Прикладная статистика 5. Корреляционный анализ.

11 марта 2013 г.

Задача исследования взаимосвязи между признаками

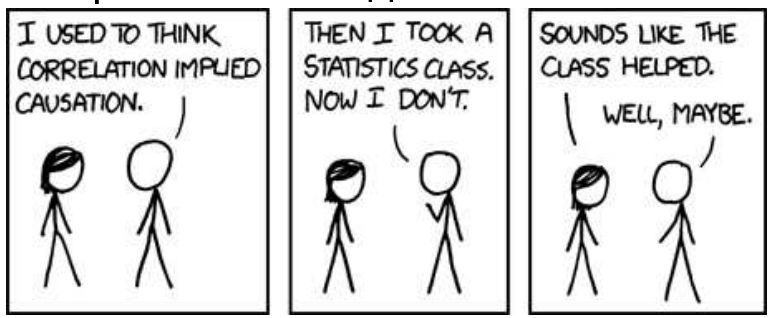
Дано: значения признаков X, Y измерены на объектах $1, \ldots, n$.
Эквивалентная формулировка: имеются связные выборки $X^n = (X_1, \ldots, X_n)$ и $Y^n = (Y_1, \ldots, Y_n)$.

Насколько сильно признаки X, Y связаны между собой?

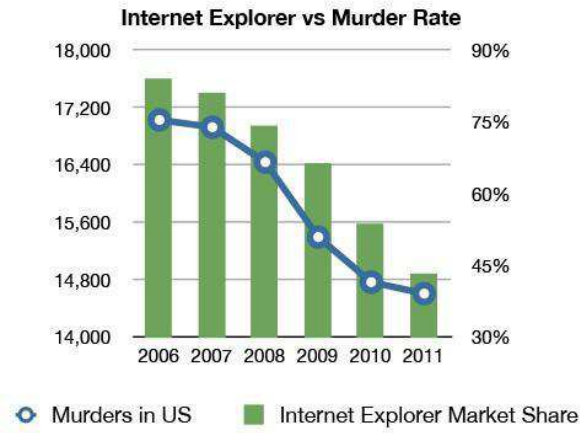
Статистическая взаимосвязь между случайными величинами — **корреляция**.

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):

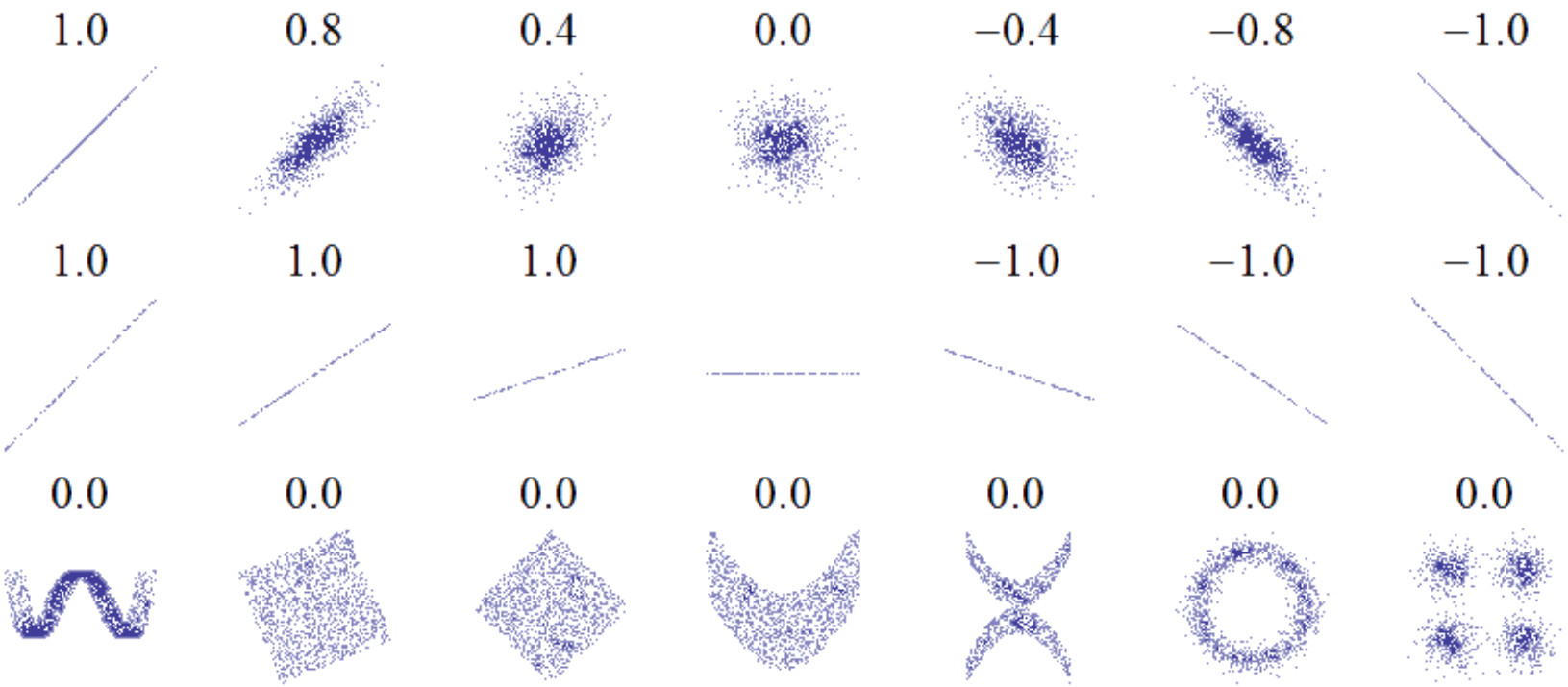
$$r_{XY} = \frac{cov\left(X, Y\right)}{\sqrt{\mathbb{D}X\mathbb{D}Y}} = \frac{\mathbb{E}\left(\left(X - \mathbb{E}X\right)\left(Y - \mathbb{E}Y\right)\right)}{\sqrt{\mathbb{D}X\mathbb{D}Y}}.$$

Выборочный коэффициент корреляции Пирсона:

$$r_{XY} = \frac{\sum\limits_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum\limits_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum\limits_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}.$$

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

Корреляция Пирсона



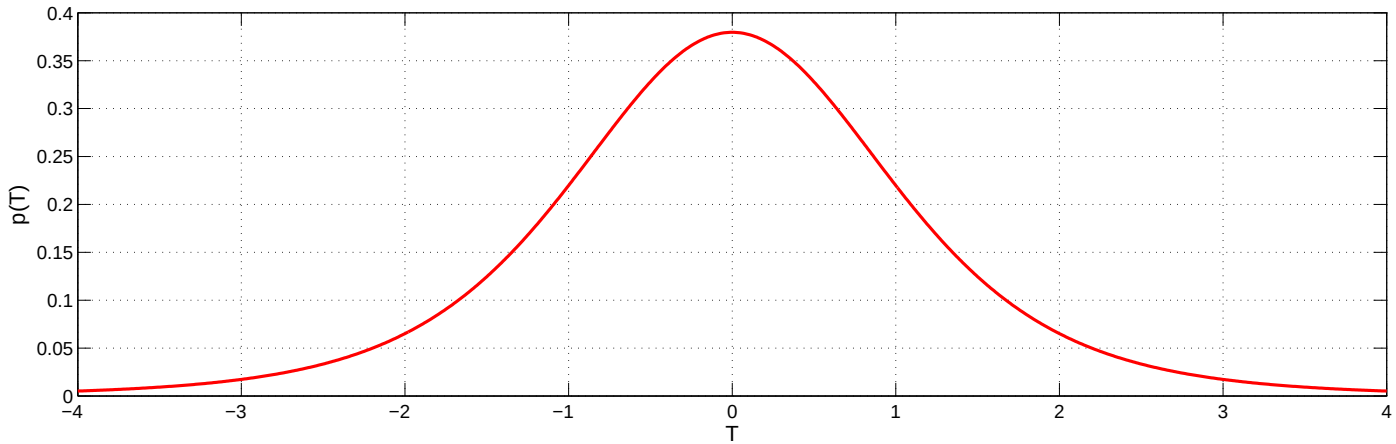
Критерий Стьюдента

выборки: $X^n = (X_1, \ldots, X_n)$,
$Y^n = (Y_1, \ldots, Y_n)$, выборки связные,
$(X, Y) \sim N(\mu, \Sigma)$;

нулевая гипотеза: $H_0 : r_{XY} = 0$;

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

статистика: $T(X^n, Y^n) = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}}$;
$T(X^n, Y^n) \sim St(n-2)$ при H_0;



достигаемый уровень значимости:

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

Критерий Стьюдента

Доверительный интервал для коэффициента корреляции:

$$r_{XY} \pm \frac{t_{n-2,\alpha/2}\left(1-r_{XY}^2\right)}{\sqrt{n}}.$$

С использованием преобразования Фишера:

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

Критерий Стьюдента

Пример: для двух марок зубной пасты, одна из которых рекламируется по телевизору, а другая нет, участники опроса 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?

$H_0 : r_{XY} = 0.$
$H_1 : r_{XY} \neq 0 \quad \Rightarrow \quad p = 0.0979.$

Перестановочный критерий

выборки: $X^n = (X_1, \ldots, X_n)$,
$Y^n = (Y_1, \ldots, Y_n)$, выборки связные;
нулевая гипотеза: $H_0 \colon r_{XY} = 0$;
альтернатива: $H_1 \colon r_{XY} <\neq> 0$;
статистика: $T(X^n, Y^n) = r_{XY}$.

Распределение $T(X^n, Y^n)$ при H_0 порождается группой перестановок

$$G = \{g \colon gY^n = (Y_{\pi_1}, \ldots, Y_{\pi_n})\},$$

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

$$|G| = n!.$$

Достигаемый уровень значимости:

$$p(t) = \begin{cases} \dfrac{\sum\limits_{g \in G} [r(X^n, gY^n) \leq \geq r(X^n, Y^n)]}{n!}, & H_1 \colon r_{XY} <> 0, \\ \dfrac{\sum\limits_{g \in G} [|r(X^n, gY^n)| \geq |r(X^n, Y^n)|]}{n!}, & H_1 \colon r_{XY} \neq 0. \end{cases}$$

Перестановочный критерий

Перестановочный доверительный интервал для коэффициента корреляции образован выборочными квантилями порядка $\alpha/2$ и $1-\alpha/2$ перестановочного распределения $r\left(X^n, gY^n\right)$.

