Прикладная статистика 6. Анализ зависимостей.

11 октября 2013 г.

Задача исследования взаимосвязи между признаками

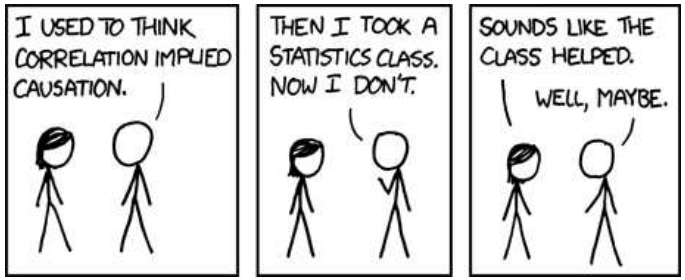
Дано: значения признаков X_1, X_2 измерены на объектах $1, \ldots, n$.
Эквивалентная формулировка: имеются связные выборки $X_1^n = (X_{11}, \ldots, X_{1n})$ и $X_2^n = (X_{21}, \ldots, X_{2n})$.

Насколько сильно признаки X_1, X_2 связаны между собой?

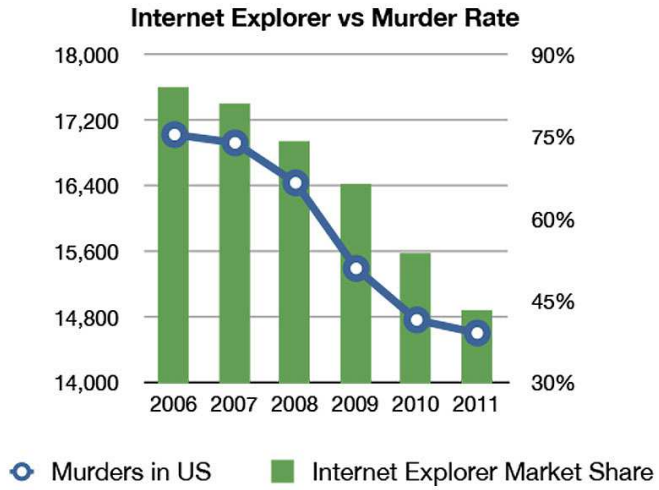
Статистическая взаимосвязь между случайными величинами — **корреляция**.

Корреляция и причинность

Корреляция — мера ассоциативной связи (одновременная встречаемость событий, сходство паттернов).
Никакого отношения к причинно-следственной связи она не имеет!



Пример:



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

Корреляция Пирсона

Корреляция Пирсона (Pearson product-moment correlation coefficient):

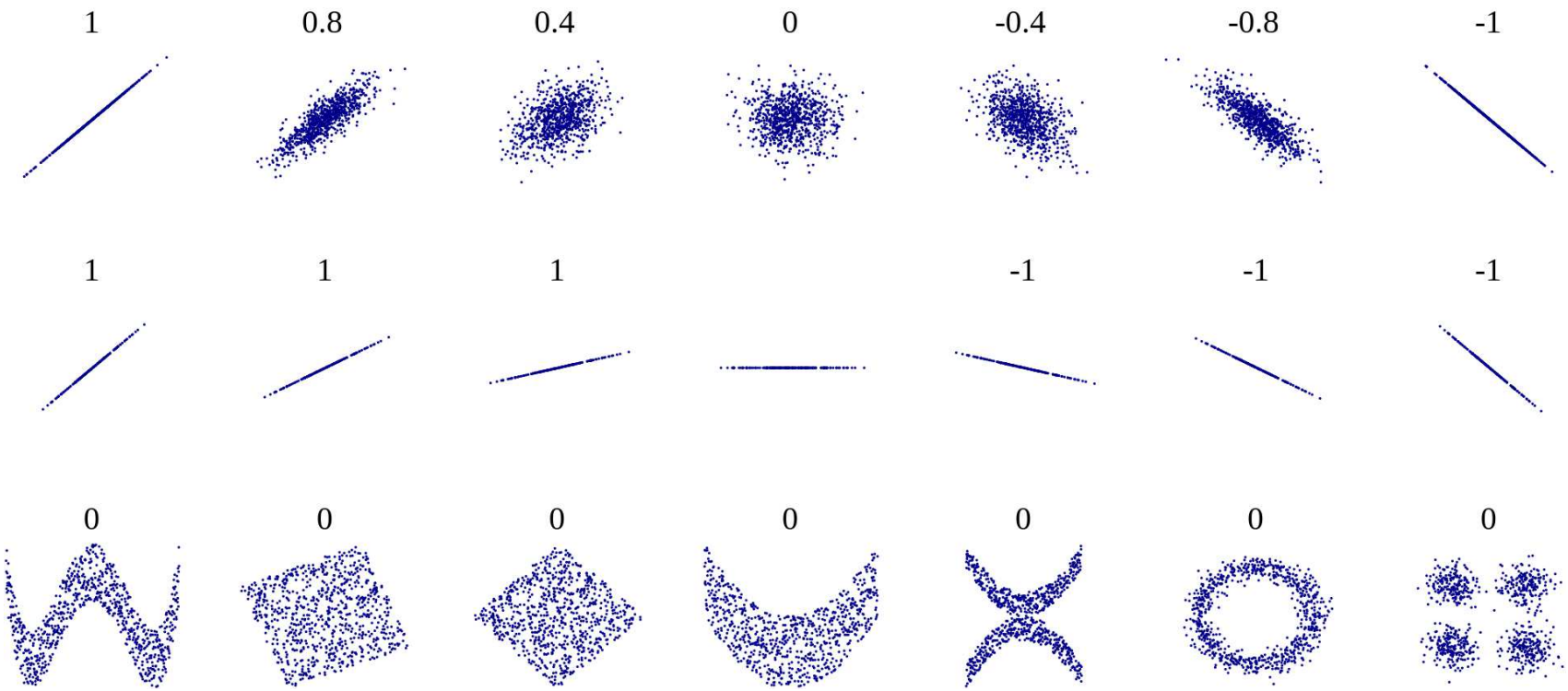
$$r_{XY} = \frac{\operatorname{cov}(X_1, X_2)}{\sqrt{\mathbb{D}X_1 \mathbb{D}X_2}} = \frac{\mathbb{E}\left((X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)\right)}{\sqrt{\mathbb{D}X_1 \mathbb{D}X_2}}.$$

Выборочный коэффициент корреляции Пирсона:

$$r_{X_1X_2} = \frac{\sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)\left(X_{2i} - \bar{X}_2\right)}{\sqrt{\sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)^2 \sum_{i=1}^{n} \left(X_{2i} - \bar{X}_2\right)^2}}.$$

$r_{X_1X_2} \in [-1, 1]$ — мера **линейной** связи.

Корреляция Пирсона



Критерий Стьюдента

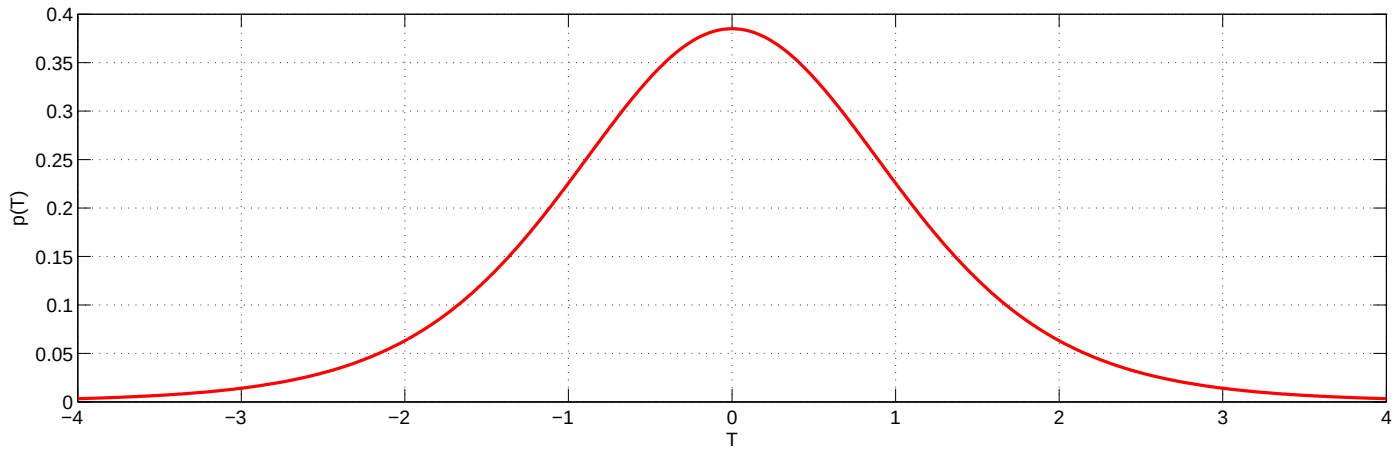
выборки: $X_1^n = (X_{11}, \ldots, X_{1n})$,
$X_2^n = (X_{21}, \ldots, X_{2n})$, выборки связные,
$(X_{1i}, X_{2i}) \sim N(\mu, \Sigma)$;

нулевая гипотеза: $H_0 : r_{X_1X_2} = 0$;

альтернатива: $H_1 : r_{X_1X_2} <\neq> 0$;

статистика: $T(X_1^n, X_2^n) = \frac{r_{X_1X_2}\sqrt{n-2}}{\sqrt{1-r_{X_1X_2}^2}}$;

$T(X_1^n, X_2^n) \sim St(n-2)$ при H_0;



достигаемый уровень значимости:

$$p(t) = \begin{cases} 1 - tcdf(t, n-2), & H_1 : r_{X_1X_2} > 0, \\ tcdf(t, n-2), & H_1 : r_{X_1X_2} < 0, \\ 2\left(1 - tcdf(|t|, n-2)\right), & H_1 : r_{X_1X_2} \neq 0. \end{cases}$$

Критерий Стьюдента

Доверительный интервал для коэффициента корреляции Пирсона:

$$\left[r_{X_1X_2} + \frac{t_{n-2,\alpha/2}\left(1 - r^2_{X_1X_2}\right)}{\sqrt{n}}, r_{X_1X_2} - \frac{t_{n-2,\alpha/2}\left(1 - r^2_{X_1X_2}\right)}{\sqrt{n}}\right].$$

С использованием преобразования Фишера:

$$\left[\operatorname{tanh}\left(\operatorname{arctanh} r_{X_1X_2} + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right), \operatorname{tanh}\left(\operatorname{arctanh} r_{X_1X_2} - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right)\right].$$

Критерий Стьюдента

Пример: для двух марок зубной пасты, одна из которых рекламируется по телевизору, а другая нет, участники опроса (30 человек) выставляют оценки в баллах от 1 до 20 в соответствии со своими предпочтениями. Коэффициент корреляции Пирсона между оценками двух марок составляет 0.32, значимо ли эта величина отличается от нуля?

