



*ОСНОВЫ
корреляционного анализа*

Многие объекты исследования характеризуются не одним, а множеством параметров, которые формируются в виде матрицы:

$$X_{n \times k} = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1k} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ik} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nk} \end{pmatrix}, \text{ где}$$

x_{ij} — i -ое наблюдение j -ого фактора.

Основная задача — выявление взаимосвязи между случайными переменными.

Задачи:

- отбор факторов, оказывающих наиболее существенное влияние на результативный признак;
- обнаружение ранее неизвестных причинных связей.

Основные средства анализа данных:

- парные коэффициенты корреляции,
- частные коэффициенты корреляции и
- множественные коэффициенты корреляции.



Парный корреляционный анализ

Если X и Y – СВ, то *теоретическая ковариация* :

$$\text{cov}(X, Y) = M[(X - M(X))(Y - M(Y))]$$

Если X и Y – независимы, то $\text{cov}(X, Y) = 0$.

Выборочная ковариация – статистическая мера взаимосвязи двух переменных.

При наличии n наблюдений:

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \text{ где}$$

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ – фактические значения X и Y ,

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

$$\boxed{\text{cov}(X, Y) = \overline{xy} - \bar{x} \cdot \bar{y}}.$$

Коэффициент парной корреляции и проверка его значимости

Для X и Y *теоретический коэффициент корреляции:*

$$\rho_{x,y} = \frac{\text{cov}(X, Y)}{\sqrt{\sigma_x^2 \sigma_y^2}} = \frac{\text{cov}(X, Y)}{\sigma_x \cdot \sigma_y}, \text{ где}$$

σ_x^2 , σ_y^2 - дисперсии СВ X и Y , $\text{cov}(X, Y)$ – их ковариация.

Свойства:

1. $-1 \leq \rho_{x,y} \leq 1$.
2. При $\rho_{xy} = \pm 1$ СВ X и Y связаны линейной зависимостью, т.е. $Y = \alpha X + \beta$.
3. При $\rho_{xy} = 0$ линейная корреляционная связь отсутствует.

Оценка коэффициента корреляции ρ – *выборочный парный коэффициент корреляции r* :

$$r = \frac{\text{cov}(X, Y)}{S_x \cdot S_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}} \quad \text{или}$$

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} \cdot \sqrt{\overline{y^2} - (\bar{y})^2}}, \text{ где}$$