Перестановочный критерий

Пример: в предыдущем примере

$H_0 : r_{XY} = 0.$
$H_1 : r_{XY} \neq 0 \quad \Rightarrow \quad p = 0.0715.$

Недостатки

Недостатки выборочного r:

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

Корреляция Спирмена

Коэффициент корреляции Спирмена — коэффициент корреляции Пирсона рангов наблюдений в выборках X^n, Y^n:

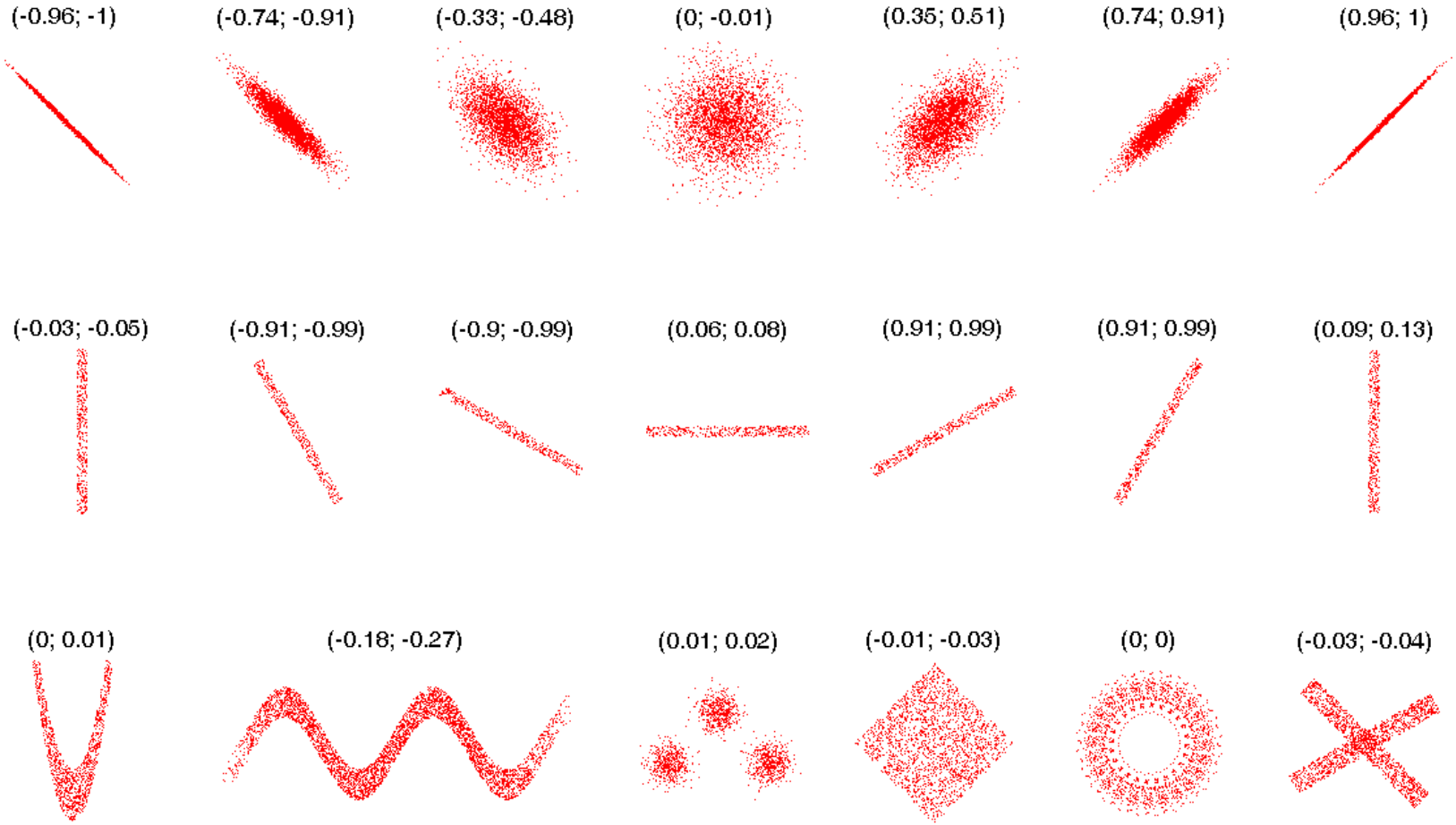
$$\rho_{XY} = \frac{\sum_{i=1}^{n} \left(r\left(X_i\right) - \frac{n+1}{2}\right)\left(r\left(Y_i\right) - \frac{n+1}{2}\right)}{\frac{1}{12}\left(n^3 - n\right)} =$$

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

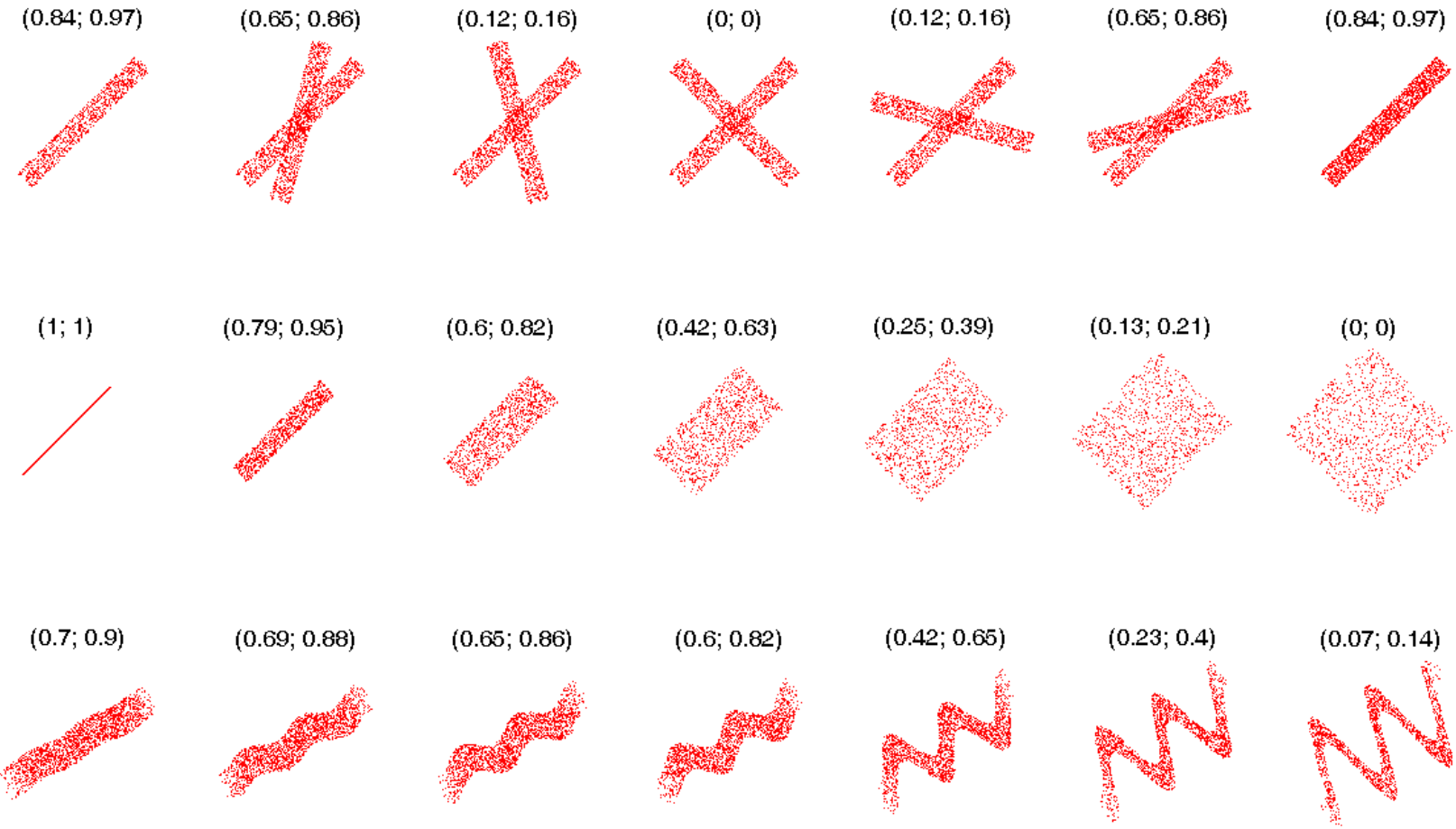
где $r\left(X_i\right), r\left(Y_i\right)$ — ранги i-х наблюдений в соответствующих выборках.

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

Корреляция Спирмена



Корреляция Спирмена



Критерий Стьюдента

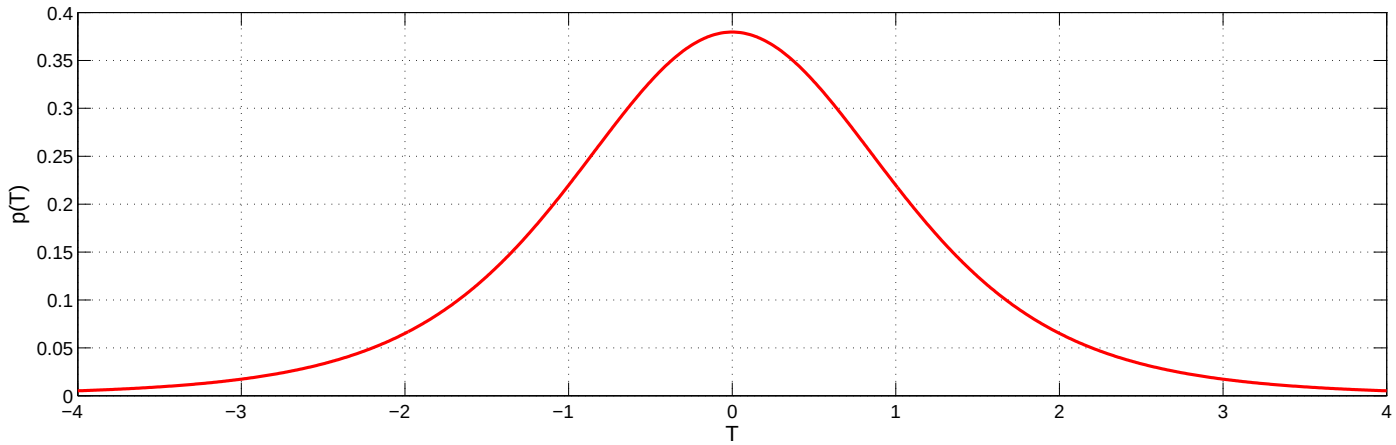
выборки: $X^n = (X_1, \ldots, X_n)$,
$Y^n = (Y_1, \ldots, Y_n)$, выборки связные;

