Корреляционный анализ

Задачи корреляционного анализа:

а) Измерение степени связности (тесноты, силы, строгости, интенсивности) двух и более явлений;

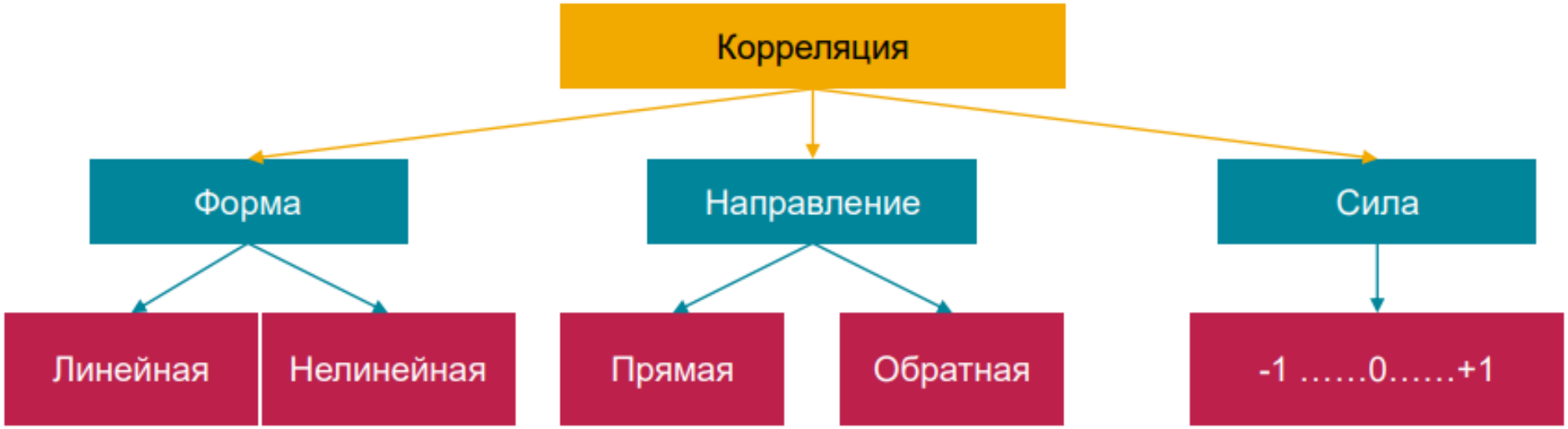
б) Отбор факторов, оказывающих наиболее существенное влияние на результативный признак, на основании измерения степени связности между явлениями. Существенные в данном аспекте факторы используют далее в регрессионном анализе;

в) Обнаружение неизвестных причинных связей.

Связь называется корреляционной, если каждому значению факторного признака соответствует вполне определенное неслучайное значение результативного признака.

Корреляционный анализ

Характер связи между переменными



- При **положительной линейной корреляции** более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака – низкие значения другого.
- При **отрицательной линейной корреляции** более высоким значениям одного признака соответствуют более низкие значения другого, а более низким значениям одного признака – высокие значения другого.

Корреляционный анализ

Устанавливает зависимость между с.в. и количественную оценку степени неслучайности их совместного изменения.

Изменение с.в. *y*, разбивается на две составляющие — стохастическую, связанную с неслучайной зависимостью *y* от *x*, и случайную, связанную со случайным характером поведения самих *y* и *x*. Стохастическая составляющая связи между *y* и *x* характеризуется коэффициентом корреляции

$$\rho = \frac{\mathbf{M}\{[x - \mathbf{M}(x)][y - \mathbf{M}(y)]\}}{\sqrt{\mathbf{D}(x)\mathbf{D}(y)}}$$

Коэффициент показывает отсутствие (ρ=0) или наличие строгой линейной (ρ=$\pm$1) связи между *x* и *y*.

Коэффициент корреляции ρ не учитывает возможной криволинейной связи между с.в.

