Корреляционный анализ

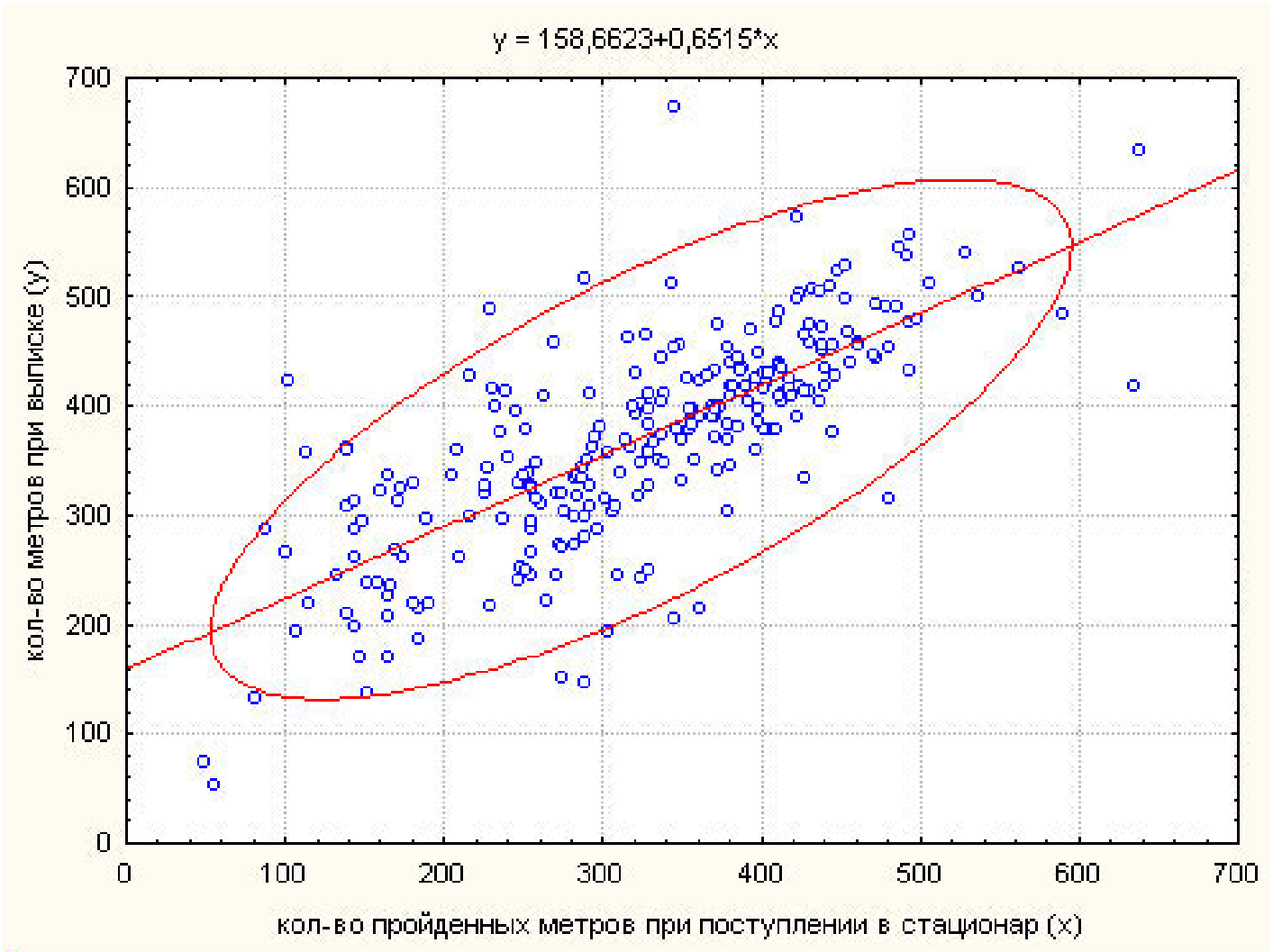
Н.П. Алексеева,

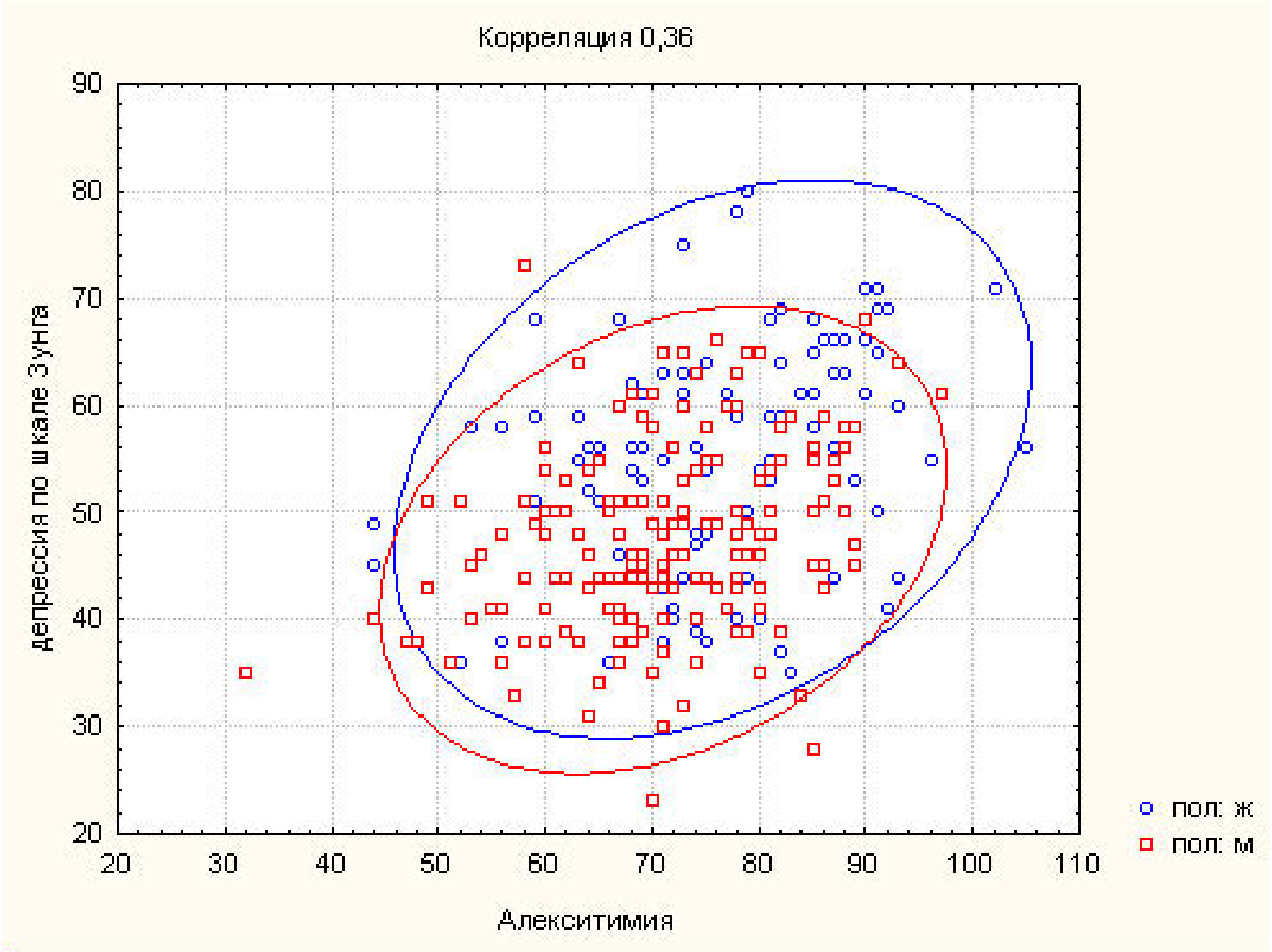
СПбГУ, математико-механический факультет

2020 г.

Корреляционный анализ

Пусть имеется n наблюдений парной выборки $(x_1, y_1), \ldots, (x_n, y_n)$ двух метрических признаков X и Y.




Корреляция 0,36
Депрессия по шкале Зунга
Алекситимия
90
80
70
60
50
40
30
20
20
30
40
50
60
70
80
90
100
110
пол: ж
пол: м

