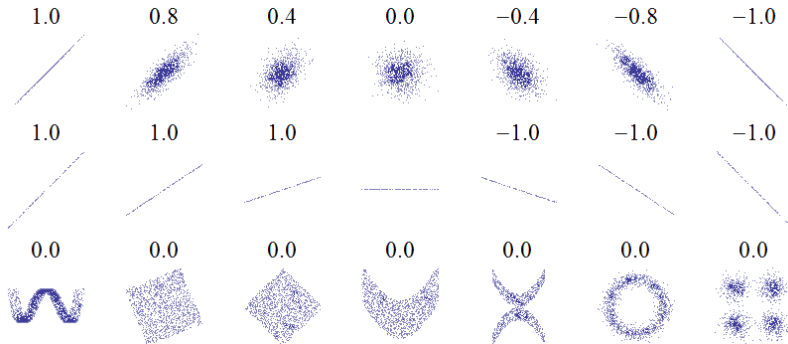


Прикладная статистика. Занятие 6. Корреляционный анализ.

27 марта 2012 г.

Корреляция Пирсона



Корреляции Кендалла и Спирмена

(-0.96; -1)



(-0.74; -0.91)



(-0.33; -0.48)



(0; -0.01)



(0.35; 0.51)



(0.74; 0.91)



(0.96; 1)



(-0.03; -0.05)



(-0.91; -0.99)



(-0.9; -0.99)



(0.06; 0.08)



(0.91; 0.99)



(0.91; 0.99)



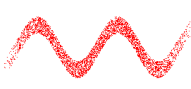
(0.09; 0.13)



(0; 0.01)



(-0.18; -0.27)



(0.01; 0.02)



(-0.01; -0.03)



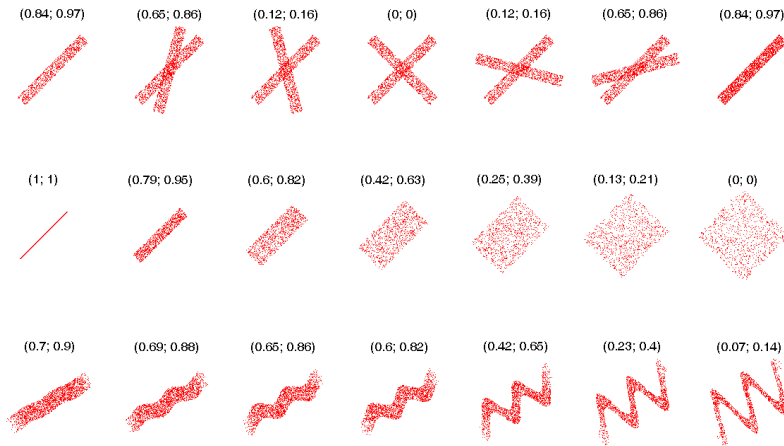
(0; 0)



(-0.03; -0.04)



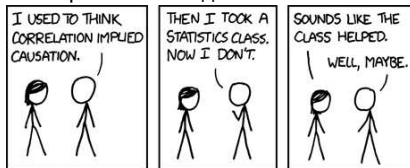
Корреляции Кендалла и Спирмена



Корреляция и причинность

Корреляция — мера ассоциативной связи (одновременная встречаемость событий, сходство паттернов).

Никакого отношения к причинно-следственной связи она не имеет!



Пример: сильная положительная корреляция между объёмом продаж мороженого на пляже и числом утонувших на нём людей.



Статистика не занимается и не имеет средств для того, чтобы заниматься причинно-следственными связями.

Корреляция и регрессия

$$\rho^2(x, y) = \sigma_{\hat{y}}^2 / \sigma_y^2$$

$$\hat{y} = a_{xy}x + b_{xy}, \quad \hat{x} = a_{yx}y + b_{yx} \Rightarrow \rho(x, y) = \sqrt{a_{xy}a_{yx}}$$

Случай бинарных переменных

		X		Σ
		0	1	
Y	0	p_{00}	p_{01}	$p_{0\cdot}$
	1	p_{10}	p_{11}	$p_{1\cdot}$
Σ		$p_{\cdot 0}$	$p_{\cdot 1}$	1

