Прикладная статистика 6. Анализ зависимостей.

24 марта 2014 г.

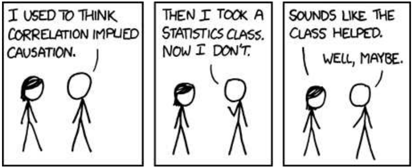
Задача исследования взаимосвязи между признаками

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?

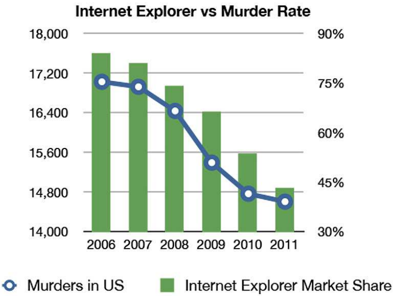
Статистическая взаимосвязь между случайными величинами — **корреляция**.

Корреляция и причинность



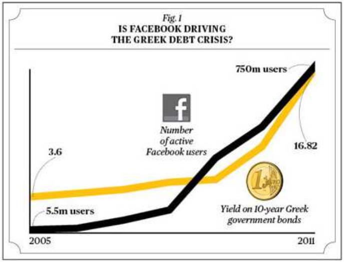
Корреляция — мера ассоциативной связи (одновременная встречаемость событий, сходство паттернов). Никакого отношения к причинно-следственной связи она не имеет!

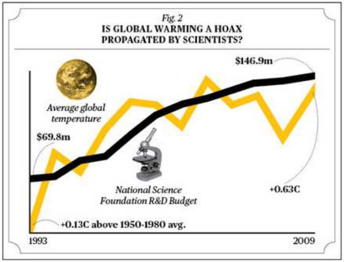
Пример:



Корреляция и причинность

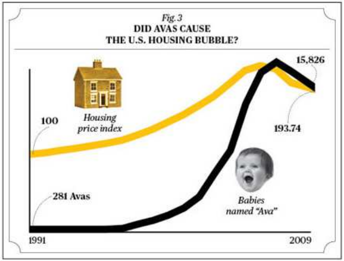
Другие примеры:

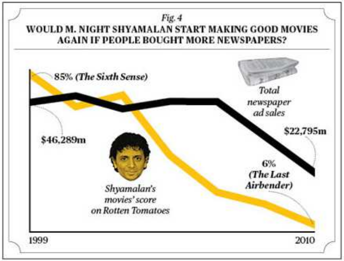




Корреляция и причинность

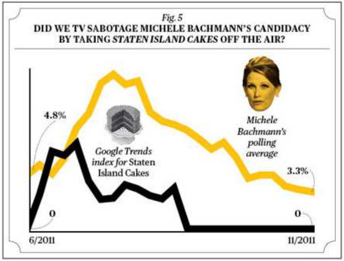
Другие примеры:

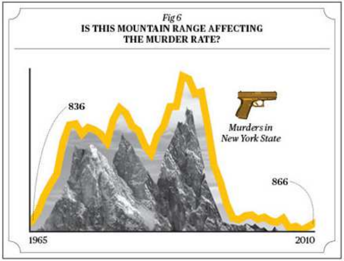




Корреляция и причинность

Другие примеры:





Корреляция Пирсона

Корреляция Пирсона (Pearson product-moment correlation coefficient):

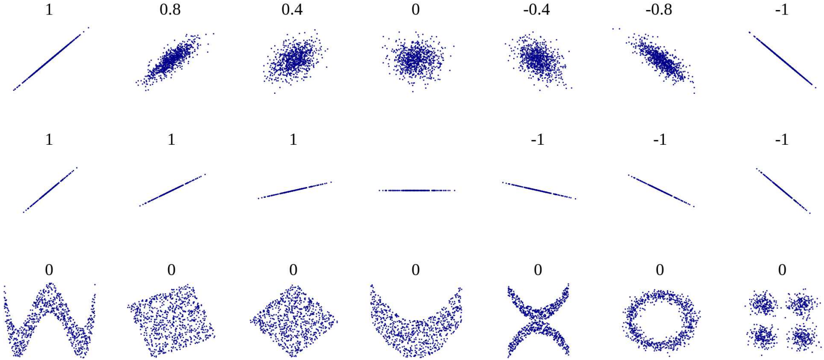
$$r_{XY} = \frac{\operatorname{cov}(X_1, X_2)}{\sqrt{\mathbb{D}X_1 \mathbb{D}X_2}} = \frac{\mathbb{E}\left(\left(X_1 - \mathbb{E}X_1\right)\left(X_2 - \mathbb{E}X_2\right)\right)}{\sqrt{\mathbb{D}X_1 \mathbb{D}X_2}}.$$

Выборочный коэффициент корреляции Пирсона:

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

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

Корреляция Пирсона



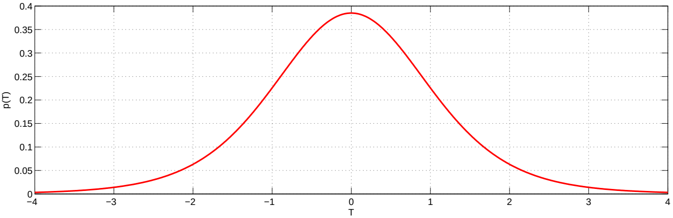
Критерий Стьюдента

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

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

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

статистика: $T(X_1^n, X_2^n) = \frac{r_{X_1X_2}\sqrt{n-2}}{\sqrt{1-r^2_{X_1X_2}}}$;
$T(X_1^n, X_2^n) \sim St(n-2)$ при H_0;



достигаемый уровень значимости:

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

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:

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

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?