$H_0 : r_{X_1X_2} = 0.$
$H_1 : r_{X_1X_2} \neq 0 \Rightarrow p = 0.0847.$

Доверительный интервал: $[-0.0157, 0.6557]$.
С использованием преобразования Фишера: $[-0.0455, 0.6100]$.

Перестановочный критерий

выборки: $X_1^n = (X_{11}, \ldots, X_{1n})$,
$X_2^n = (X_{21}, \ldots, X_{2n})$, выборки связные;
нулевая гипотеза: $H_0 \colon r_{X_1X_2} = 0$;
альтернатива: $H_1 \colon r_{X_1X_2} <\neq> 0$;
статистика: $T(X_1^n, X_2^n) = r_{X_1X_2}$.

Нулевое распределение $T(X_1^n, X_2^n)$ порождается группой перестановок

$$G = \{g \colon gX_2^n = (X_{2\pi_1}, \ldots, X_{2\pi_n})\},$$

где $\pi_1, \ldots, \pi_n$ — перестановка индексов $1, \ldots, n$;

$$|G| = n!$$

Достигаемый уровень значимости:

$$p(t) = \begin{cases} \dfrac{\sum\limits_{g \in G} \left[T\left(X_1^n, gX_2^n\right) \leq \geq T\left(X_1^n, X_2^n\right)\right]}{n!}, & H_1 \colon r_{X_1X_2} <> 0, \\ \dfrac{\sum\limits_{g \in G} \left[\left|T\left(X_1^n, gX_2^n\right)\right| \geq \left|T\left(X_1^n, X_2^n\right)\right|\right]}{n!}, & H_1 \colon r_{X_1X_2} \neq 0. \end{cases}$$

Перестановочный критерий

Перестановочный $100(1-\alpha)$-% доверительный интервал для коэффициента корреляции образован выборочными квантилями порядка $\alpha/2$ и $1-\alpha/2$ перестановочного распределения $T\left(X_1^n, gX_2^n\right)$.

Пример: в предыдущем примере

$H_0: r_{X_1X_2} = 0.$
$H_1: r_{X_1X_2} \neq 0 \Rightarrow p = 0.0564.$

Недостатки

Недостатки выборочного коэффициента Пирсона:

- служит мерой только линейной взаимосвязи;
- неустойчив к выбросам;
- для распределений, отличных от нормального, выборочный коэффициент корреляции перестаёт быть эффективной оценкой популяционного.

Корреляция Спирмена

Коэффициент корреляции Спирмена — коэффициент корреляции Пирсона рангов наблюдений в выборках X_1^n, X_2^n :

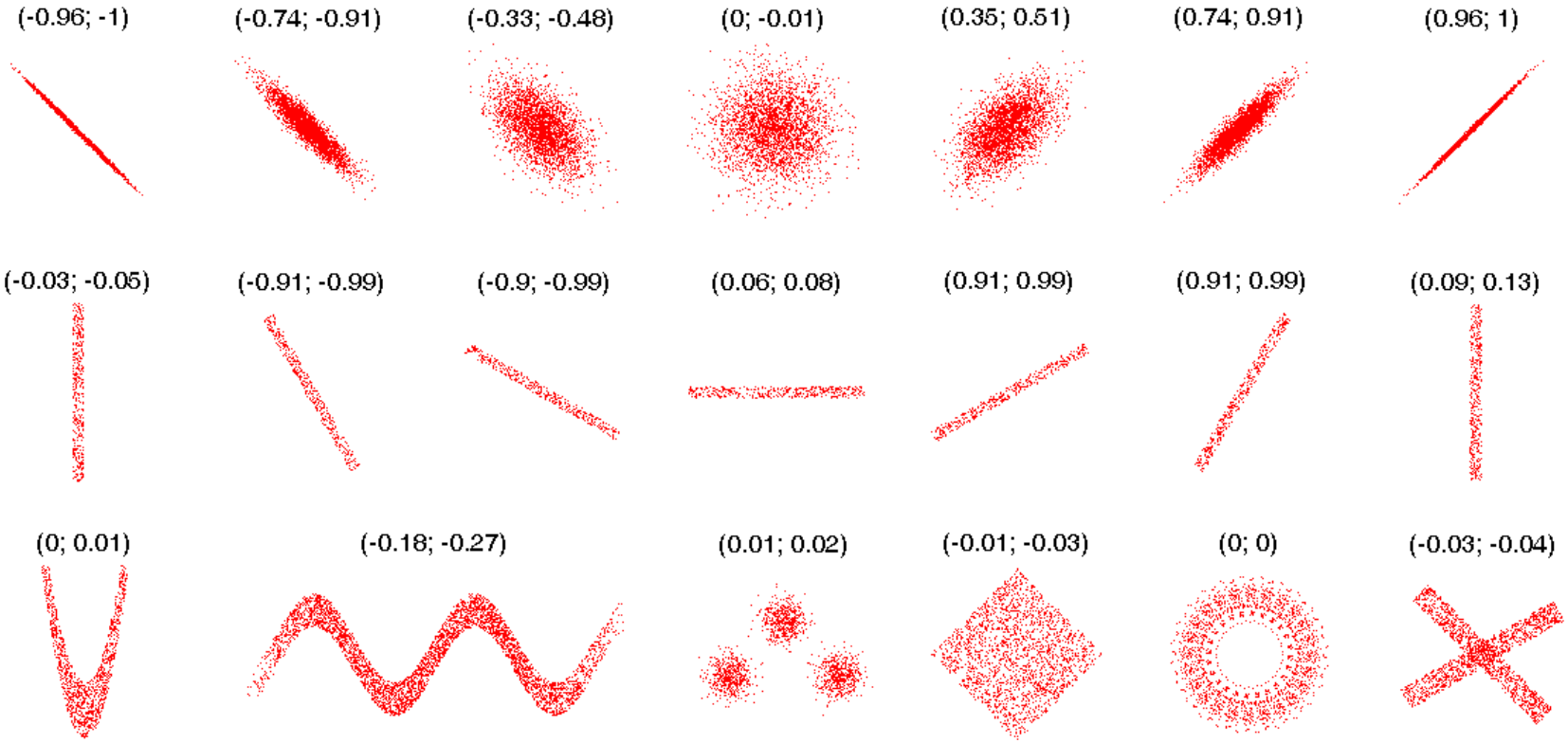
$$\rho_{X_1X_2} = \frac{\sum\limits_{i=1}^{n} \left(r\left(X_{1i}\right) - \frac{n+1}{2}\right)\left(r\left(X_{2i}\right) - \frac{n+1}{2}\right)}{\frac{1}{12}\left(n^3 - n\right)} =$$

$$= 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}\left(r\left(X_{1i}\right) - r\left(X_{2i}\right)\right)^2,$$

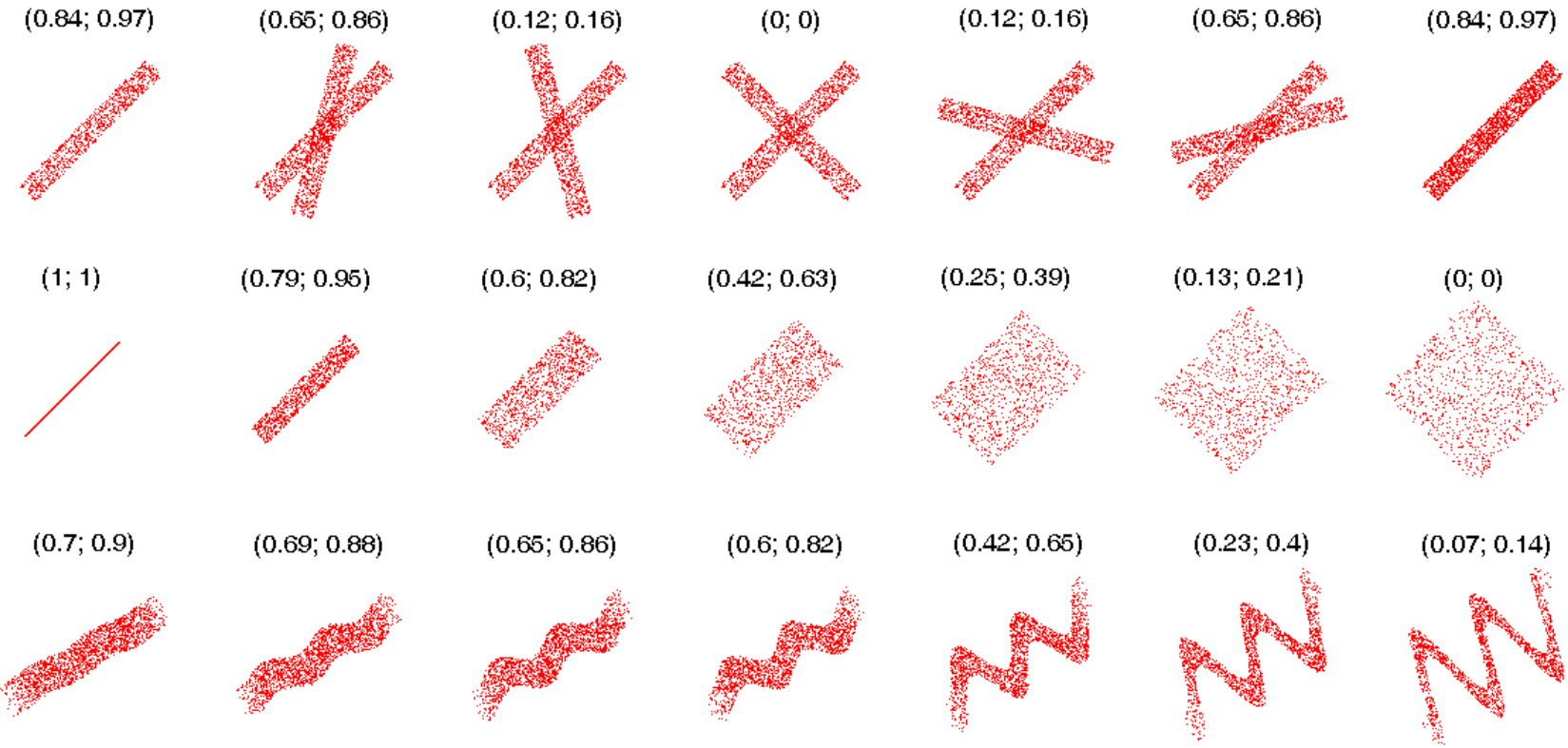
где $r\left(X_{1i}\right), r\left(X_{2i}\right)$ — ранги i-х наблюдений в соответствующих выборках.