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

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

статистика: $T(X^n, Y^n) = \frac{\rho_{XY}\sqrt{n-2}}{\sqrt{1-\rho_{XY}^2}}$;

$T(X^n, Y^n) \sim St(n-2)$ при H_0;



достигаемый уровень значимости:

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

Критерий Стьюдента

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

$H_0 : \rho_{XY} = 0.$
$H_1 : \rho_{XY} \neq 0 \quad \Rightarrow \quad p = 0.0024.$

Корреляция Кендалла

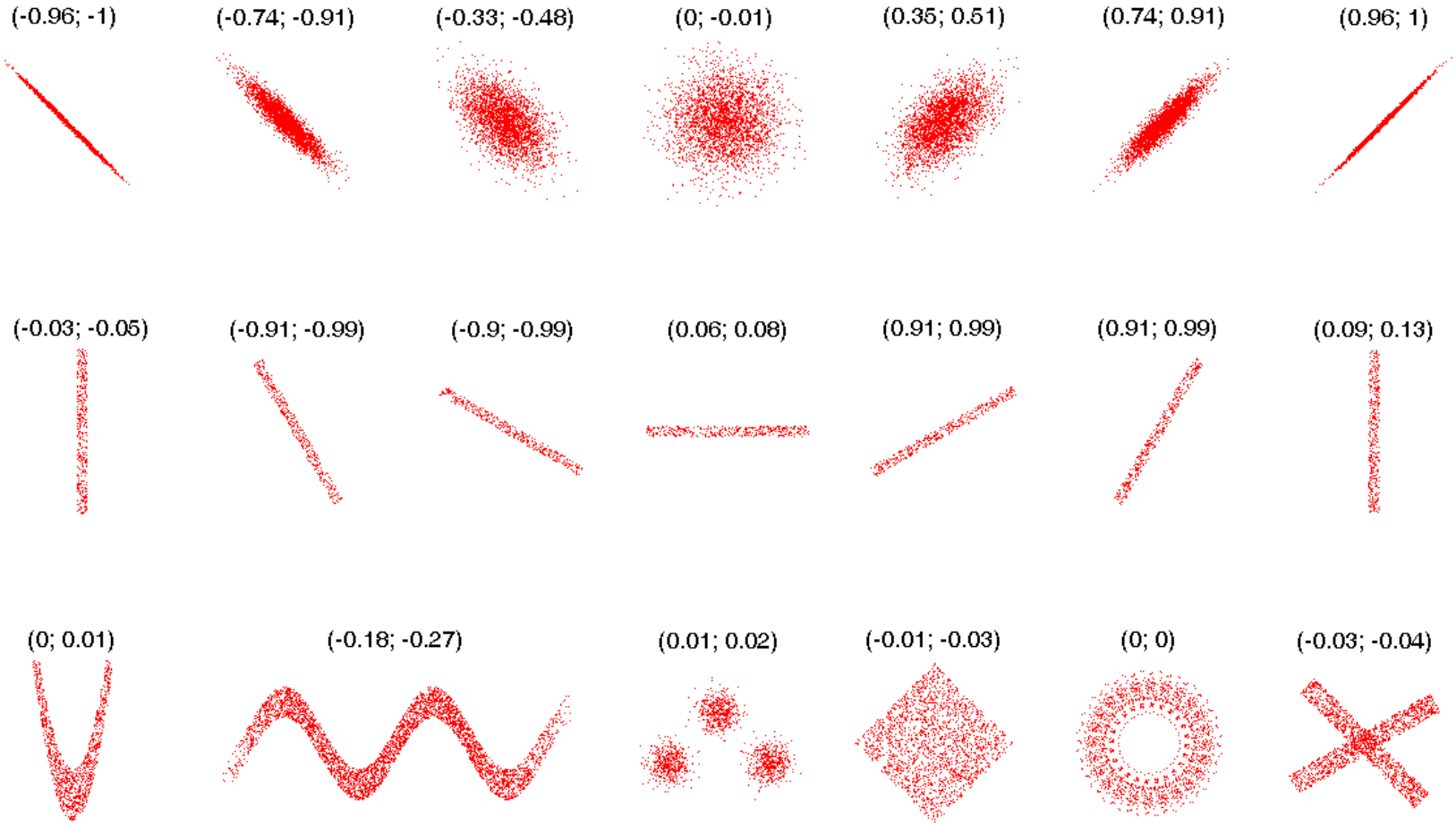
Коэффициент корреляции Кендалла — мера взаимной неупорядоченности X^n и Y^n:

$$\tau_{XY} = 1 - \frac{4}{n\,(n-1)} \sum_{i=1}^{n-1} \sum_{j=1}^{n} [[X_i < X_j] \neq [Y_i < Y_j]] = \frac{S-R}{S+R},$$

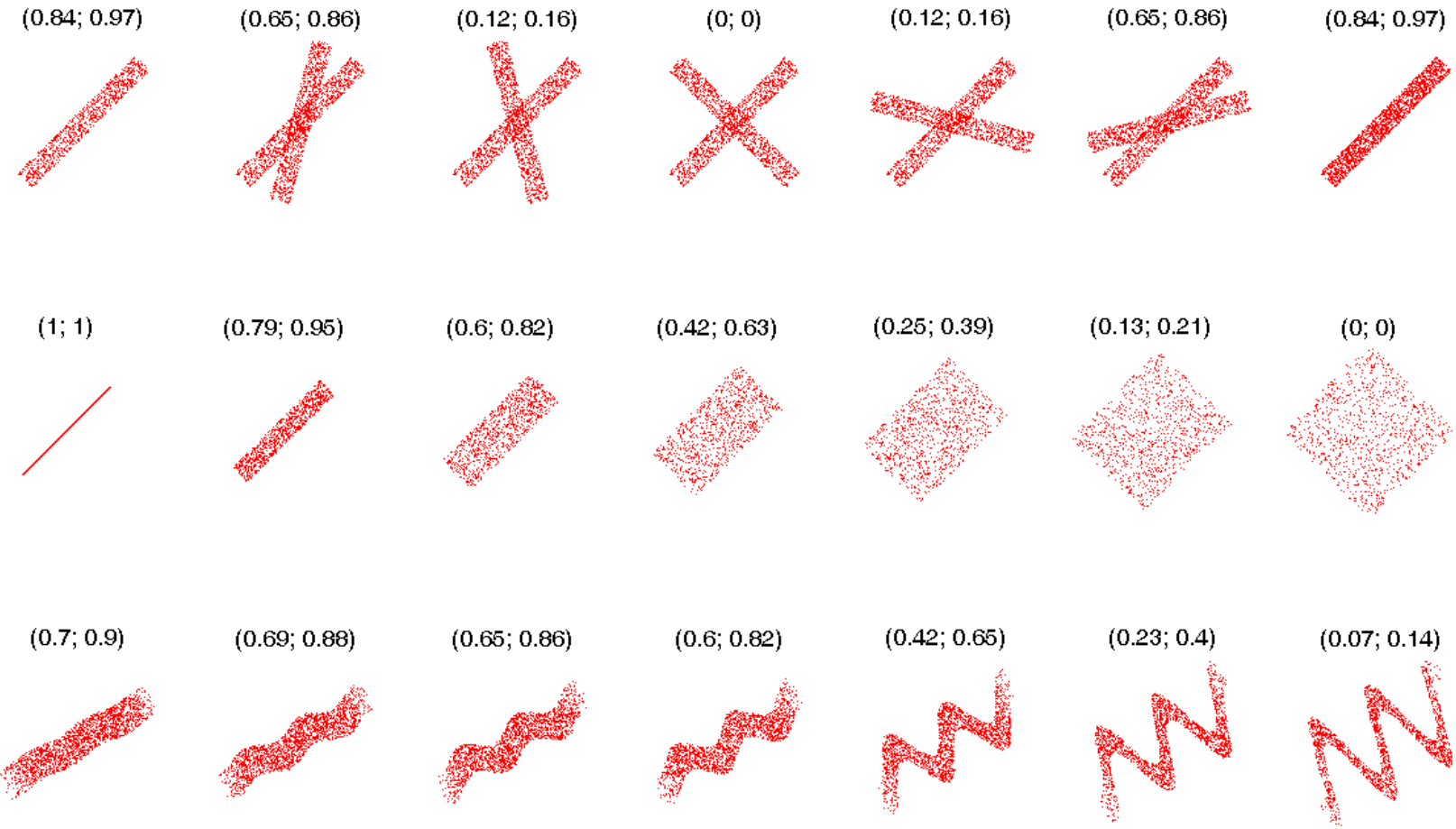
где R — число несогласованных пар, S — число согласованных.

$\tau_{XY} \in [-1, 1]$ — сила **монотонной** связи.

Корреляция Кендалла



Корреляция Кендалла



Критерий без названия

выборки: $X^n = (X_1, \ldots, X_n)$,
$Y^n = (Y_1, \ldots, Y_n)$, выборки связные;
нулевая гипотеза: $H_0 \colon \tau_{XY} = 0$;
альтернатива: $H_1 \colon \tau_{XY} <\neq> 0$;
статистика: τ_{XY};
τ_{XY} имеет табличное распределение при H_0.

При справедливости H_0

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

Для $n > 10$ справедлива аппроксимация:

$$\frac{\tau_{XY}}{\sqrt{\mathbb{D}\tau_{XY}}} \sim N(0,1).$$

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 $R = 5, S = 38, \tau_{XY} = 0.7821$.