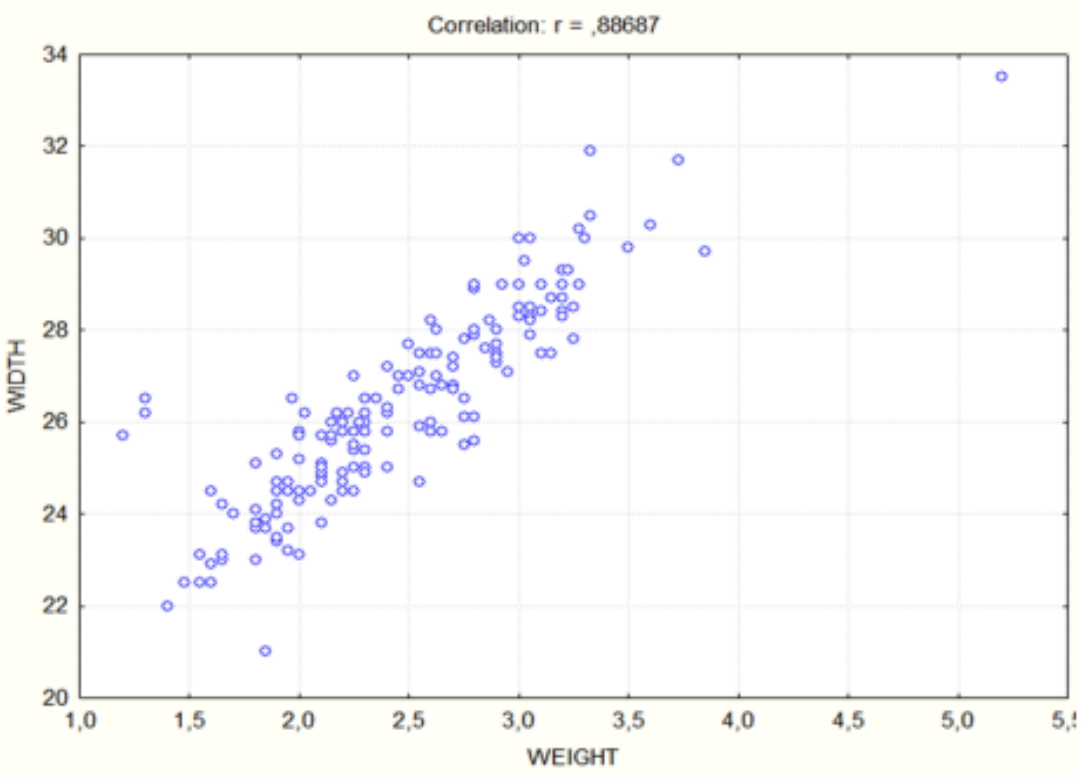
Корреляционный анализ

Поле корреляции - *это поле точек, координаты которых (x; y) определяются значениями факторного и результативного признаков.*