$H_0: r_{X_1 X_2} = 0.$
$H_1: r_{X_1 X_2} \neq 0 \Rightarrow p = 0.0847.$

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

Перестановочный критерий

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

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

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

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

$$|G| = n!$$

Достигаемый уровень значимости:

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

Перестановочный критерий

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

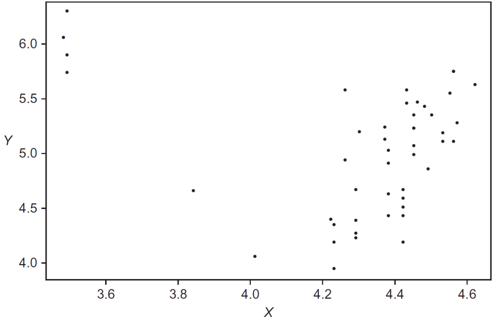
Пример: в предыдущем примере

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

Недостатки

Недостатки выборочного коэффициента Пирсона:

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



Корреляция между логарифмами эффективной температуры на поверхности звезды (X) и интенсивности её света (Y) получается отрицательной $(r_{XY} = -0.21)$ из-за наличия в выборке красных гигантов.

Корреляция Спирмена

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

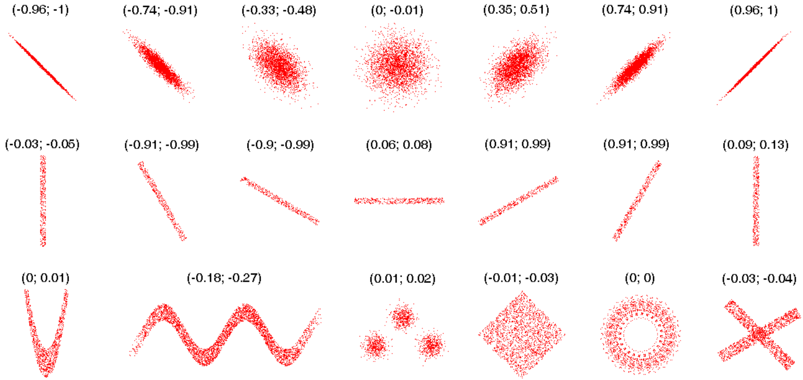
$$\rho_{X_1 X_2} = \frac{\sum\limits_{i=1}^{n} \left(r\left(X_{1i}\right) - \frac{n+1}{2}\right)\left(r\left(X_{2i}\right) - \frac{n+1}{2}\right)}{\frac{1}{12}\left(n^3 - n\right)} =$$

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

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

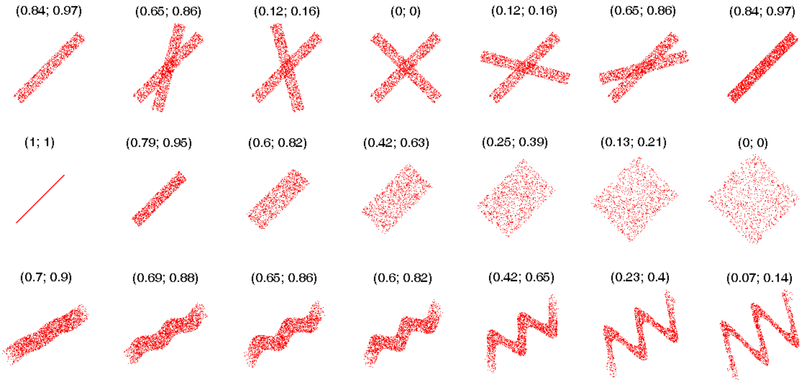
$\rho_{X_1 X_2} \in [-1, 1]$ — мера **монотонной** связи.

Корреляция Спирмена



Корреляция Спирмена — второе число в скобках.

Корреляция Спирмена



Корреляция Спирмена — второе число в скобках.