$\rho_{X_1X_2} \in [-1, 1]$ — мера **монотонной** связи.

Корреляция Спирмена



Корреляция Спирмена



Критерий Стьюдента

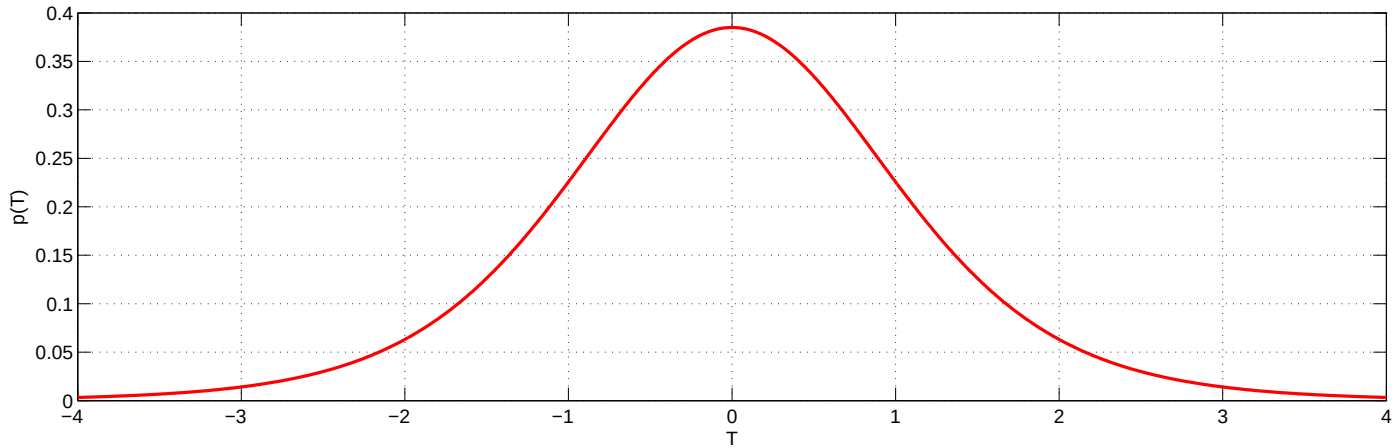
выборки: $X_1^n = (X_{11}, \ldots, X_{1n})$,
$X_2^n = (X_{21}, \ldots, X_{2n})$, выборки связные;

нулевая гипотеза: $H_0 \colon \rho_{X_1 X_2} = 0$;

альтернатива: $H_1 \colon \rho_{X_1 X_2} <\neq> 0$;

статистика: $T(X_1^n, X_2^n) = \frac{\rho_{X_1 X_2}\sqrt{n-2}}{\sqrt{1-\rho^2_{X_1 X_2}}}$;

$T(X_1^n, X_2^n) \sim St(n-2)$ при H_0;



достигаемый уровень значимости:

$$p(t) = \begin{cases} 1 - tcdf(t, n-2), & H_1 \colon \rho_{X_1 X_2} > 0, \\ tcdf(t, n-2), & H_1 \colon \rho_{X_1 X_2} < 0, \\ 2\left(1 - tcdf(|t|, n-2)\right), & H_1 \colon \rho_{X_1 X_2} \neq 0. \end{cases}$$

Критерий Стьюдента

Пример: выборка из 11 потребителей вегетариантских сосисок оценивает качество двух брендов. Если целевая аудитория двух брендов совпадает, то их рекламу можно давать совместно. Корреляция Спирмена оценок потребителей равна -0.854

$H_0 : \rho_{X_1 X_2} = 0.$
$H_1 : \rho_{X_1 X_2} \neq 0 \Rightarrow p = 0.0024.$

Корреляция Кендалла

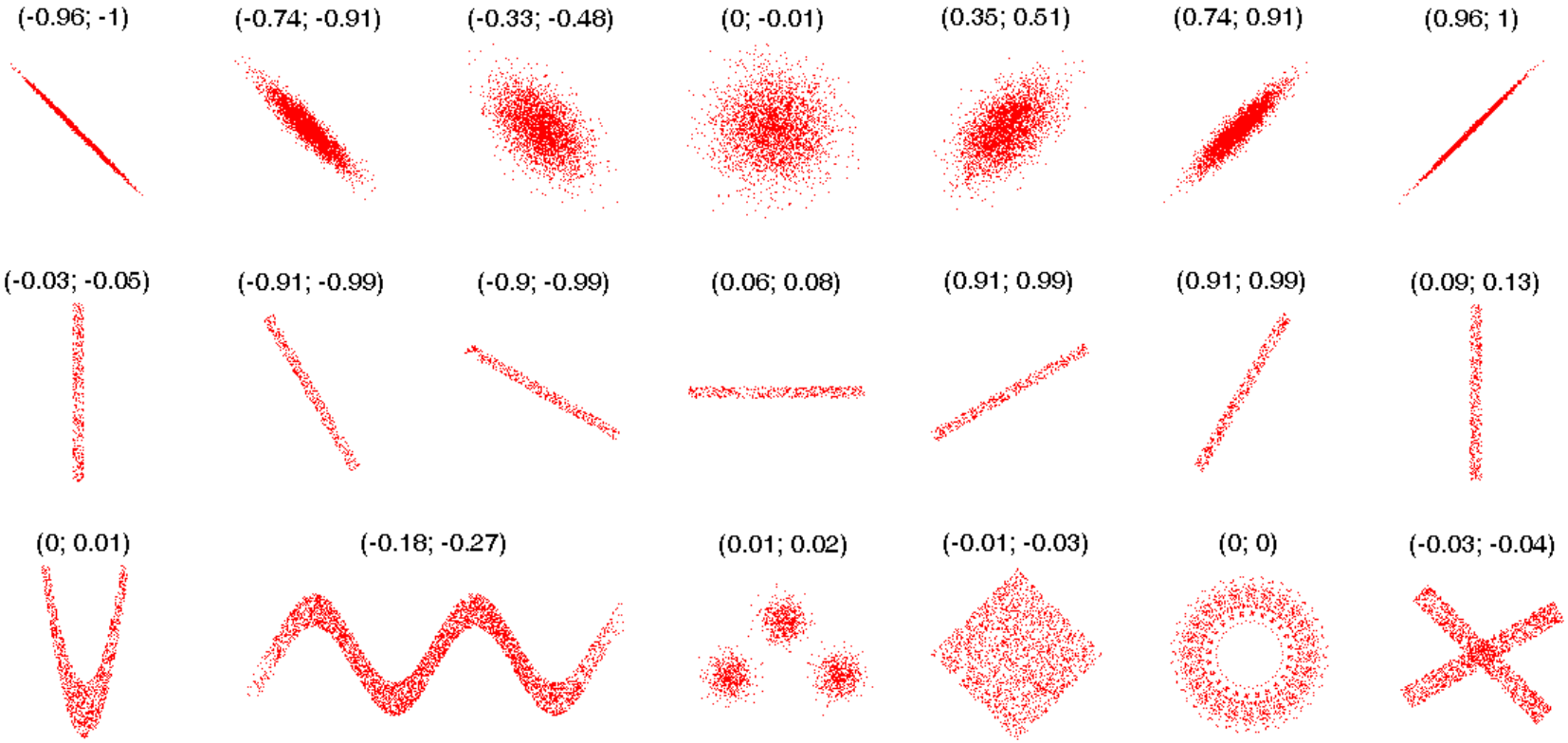
Коэффициент корреляции Кендалла — мера взаимной неупорядоченности X_1^n и X_2^n:

$$\tau_{X_1 X_2} = 1 - \frac{4}{n\,(n-1)} \sum_{i=1}^{n-1} \sum_{j=1}^{n} [[X_{1i} < X_{1j}] \neq [X_{2i} < X_{2j}]] = \frac{C-D}{C+D},$$

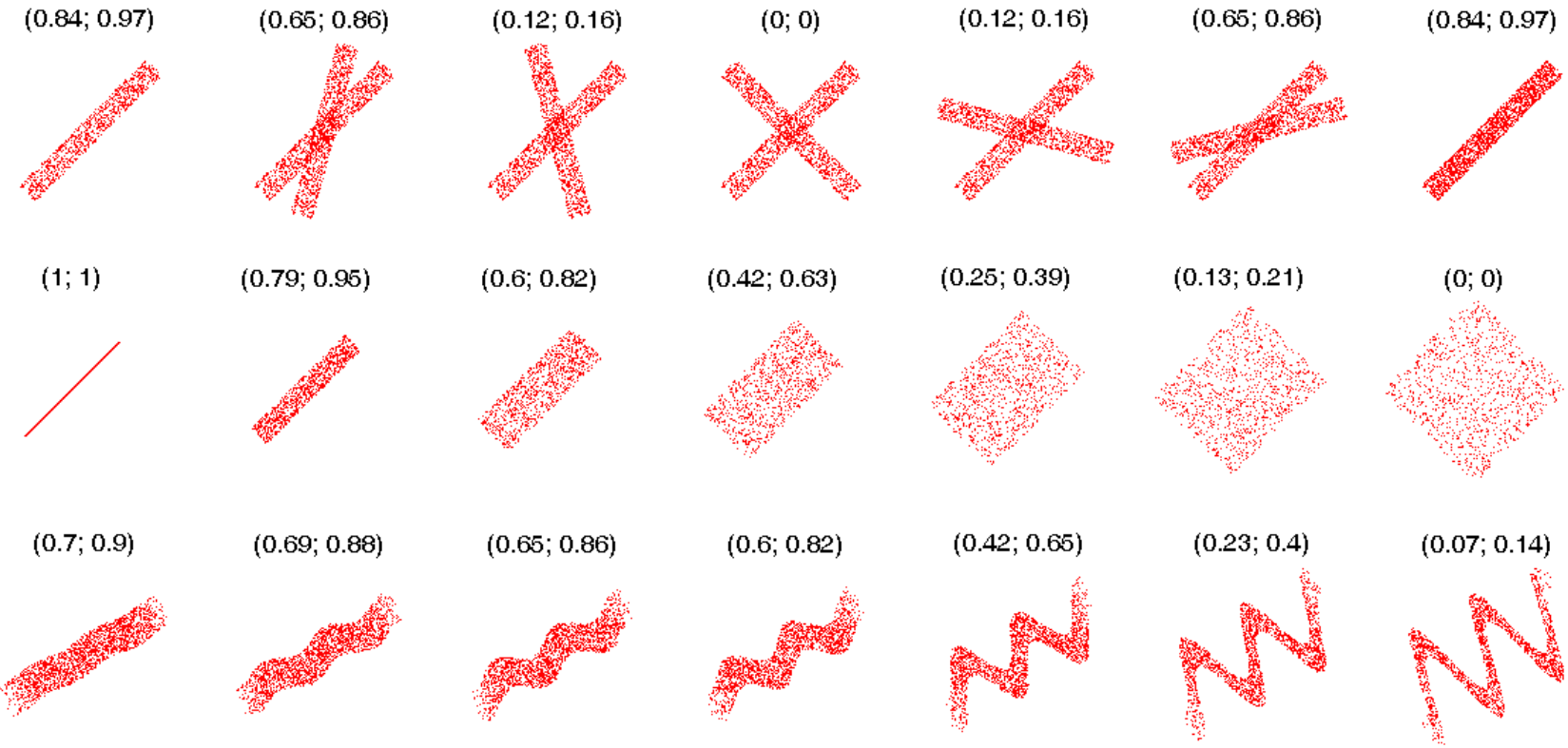
где C — число согласованных пар, D — число несогласованных пар.