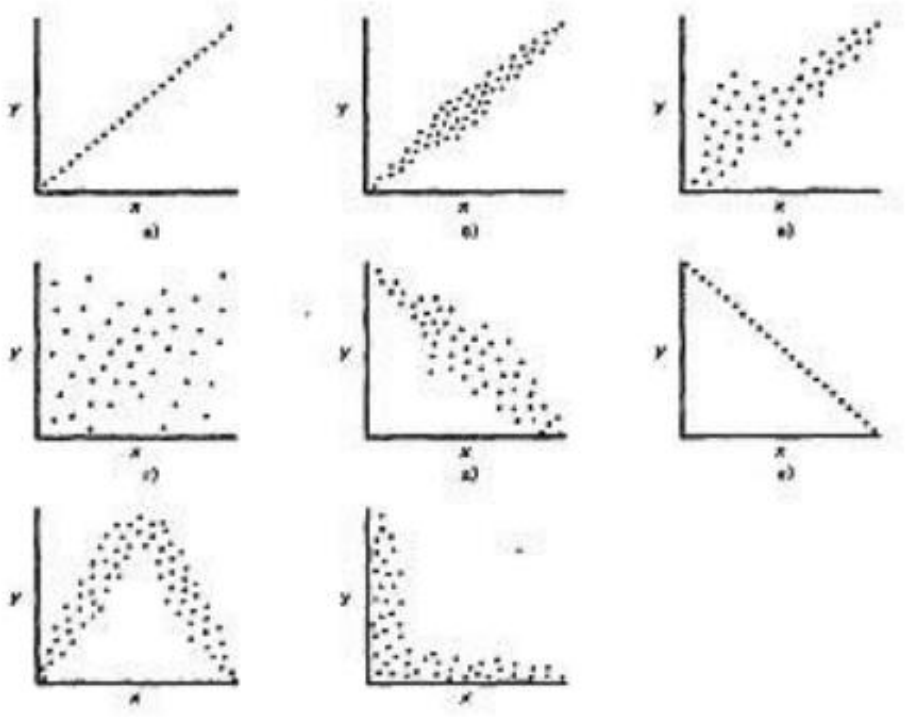
Позволяет судить о наличии и о характере связи (нелинейная, а если линейная, то и о направлении (прямая или обратная)).

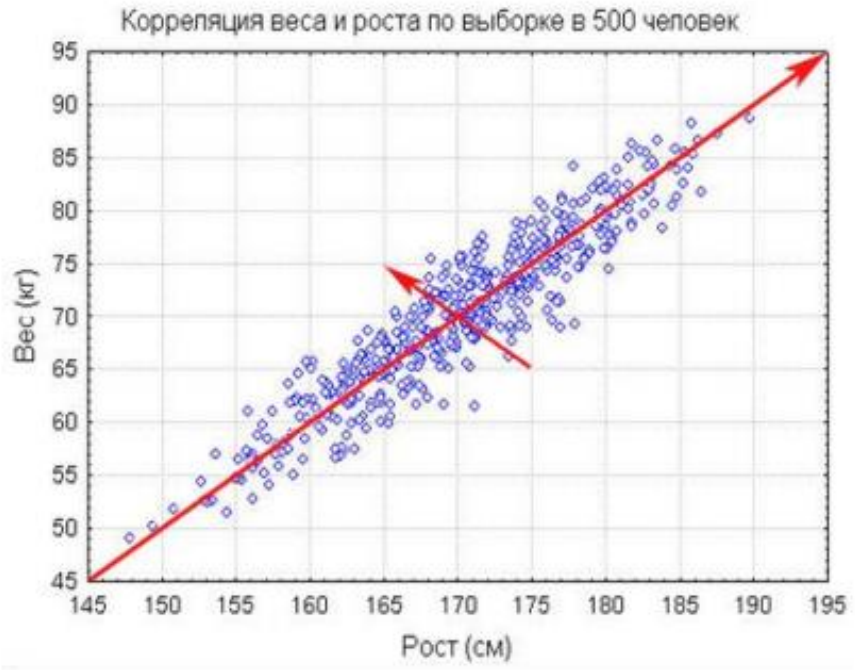
Характеристики диаграммы:

- наклон (направление связи)
- ширина (сила, теснота связи).



Корреляционный анализ





а) строгая положительная корреляция
б) положительная корреляция
в) слабая положительная корреляция
г) нулевая корреляция
д) отрицательная корреляция
е) строгая отрицательная корреляция
ж) нелинейная корреляция
з) нелинейная корреляция

Корреляционный анализ

Коэффициент корреляции Пирсона

x и y — нормально распределенные с.в. (x_1, y_1), (x_2, y_2),…, (x_n, y_n). Выборочной оценкой коэффициента корреляции ρ является

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}},$$

где n – объем выборки, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i;\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

при $n<15$ $r^* = r\left[1 + \frac{1 - r^2}{2(n-3)}\right]$

Корреляционный анализ

Задача проверки гипотезы о значимости корреляционной связи между с.в., т.е. значимости отклонения коэффициента корреляции ρ от нуля.