Пример отрицательной зависимости

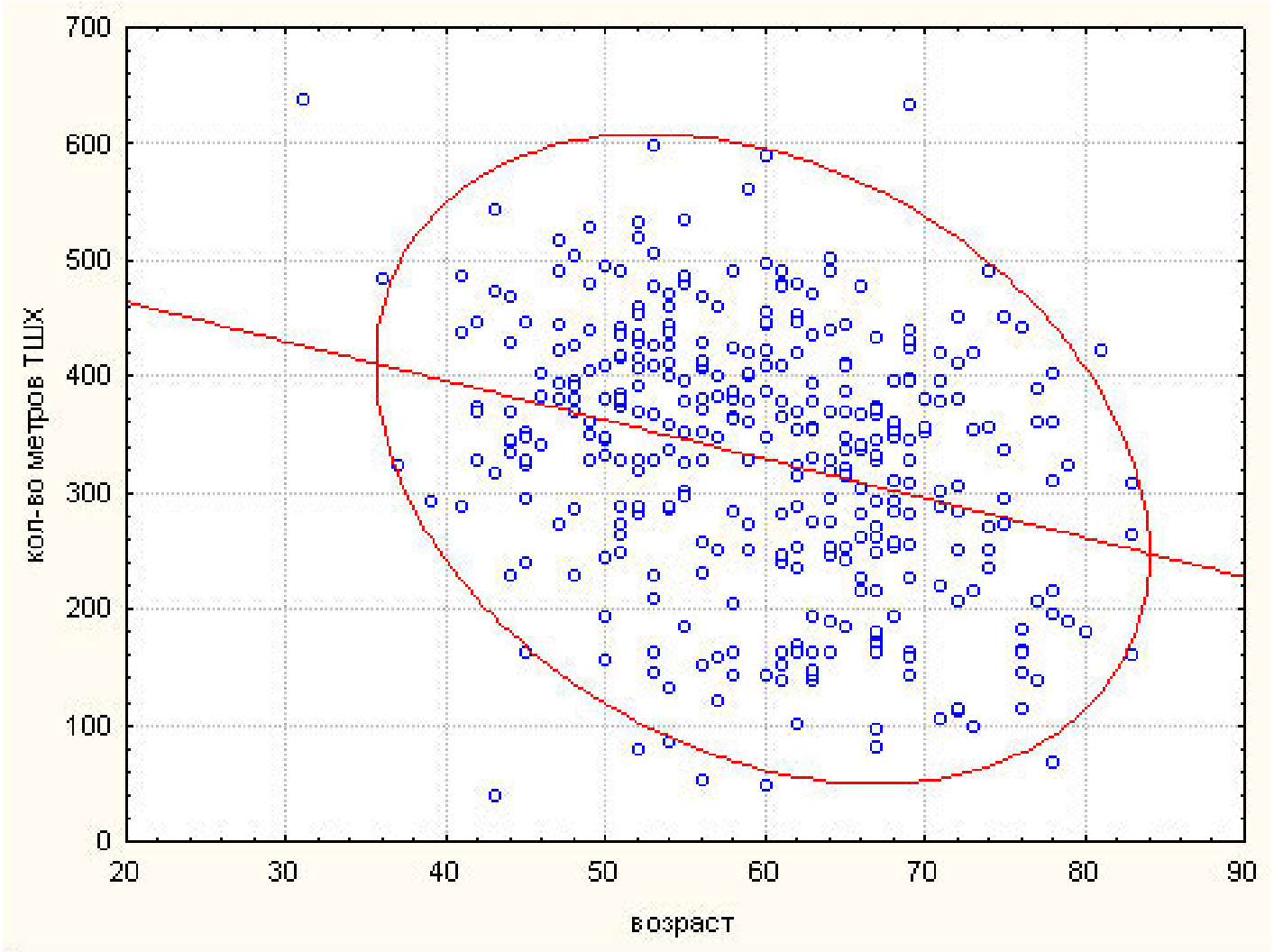


Рис.: Отрицательная зависимость между возрастом и результатом теста шестиминутной ходьбы $R = -0.29,\ P < 0.01$.

Коэффициент корреляции Пирсона

Численно уровень линейной зависимости между признаками – генеральный коэффициент корреляции

$$\varrho = \frac{\mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta)}{\sqrt{\mathbb{E}(\xi - \mathbb{E}\xi)^2}\sqrt{\mathbb{E}(\eta - \mathbb{E}\eta)^2}}$$

Частные случаи:

- Если ξ и η независимы, то $\varrho = 0$
- Если существует линейная зависимость $\eta = \mathrm{a}\xi + \mathrm{b}$, то $\varrho = 1$ при $\mathrm{a} > 0$ и $\varrho = -1$ при $\mathrm{a} < 0$.