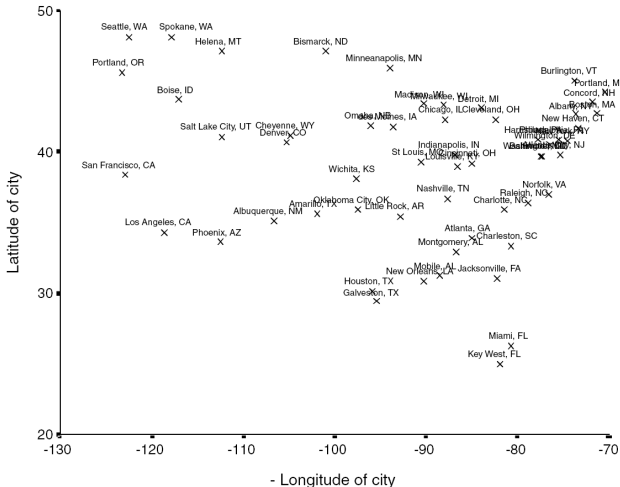
Коэффициент фи (коэффициент среднеквадратичной сопряжённости, в машинном обучении — коэффициент корреляции Мэтьюса):

$$\phi = \frac{p_{11}p_{00} - p_{10}p_{01}}{\sqrt{p_{\cdot 1}p_{\cdot 0}p_{0\cdot}p_{1\cdot}}}$$

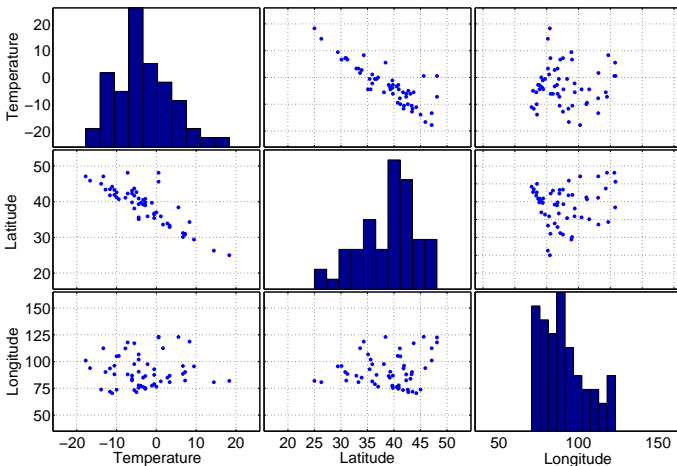
$$\phi^2 = \frac{\chi^2}{n}$$

Температура воздуха и географическое положение

По 56 городам США известны средняя минимальная температура января и географические координаты (широта, долгота). Требуется исследовать характер зависимости между переменными.



Температура воздуха и географическое положение



Температура воздуха и географическое положение

T — температура, λ — долгота, ϕ — широта;
 r — корреляция Пирсона, τ — Кендалла, ρ — Спирмена.

Коэффициенты корреляции:

τ	T	ϕ	λ
T	1.000	-0.848	0.024
ϕ	-0.848	1.000	0.145
λ	0.024	0.145	1.000

τ	T	ϕ	λ
T	1.000	-0.683	0.030
ϕ	-0.683	1.000	-0.011
λ	0.030	-0.011	1.000

ρ	T	ϕ	λ
T	1.000	-0.815	0.030
ϕ	-0.815	1.000	0.023
λ	0.030	0.023	1.000

Достигаемые уровни значимости:

τ	T	ϕ	λ
T	0.000	0.000	0.861
ϕ	0.000	0.000	0.287
λ	0.861	0.287	0.000

τ	T	ϕ	λ
T	0.000	0.000	0.756
ϕ	0.000	0.000	0.910
λ	0.756	0.910	0.000

ρ	T	ϕ	λ
T	0.000	0.000	0.829
ϕ	0.000	0.000	0.865
λ	0.829	0.865	0.000

Температура воздуха и географическое положение

T — температура, λ — долгота, ϕ — широта;
 r — частная корреляция Пирсона, ρ — Спирмена.