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

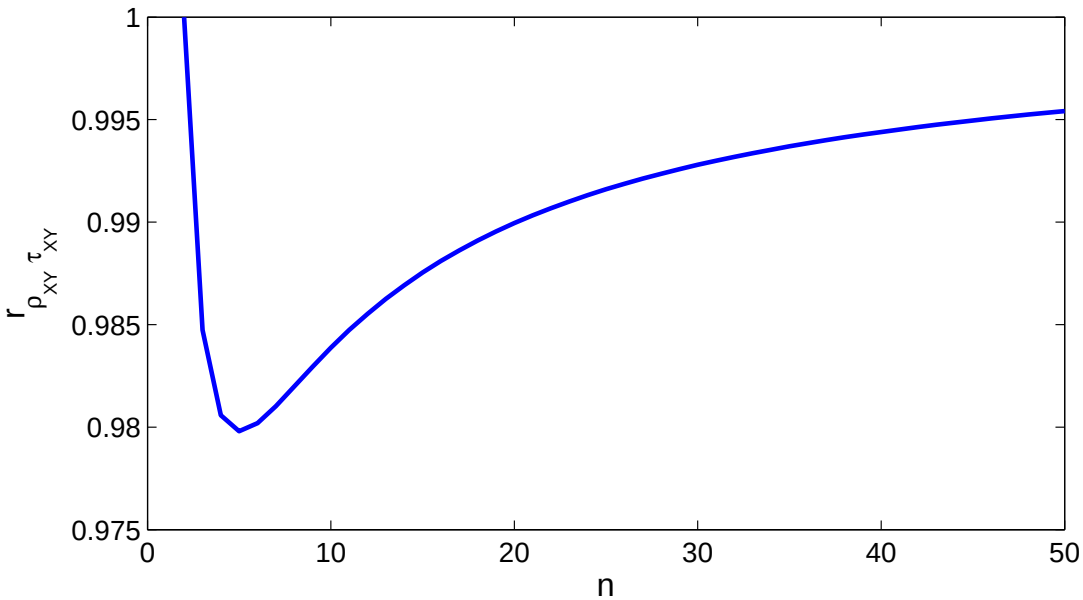
Связь коэффициентов корреляции

Если $(X, Y) \sim N(\mu, \Sigma)$, то

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

При справедливости H_0 (отсутствии монотонной зависимости)

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



Частная корреляция

Если мы подозреваем, что наблюдаемая линейная взаимосвязь между признаками X и Y вызвана влиянием третьей переменной Z, можно попытаться её исключить.
Частная корреляция:

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}.$$

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

$$r_{XY|ZV} = \frac{r_{XY|V} - r_{XZ|V}r_{YZ|V}}{\sqrt{\left(1 - r_{XZ|V}^2\right)\left(1 - r_{YZ|V}^2\right)}}.$$

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

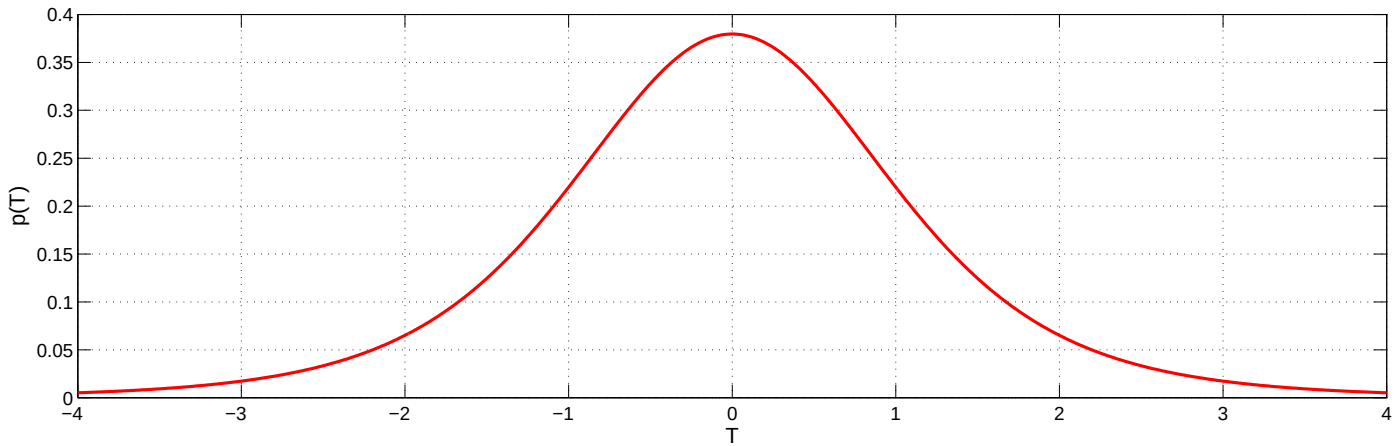
выборки: $X^n = (X_1, \ldots, X_n), \quad Y^n = (Y_1, \ldots, Y_n),$
$Z^n = (Z_1, \ldots, Z_n), \quad Z_i \in \mathbb{R}^M, \quad (X, Y, Z) \sim N(\mu, \Sigma);$

нулевая гипотеза: $H_0 : r_{XY|Z} = 0;$

альтернатива: $H_1 : r_{XY|Z} <\neq> 0;$

статистика: $T(X^n, Y^n, Z^n) = \frac{r_{XY|Z}\sqrt{n-M-2}}{\sqrt{1-r^2_{XY|Z}}};$

$T(X^n, Y^n, Z^n) \sim St(n-M-2)$ при $H_0;$



достигаемый уровень значимости:

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

Множественная корреляция

Для того, чтобы оценить силу линейной взаимосвязи одной переменной с несколькими другими, используется множественная корреляция:

$$r^2_{X,YZ} = \frac{r^2_{XY} + r^2_{XZ} - 2r_{XY}r_{XZ}r_{YZ}}{1 - r^2_{YZ}}.$$

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

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

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

Критерий Фишера

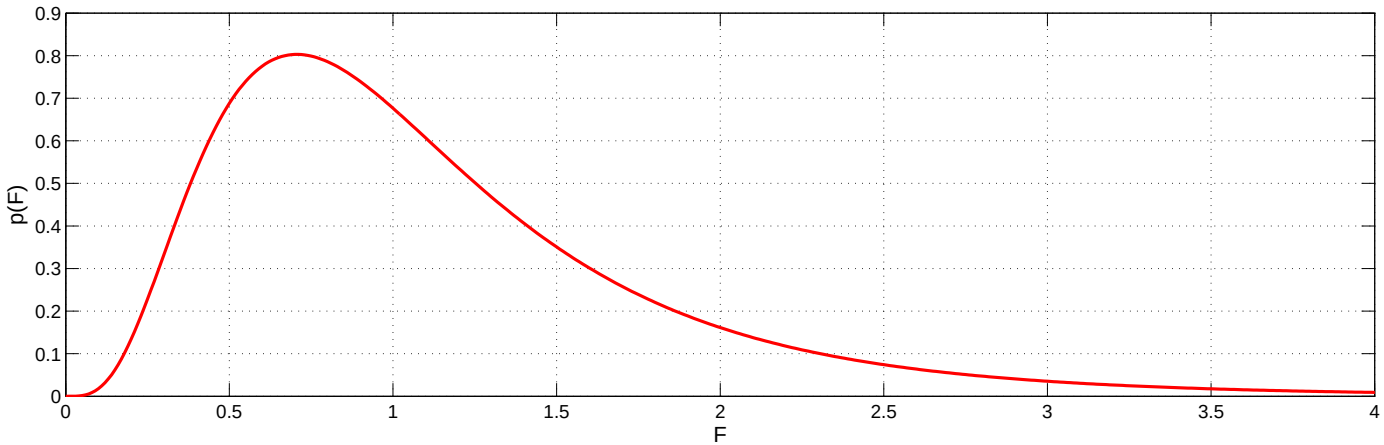
выборки: $X^n = (X_1, \ldots, X_n)$,
$Y^n = (Y_1, \ldots, Y_n)$, $Y_i \in \mathbb{R}^M$, $(X, Y) \sim N(\mu, \Sigma)$;

нулевая гипотеза: $H_0 : r_{X,Y} = 0$;

альтернатива: $H_1 : r_{X,Y} \neq 0$;

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

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

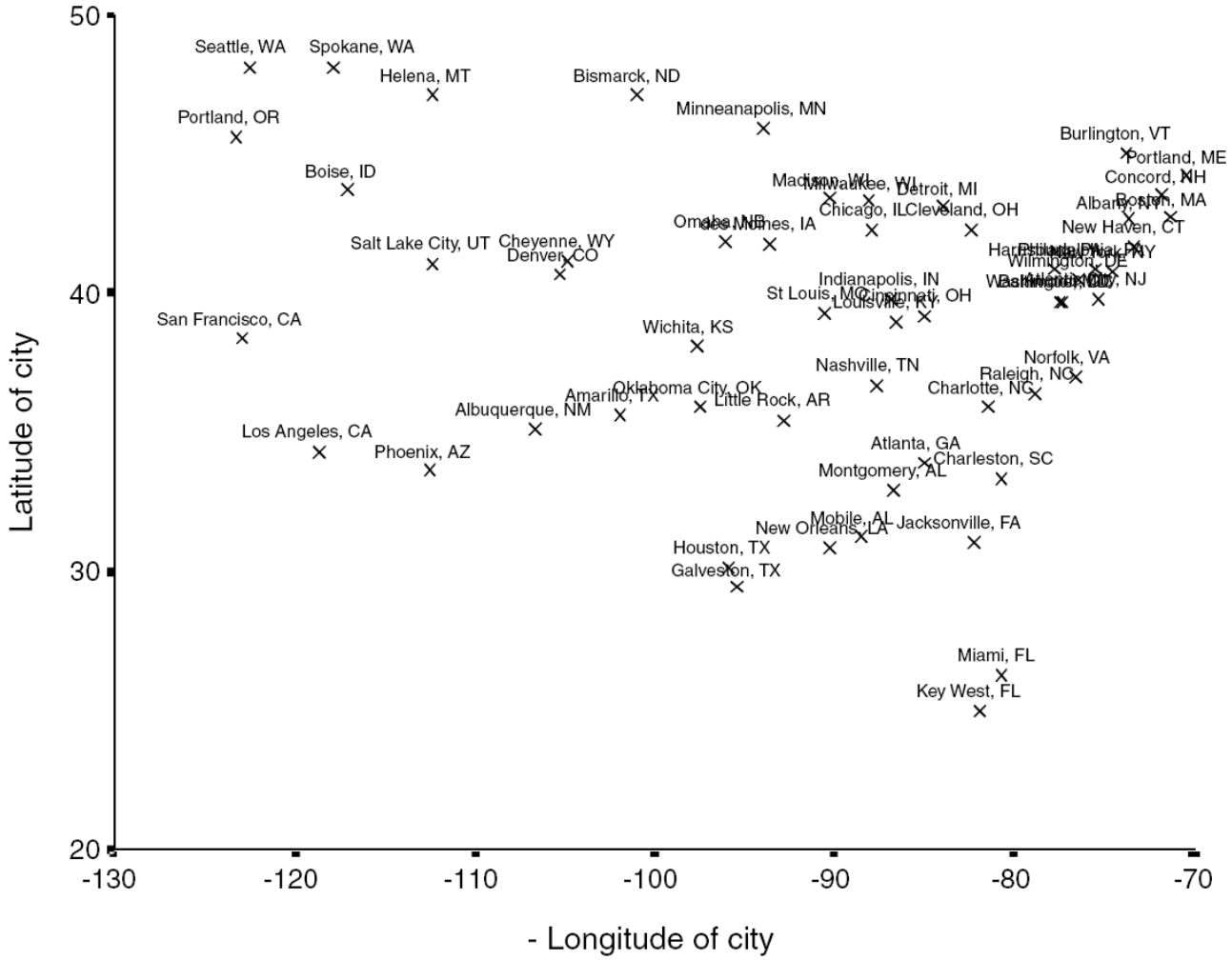


достигаемый уровень значимости:

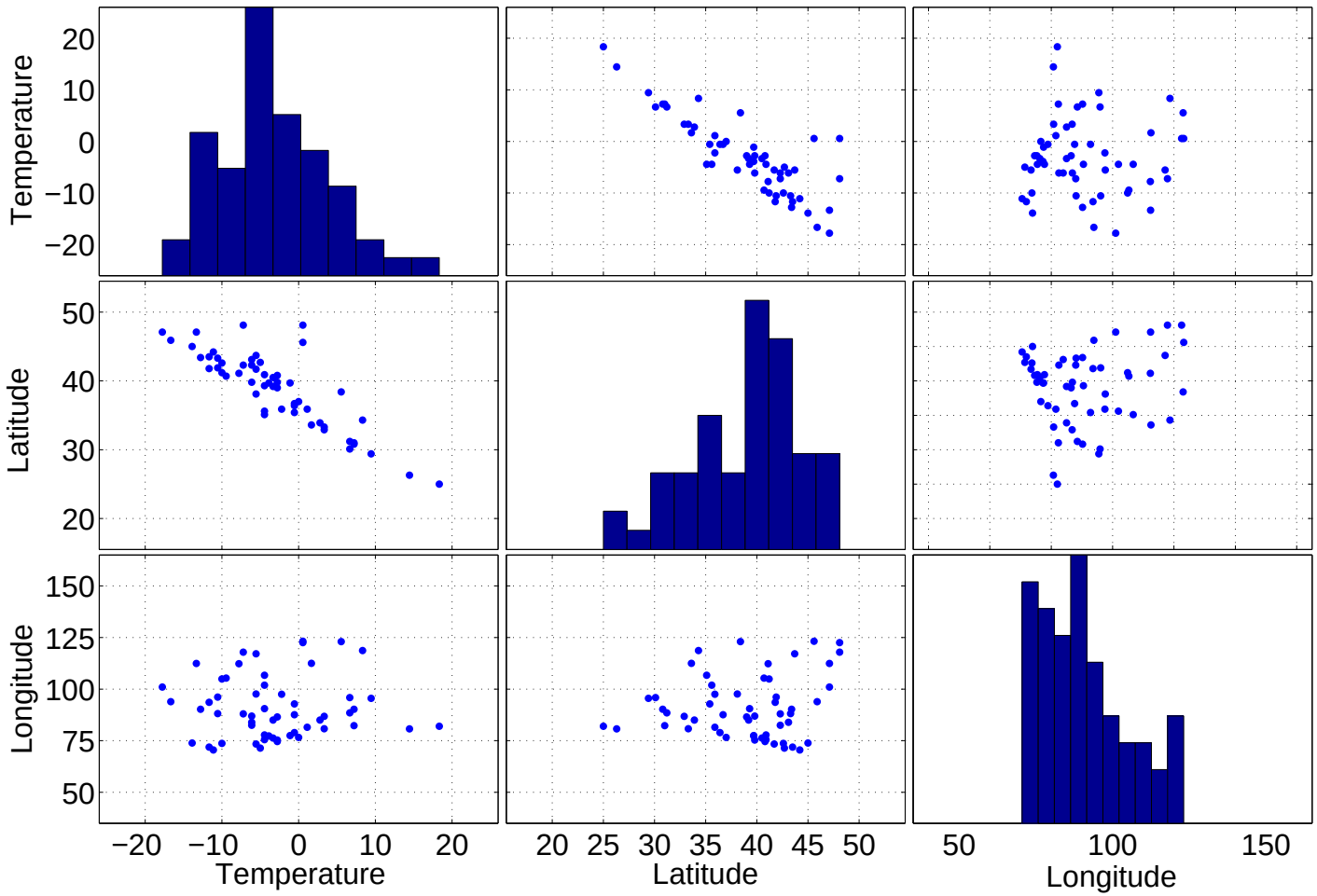
$$p(f) = 1 - fcdf(f, M-2, n-M-1).$$

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, τ — Кендалла, ρ — Спирмена.

Коэффициенты корреляции:

r	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:

r	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

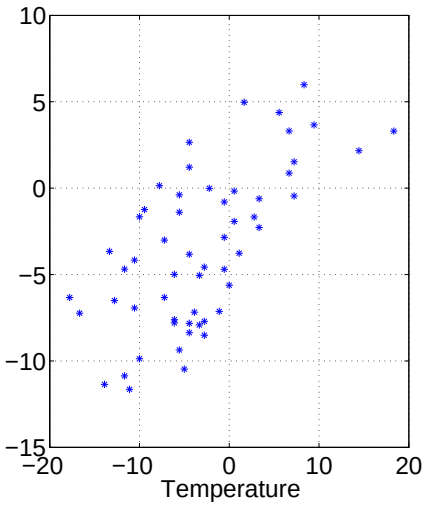
ρ	T	ϕ	λ
T	0.000	0.000	0.543
ϕ	0.000	0.000	0.552
λ	0.543	0.552	0.000

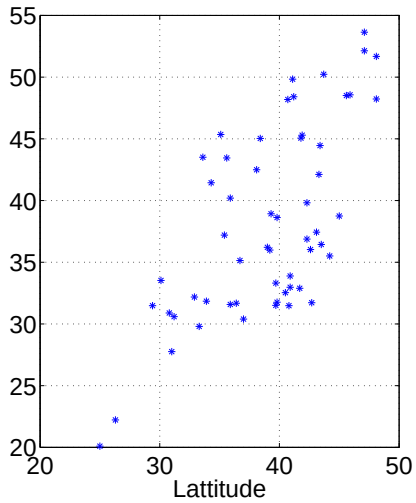
Температура воздуха и географическое положение

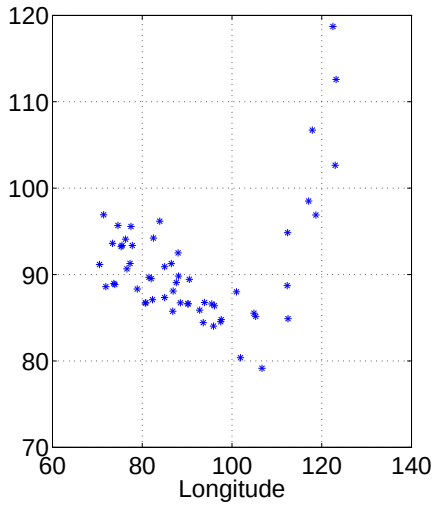
T — температура, λ — долгота, ϕ — широта;
R — множественная корреляция.

Коэффициенты множественной корреляции:

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

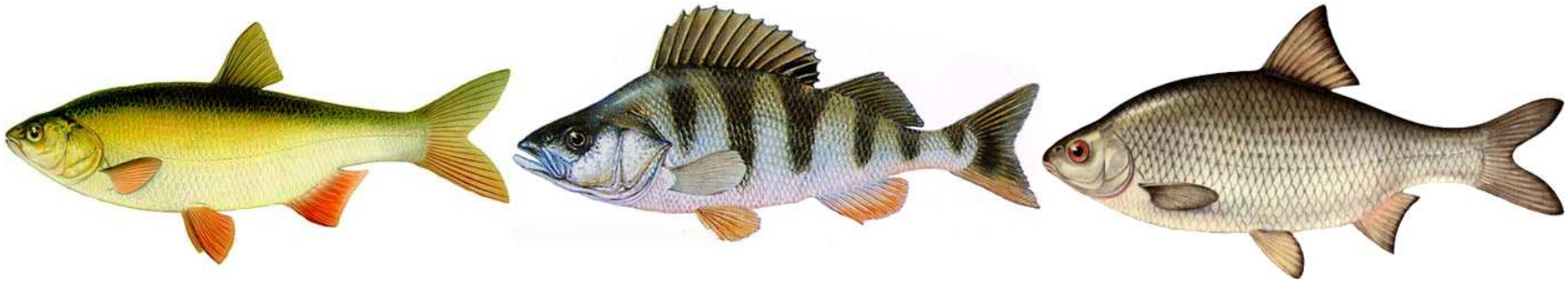




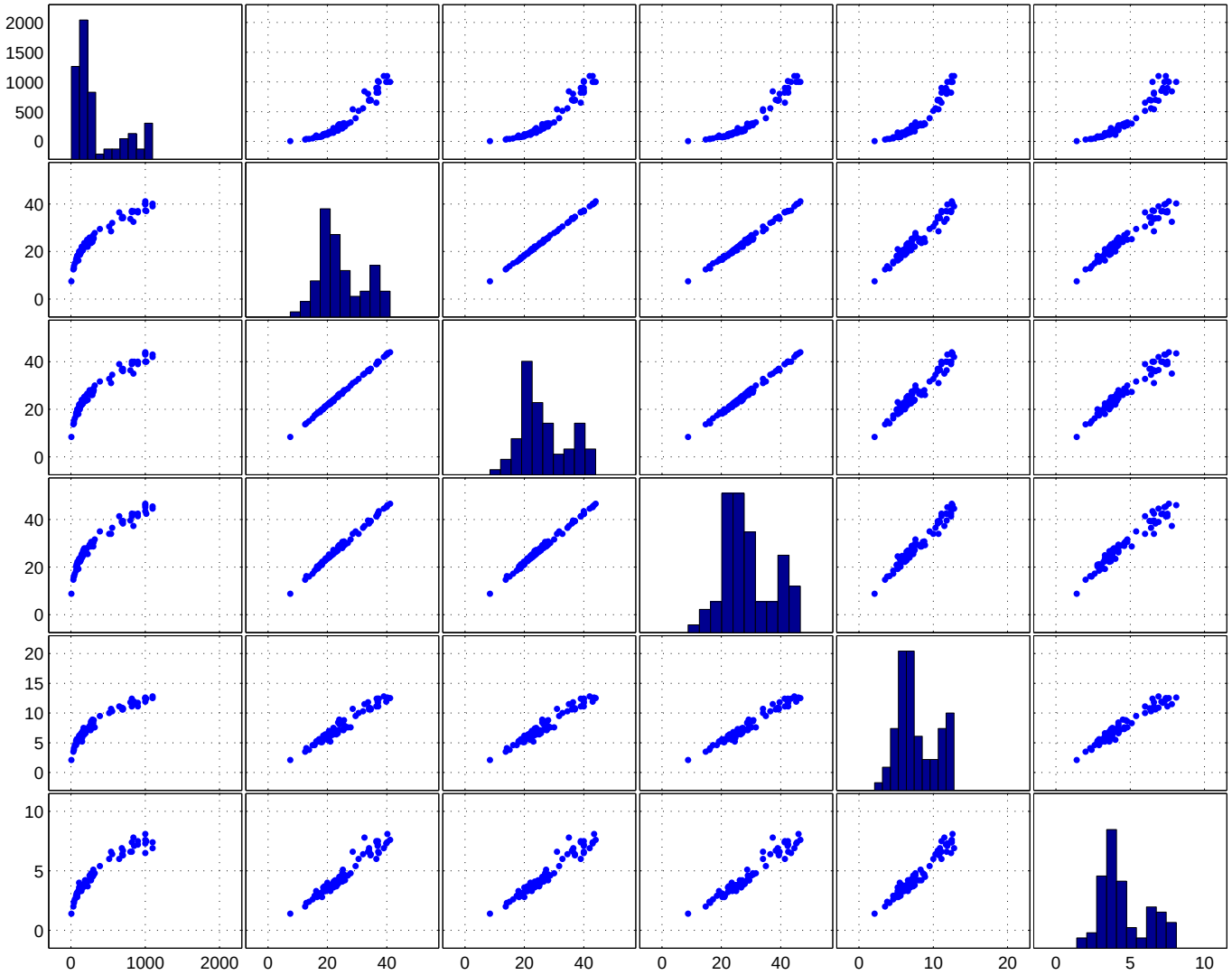


Вес и линейные размеры рыб

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

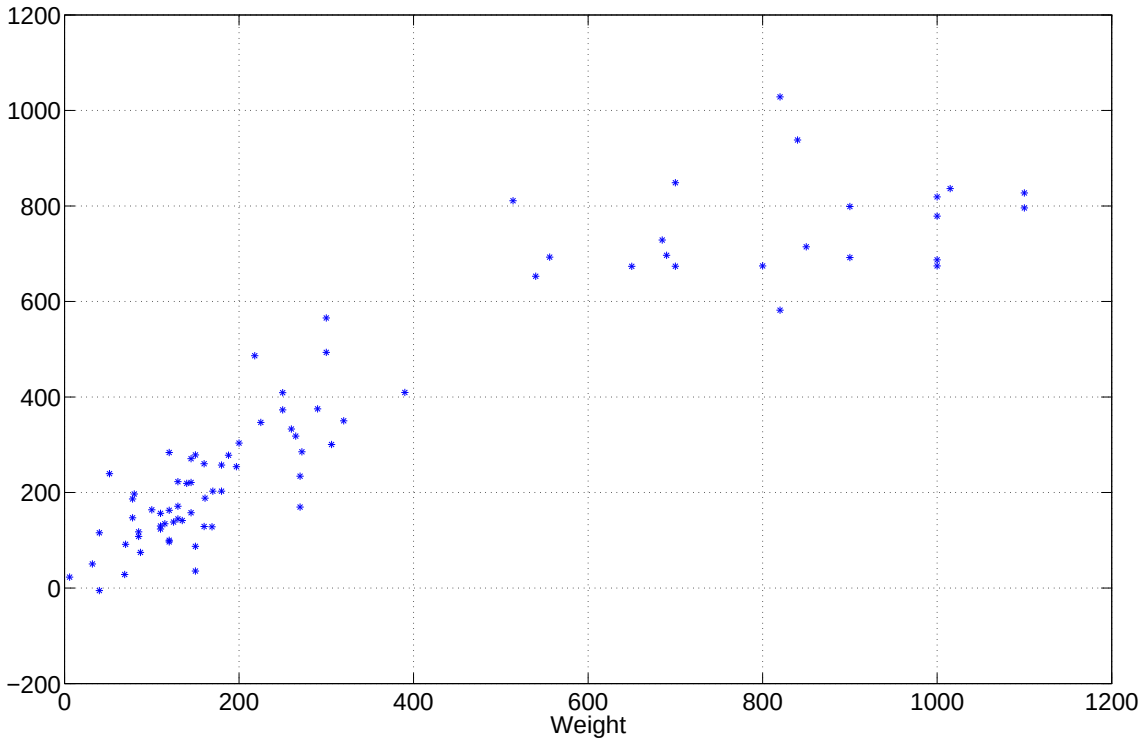


Таблица сопряжённости $K \times L$

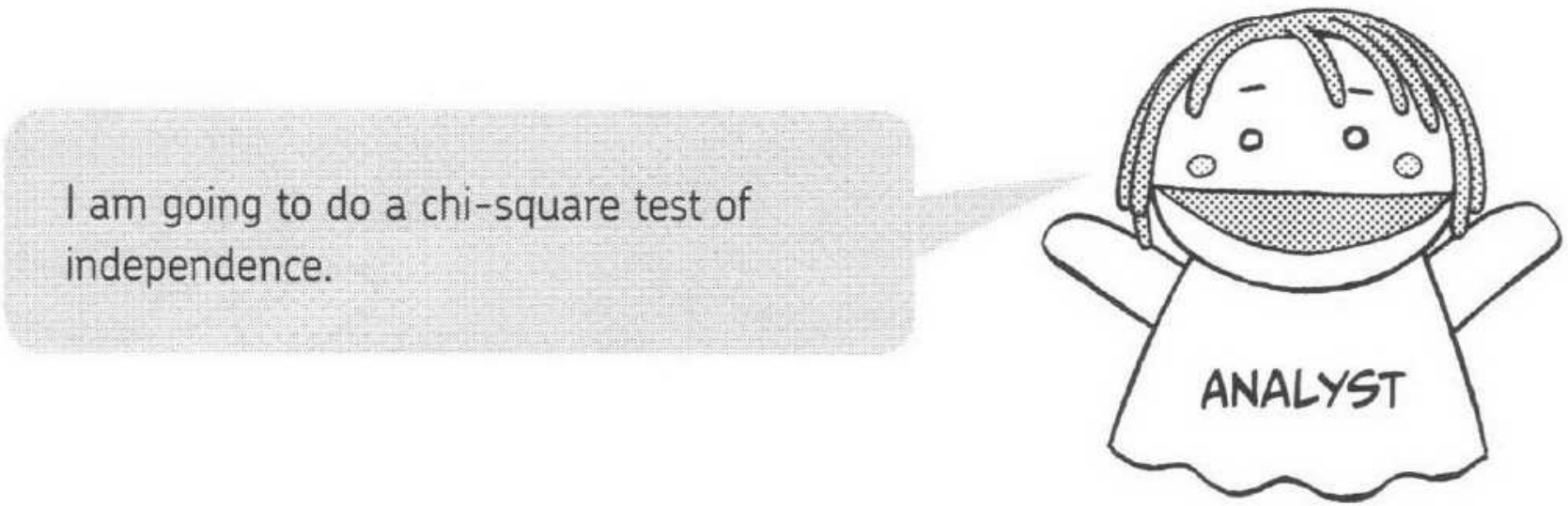
Пусть X и Y — категориальные переменные,
$X \in \{1, \ldots, K\}$, $Y \in \{1, \ldots, L\}$.

X \ Y	1	$\ldots$	l	$\ldots$	L	Σ
1						
$\vdots$						
k			n_{kl}			n_k
$\vdots$						
K						
Σ			n_l			n

$$n_{kl} = \sum_{i=1}^{n} [X_i = k] [Y_i = l],$$

$$n_k = \sum_{l=1}^{L} n_{kl}, \quad n_l = \sum_{k=1}^{K} n_{kl}, \quad n = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{kl}.$$

Критерий хи-квадрат



Критерий хи-квадрат

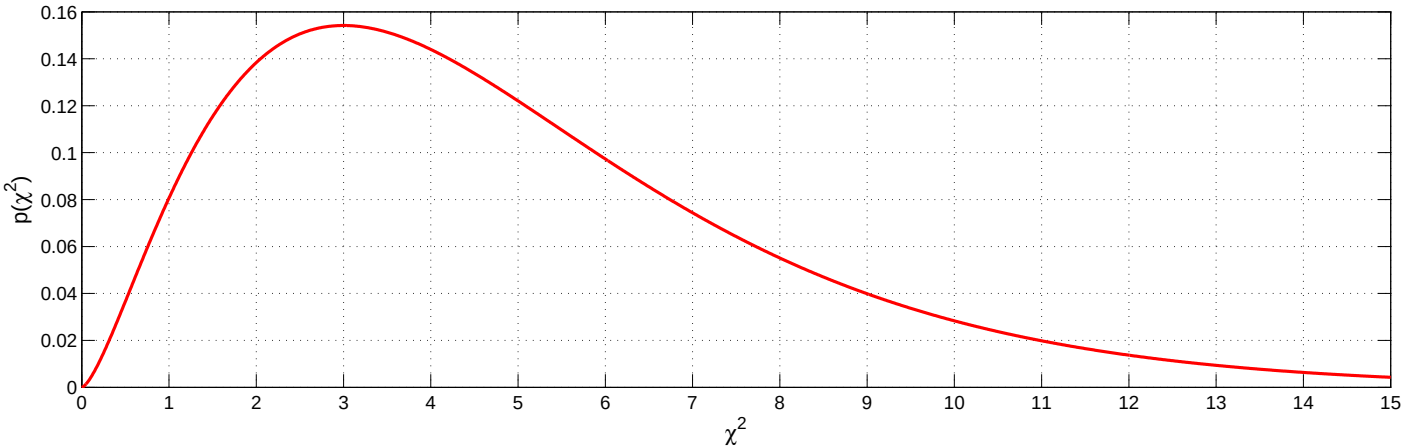
выборки: $X^n = (X_1, \ldots, X_n), \;\; X_i \in \{1, \ldots, K\},$
$Y^n = (Y_1, \ldots, Y_n), \;\; Y_i \in \{1, \ldots, L\},$ выборки связные;

нулевая гипотеза: H_0: X и Y независимы;

альтернатива: H_1: H_0 неверна;

статистика: $\chi^2(X^n, Y^n) = \sum\limits_{k=1}^{K} \sum\limits_{l=1}^{L} \frac{\left(n_{kl} - \frac{n_k n_l}{n}\right)^2}{\frac{n_k n_l}{n}} = n\left(\sum\limits_{k=1}^{K} \sum\limits_{l=1}^{L} \frac{n_{kl}^2}{n_k n_l} - 1\right);$

$\chi^2(X^n, Y^n) \sim \chi^2_{(K-1)(L-1)}$ при H_0;



достигаемый уровень значимости:

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

Критерий хи-квадрат

Условия применимости критерия:

- $n \geq 40$;
- $\frac{n_k n_l}{n} < 5$ не более чем в 20% ячеек.