$\tau_{X_1 X_2} \in [-1, 1]$ — мера **монотонной** связи.

Корреляция Кендалла



Корреляция Кендалла



Критерий без названия

выборки: $X_1^n = (X_{11}, \ldots, X_{1n})$,
$X_2^n = (X_{21}, \ldots, X_{2n})$, выборки связные;
нулевая гипотеза: $H_0 : \tau_{X_1 X_2} = 0$;
альтернатива: $H_1 : \tau_{X_1 X_2} <\neq> 0$;
статистика: $\tau_{X_1 X_2}$ имеет табличное распределение при H_0.

При справедливости H_0

$$\mathbb{E}\tau_{X_1 X_2} = 0, \quad \mathbb{D}\tau_{X_1 X_2} = \frac{2\,(2n+5)}{9n\,(n-1)}.$$

Для $n > 10$ справедлива аппроксимация нормальным распределением.

Критерий без названия

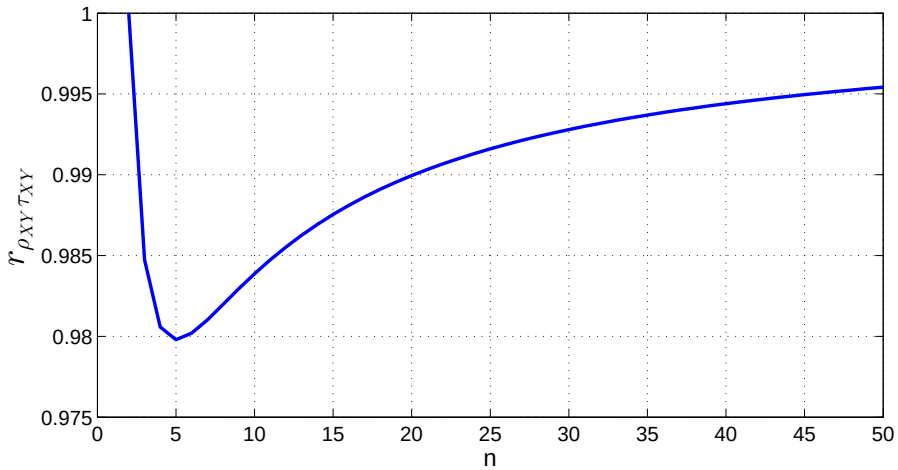
Пример: налоговый инспектор хочет проверить наличие взаимосвязи между величинами общего дохода от инвестиций и общего объёма дополнительных доходов. На выборке из 10 налоговых деклараций он получил $D = 5,\ S = 38,\ \tau_{X_1X_2} = 0.7821.$

$H_0 : \tau_{X_1X_2} = 0.$
$H_1 : \tau_{X_1X_2} \neq 0 \Rightarrow p = 0.0027.$

Связь между коэффициентами корреляции

При справедливости H_0 (отсутствии монотонной зависимости):

$$r_{\rho_{X_1X_2}\tau_{X_1X_2}} = \frac{2n+2}{\sqrt{4n^2+10n}}.$$



Кендалла vs. Спирмена: `http://youtu.be/D56dvoVrBBE`
Корреляция Кендалла:

- менее чувствительна к большим различиям между рангами наблюдений;
- точнее оценивается по выборке небольших объёмов;
- обычно меньше по модулю, чем корреляция Спирмена.

Связь между коэффициентами корреляции

Если $(X_{1i}, X_{2i}) \sim N(\mu, \Sigma)$, то

$$\lim_{n\to\infty} \mathbb{E}\tau_{X_1X_2} = \lim_{n\to\infty} \mathbb{E}\rho_{X_1X_2} = \frac{2}{\pi} \arcsin r_{X_1X_2}.$$

Частная корреляция

Если мы подозреваем, что наблюдаемая линейная взаимосвязь между признаками X_1 и X_2 вызвана влиянием третьего признака X_3, можно попытаться его снять.
Частная корреляция:

$$r_{X_1X_2|X_3} = \frac{r_{X_1X_2} - r_{X_1X_3}r_{X_2X_3}}{\sqrt{\left(1 - r^2_{X_1X_3}\right)\left(1 - r^2_{X_2X_3}\right)}}.$$

Если нужно снять влияние нескольких признаков, можно пользоваться рекуррентной формулой:

$$r_{X_1X_2|X_3X_4} = \frac{r_{X_1X_2|X_4} - r_{X_1X_3|X_4}r_{X_2X_3|X_4}}{\sqrt{\left(1 - r^2_{X_1X_3|X_4}\right)\left(1 - r^2_{X_2X_3|X_4}\right)}}.$$

Другой вариант: если M — множество признаков, Ω — обратимая матрица их корреляций, $R = \Omega^{-1}$, то

$$r_{X_iX_j|M\setminus\{X_i,X_j\}} = -\frac{r_{ij}}{\sqrt{r_{ii}r_{jj}}}.$$

Критерий Стьюдента

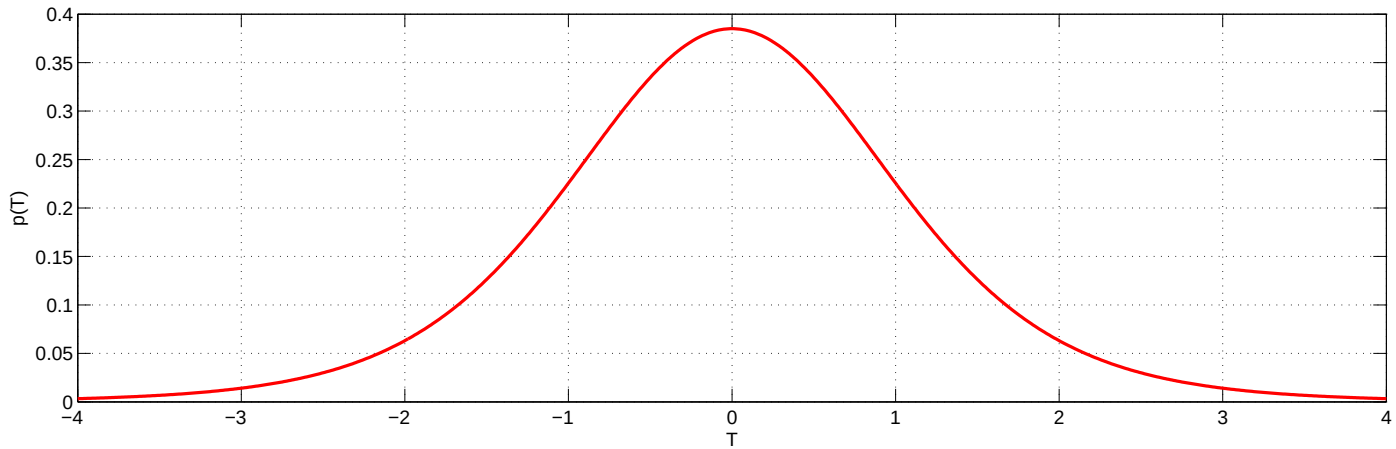
выборки: $X_1^n = (X_{11}, \ldots, X_{1n}), X_2^n = (X_{21}, \ldots, X_{2n}),$
$X_3^n = (X_{31}, \ldots, X_{3n}), X_{3i} \in \mathbb{R}^M,$
$(X_{1i}, X_{2i}, X_{3i}) \sim N(\mu, \Sigma);$

нулевая гипотеза: $H_0 : r_{X_1 X_2 | X_3} = 0;$

альтернатива: $H_1 : r_{X_1 X_2 | X_3} <\neq> 0;$

статистика: $T(X_1^n, X_2^n, X_3^n) = \frac{r_{X_1 X_2 | X_3} \sqrt{n-M-2}}{\sqrt{1 - r^2_{X_1 X_2 | X_3}}};$

$T(X_1^n, X_2^n, X_3^n) \sim St(n - M - 2)$ при $H_0;$



достигаемый уровень значимости:

$$p(t) = \begin{cases} 1 - tcdf(t, n - M - 2), & H_1 : r_{X_2 X_2 | X_3} > 0, \\ tcdf(t, n - M - 2), & H_1 : r_{X_2 X_2 | X_3} < 0, \\ 2\left(1 - tcdf(|t|, n - M - 2)\right), & H_1 : r_{X_2 X_2 | X_3} \neq 0. \end{cases}$$

Множественная корреляция

Для того, чтобы оценить силу линейной взаимосвязи одной переменной (X_1) с несколькими другими (X_2, X_3), используется множественная корреляция:

$$r_{X_1,X_2,X_3} = \frac{r^2_{X_1X_2} + r^2_{X_1X_3} - 2r_{X_1X_2}r_{X_1X_3}r_{X_2X_3}}{1 - r^2_{X_2X_3}}.$$

Для большего числа признаков: пусть M — множество дополнительных признаков, Ω — обратимая матрица их корреляций, $R = \Omega^{-1}$, c — вектор корреляций основного признака X с дополнительными; тогда

$$r^2_{X,M} = c^T Rc.$$

Фактически, находится такая линейная комбинация признаков из M, что корреляция X с ней максимальна.