Гипотезы: $H_0 : |\rho| = 0$, $H_1 : |\rho| \neq 0$.

Для этого сравнивают r с критическим значением r_α - α-квантиль распределения r при $\rho = 0$. Корреляция между случайными величинами признается значимой, если $|r| \geq r_\alpha$. Критические значения r_α приведены в таблице.

Корреляционный анализ

Критические значения r_α выборочного коэффициента корреляции для $\rho = 0$

n	Доверительная вероятность α			n	Доверительная вероятность α		
	0,90	0,95	0,99		0,90	0,95	0,99
3	0,988	0,997	1,000	13	0,476	0,553	0,684
4	0,900	0,950	0,990	14	0,457	0,532	0,661
5	0,805	0,878	0,959	15	0,441	0,514	0,641
6	0,729	0,811	0,917	16	0,426	0,497	0,623
7	0,669	0,754	0,874	17	0,412	0,482	0,606
8	0,621	0,707	0,834	18	0,400	0,468	0,590
9	0,582	0,666	0,798	19	0,389	0,456	0,575
10	0,549	0,632	0,765	20	0,378	0,444	0,561
11	0,521	0,602	0,735	21	0,369	0,433	0,549
12	0,497	0,576	0,708	22	0,360	0,423	0,537

Корреляционный анализ

Для определения критического значения r_α можно использовать аппроксимации

При $n > 5$

$$r_\alpha = \frac{\exp\left(\frac{2}{\sqrt{n-3}} u_{\frac{1+\alpha}{2}}\right) - 1}{\exp\left(\frac{2}{\sqrt{n-3}} u_{\frac{1+\alpha}{2}}\right) + 1};$$

при $n > 10$

$$r_\alpha = \sqrt{\frac{t^2_{\frac{1+\alpha}{2}}}{n - 2 + t^2_{\frac{1+\alpha}{2}}}};$$

при $n > 200$

$$r_\alpha = \frac{1}{\sqrt{n-1}} u_{\frac{1+\alpha}{2}}.$$

где u_α и t_α - α-квантиль соответственно стандартного нормального распределения и распределения Стьюдента с $f = n - 2$ степенями свободы.

Корреляционный анализ

Линейный коэффициент корреляции принимает значения от –1 до +1.

Сила связи оцениваются по шкале Чеддока:

$0.1 < r_{xy} < 0.3$: слабая;

$0.3 < r_{xy} < 0.5$: умеренная;

$0.5 < r_{xy} < 0.7$: заметная;

$0.7 < r_{xy} < 0.9$: высокая;

$0.9 < r_{xy} < 1$: весьма высокая.

Корреляционный анализ

Пример. В результате наблюдений над случайными величинами *x* и *y* получена следующая совокупность данных (*n* = 10):

x	2	4	1	7	3	11	14	15	21	4
y	7	6	4	11	2	21	31	23	40	15

Проверить гипотезу о наличии корреляции между с.в. *x* и *y* с достоверностью α = 0,95.

$$\bar{x} = \frac{1}{10} \cdot \sum_{i=1}^{10} x_i = 8{,}2; \quad \sum_{i=1}^{10} (x_i - \bar{x})^2 = 405{,}6; \quad \bar{y} = \frac{1}{10} \cdot \sum_{i=1}^{10} y_i = 16{,}0;$$

$$\sum_{i=1}^{10} (y_i - \bar{y})^2 = 1422; \quad \sum_{i=1}^{10} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = 723.$$

Корреляционный анализ

$$r = \frac{723}{\sqrt{405{,}6 \cdot 1422}} = 0{,}952; \quad r^* = 0{,}952 \cdot \left(1 + \frac{1 - 0{,}952^2}{2 \cdot 7}\right) = 0{,}958.$$

Из таблицы для n = 10 и α = 0,95 находим $r_{0,95}$ = 0,632.