Для независимых величин

$$p_{ij} = P\{\xi = x_i, \eta = y_j\} = P\{\xi = x_i\}P\{\eta = y_j\} = p_{i*}p_{*j},$$
$$i = 1, 2, \ldots, r,\ \ j = 1, 2, \ldots, s$$

$$\mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta) = \sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij}(x_i - \mathbb{E}\xi)(y_j - \mathbb{E}\eta) =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s} p_{i*}p_{*j}(x_i - \mathbb{E}\xi)(y_j - \mathbb{E}\eta) =$$

$$\sum_{i=1}^{r} p_{i*}(x_i - \mathbb{E}\xi)\sum_{j=1}^{s} p_{*j}(y_j - \mathbb{E}\eta) =$$

$$= \left(\sum_{i=1}^{r} p_{i*}x_i - \mathbb{E}\xi)\right)\left(\sum_{j=1}^{s} p_{*j}y_j - \mathbb{E}\eta\right) = 0\,.$$

В случае $\eta = \mathrm{a}\xi + \mathrm{b}$ имеем $\mathbb{E}\eta = \mathbb{E}(\mathrm{a}\xi + \mathrm{b}) = \mathrm{a}\mathbb{E}\xi + \mathrm{b}$

$$\mathbb{E}(\eta - \mathbb{E}\eta)^2 = \mathbb{E}(\mathrm{a}\xi + \mathrm{b} - \mathbb{E}(\mathrm{a}\xi + \mathrm{b}))^2 = \\ = \mathrm{a}^2\mathbb{E}(\xi - \mathbb{E}\xi)^2$$

$$\mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta) = \mathbb{E}(\xi - \mathbb{E}\xi)(\mathrm{a}\xi + \mathrm{b} - \mathbb{E}(\mathrm{a}\xi + \mathrm{b})) = \\ \mathrm{a}\mathbb{E}(\xi - \mathbb{E}\xi)^2$$

$$\varrho = \frac{\mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta)}{\sqrt{\mathbb{E}(\xi - \mathbb{E}\xi)^2}\sqrt{\mathbb{E}(\eta - \mathbb{E}\eta)^2}} =$$

$$= \frac{\mathrm{a}\mathbb{E}(\xi - \mathbb{E}\xi)^2}{\sqrt{\mathbb{E}(\xi - \mathbb{E}\xi)^2}\sqrt{\mathrm{a}^2\mathbb{E}(\xi - \mathbb{E}\xi)^2}} = \frac{\mathrm{a}}{|\mathrm{a}|} = \begin{cases} 1, & \mathrm{a} > 0 \\ -1, & \mathrm{a} < 0 \end{cases}$$

Выборочный коэффициент корреляции Пирсона

$$R = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

$(x_i, y_i), i = 1, 2, \ldots, n$, парная выборка массы тела до и после диеты, $\bar{x} = 74.7, \bar{y} = 72.1$, $R = 0.88$.

	1	2	3	4	5	6	7	8	9	10
x_i	68	80	92	81	70	79	78	66	57	76
y_i	60	84	87	79	74	71	72	67	57	70
$x_i y_i$	4080	6720	8004	6399	5180	5609	5616	4422	3249	5320
x_i^2	4624	6400	8464	6561	4900	6241	6084	4356	3249	5776
y_i^2	3600	7056	7569	6241	5476	5041	5184	4489	3249	4900

Вычисление в Excel

Например, значения признака X находятся в ячейках от A2 до A101, признака Y в ячейках от B2 до B101. Функция, вычисляющая коэффициент корреляции, имеет вид: КОРРЕЛ(A2 : A101; B2 : B101).
Заметим, что количество ячеек в этих массивах должно совпадать, иначе функция выдает ошибку.

Значимость коэффициента корреляции

При $H_0 : \varrho = 0$ статистика критерия вида:

$$\tau = \frac{R}{\sqrt{1 - R^2}}\sqrt{n - 2} \sim T(n - 2) \qquad (1)$$

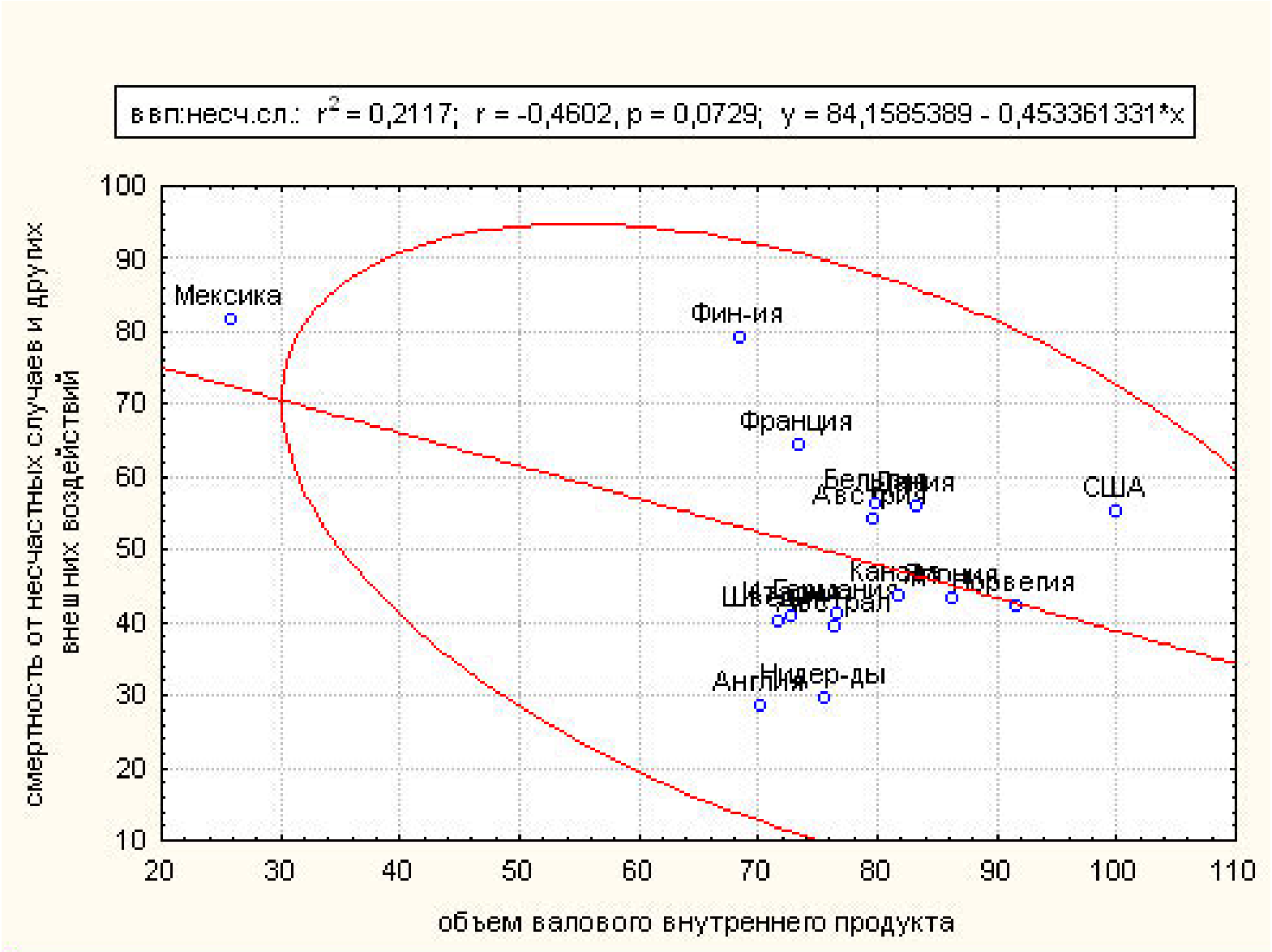
имеет распределение Стьюдента с $n - 2$ степенями свободы.

$$T(0.74; 255) = 17.5,$$
$$P = P\{|\tau| > 17.5\} = \text{СТЬЮДРАСП}(17.5; 253; 2) = 2 \cdot 10^{-45},$$

$$T(0.36; 255) = 6.16,$$
$$P = P\{|\tau| > 6.16\} = \text{СТЬЮДРАСП}(17.5; 253; 2) = 3 \cdot 10^{-9}.$$

P-значения меньше 10^{-8}, гипотеза $H_0 : \varrho = 0$ отвергается в пользу альтернативной $H_1 : \varrho \neq 0$, отсюда с вероятностью $1 - P$, можно говорить о том, что между признаками есть зависимость.