$r_{X,M} \in [0, 1]$.

Критерий Фишера

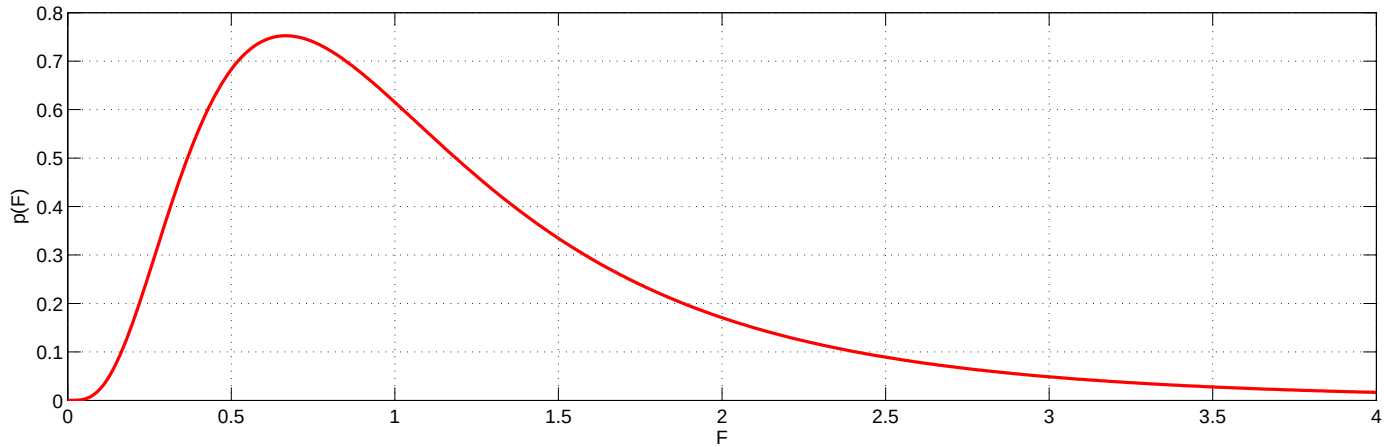
выборки: $X_1^n = (X_{11}, \ldots, X_{1n})$,
$X_2^n = (X_{21}, \ldots, X_{2n}), X_{2i} \in \mathbb{R}^M$,
$(X_{1i}, X_{2i}) \sim N(\mu, \Sigma)$;

нулевая гипотеза: $H_0 : r_{X_1,X_2} = 0$;

альтернатива: $H_1 : r_{X_1,X_2} > 0$;

статистика: $F(X_1^n, X_2^n) = \frac{r^2_{X_1,X_2}}{1-r^2_{X_1,X_2}} \frac{n-M-1}{M-2}$;

$F(X_1^n, X_2^n) \sim F(M-2, n-M-1)$ при H_0;

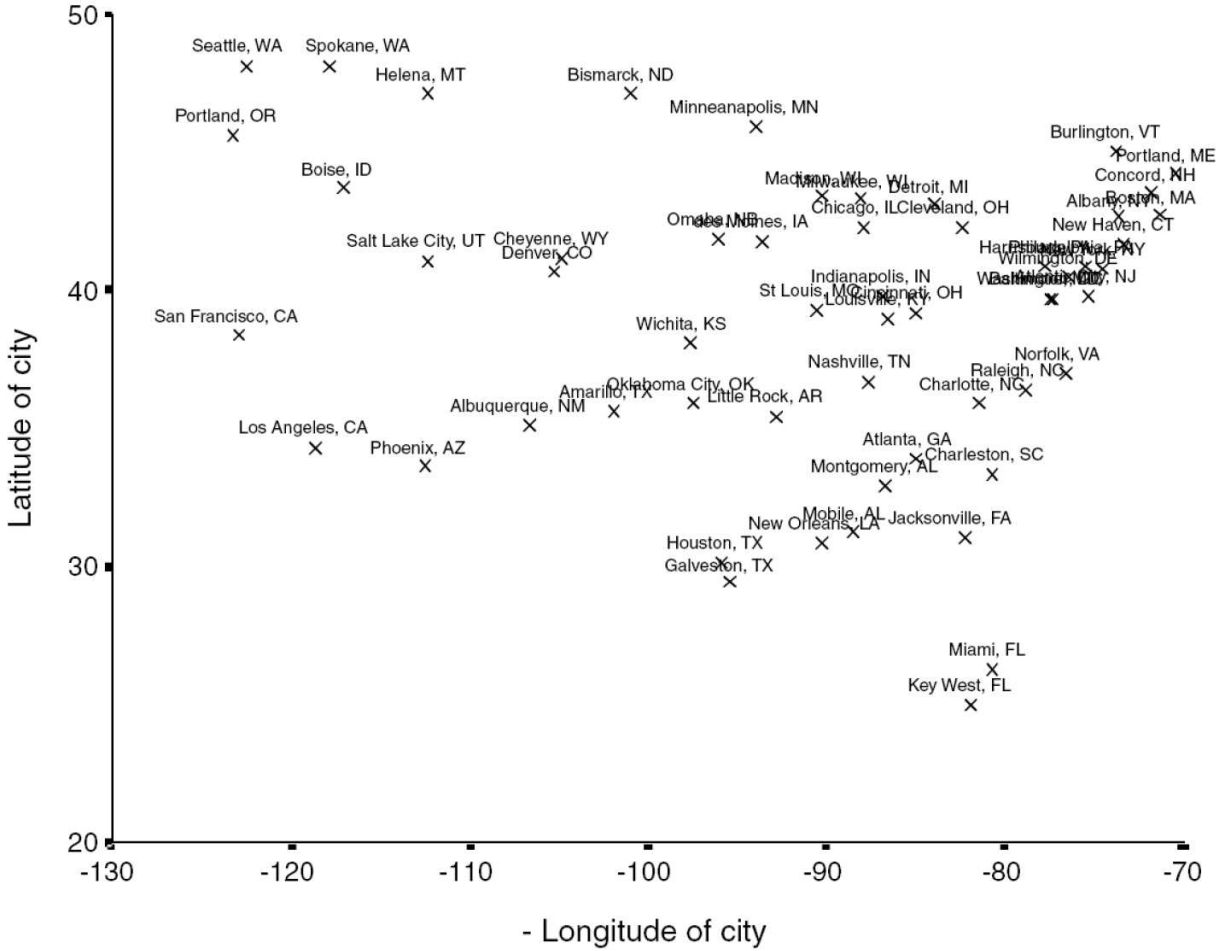


достигаемый уровень значимости:

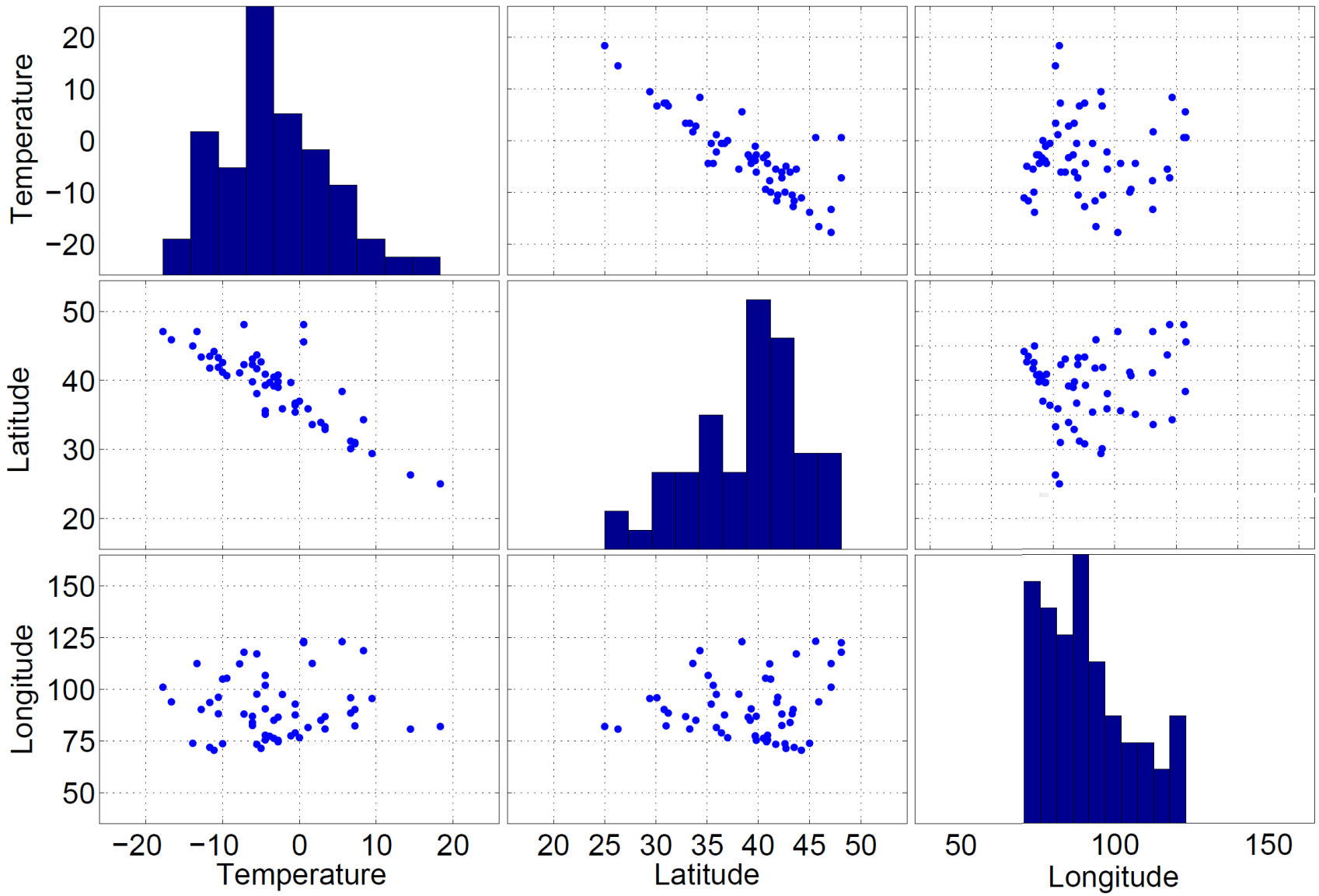
$$p(f) = 1 - fcdf(f, M-2, n-M-1).$$

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, ρ — Спирмена, τ — Кендалла.