Так как $r(r^*)$ = 0,952 (0,958) > $r_{0,95}$ = 0,632, то наличие зависимости между с.в. x и y следует признать значимой с достоверностью α = 0,95.

По шкале Чеддока сила связи – весьма высокая.

Корреляционный анализ

Оценка корреляционного отношения

Имеется n значений с.в. y: y_1, y_2,…,y_k. При $y = y_i$ наблюдаются n_i значений с.в. x. Если x_{ij} – j-е значение величины x, наблюдаемое при $y = y_i$ (j = 1, 2, …, n_i), то выборочная оценка корреляционного отношения x по y равна

$$\eta^2_{xy} = \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i^2 - n\bar{x}^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - n\bar{x}^2}.$$

$$n = \sum_{i=1}^{k} n_i;$$

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij};$$

$$\bar{x} = \frac{1}{k}\sum_{i=1}^{k} \bar{x}_i,$$

Корреляционный анализ

Оценка корреляционного отношения

Гипотеза H_0: $\eta^2 = 0$, $\eta^2 \neq 0$.

Статистика $$l = \frac{\eta^2 (n-k)}{(k-1)(1-\eta^2)}$$

Если $l \geq F_\alpha(f_1, f_2)$, то нулевая гипотеза отклоняется с достоверностью α. $F_\alpha(f_1, f_2)$ - α- квантиль F-распределения с $f_1 = k - 1$ и $f_2 = n - k$ степенями свободы.

Корреляционный анализ

Оценка корреляционного отношения

Пример. Проверить линейность корреляционной связи для выборки

x	2	4	9	13	15
y	1,3,4	7,8,12	14,19,21	11,9,6	8,7,3

с доверительной вероятностью α = 0,95.

$k = 5$, $n_i = 3$, $n = 15$

$$\bar{x}_1 = \frac{1+3+4}{3} = 2{,}66; \quad \bar{x}_2 = 9; \quad \bar{x}_3 = 18; \quad \bar{x}_4 = 8{,}67; \quad \bar{x}_5 = 6; \quad \sum_{i=1}^{5}\sum_{j=1}^{3} x_{ij}^2 = 1641;$$

$$\bar{x} = \frac{2{,}67+9+18+8{,}67+6}{5} = 8{,}864; \quad \sum_{i=1}^{5} n_i \cdot \bar{x}_i^2 = 3 \cdot \left(2{,}66^2 + 9^2 + \ldots + 6^2\right) = 1569{,}2136.$$

$$\eta^2_{xy} = \frac{1569{,}2136 - 15 \cdot 8{,}864^2}{1641 - 15 \cdot 8{,}864^2} = 0{,}8448.$$

Корреляционный анализ

Оценка корреляционного отношения

$F_{0,95}(f_1, f_2) = F_{0,95}(5 - 1, 15 - 5) = F_{0,95}(4, 10) = 3,5.$

$$l = \frac{\eta^2 \cdot (n-k)}{(k-1) \cdot (1-\eta^2)} = \frac{0,845 \cdot 10}{4 \cdot (1-0,845)} = 13,629.$$

13,629 > 3,5, следовательно, необходимо признать наличие существенной нелинейной связи между *x* и *y*.

Корреляционный анализ

Частная и множественная корреляция

Для исследования связи между ≥3 с.в. используют частные и множественные коэффициенты корреляции.

Для трех с.в. *x*, *y* и *z* зависимость между двумя *x* и *y* при фиксированной *z* оценивается с помощью частного коэффициента корреляции $\rho_{xy,z}$. Аналогично определяют $\rho_{xz,y}$ и $\rho_{zy,x}$.