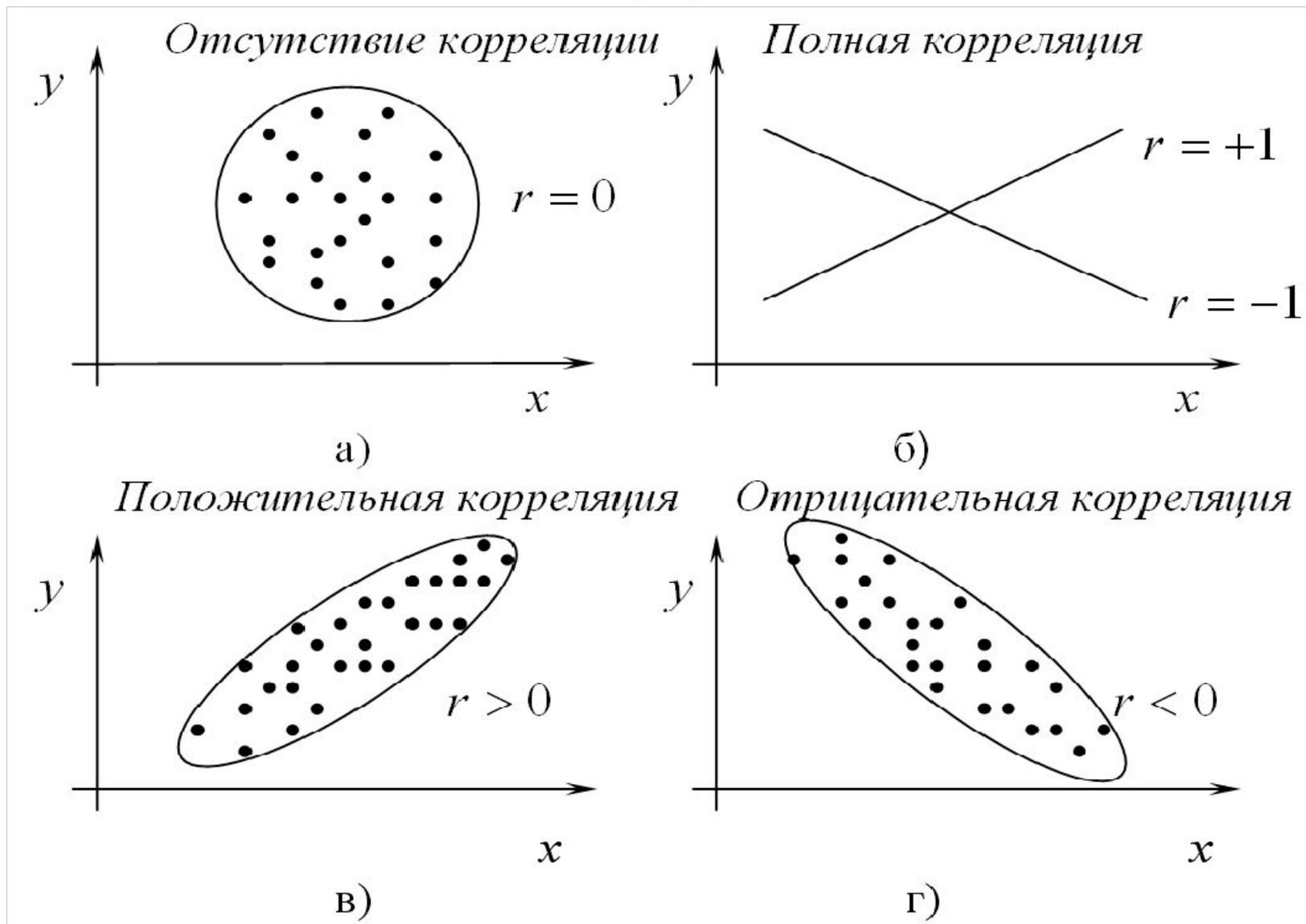
$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \quad - \text{выборочные}$$

дисперсии величин X и Y .

Оценка тесноты связи можно переменными

Теснота связи	Величина коэффициента корреляции при наличии	
	прямой связи (+)	обратной связи (-)
<i>Связь отсутствует</i>	$r = 0$	$r = 0$
<i>Связь очень слабая</i>	$0 < r \leq 0,3$	$-0,3 \leq r < 0$
<i>Связь слабая</i>	$0,3 < r \leq 0,5$	$-0,5 \leq r < -0,3$
<i>Связь средняя</i>	$0,5 < r \leq 0,7$	$-0,7 \leq r < -0,5$
<i>Связь сильная</i>	$0,7 < r \leq 0,9$	$-0,9 \leq r < -0,7$
<i>Связь очень сильная</i>	$0,9 < r < 1$	$-1 < r < -0,9$
<i>Полная функциональная</i>	$r = 1$	$r = -1$

Наиболее простым, приближенным способом выявления корреляционной связи является графический метод.



ОЦЕНКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТА ПАРНОЙ КОРРЕЛЯЦИИ

$H_0 : \rho = 0$ – значение коэффициента корреляции для генеральной совокупности равно нулю, т.е. *в генеральной совокупности отсутствует корреляция.*

$$H_1 : \rho \neq 0 .$$

Оценка значимости осуществляется с помощью t -критерия Стьюдента:

$$t_{\text{набл}} = \frac{r}{S_r} = r \sqrt{\frac{n-2}{1-r^2}} , \text{ где}$$

$$S_r = \sqrt{\frac{1-r^2}{n-2}} \quad - \text{ ошибка коэффициента корреляции.}$$

Строится двусторонняя критическая область, границы критической области которой находят из условия:

$$P\left(t_{\text{набл}} < -t_{\frac{\alpha}{2}, n-2}\right) = P\left(t_{\text{набл}} > t_{\frac{\alpha}{2}, n-2}\right) = \frac{\alpha}{2}.$$

$t_{\text{набл}}$ сравнивается с $t_{\text{кр}}(\alpha; \nu)$ для двусторонней критической области, которое берется с учетом заданного уровня значимости α и числа степеней свободы $\nu = n - 2$ (функция Excel СТЮДРАСПОБР ($\alpha; \nu = n - 2$)).

