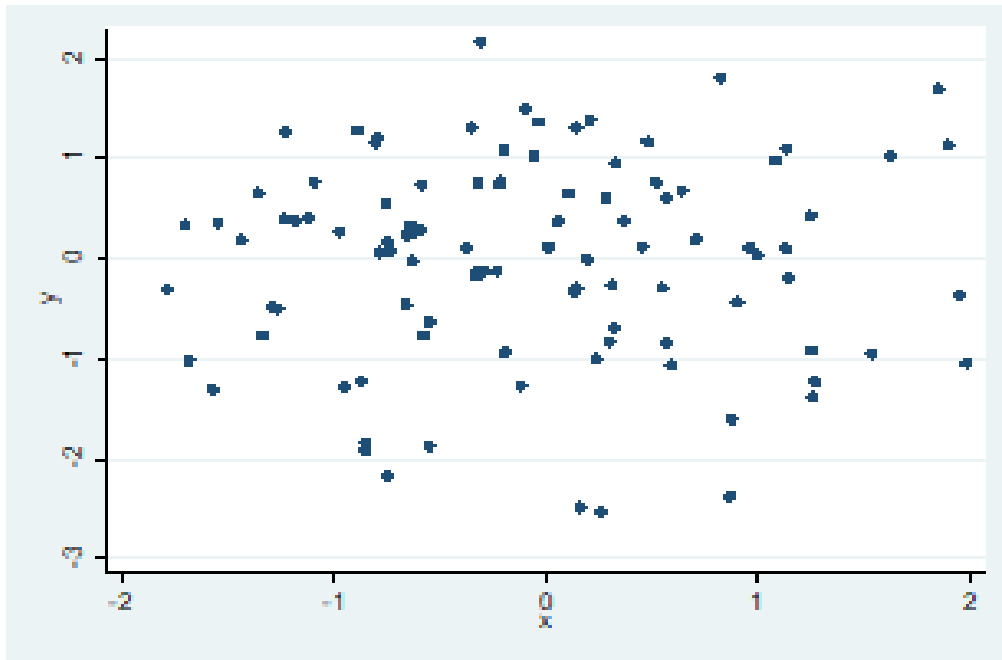


Ранги:

$$r_{price, dist}^S = -0.813$$

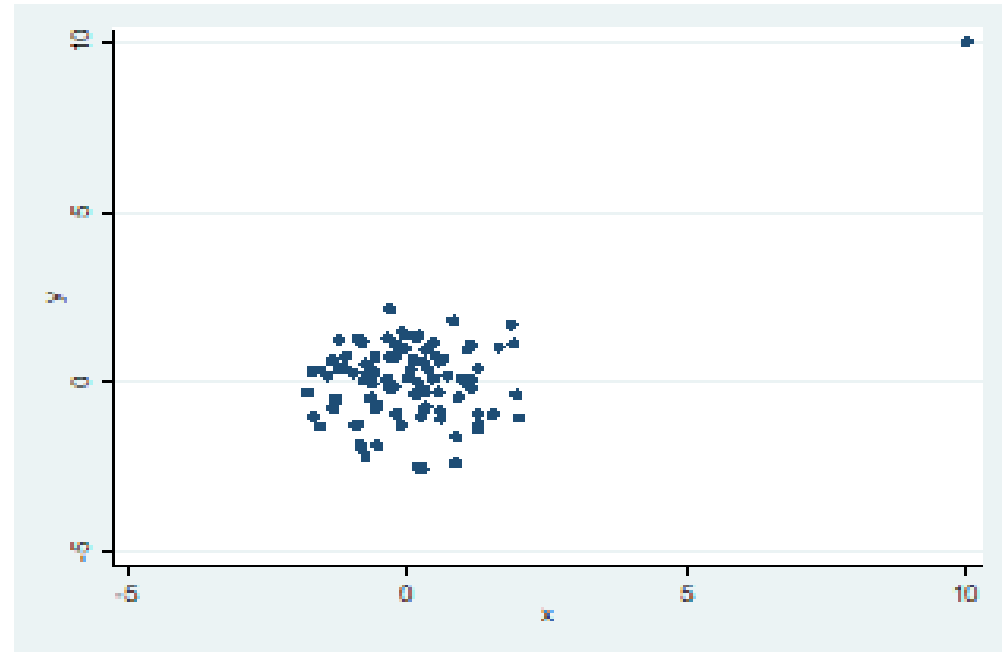


Корреляции и выбросы



$$r = 0.018, \quad p\text{-value} = 0.854$$

$$r^S = 0.007, \quad p\text{-value} = 0.949$$



$$r = 0.531, \quad p\text{-value} = 0.000$$

$$r^S = 0.036, \quad p\text{-value} = 0.723$$

Коэффициент Спирмена не чувствителен к выбросам

Подытожим

Коэффициент корреляции Пирсона r

- ▶ характеризует тесноту и направление линейной связи между признаками;
- ▶ $r = \pm 1$ при строгой линейной зависимости;
- ▶ не меняется при линейных преобразованиях (может менять знак);
- ▶ применим к количественным признакам.

Коэффициент ранговой корреляции Спирмена r^S

- ▶ характеризует тесноту и направление монотонной связи между признаками;
- ▶ $r^S = \pm 1$ при строгой монотонной зависимости;
- ▶ не меняется при монотонных преобразованиях (может менять знак);
- ▶ нечувствителен к выбросам;
- ▶ применим к количественным и порядковым признакам.

Ни один из этих коэффициентов не позволяет измерить немонотонную связь.

Темы поболтать:

- статистическая и причинно-следственная связь;
- коэффициент Спирмена и дисперсионный анализ.

Следующая лекция

Корреляционный анализ, часть II:

таблицы сопряжённости, критерий независимости хи-квадрат и коэффициент Крамера.