Непараметрические коэффициенты корреляции



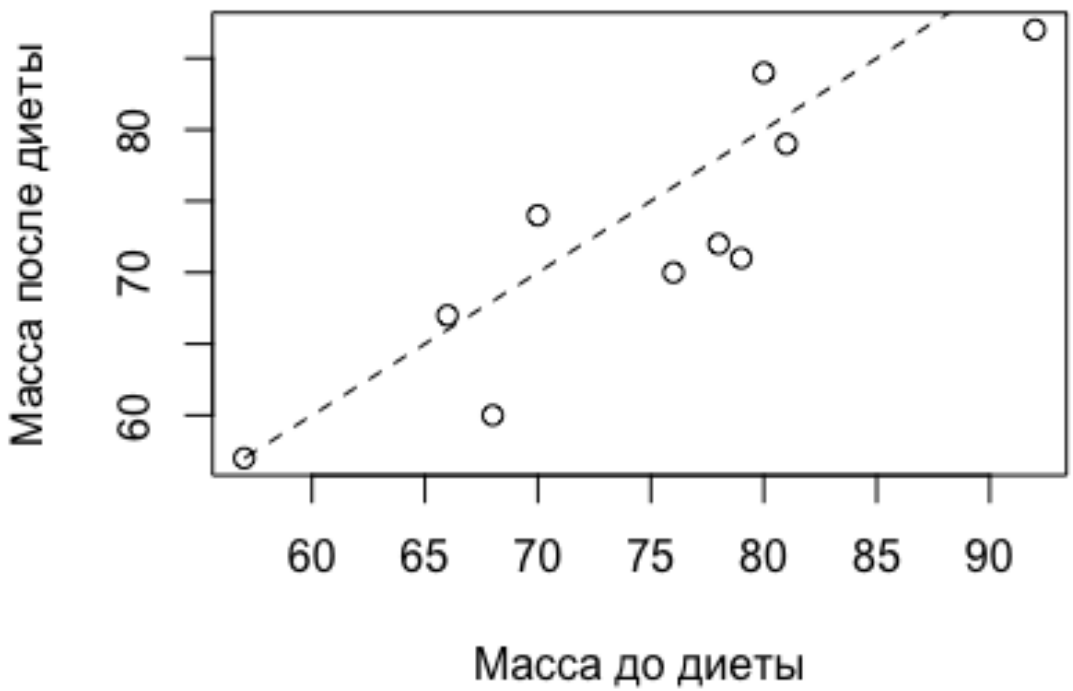
Корреляция Спирмена

Пусть $x'_1, \ldots, x'_n$ и $y'_1, \ldots, y'_n$ – ранги (порядковый номер элемента в упорядоченной выборке) индивидов по признакам X и Y,

k_i – количество индивидов, которых превосходит x_i,
l_i– количество индивидов, которые превосходят x_i,
$x_{i*} = k_i - l_i = 2(x'_i - \frac{n+1}{2})$.

$$k_3 = 2, l_3 = 7, x_{3*} = 2 - 7 = -5 = 2\left(3 - \tfrac{11}{2}\right) = 2 \cdot (-2.5)$$

ранг	1	2	3	4	5	6	7	8	9	10
Масса до диеты	57	66	68	70	76	78	79	80	81	92
x_{i*}	−9	−7	−5	−3	−1	1	3	5	7	9
Масса после диеты	57	67	60	74	70	72	71	84	79	87
k_i					3					
l_i					6					
$y_{i*} = k_i - l_i$					−3					


Масса после диеты
60
70
80
60
65
70
75
80
85
90
Масса до диеты

ранг	1	2	3	4	5	6	7	8	9	10
Масса до диеты	57	66	68	70	76	78	79	80	81	92
x_{i*}	-9	-7	-5	-3	-1	1	3	5	7	9
Масса после диеты	57	67	60	74	70	72	71	84	79	87
k_i	0	2	1	6	3	5	4	8	7	9
l_i	9	7	8	3	6	4	5	1	2	0
$y_{i*} = k_i - l_i$	-9	-5	-7	3	-3	1	-1	7	5	9
$x_{i*}y_{i*}$	81	35	35	-9	3	1	-3	35	35	81

$$Q_{xy} = \sum_{i=1}^{n} x_{i*}y_{i*} = 294, \quad Q_{xx} = \sum_{i=1}^{n} x_{i*}^2 = 330, \quad Q_{yy} = \sum_{i=1}^{n} y_{i*}^2 = 330,$$

$$R_s = \frac{Q_{xy}}{\sqrt{Q_{xx}Q_{yy}}} = 0.8909$$

Коэффициент корреляции Спирмена R_S

$$x_{i*} = k_i - l_i = 2(x'_i - \tfrac{n+1}{2}),\ y_{i*} = 2(y'_i - \tfrac{n+1}{2}).$$

$$Q_{xy} = \sum_{i=1}^{n} x_{i*}y_{i*},\ Q = Q_{xx} = \sum_{i=1}^{n} x_{i*}^2 = Q_{yy} = \sum_{i=1}^{n} y_{i*}^2$$

$$R_S = \frac{Q_{xy}}{Q} = \frac{Q_{xy}}{\sqrt{Q_{xx}}\sqrt{Q_{yy}}}$$

При полной положительной коррелиованности $Q_{xy} = Q$, $R_S = 1$.

$$\begin{array}{c} X \\ Y \end{array} \begin{pmatrix} 1 & 2 & \dots & n-1 & n \\ 1 & 2 & \dots & n-1 & n \end{pmatrix}$$

При полной отрицательной коррелиованности $Q_{xy} = -Q$, $R_S = -1$.