Если $|t_{\text{набл}}| > t_{\text{кр}}(\alpha, \nu = n - 2)$, то полученное значение коэффициента корреляции признается *значимым* (т.е. H_0 отвергается).

Вывод: с доверительной вероятностью $\gamma = 1 - \alpha$ можно утверждать, что между исследуемыми переменными *есть линейная статистическая зависимость*.

КОРРЕЛЯЦИОННОЕ ОТНОШЕНИЕ И ПРОВЕРКА ЕГО ЗНАЧИМОСТИ

В случае нелинейной зависимости тесноту связи между величинами оценивают по величине *корреляционного отношения*:

$$\eta = \sqrt{1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (0 \leq \eta \leq 1), \text{ где}$$

y_i – наблюдаемые значения,

\hat{y}_i – расчетные значения результивной переменной.

Величина η^2 , называемая *коэффициентом детерминации*, показывает, какая часть общей вариации Y обусловлена вариацией X .

$H_0: \eta = 0$ – в генеральной совокупности отсутствует корреляция.

$H_1: \eta \neq 0$.

Для проверки гипотезы вычисляется статистика:

$$F = \frac{\eta^2 (n - k)}{(1 - \eta^2)(k - 1)} \sim F(\alpha, v_1 = k - 1, v_2 = n - k), \text{ где}$$

k – число факторов, n – количество наблюдений).

Строится критическая область $(F_{кр, \alpha}; \infty)$, для этого границы критической области находят из условия:

$$P(F_{набл} > F_{\alpha, v_1, v_2}) = \alpha$$

(функция Excel ФРАСПОБР $(\alpha, v_1 = k - 1, v_2 = n - k)$).

Можно утверждать с доверительной вероятностью $\gamma = 1 - \alpha$, что корреляционное отношение η значительно отличается от нуля, если $F_{набл} > F_{кр}(\alpha, v_1 = k - 1, v_2 = n - k)$.



Множественный корреляционный анализ

Для измерения силы линейных связей одной переменной X_i с совокупность других $(k-1)$ переменных из их множества (X_1, \dots, X_n) также используются коэффициенты парной корреляции.

Матрица коэффициентов парной корреляции R :

$$R = \begin{pmatrix} 1 & r_{x_1x_2} & r_{x_1x_3} & \dots & r_{x_1x_k} \\ r_{x_2x_1} & 1 & r_{x_2x_3} & \dots & r_{x_2x_k} \\ r_{x_3x_1} & r_{x_3x_2} & 1 & \dots & r_{x_3x_k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{x_kx_1} & r_{x_kx_2} & r_{x_kx_3} & \dots & 1 \end{pmatrix} \text{ или } R = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1k} \\ r_{21} & 1 & r_{22} & \dots & r_{2k} \\ r_{31} & r_{32} & 1 & \dots & r_{3k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & r_{k3} & \dots & 1 \end{pmatrix}, \text{ где}$$

r_{ij} – выборочный парный коэффициент корреляции, характеризующий тесноту линейной связи между показателями X_i и X_j .

Многомерный корреляционный анализ решает *две задачи*:

1. Определение тесноты связи одной переменной с совокупностью остальных $(k - 1)$ величин, включенных в анализ;
2. Определение тесноты связи между переменными при фиксировании (исключении) влияния остальных.

Эти задачи решаются с помощью коэффициентов множественной и частной корреляции, соответственно.

Выборочный коэффициент множественной корреляции и проверка его значимости

Выборочный коэффициент множественной корреляции

$$R_{j,1,2,\dots,j-1,j+1,\dots,k} = \sqrt{1 - \frac{|R|}{R_{jj}}}, \text{ где}$$

$|R|$ – определитель корреляционной матрицы R ;

R_{jj} – алгебраическое дополнение элемента r_{jj} матрицы R .