Коэффициенты корреляции:

r	T	ϕ	λ
T	–	**−0.848**	0.024
ϕ	**−0.848**	–	0.145
λ	0.024	0.145	–

τ	T	ϕ	λ
T	–	**−0.683**	0.030
ϕ	**−0.683**	–	−0.011
λ	0.030	−0.011	–

ρ	T	ϕ	λ
T	–	**−0.815**	0.030
ϕ	**−0.815**	–	0.023
λ	0.030	0.023	–

Достигаемые уровни значимости:

r	T	ϕ	λ
T	–	**0.000**	0.861
ϕ	**0.000**	–	0.287
λ	0.861	0.287	–

τ	T	ϕ	λ
T	–	**0.000**	0.756
ϕ	**0.000**	–	0.910
λ	0.756	0.910	–

ρ	T	ϕ	λ
T	–	**0.000**	0.829
ϕ	**0.000**	–	0.865
λ	0.829	0.865	–

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	–	**−0.861**	**0.280**
ϕ	**−0.861**	–	**0.312**
λ	**0.280**	**0.312**	–

ρ	T	ϕ	λ
T	–	**−0.817**	0.084
ϕ	**−0.817**	–	0.082
λ	0.084	0.082	–

Достигаемые уровни значимости:

r	T	ϕ	λ
T	–	**0.000**	**0.039**
ϕ	**0.000**	–	**0.021**
λ	**0.039**	**0.021**	–

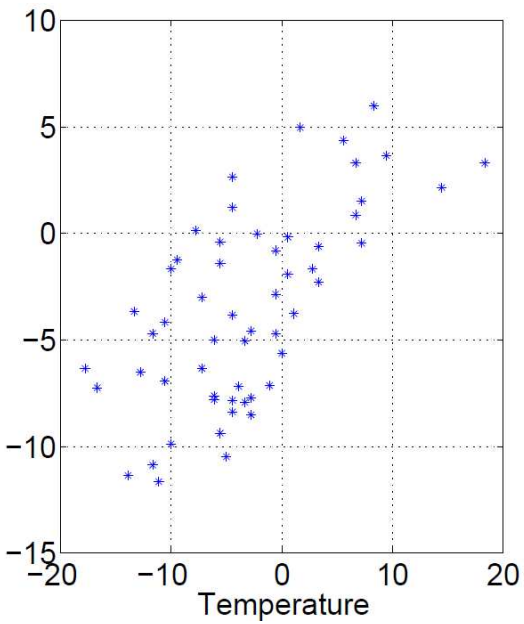
ρ	T	ϕ	λ
T	–	**0.000**	0.543
ϕ	**0.000**	–	0.552
λ	0.543	0.552	–

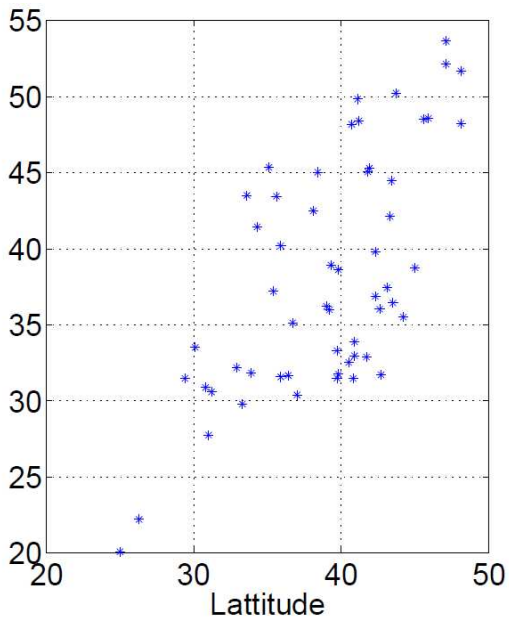
Температура воздуха и географическое положение

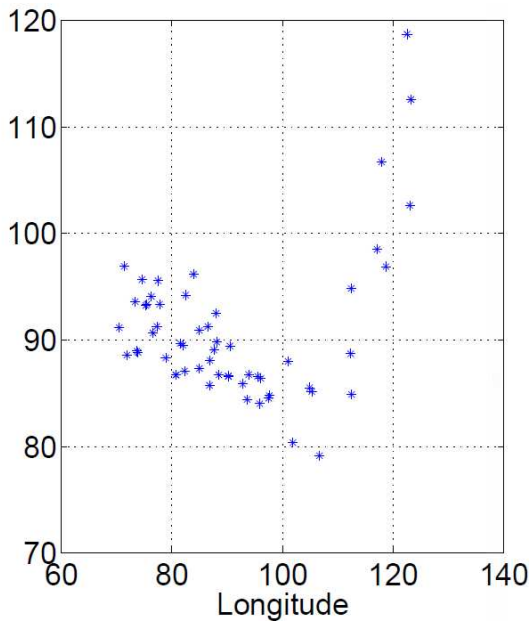
T — температура, λ — долгота, ϕ — широта;
R — множественная корреляция.

Коэффициенты множественной корреляции:

	T	ϕ	λ
R	0.659	0.667	0.312
p	6.0347×10^{-8}	3.6481×10^{-8}	0.0216
with	$0.235 \cdot \lambda - 0.638 \cdot \phi$	$0.397 \cdot \lambda - 0.678 \cdot T$	$1.542 \cdot T + 2.450 \cdot \phi$



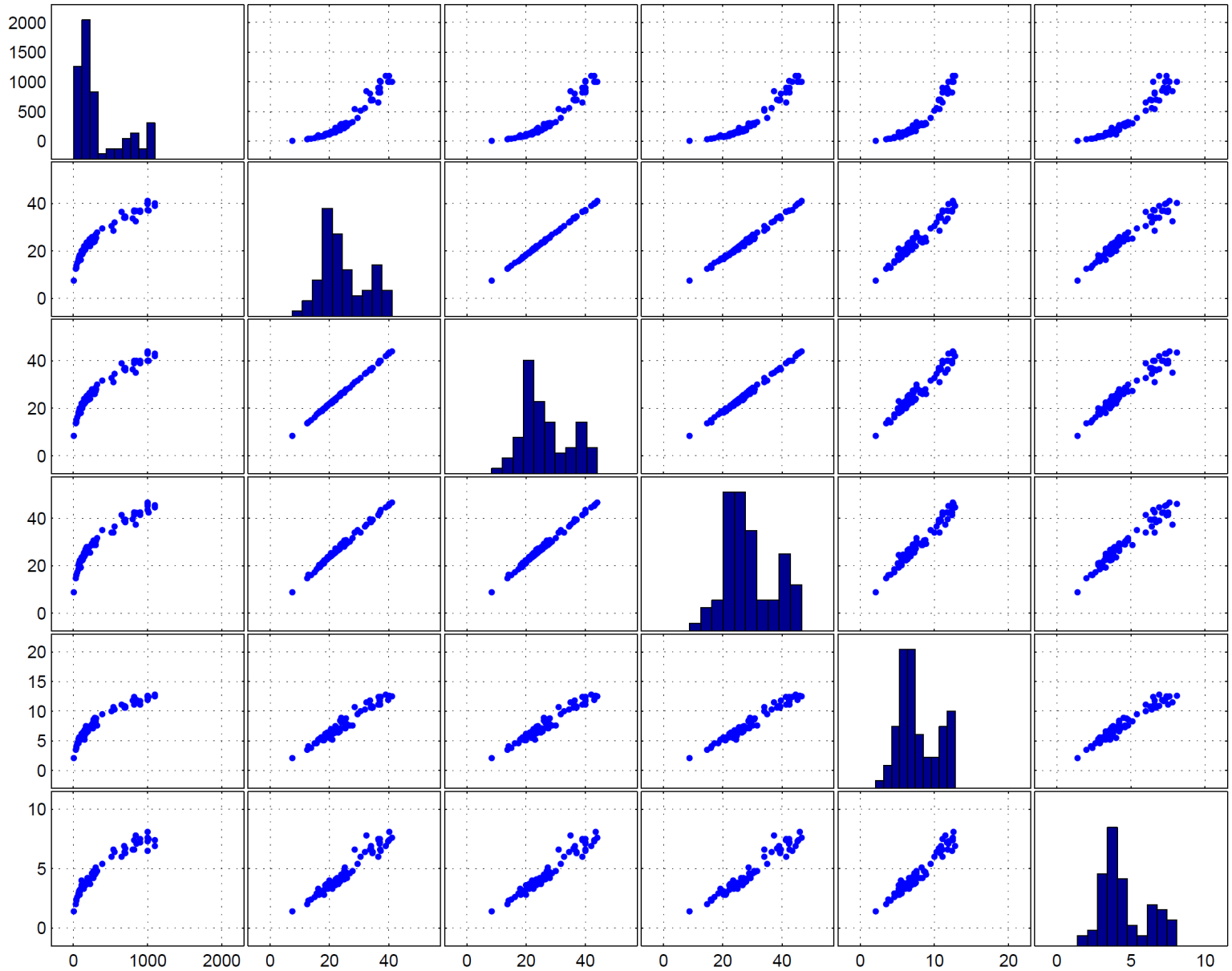




Вес и линейные размеры рыб

В 1917 году в финском озере Längelmävesi 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:

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

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

Вес и линейные размеры рыб

Попарные корреляции Кендалла:

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

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

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

Вес и линейные размеры рыб

Частные корреляции Пирсона:

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

Достигаемые уровни значимости:

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

Вес и линейные размеры рыб

Частные корреляции Спирмена:

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

Достигаемые уровни значимости:

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

Вес и линейные размеры рыб

Множественная корреляция всех признаков с весом: $R = 0.9207$ $(p \approx 0)$.
Максимизирующая корреляцию линейная комбинация:
$292.2458 \cdot Length1 - 151.1554 \cdot Length2 - 151.0027 \cdot Length3 + 148.8896 \cdot Width + 83.4345 \cdot Thickness$.