G-критерий

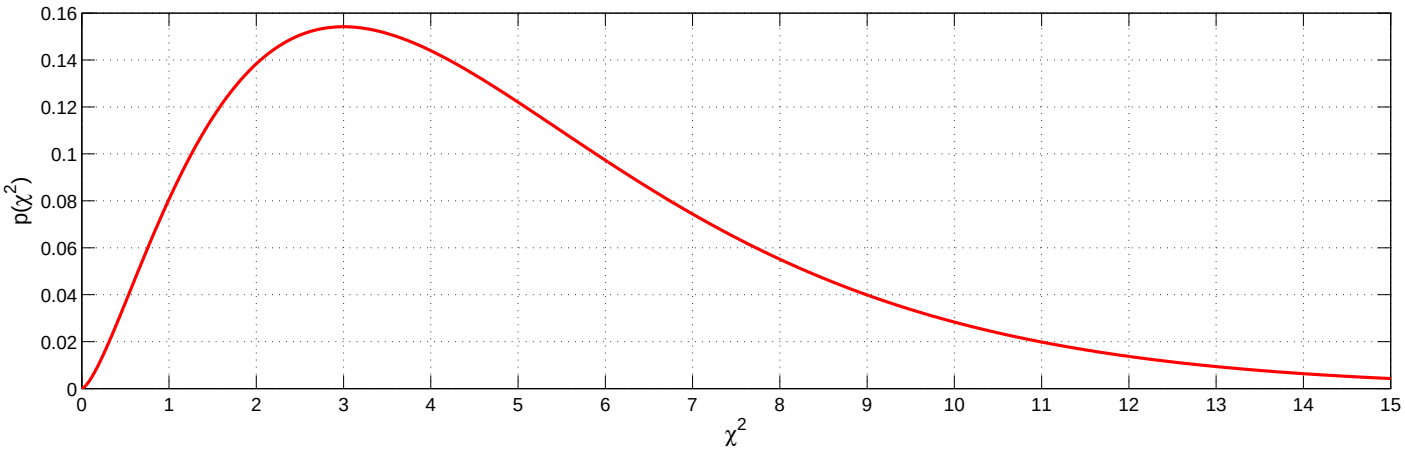
выборки: $X^n = (X_1, \ldots, X_n), \;\; X_i \in \{1, \ldots, K\},$
$Y^n = (Y_1, \ldots, Y_n), \;\; Y_i \in \{1, \ldots, L\},$ выборки связные;

нулевая гипотеза: H_0: X и Y независимы;

альтернатива: H_1: H_0 неверна;

статистика: $G^2\left(X^n, Y^n\right) = 2 \sum\limits_{k=1}^{K} \sum\limits_{l=1}^{L} n_{kl} \ln \frac{n_{kl} n}{n_k n_l}$;

$G^2\left(X^n, Y^n\right) \sim \chi^2_{(K-1)(L-1)}$ при H_0;



достигаемый уровень значимости:

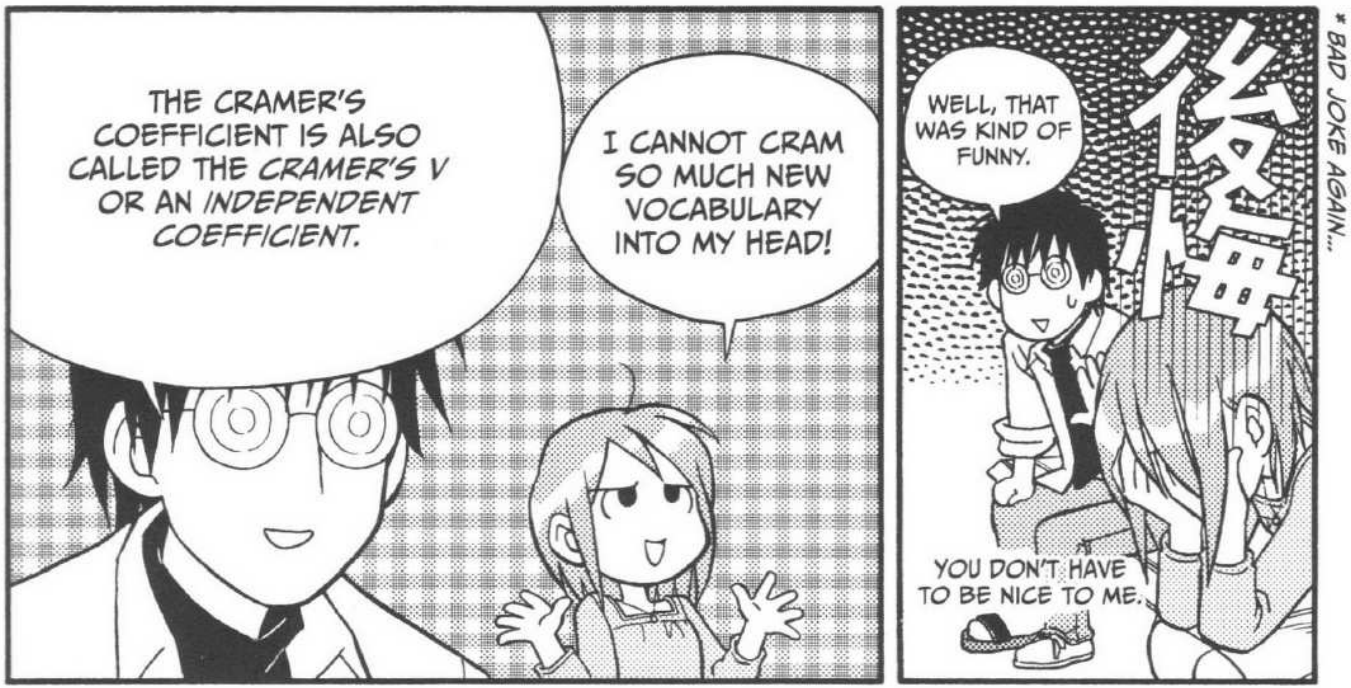
$$p\left(G^2\right) = 1 - chi2cdf\left(g^2, (K-1)(L-1)\right).$$

Коэффициент V Крамера

Мера взаимосвязи между двумя категориальными переменными — коэффициент V Крамера:

$$\phi_c\left(X^n, Y^n\right) = \sqrt{\frac{\chi^2\left(X^n, Y^n\right)}{n\left(\min\left(K, L\right) - 1\right)}}.$$

$\phi_c \in [0, 1]$; 0 соответствует полному отсутствию взаимосвязи, 1 — равенству переменных.



Смертность среди королевских пингвинов

Descamps et al., Relating demographic performance to breeding-site location in the king penguin, 2009: было помечено 50 королевских пингвинов в каждой из трёх областей гнездования на острове Владение архипелага Крозе; через год были собраны данные о том, сколько пингвинов погибло в каждой из групп.

Одинакова ли смертность пингвинов, обитающих в различных областях гнездования?

Смертность среди королевских пингвинов

	Выжило	Погибло
Lower area	43	7
Middle area	44	6
Upper area	49	1

Коэффициент Крамера: $V = 0.1804$.

Критерий хи-квадрат: $p = 0.0872$.
Ожидаемое число наблюдений во всех ячейках второго столбца меньше 5 $\Rightarrow$ применимость критерия под вопросом.

G-критерий: $p = 0.04827$.

Таблица сопряжённости 2×2

Пусть X и Y принимают значения 0 и 1.

X \ Y	0	1	Σ
0	a	b	$a+b$
1	c	d	$c+d$
Σ	$a+c$	$b+d$	n

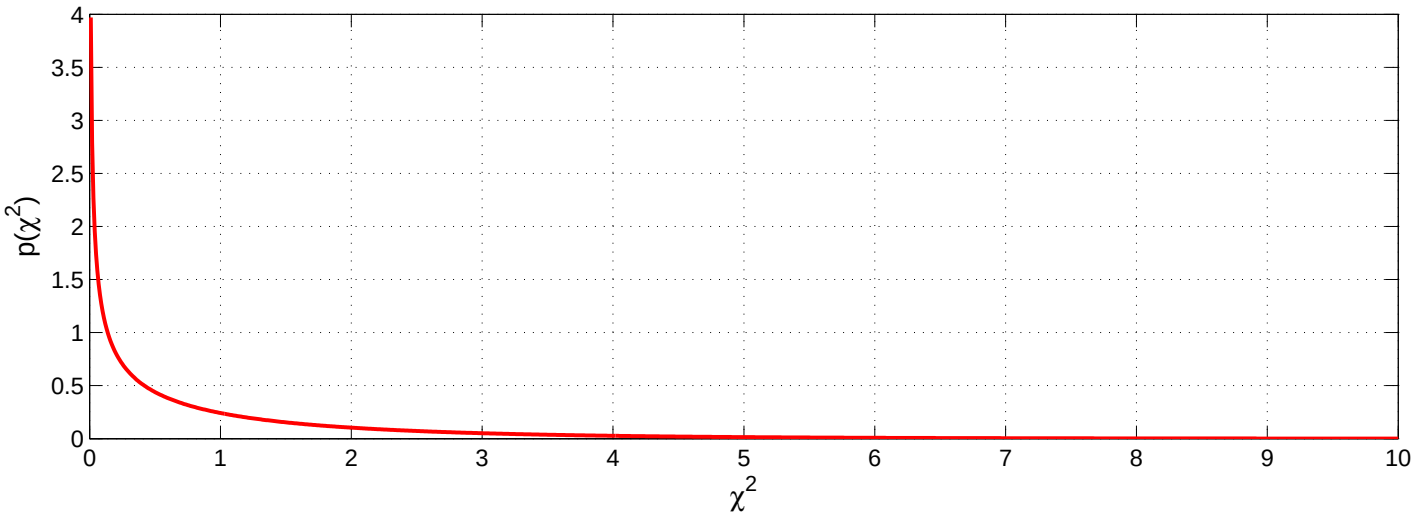
Критерий хи-квадрат

выборки: $X^n = (X_1, \ldots, X_n), \;\; X_i \in \{0, 1\},$
$Y^n = (Y_1, \ldots, Y_n), \;\; Y_i \in \{0, 1\},$ выборки связные;

нулевая гипотеза: H_0: X и Y независимы;

альтернатива: H_1: H_0 неверна;

статистика: $\chi^2(X^n, Y^n) = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$;
$\chi^2(X^n, Y^n) \sim \chi^2_1$ при H_0;



достигаемый уровень значимости:

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

Критерий хи-квадрат

Условия применимости критерия:

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

Критерий хи-квадрат

Пример: собраны данные по двум группам кандидатов в пилоты, для каждого из них имеются результаты двух способов тестирования скорости реакции. Есть ли связь между способом тестирования и числом отобранных кандидатов?
$a = 15, b = 85, c = 4, d = 77.$

H_0: между способом тестирования и числом отобранных кандидатов нет связи.
H_1: способ тестирования как-то связан с числом отобранных кандидатов
$\Rightarrow p = 0.0286.$

Точный критерий Фишера

выборки: $X^n = (X_1, \ldots, X_n)\,,\ \ X_i \in \{0,1\}\,,$
$Y^n = (Y_1, \ldots, Y_n)\,,\ \ Y_i \in \{0,1\}\,,$ выборки связные;

нулевая гипотеза: H_0: X и Y независимы;

альтернатива: H_1: H_0 неверна.