Коэффициенты частной корреляции:

r	T	ϕ	λ
T	1.000	-0.861	0.280
ϕ	-0.861	1.000	0.312
λ	0.280	0.312	1.000

ρ	T	ϕ	λ
T	1.000	-0.817	0.084
ϕ	-0.817	1.000	0.082
λ	0.084	0.082	1.000

Достигаемые уровни значимости:

r	T	ϕ	λ
T	0.000	0.000	0.039
ϕ	0.000	0.000	0.021
λ	0.039	0.021	0.000

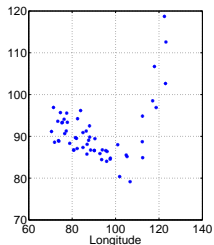
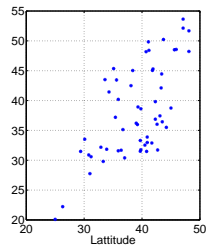
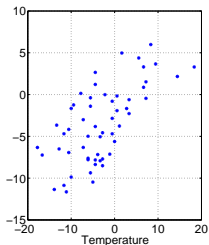
ρ	T	ϕ	λ
T	0.000	0.000	0.543
ϕ	0.000	0.000	0.552
λ	0.543	0.552	0.000

Температура воздуха и географическое положение

T — температура, λ — долгота, ϕ — широта;
 R — множественная корреляция.

Коэффициенты множественной корреляции:

	T	ϕ	λ
R	0.659	0.667	0.312
with	$0.235 \cdot \lambda - 0.638 \cdot \phi$	$0.397 \cdot \lambda - 0.678 \cdot T$	$1.542 \cdot T + 2.450 \cdot \phi$

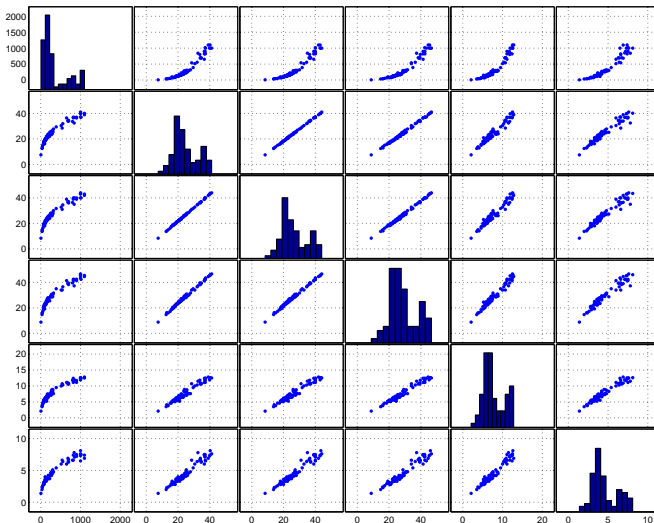


Вес и линейные размеры рыб

В 1917 году в финском озере Langelmavesi исследователи поймали и измерили 81 рыбу трёх схожих видов. Известны: вес, длина от носа до начала хвоста, длина от носа до развилки хвоста, длина от носа до кончика хвоста, наибольшая высота, наибольшая толщина. Исследовать взаимосвязи между переменными.



Вес и линейные размеры рыб



Вес и линейные размеры рыб

Попарные корреляции Пирсона:

ρ	Weight	Length1	Length2	Length3	Width	Thickness
Weight	1.0000	0.9575	0.9581	0.9545	0.9527	0.9584
Length1	0.9575	1.0000	0.9996	0.9974	0.9753	0.9718
Length2	0.9581	0.9996	1.0000	0.9973	0.9754	0.9724
Length3	0.9545	0.9974	0.9973	1.0000	0.9822	0.9707
Width	0.9527	0.9753	0.9754	0.9822	1.0000	0.9734
Thickness	0.9584	0.9718	0.9724	0.9707	0.9734	1.0000

Наибольший достигаемый уровень значимости: $p = 2.8772 \times 10^{-43}$.

Вес и линейные размеры рыб

Попарные корреляции Кендалла:

ρ	Weight	Length1	Length2	Length3	Width	Thickness
Weight	1.0000	0.9125	0.9162	0.9177	0.8805	0.8827
Length1	0.9125	1.0000	0.9834	0.9609	0.8467	0.8558
Length2	0.9162	0.9834	1.0000	0.9520	0.8444	0.8631
Length3	0.9177	0.9609	0.9520	1.0000	0.8720	0.8540
Width	0.8805	0.8467	0.8444	0.8720	1.0000	0.8223
Thickness	0.8827	0.8558	0.8631	0.8540	0.8223	1.0000

Наибольший достигаемый уровень значимости: $p = 1.3078 \times 10^{-26}$.