$R^2_{j,1,2,\dots,j-1,j+1,\dots,k}$ называют *выборочным множественным коэффициентом детерминации*, который показывает, какую долю вариации исследуемой j -ой величины объясняет вариация остальных $(k-1)$ величин.

Значимость множественного коэффициента корреляции проверяется по F – критерию Фишера.

$H_0 : R^2 = 0$, т.е. в генеральной совокупности отсутствует корреляция.

$H_1 : R^2 \neq 0$.

Для проверки гипотезы вычисляется статистика:

$$F_{\text{набл}} = \frac{R^2 / (k - 1)}{(1 - R^2) / (n - k)} \quad \text{или} \quad \boxed{F_{\text{набл}} = \frac{R^2 \cdot (n - k)}{(1 - R^2) \cdot (k - 1)}}.$$

Для критической области $(F_{\text{кр},\alpha}; \infty)$ критические значения статистики находят из условия:

$$P(F_{\text{набл}} > F_{\text{кр}}(\alpha, v_1 = k - 1, v_2 = n - k)) = \alpha$$

(функция Excel **FRASПОБР** ($\alpha, v_1 = k - 1, v_2 = n - k$)).

Множественный коэффициент корреляции считается значимым, если выполняется неравенство: $F_{\text{набл}} > F_{\text{кр}}$.

Частный коэффициент корреляции и проверка его значимости

Выборочный частный коэффициент корреляции X_j и X_i , при фиксированных значениях остальных переменных ($k-2$) определяется по формуле:

$$r_{ji \cdot 1, 2, \dots, k} = - \frac{R_{ji}}{\sqrt{R_{jj} R_{ii}}}, \text{ где}$$

R_{ji}, R_{jj}, R_{ii} – алгебраические дополнения к соответствующим элементам корреляционной матрицы R .

$R_{ji} = (-1)^{j+i} \cdot M_{ji}$, где M_{ji} – минор элемента r_{ji} (определитель матрицы, получаемой путем вычеркивания j -й строки и i -го столбца из матрицы R).

Частный коэффициент корреляции, так же как и парный коэффициент корреляции изменяется от -1 до $+1$.

Для вычисления коэффициентов частной корреляции можно использовать рекуррентную формулу:

$$r_{yx_i \cdot x_1 \dots x_k} = \frac{r_{yx_i \cdot x_1 \dots x_{k-1}}^{\text{без } x_i} - \overbrace{r_{yx_k \cdot x_1 \dots x_{k-1}}^A}^{\text{без } x_i} \overbrace{r_{x_i x_k \cdot x_1 \dots x_{k-1}}^B}^{\text{без } x_i}}{\sqrt{(1 - A^2)(1 - B^2)}}$$

В частности, для случая трех переменных, выборочный частный коэффициент корреляции между переменными X и Y при фиксированных значениях переменной Z равен:

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

Оценка значимости коэффициента частной корреляции осуществляется с помощью *t*-критерия Стьюдента:

$$t_{\text{набл}} = \frac{r}{S_r} = \sqrt{\frac{r^2}{1-r^2} \cdot (n-l-2)} = r \sqrt{\frac{n-l-2}{1-r^2}}, \text{ где}$$

r – соответственно оценка частного коэффициент корреляции;

l – число фиксируемых факторов.

$H_0 : R^2 = 0$ – в генеральной совокупности отсутствует корреляция.

$$H_1 : R^2 \neq 0.$$

Строится критическая область, границы которой находят из условия:

$$P\left(t_{\text{набл}} < -t_{\frac{\alpha}{2}, n-2}\right) = P\left(t_{\text{набл}} > t_{\frac{\alpha}{2}, n-2}\right) = \frac{\alpha}{2}.$$

$t_{\text{набл}}$ сравнивается с критическим $t_{\text{кр}}(\alpha; \nu)$ для двусторонней критической области (функция Excel **СТЮДРАСПОБР** ($\alpha; \nu = n - l - 2$)).

Если $|t_{\text{набл}}| > t_{\text{кр}}(\alpha, \nu)$, то коэффициент корреляции признается *значимым* (т.е. H_0 отвергается).

Вывод: с вероятностью $\gamma = 1 - \alpha$ можно утверждать, что между исследуемыми переменными есть линейная статистическая зависимость.