$$\begin{array}{c} X \\ Y \end{array} \begin{pmatrix} 1 & 2 & \dots & n-1 & n \\ n & n-1 & \dots & 2 & 1 \end{pmatrix}$$

Коэффициент корреляции Спирмена как выборочная корреляция между рангами

$$R_S = \frac{Q_{xy}}{\sqrt{Q_{xx}}\sqrt{Q_{yy}}} = \frac{\sum\limits_{i=1}^{n} 2(x'_i - \frac{n+1}{2})2(y'_i - \frac{n+1}{2})}{\sqrt{\sum\limits_{i=1}^{n}(2(x'_i - \frac{n+1}{2}))^2}\sqrt{\sum\limits_{i=1}^{n}(2(y'_i - \frac{n+1}{2}))^2}}$$

Так как $1 + 2 + \ldots + n = \frac{n(n+1)}{2}$, то среднее арифметическое последовательности $1, 2, \ldots, n$ равно $\frac{n+1}{2}$.

Значимость коэффициента корреляции Спирмена

$H_0 : R_S = 0$

$$\tau = \frac{R_S}{\sqrt{1 - R_S^2}}\sqrt{n-2} \sim T(n-2)$$

$\tau_* = 5.548157$, $p = P\{|\tau| > \tau_*\} = 0.0005421442$. Гипотеза $H_0 : \varrho = 0$ отвергается в пользу альтернативной $H_1 : \varrho \neq 0$.

Масса до диеты	68	80	92	81	70	79	78	66	57	76
Масса после диеты	60	84	87	79	74	71	72	67	57	70
Ранги до диеты	3	8	10	9	4	7	6	2	1	5
Ранги после диеты	2	9	10	8	7	5	6	3	1	4

cor.test(x,y,method="spearman exact=FALSE)

Spearman's rank correlation rho
data: x and y
S = 18, p-value = 0.0005421
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.8909091

Коэффициент корреляция Кендала

$x_1, \ldots, x_n$ и $y_1, \ldots, y_n$

$$x_{ik} = \begin{cases} 1, & x_i < x_k, \\ 0, & x_i = x_k, \\ -1, & x_i > x_k \end{cases} \quad \text{и} \quad y_{ik} = \begin{cases} 1, & y_i < y_k, \\ 0, & y_i = y_k, \\ -1, & y_i > y_k \end{cases}$$

и введем показатель однонаправленности неравенств

$$S = \sum_{i=1}^{n} \sum_{k=i+1}^{n} x_{ik} y_{ik} \, .$$

Поскольку число всевозможных пар равно $N = C_n^2 = \frac{n(n-1)}{2}$, то доля однонаправленных пар, называемая коэффициентом корреляции Кендалла, равна $T = \frac{S}{N}$.

Пример корреляции Кендала

$$\begin{matrix} X \\ Y \end{matrix} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 1 & 3 & 2 \end{pmatrix}.$$

ik	12	13	14	15	23	24	25	34	35	45
x_{ik}	1	1	1	1	1	1	1	1	1	1
y_{ik}	-1	-1	-1	-1	-1	-1	-1	1	1	-1
$x_{ik}y_{ik}$	-1	-1	-1	-1	-1	-1	-1	1	1	-1

$S_+ = 2$ и $S_- = 8$ количество 1 и -1 соответственно

$$S = 1 \cdot S_+ + (-1) \cdot S_- = S_+ - S_- = S_+ - (N - S_+) = 2S_+ - N.$$

S_+ равен количеству элементов выборки Y, больших y_m по всем m при упорядочивании их по признаку X.

$$T = \frac{S}{N} = \frac{2S_+ - N}{N} = \frac{2S_+}{N} - 1\,.$$

Значимость корреляции Кендала

Для проверки значимости отклонения от нуля коэффициента корреляции Кендалла используется то, что при независимости признаков

$$\mathrm{T} \sim \mathcal{N}(0, \sigma), \text{ где } \sigma^2 = \frac{2(2\mathrm{n}+5)}{9\mathrm{n}(\mathrm{n}-1)}.$$

Например, $\mathrm{T}_* = -0.6$, $\mathrm{n} = 5$, $\sigma^2 = \frac{2(2\cdot 5+5)}{9\cdot 5\cdot 4} = \frac{1}{6}$, $\sigma \approx 0.4$. При альтернативе отрицательной корреляции
$\alpha_* = \mathrm{P}\{\mathrm{T} < \mathrm{T}_*\} = \mathrm{pnorm}(-0.6; 0; 0.4) = 0.07$, отсюда отрицательную зависимость нельзя объяснить случайностью при $\alpha > 0.07$.