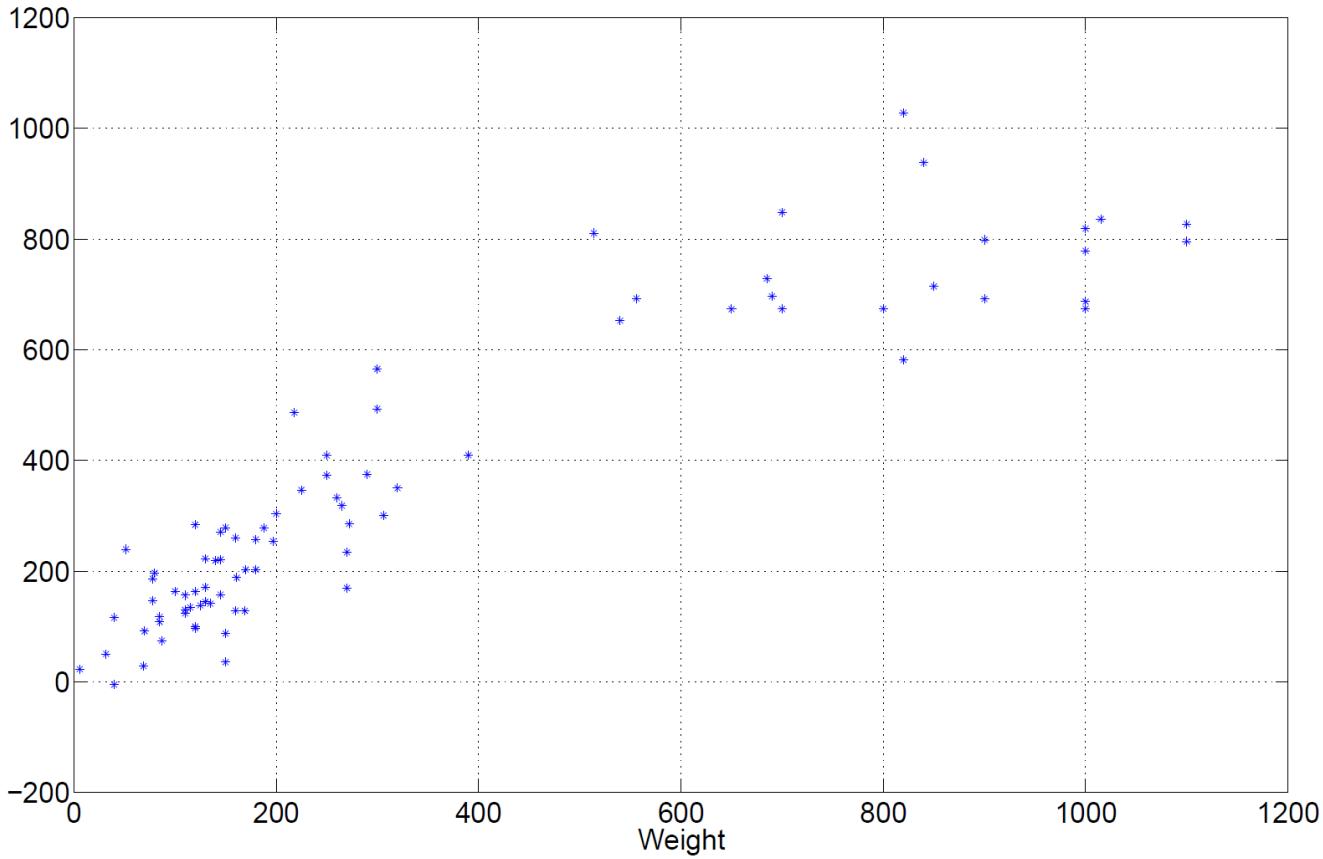


Таблица сопряжённости 2×2

Имеются связные выборки $X_1^n = (X_{11}, \ldots, X_{1n})$ и $X_2^n = (X_{21}, \ldots, X_{2n})$.
Пусть X_{1i} и X_{2i} принимают значения 0 и 1.

№	X_1	X_2
1	0	0
2	1	1
3	0	1
$\vdots$	$\vdots$	$\vdots$
n	1	0

$\Rightarrow$

X_1 \ X_2	0	1	$\sum$
0	a	b	$a+b$
1	c	d	$c+d$
$\sum$	$a+c$	$b+d$	n

Корреляция Мэтьюса

Мера взаимосвязи между двумя бинарными переменными — коэффициент корреляции Мэтьюса:

$$MCC = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}.$$

$MCC \in [-1, 1]$; 0 соответствует полному отсутствию взаимосвязи, 1 — нулям на побочной диагонали, -1 — нулям на главной диагонали.

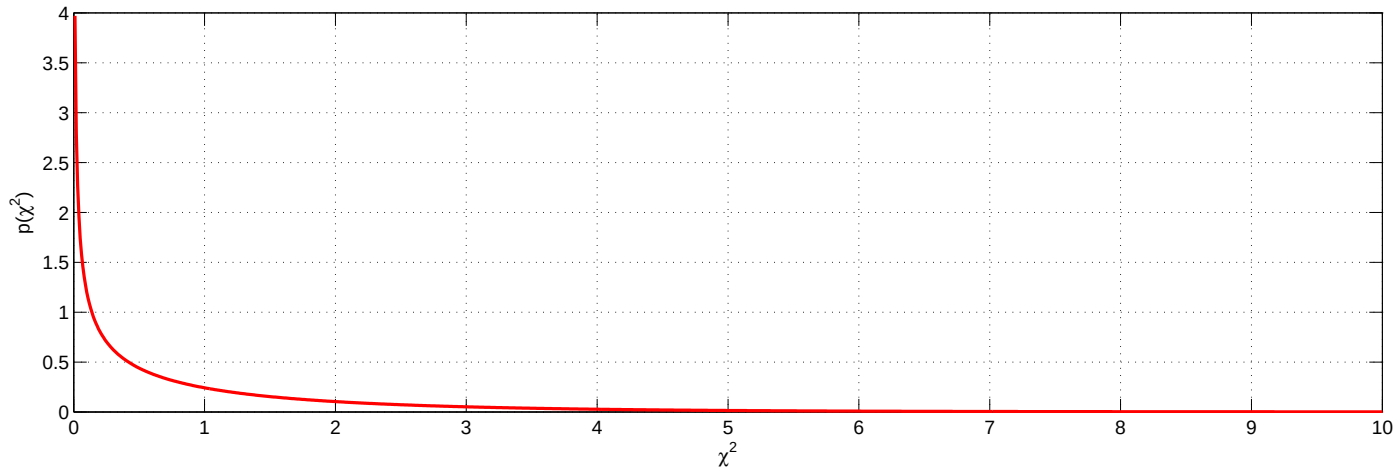
Критерий Мак-Нимара

выборки: $X_1^n = (X_{11}, \ldots, X_{1n}), \;\; X_{1i} \sim Ber(p_1),$
$X_2^n = (X_{21}, \ldots, X_{2n}), \;\; X_{2i} \sim Ber(p_2),$ выборки связные;

нулевая гипотеза: H_0: X_1 и X_2 независимы

альтернатива: H_1: H_0 неверна;

статистика: $\chi^2\left(X_1^n, X_2^n\right) = \frac{n(|b-c|-0.5)^2}{b+c}$;
$\chi^2\left(X_1^n, X_2^n\right) \sim \chi_1^2$ при H_0;



достигаемый уровень значимости:

$$p\left(\chi^2\right) = 1 - chi2cdf(\chi^2, 1).$$

Критерий Мак-Нимара

Условия применимости критерия:

- $n \geq 40$;
- $a, b, c, d > 5$.

Критерий Мак-Нимара

Пример: исследуется влияние препарата на симптом некоторого заболевания.

До лечения \ После лечения	Есть	Нет
Есть	101	121
Нет	59	33

$b = 121$ — испытуемые, у которых симптом после приёма препарата исчез,
$c = 59$ — испытуемые, у которых симптом появился.
$MCC = -0.1673$.

H_0: препарат не влияет на наличие симптома.
H_1: препарат влияет на наличие симптома $\Rightarrow p = 4.5689 \times 10^{-6}$.

Анализ зависимостей независимых выборок

Имеются независимые выборки $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1})$ и $X_2^{n_2} = (X_{21}, \ldots, X_{2n_2})$, $n_1 + n_2 = n$, X_{1i} и X_{2i} принимают значения 0 и 1. Данные можно представить сходным образом:

№	$\in X_1$	Значение
1	1	0
2	0	1
3	0	0
⋮	⋮	⋮
n	1	1

$\Rightarrow$

Выборка \ Значение	0	1	$\sum$
X_1	a	b	$n_1 = a + b$
X_2	c	d	$n_2 = c + d$
$\sum$	$a + c$	$b + d$	n

Признаки "выборка"и "значение"независимы $\Leftrightarrow$ выборки X_1 и X_2 однородны.

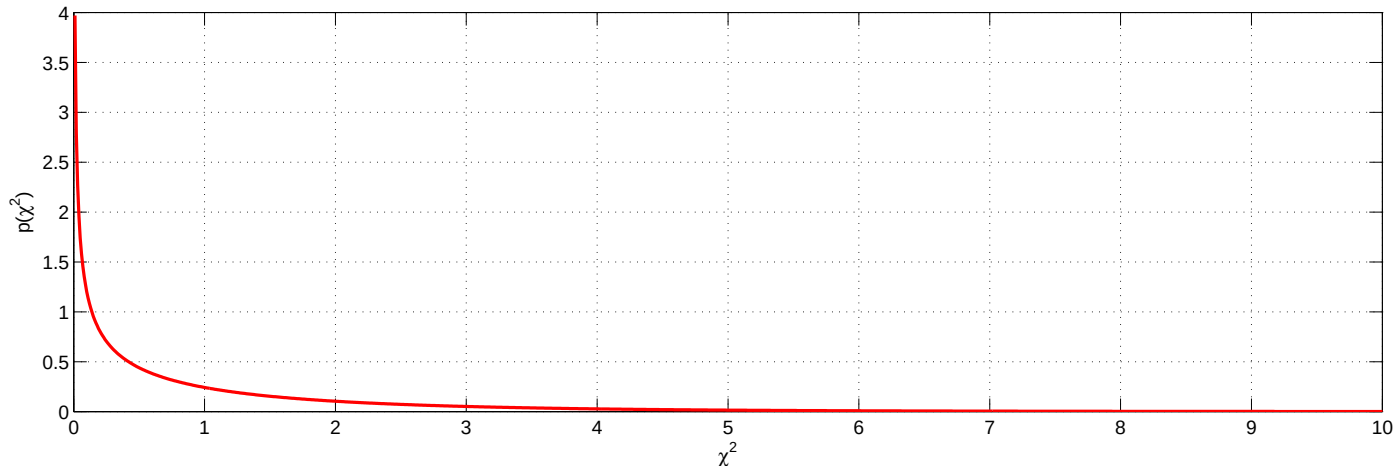
Критерий хи-квадрат

выборки: $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1}), X_{1i} \in \{0,1\}$,
$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2}), X_{2i} \in \{0,1\}$,
выборки независимые;

нулевая гипотеза: H_0: X_{1i} и X_{2j} однородны,

альтернатива: H_1: H_0 неверна;

статистика: $\chi^2\left(X_1^{n_1}, X_2^{n_2}\right) = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$;
$\chi^2\left(X_1^{n_1}, X_2^{n_2}\right) \sim \chi_1^2$ при H_0;



достигаемый уровень значимости:

$$p\left(\chi^2\right) = 1 - chi2cdf(\chi^2, 1).$$

Критерий хи-квадрат

Критерий эквивалентен Z-критерию для сравнения двух пропорций.