Выборочные *частные* (парные) коэффициенты корреляции

$$r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}; \quad r_{xz,y} = \frac{r_{xz} - r_{xy}r_{zy}}{\sqrt{(1-r_{xy}^2)(1-r_{zy}^2)}}; \quad r_{zy,x} = \frac{r_{zy} - r_{zx}r_{yx}}{\sqrt{(1-r_{zx}^2)(1-r_{yx}^2)}};$$

$$r_{yz,x} = r_{ez,x}; \quad r_{xy,z} = r_{yx,z}; \quad r_{xz,y} = r_{zx,y}.$$

Парные коэффициенты принимают значения от −1 до +1.

Корреляционный анализ

Частная и множественная корреляция

Гипотеза H_0: $\rho_{xy,z} = 0$ проверяется с помощью статистики

$$t = \frac{\sqrt{n-k}\, r_{xy,z}}{\sqrt{1 - r^2_{xy,z}}},$$

где k — число переменных (в нашем случае k =3).

При справедливости H_0 величина t распределена в соответствии с распределением Стьюдента при $f = n - k$ степенях свободы.

При $|t| > t_{\frac{1+\alpha}{2}}(n-k)$ — H_0 отклоняется с вероятностью α.

Корреляционный анализ

Частная и множественная корреляция

Множественная корреляция исследуется в случае, когда необходимо установить существенность взаимосвязи одной переменной с совокупностью остальных.

Выборочные *множественные* коэффициенты корреляции обозначаются $r_{x,yz}$, $r_{y,xz}$ и $r_{z,xy}$.

$$r^2_{x,yz} = \frac{r^2_{xy} + r^2_{xz} - 2r_{xy}r_{zx}r_{yz}}{1 - r^2_{yz}}; \qquad r^2_{y,xz} = \frac{r^2_{yx} + r^2_{yz} - 2r_{yx}r_{yz}r_{zx}}{1 - r^2_{xz}};$$

$$r^2_{z,xy} = \frac{r^2_{zx} + r^2_{zy} - 2r_{zx}r_{zy}r_{xy}}{1 - r^2_{xy}}.$$

$$r^2_{x,yz} = 1 - (1 - r^2_{xz})(1 - r^2_{xy,z}) = 1 - (1 - r^2_{xy})(1 - r^2_{xz,y});$$
$$r^2_{y,xz} = 1 - (1 - r^2_{yz})(1 - r^2_{yx,z}) = 1 - (1 - r^2_{yx})(1 - r^2_{yz,x});$$
$$r^2_{z,xy} = 1 - (1 - r^2_{zy})(1 - r^2_{zx,y}) = 1 - (1 - r^2_{zx})(1 - r^2_{zy,x}).$$

Корреляционный анализ

Частная и множественная корреляция

Гипотеза H_0: $\rho_{x,yz} = 0$ проверяется с помощью статистики

$$F = \frac{r^2_{x,yz}}{1 - r^2_{x,yz}} \frac{n-k}{k-1},$$

имеющая при справедливости H_0 F-распределение с $f_1 = k - 1$ и $f_2 = n - k$ степенями свободы (k = 3).

Если $F > F_\alpha(f_1, f_2)$, то соответствующая корреляция признается значимой. Критическое значение коэффициента множественной корреляции равно

$$r_{x.yz}(\alpha) = \sqrt{\frac{(k-1)\,F_\alpha(f_1, f_2)}{n-k+(k-1)\,F_\alpha(f_1, f_2)}}.$$

Корреляция признается значимой при $r_{x,yz} \geq r_{x,yz}(\alpha)$.

Корреляционный анализ

Критические значения $r_{1,23\ldots k}$ коэффициента множественной корреляции (k — число переменных, n — объем выборки)