Пример с коррелированностью массы до и после диеты. Коэффициент Кендалла равен $\mathrm{T} = 0.7333$, $\mathrm{p} = 0.003$. Зависимость значимая.

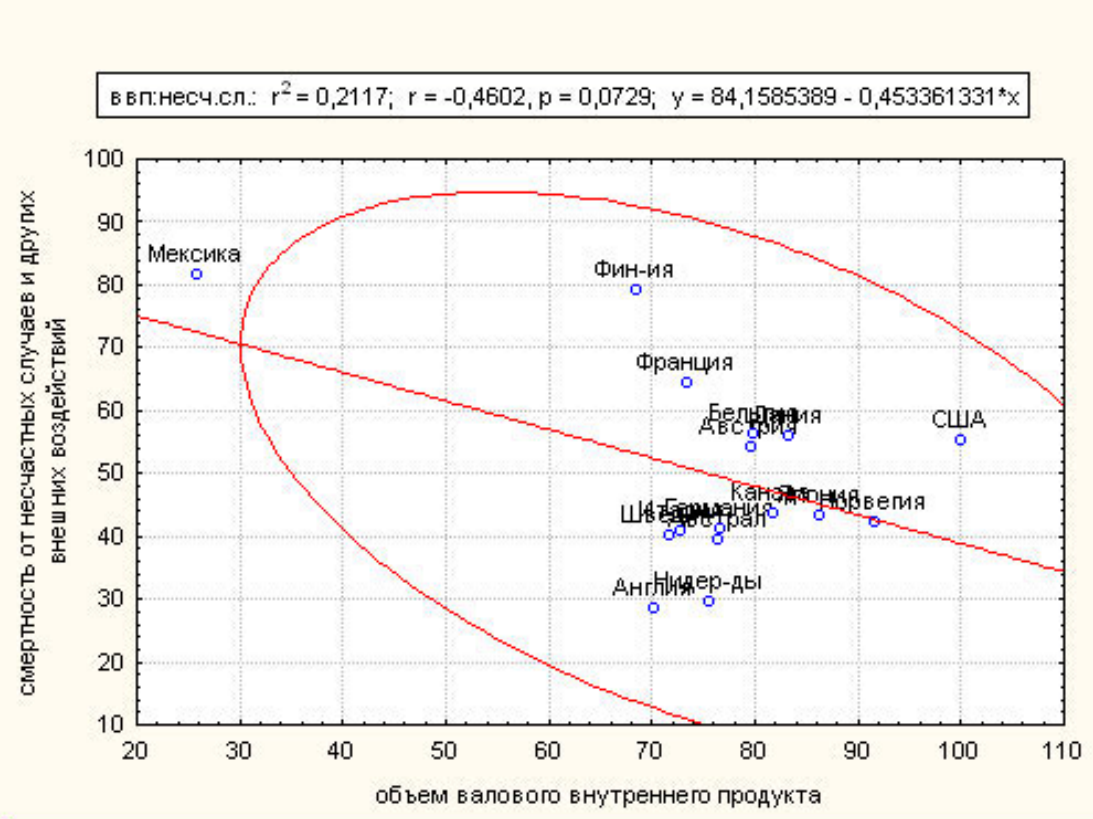


Рис.: $\mathrm{R} = -0.46,\ \mathrm{P} = 0.07.$

Коэффициент Спирмена $\mathrm{R_S} = 0.03$, $\mathrm{p} = 0.91$. Наблюдаемое уменьшение смертности при увеличении объема ВВП является не систематическим, а случайным за счет одной страны (Мексики).

Корреляционный анализ в BOA

Вопросы по корреляционному анализу	да-нет	r	p	зн./незн.
1. Продолжительность употребления героина (asid3 dyr) уменьшает индекс общего функционирования (gaf).	да	-0.17	0.0039	знач.
4. Ухудшение социального статуса asi2 emp приводит к увеличению тревожности у мужчин (sstati).	нет	-0.13	0.05898	тенд.

Корреляционный анализ в BOA

Вопросы по корреляционному анализу	да-нет	r	p	зн./незн.
1. При увеличении обычной дозы употребления героина (tlfbh2) увеличивается уровень депрессии (bdi).				
2. Плохие семейные отношения (asi6 fam) увеличивают уровень тревожности (sstati).				
3. Ухудшение социального статуса asi2 emp приводит к увеличению тревожности у женщин(sstati).				
4. У мужчин на повышении депрессии сказываются семейные отношения				