Условия применимости критерия:

- $n \geq 40$;
- $a, b, c, d > 5$.

Критерий хи-квадрат

Пример: исследуется влияние препарата на некоторое заболевание. Часть испытуемых принимает препарат, часть — плацебо; по окончании курса определяется, произошло ли выздоровление.

	Выздоровели	Нет
Препарат	850	870
Плацебо	380	410

$MCC = 0.0122.$

H_0: препарат неотличим от плацебо.
H_1: эффект препарата отличается от эффекта плацебо $\Rightarrow p = 0.5398.$

Точный критерий Фишера

выборки: $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1}), X_{1i} \in \{0,1\}$,
$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2}), X_{2i} \in \{0,1\}$,
выборки независимые;

нулевая гипотеза: H_0: X_{1i} и X_{2j} однородны,

альтернатива: H_1: H_0 неверна.

Пусть в таблице сопряжённости суммы по строкам и столбцам фиксированы, тогда вероятность появления наблюдаемой таблицы равна

$$P\left(X_1^{n_1}, X_2^{n_2}\right) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!b!c!d!n!}.$$

Достигаемый уровень значимости определяется как сумма по всем возможным вариантам таблицы с такими же суммами по строкам и столбцам, имеющим вероятность не более $P\left(X_1^{n_1}, X_2^{n_2}\right)$.

Для односторонней альтернативы ($ad \ll bc$) достигаемый уровень значимости можно определить через гипергеометрическое распределение:

$$p = \sum_{i=0}^{a} \frac{C_{a+b}^{i} C_{c+d}^{a+c-i}}{C_n^{a+c}}.$$

Точный критерий Фишера

Пример: для 24 опрошенных известен пол и сидят ли они на диете. Есть ли связь между этими признаками?

	М	Ж
На диете	1	9
Не на диете	13	3

$MCC = -0.6953.$

H_0: связи нет.
H_1: признаки связаны $\Rightarrow p = 0.0014.$

Парадокс хи-квадрат (Симпсона)

Эксперимент: пациенты принимают препарат или плацебо, по окончании курса определяется, выздоровели они или нет.
Есть ли связь между выздоровлением и приёмом препарата?

Мужчины	Выздоровели	Нет
Препарат	700	800
Плацебо	80	130

Женщины	Выздоровели	Нет
Препарат	150	70
Плацебо	300	280

Для мужчин: $\chi^2 = 5.456,\ p = 0.0195.$
Для женщин: $\chi^2 = 17.555,\ p = 2.7914 \times 10^{-5}.$

М+Ж	Выздоровели	Нет
Препарат	850	870
Плацебо	380	410

Суммарно: $\chi^2 = 0.376,\ p = 0.5398.$

Парадокс хи-квадрат (Симпсона)

Причины несогласованности выводов — большие отличия в размерах групп пациентов, принимающих плацебо и препарат: основной вклад в выводы вносят женщины, принимавшие плацебо, и мужчины, принимавшие препарат.

Чтобы такого не происходило, плацебо и препарат должны поровну распределяться по всем анализируемым подгруппам.

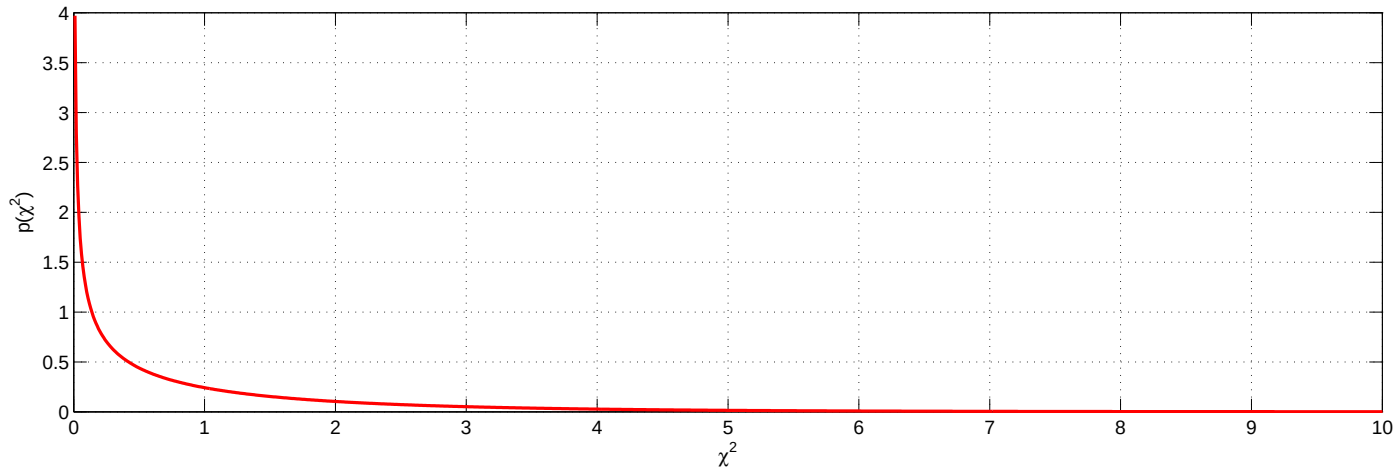
Парадокс хи-квадрат (Симпсона)

Bikel at el., Sex Bias in Graduate Admissions: Data from Berkeley, 1975.
В 1973 году на университет Беркли, Калифорния, подали в суд: доля поступивших абитуриентов мужского пола была выше, чем доля поступивших женского пола.

	Не поступили	Поступили	Доля поступивших
Мужчины	4704	3738	44.3%
Женщины	2827	1494	34.6%

Парадокс хи-квадрат (Симпсона)

Критерий хи-квадрат: $\chi^2 = 108.1,\ p \approx 0.$



	Наблюдаемые		Ожидаемые		Разности	
	-	+	-	+	-	+
Мужчины	4704	3738	4981.3	3460.7	−227.3	227.3
Женщины	2827	1494	2549.7	1771.3	227.3	−227.3

Парадокс хи-квадрат (Симпсона)

Будем искать виноватых: посмотрим детализированную статистику по 85 факультетам.

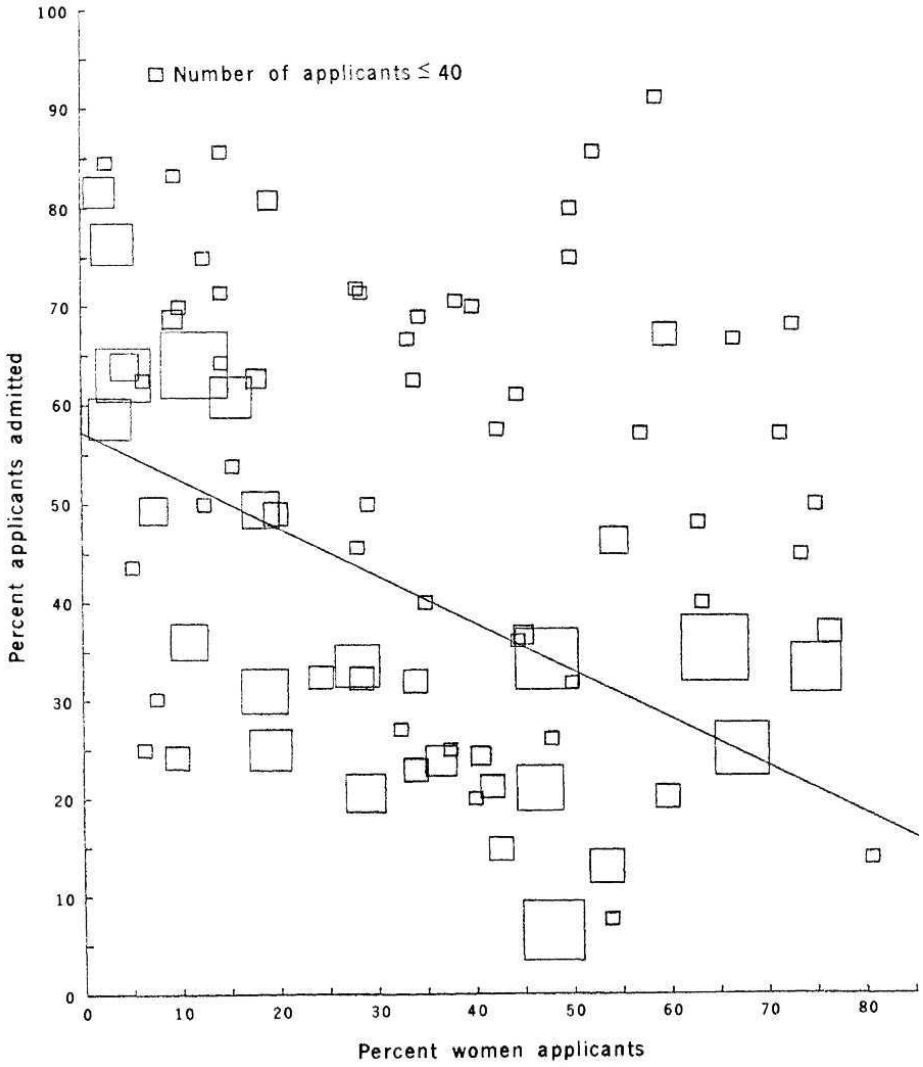
Значимо (при $\alpha = 0.05$) меньше женщин прошли отбор на 4 факультета, суммарный дефицит по ним — 26.
На 6 факультетов поступило значимо меньше мужчин, суммарный дефицит — 64.

Данные по 6 крупнейшим факультетам:

	Мужчины		Женщины	
	$\sum$	+	$\sum$	+
1	825	62%	108	**82%**
2	560	63%	25	**68%**
3	325	**37%**	593	34%
4	417	33%	375	**35%**
5	191	**28%**	393	24%
6	272	6%	341	**7%**

Парадокс хи-квадрат (Симпсона)

Ответ: женщины чаще пытались поступить на факультеты с большим конкурсом.



Прикладная статистика
6. Анализ зависимостей.

Рябенко Евгений
riabenko.e@gmail.com