$n-k$	Доверительная вероятность α							
	0,95				0,99			
	k				k			
	3	4	5	6	3	4	5	6
1	0,999	0,999	0,999	1,000	1,000	1,000	1,000	1,000
2	0,975	0,983	0,987	0,990	0,995	0,997	0,997	0,998
3	0,930	0,950	0,961	0,968	0,977	0,983	0,987	0,990
4	0,881	0,912	0,930	0,942	0,949	0,962	0,970	0,975
5	0,836	0,874	0,898	0,914	0,917	0,937	0,949	0,957
6	0,795	0,839	0,867	0,886	0,886	0,911	0,927	0,938
7	0,758	0,807	0,838	0,860	0,855	0,885	0,904	0,918
8	0,726	0,777	0,811	0,835	0,827	0,860	0,882	0,898
9	0,697	0,750	0,786	0,812	0,800	0,837	0,861	0,878
10	0,671	0,726	0,763	0,790	0,776	0,814	0,840	0,859
11	0,648	0,703	0,741	0,770	0,753	0,793	0,821	0,841
12	0,627	0,683	0,722	0,751	0,732	0,773	0,802	0,824
13	0,608	0,664	0,703	0,733	0,712	0,755	0,785	0,807
14	0,590	0,646	0,686	0,717	0,694	0,737	0,768	0,791
15	0,574	0,630	0,670	0,701	0,677	0,721	0,752	0,776
16	0,559	0,615	0,655	0,687	0,662	0,706	0,738	0,762
17	0,545	0,601	0,641	0,673	0,647	0,691	0,724	0,749
18	0,532	0,587	0,628	0,660	0,633	0,678	0,710	0,736
19	0,520	0,575	0,615	0,647	0,620	0,665	0,697	0,723
20	0,509	0,563	0,604	0,636	0,607	0,652	0,685	0,712
22	0,488	0,542	0,582	0,614	0,585	0,630	0,663	0,690
24	0,470	0,523	0,562	0,594	0,565	0,609	0,643	0,669
26	0,454	0,506	0,545	0,576	0,546	0,590	0,624	0,651
28	0,439	0,490	0,529	0,560	0,529	0,573	0,607	0,633
30	0,425	0,476	0,514	0,545	0,514	0,557	0,591	0,618
40	0,373	0,419	0,455	0,484	0,454	0,494	0,526	0,552
60	0,308	0,348	0,380	0,406	0,377	0,414	0,442	0,467

Корреляционный анализ

Пример. Вычислить коэффициенты частной и множественной корреляций и проверить их значимость при доверительной вероятности α = 0,95 для данных, приведенных ниже (n = 10, k = 3).

x	1	3	4	7	12	4	19	21	1	3
y	12	42	58	71	68	50	49	85	18	26
z	41	12	7	3	14	27	38	13	64	75

Определяем парные коэффициенты корреляции

$$\sum_{i=1}^{10}(x_i-\bar{x})^2 = 484{,}5; \quad \sum_{i=1}^{10}(y_i-\bar{y})^2 = 5118{,}9 \quad \sum_{i=1}^{10}(x_i-\bar{x})\cdot(y_i-\bar{y}) = 1102{,}5$$

$$\bar{x} = \frac{1}{10}\cdot\sum_{i=1}^{10}x_i = 7{,}5; \quad \bar{y} = \frac{1}{10}\cdot\sum_{i=1}^{10}y_i = 47{,}9; \quad r_{xy} = \frac{1102{,}5}{\sqrt{484{,}5\cdot 5118{,}9}} = 0{,}7$$

Корреляционный анализ

$$\sum_{i=1}^{10}(x_i-\bar{x})^2=484{,}5;\quad \sum_{i=1}^{10}(z_i-\bar{z})^2=5498{,}4;\quad \sum_{i=1}^{10}(x_i-\bar{x})(z_i-\bar{z})=-519;$$

$$\bar{x}=7{,}5;\quad \bar{z}=29{,}4;\quad r_{xz}=-\frac{519}{\sqrt{484{,}5\cdot 5498{,}4}}=-0{,}318.$$

$$\sum_{i=1}^{10}(y_i-\bar{y})^2=\text{ 5118,9}\quad \sum_{i=10}^{10}(z_i-\bar{z})^2=5498{,}4;\quad \sum_{i=1}^{10}(y_i-\bar{y})\cdot(z_i-\hat{z})=\text{ - 4096,6}$$

$$\bar{y}=47{,}9;\quad \bar{z}=29{,}4;\quad r_{yz}=-\frac{\text{- 4096,6}}{\sqrt{\text{5118,9}\cdot 5498{,}4}}=\text{ - 0,772}$$