Вес и линейные размеры рыб

Попарные корреляции Спирмена:

ρ	Weight	Length1	Length2	Length3	Width	Thickness
Weight	1.0000	0.9871	0.9864	0.9881	0.9738	0.9715
Length1	0.9871	1.0000	0.9984	0.9955	0.9627	0.9638
Length2	0.9864	0.9984	1.0000	0.9933	0.9594	0.9655
Length3	0.9881	0.9955	0.9933	1.0000	0.9719	0.9637
Width	0.9738	0.9627	0.9594	0.9719	1.0000	0.9440
Thickness	0.9715	0.9638	0.9655	0.9637	0.9440	1.0000

Наибольший достигаемый уровень значимости: $p = 8.4691 \times 10^{-40}$.

Вес и линейные размеры рыб

Частные корреляции Пирсона:

ρ	Weight	Length1	Length2	Length3	Width	Thickness
Weight	1.0000	0.0458	0.0880	-0.2054	0.2404	0.2372
Length1	0.0458	1.0000	0.8847	0.2814	-0.1484	0.0318
Length2	0.0880	0.8847	1.0000	0.1743	-0.0958	0.1314
Length3	-0.2054	0.2814	0.1743	1.0000	0.6557	-0.2526
Width	0.2404	-0.1484	-0.0958	0.6557	1.0000	0.4690
Thickness	0.2372	0.0318	0.1314	-0.2526	0.4690	1.0000

Достигаемые уровни значимости:

p-value	Weight	Length1	Length2	Length3	Width	Thickness
Weight	0	0.6926	0.4466	0.0731	0.0352	0.0378
Length1	0.6926	0	0.0000	0.0132	0.1976	0.7840
Length2	0.4466	0.0000	0	0.1294	0.4070	0.2546
Length3	0.0731	0.0132	0.1294	0	0.0000	0.0266
Width	0.0352	0.1976	0.4070	0.0000	0	0.0000
Thickness	0.0378	0.7840	0.2546	0.0266	0.0000	0

Вес и линейные размеры рыб

Частные корреляции Спирмена:

ρ	Weight	Length1	Length2	Length3	Width	Thickness
Weight	1.0000	0.0626	0.1041	0.0956	0.4291	0.3943
Length1	0.0626	1.0000	0.8515	0.5177	-0.0510	-0.1310
Length2	0.1041	0.8515	1.0000	-0.0940	-0.1433	0.1887
Length3	0.0956	0.5177	-0.0940	1.0000	0.4051	0.0327
Width	0.4291	-0.0510	-0.1433	0.4051	1.0000	-0.0165
Thickness	0.3943	-0.1310	0.1887	0.0327	-0.0165	1.0000

Достигаемые уровни значимости:

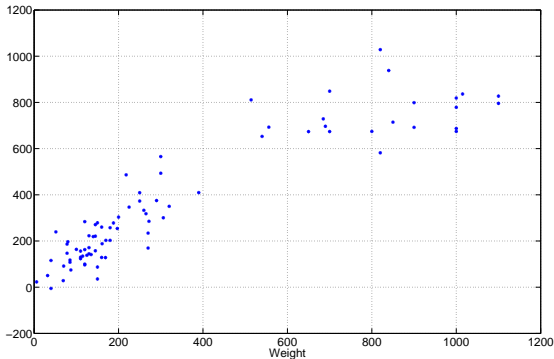
p-value	Weight	Length1	Length2	Length3	Width	Thickness
Weight	0	0.5885	0.3674	0.4083	0.0001	0.0004
Length1	0.5885	0	0.0000	0.0000	0.6598	0.2562
Length2	0.3674	0.0000	0	0.4163	0.2139	0.1003
Length3	0.4083	0.0000	0.4163	0	0.0003	0.7774
Width	0.0001	0.6598	0.2139	0.0003	0	0.8869
Thickness	0.0004	0.2562	0.1003	0.7774	0.8869	0

Вес и линейные размеры рыб

Множественная корреляция всех признаков с весом: $R = 0.9207$.

Максимизирующая корреляцию линейная комбинация:

$$292.2458 \cdot Length1 - 151.1554 \cdot Length2 - 151.0027 \cdot Length3 + 148.8896 \cdot Width + 83.4345 \cdot Thickness.$$



Прикладная статистика
Семинар 6. Корреляционный анализ.

Рябенко Евгений
riabenko.e@gmail.com