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

$$p\left(X^n, Y^n\right) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}.$$

Достигаемый уровень значимости определяется как сумма по всем возможным вариантам таблицы, имеющим вероятность не больше $p\left(X^n, Y^n\right)$.

Для односторонней альтернативы $(H_1: ad \ll bc)$ достигаемый уровень значимости легко 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: собраны данные по двум группам кандидатов в пилоты, для каждого из них имеются результаты двух способов тестирования скорости реакции. Есть ли связь между способом тестирования и числом отобранных кандидатов?
$a = 9, b = 2, c = 7, d = 6.$

H_0: между способом тестирования и числом отобранных кандидатов нет связи.
H_1: способ тестирования как-то связан с числом отобранных кандидатов
$\Rightarrow p = 0.2108.$

Корреляция Мэтьюса

Мера взаимосвязи между двумя бинарными переменными — коэффициент корреляции Мэтьюса:

$$|MCC| = \sqrt{\frac{\chi^2\left(X^n, Y^n\right)}{n}},$$

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

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

Парадокс хи-квадрат (Симпсона)

Charig et al., Comparison of treatment of renal calculi by operative surgery, percutaneous nephrolithotomy, and extracorporeal shock wave lithotripsy, 1986: по числу успешных исходов сравнивались два метода удаления камней в почках.

Пациенты с одним камнем <2 см

	Открытая хирургия	Чрескожная нефролитотомия
Успешный исход	81	234
Неудачный исход	6	36
Доля успешных исходов	**93%**	87%

Критерий хи-квадрат: $p = 0.1051$.

Пациенты с крупным камнем или несколькими камнями

	Открытая хирургия	Чрескожная нефролитотомия
Успешный исход	192	55
Неудачный исход	71	25
Доля успешных исходов	**73%**	69%

Критерий хи-квадрат: $p = 0.4580$.

Парадокс хи-квадрат (Симпсона)

Все пациенты вместе:

	Открытая хирургия	Чрескожная нефролитотомия
Успешный исход	273	289
Неудачный исход	77	61
Доля успешных исходов	78%	**83%**

Критерий хи-квадрат: $p = 0.1285$.

Парадокс хи-квадрат (Симпсона)

	Открытая хирургия	Чрескожная нефролитотомия	χ^2 p-value
Один камень <2 см	*Группа1* **93% (81/87)**	*Группа2* 87% (234/270)	0.1051
Крупный камень или несколько камней	*Группа3* **73% (192/263)**	*Группа4* 69% (55/80)	0.4580
Суммарно	78% (273/350)	**83% (289/350)**	0.1285

Причины несогласованности выводов:

- размеры групп 1-4 и 2-3 чересчур сильно отличаются, суммарный вывод определяется в основном вкладом групп 2-3;
- дополнительная переменная — размер камня — оказывает большее влияние на результат операции, чем выбор метода;
- все различия статистически незначимы, перед нами случайные колебания.

Парадокс хи-квадрат (Симпсона)

Эксперимент: пациенты принимают препарат или плацебо, по окончании курса определяется, выздоровели они или нет.
Есть ли зависимость 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: $\chi^2 = 0.3759, p = 0.5398$.

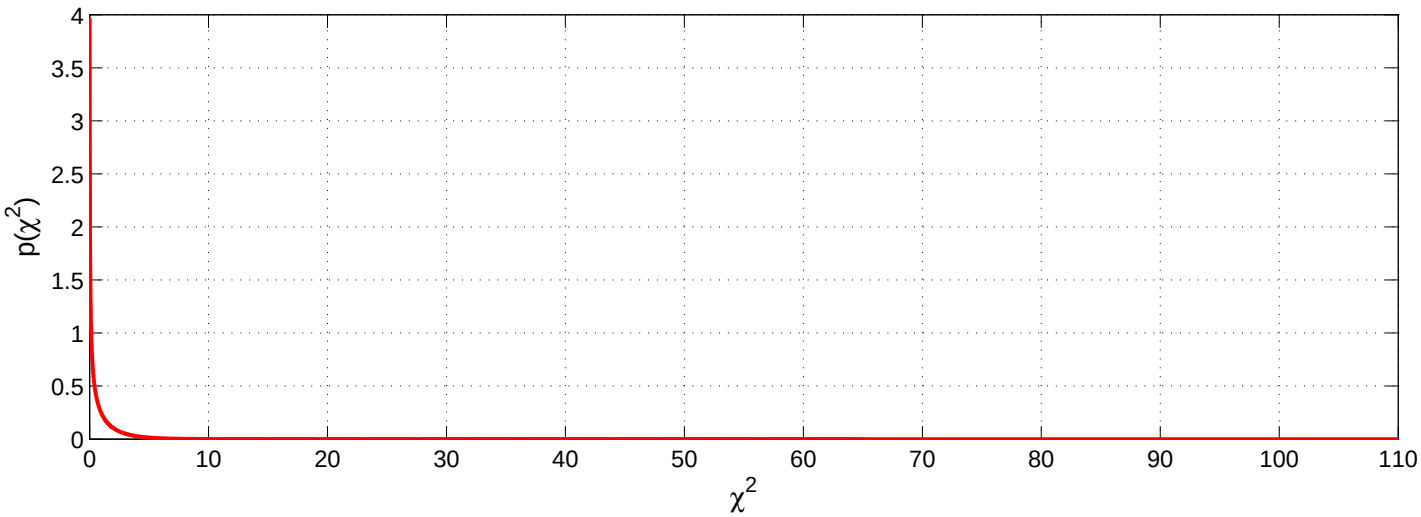
Парадокс хи-квадрат (Симпсона)

Bickel 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: $\chi^2 = 108.1, \quad p \approx 0.$



	Наблюдаемые		Ожидаемые		Разности	
	-	+	-	+	-	+
Мужчины	4704	3738	4981.3	3460.7	-277.3	277.3
Женщины	2827	1494	2549.7	1771.3	277.3	-277.3

Парадокс хи-квадрат (Симпсона)

Будем искать виноватых: посмотрим детализированную статистику по 85 факультетам.

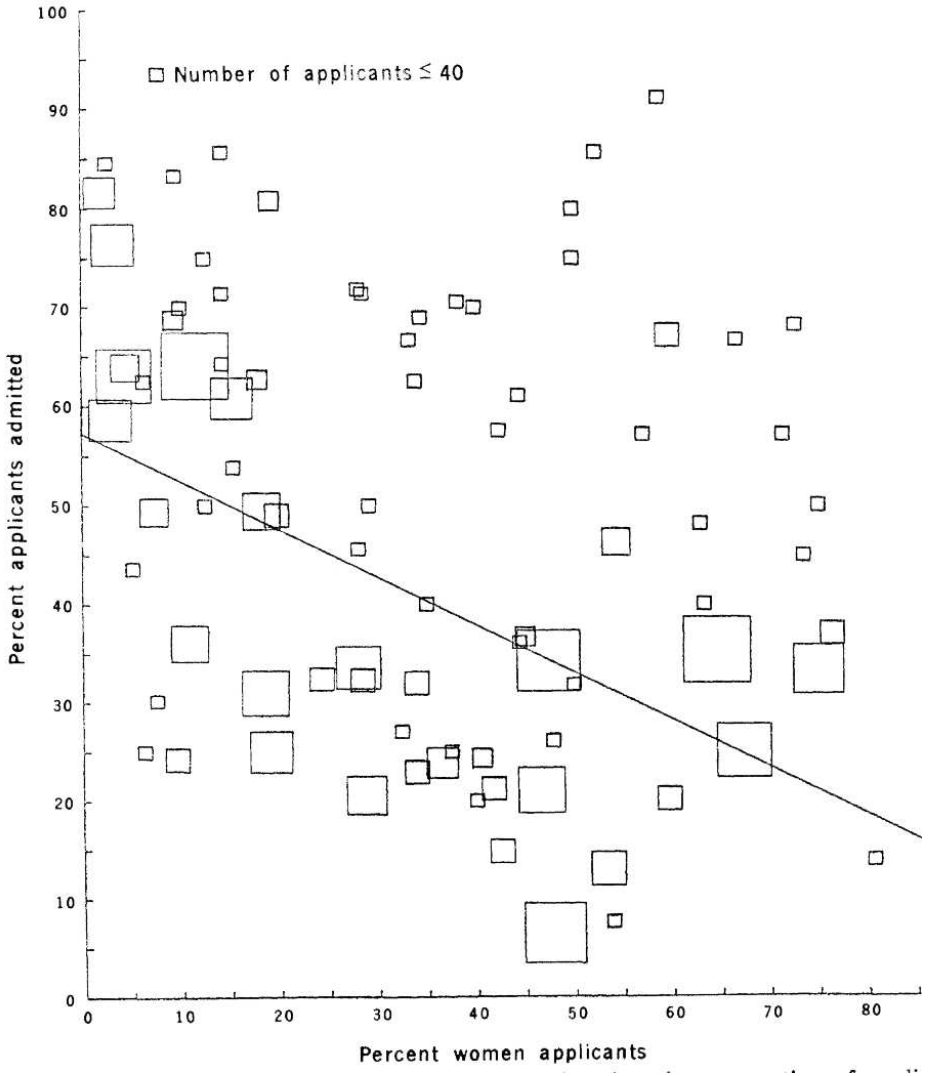
Значимо (на уровне $\alpha = 0.05$) меньше женщин прошли отбор на 4 факультета, суммарный дефицит — 26.
На 6 факультетов поступило значимо меньше мужчин, суммарный дефицит — 64.

Данные по шести крупнейшим факультетам:

	Мужчины		Женщины	
	Σ	$+$	Σ	$+$
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 пытаются поступить на факультеты с большим конкурсом.



Прикладная статистика
5. Корреляционный анализ.

Рябенко Евгений
riabenko.e@gmail.com