Частные коэффициенты корреляции

$$r_{xy,z}=\frac{r_{xy}-r_{xz}\cdot r_{yz}}{\sqrt{(1-r_{xy}^2)\cdot(1-r_{yz}^2)}}=\frac{0{,}7\quad -(-0{,}318)\cdot(\text{- 0,772})}{\sqrt{(1-0{,}318^2)\cdot(1\quad \text{- 0,772}^2)}}=\text{ 0,755}$$

$$r_{xz,y}=\frac{r_{xz}-r_{xy}\cdot r_{zy}}{\sqrt{(1-r_{xy}^2)\cdot(1-r_{zy}^2)}}=\frac{-0{,}318-0{,}7\quad \cdot(\text{- 0,772})}{\sqrt{(1-\quad 0{,}7^2)\cdot(1\quad \text{- 0,772}^2)}}=\text{ 0,491}$$

$$r_{zy,x}=\frac{r_{xy}-r_{zx}\cdot r_{yx}}{\sqrt{(1-r_{zx}^2)\cdot(1-r_{yx}^2)}}=\frac{\text{- 0,772}\quad -(-0{,}318)\cdot 0{,}7}{\sqrt{(1-0{,}318^2)\cdot(1-\quad 0{,}7^2)}}=\text{ -0,811}$$

Корреляционный анализ

Множественные коэффициенты корреляции

$$r^2_{x,yz} = \frac{0{,}7^2 + 0{,}318^2 - 2 \cdot 0{,}7 \cdot (-0{,}18) \cdot (-0{,}772)}{1 - (-0{,}772)^2} = 0{,}613 \quad (r_{x,yz} = 0{,}783)$$

$$r^2_{y,xz} = \frac{0{,}7^2 + (-0{,}772)^2 - 2 \cdot 0{,}7 \cdot (-0{,}772) \cdot (-0{,}318)}{1 - 0{,}318^2} = 0{,}826 \quad (r_{y,xz} = 0{,}909)$$

$$r^2_{z,xy} = \frac{0{,}318^2 + (-0{,}772)^2 - 2 \cdot (-0{,}318) \cdot 0{,}7 \cdot (-0{,}772)}{1 - 0{,}7^2} = 0{,}693 \quad (r_{z,xy} = 0{,}833)$$

t-статистики для проверки значимости частных коэффициентов корреляции

$$r_{xy,z}: t_{xy,z} = \frac{\sqrt{10-3} \cdot 0{,}755}{\sqrt{1 - 0{,}755^2}} = 3{,}04$$

$$r_{xz,y}: t_{xz,y} = \frac{\sqrt{7} \cdot 0{,}491}{\sqrt{1 - 0{,}491^2}} = 1{,}48$$

$$r_{zy,x}: t_{zy,x} = \frac{\sqrt{7} \cdot (-0{,}811)}{\sqrt{1 - (-0{,}811)^2}} = -3{,}68$$

Корреляционный анализ

Для α = 0,95 и $f = n - k$ = 7 определяем статистику t-распределения

$$t_{\frac{1+0.95}{2}} = t_{0.975}(t) = 2.37$$

$|t_{xz,y}| < 2,37$, следовательно, наличие частной корреляции отклоняется с достоверностью α = 0,95, $|t_{xy,z}|$, $|t_{zy,x}| > 2,37$ зависимость следует признать значимой.

Для коэффициентов множественной корреляции находим критическое значение из таблицы при k = 3, $n - k$ = 7 и α = 0,95. Имеем $r_{1,23}(0,95)$ =0,758.

Все множественные коэффициенты корреляции (0,783, 0,909 и 0,833) превышают критическое значение 0,758, поэтому наличие множественной корреляции следует признать значимой с достоверностью 0,95.