Критерий Стьюдента

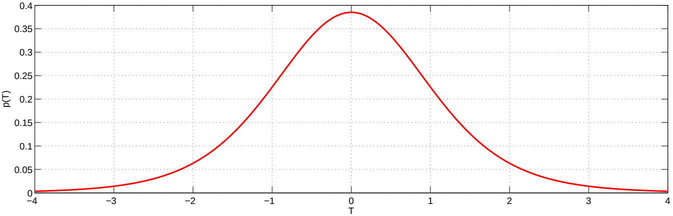
выборки: $X_1^n = (X_{11}, \dots, X_{1n})$,
$X_2^n = (X_{21}, \dots, X_{2n})$, выборки связные;

нулевая гипотеза: $H_0\colon \rho_{X_1X_2} = 0$;

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

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

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



достигаемый уровень значимости:

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

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

$H_0\colon \rho_{X_1X_2} = 0.$
$H_1\colon \rho_{X_1X_2} \neq 0 \Rightarrow p = 0.0024.$

Корреляция Кендалла

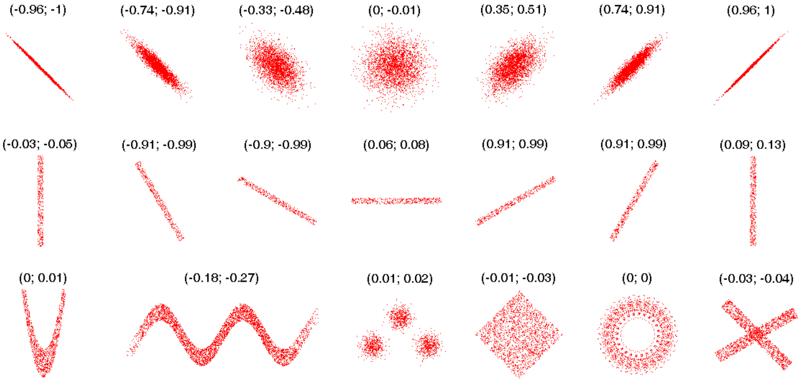
Коэффициент корреляции Кендалла — мера взаимной неупорядоченности X_1^n и X_2^n:

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

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

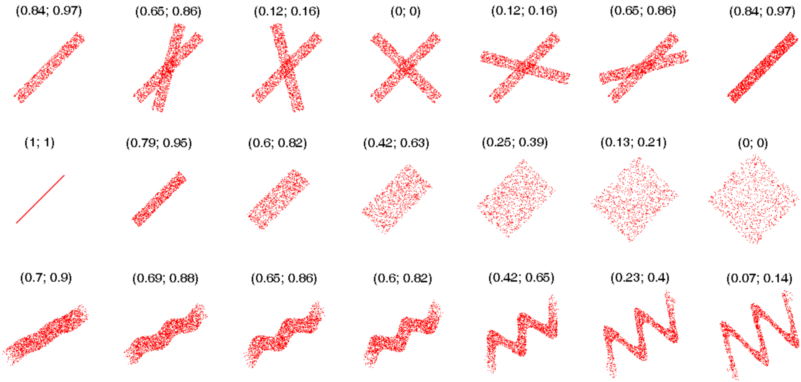
$\tau_{X_1 X_2} \in [-1, 1]$ — мера **монотонной** связи.

Корреляция Кендалла



Корреляция Кендалла — первое число в скобках.

Корреляция Кендалла



Корреляция Кендалла — первое число в скобках.

Критерий без названия

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

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

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

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

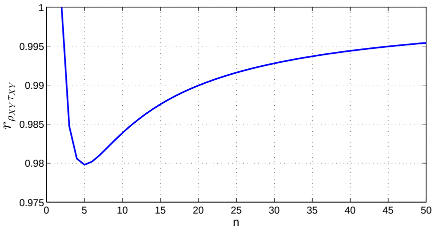
Пример: налоговый инспектор хочет проверить наличие взаимосвязи между величинами общего дохода от инвестиций и общего объёма дополнительных доходов. На выборке из 10 налоговых деклараций он получил $D = 5,\ C = 38,\ \tau_{X_1 X_2} = 0.7821.$

$H_0\colon \tau_{X_1 X_2} = 0.$
$H_1\colon \tau_{X_1 X_2} \neq 0 \Rightarrow p = 0.0027.$

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 выборочных 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

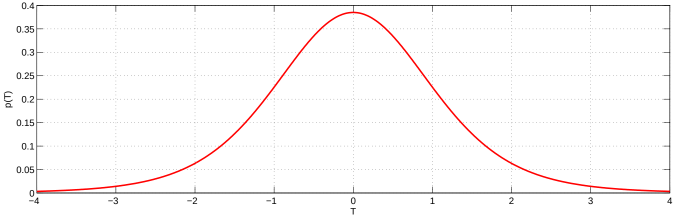
выборки: $X_1^n = (X_{11}, \dots, X_{1n}), X_2^n = (X_{21}, \dots, X_{2n}),$
$X_3^n = (X_{31}, \dots, X_{3n}), X_{3i} \in \mathbb{R}^M,$
$(X_{1i}, X_{2i}, X_{3i}) \sim N(\mu, \Sigma);$

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

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

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:

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

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 выборочных корреляций, $R = \Omega^{-1}$, c — вектор корреляций основного признака X с дополнительными; тогда

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

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

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

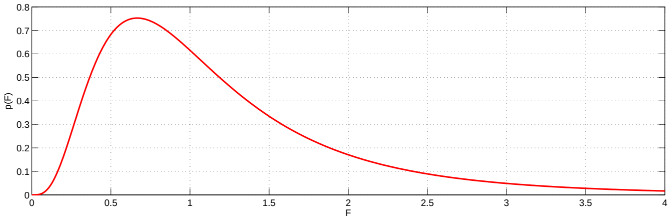
Критерий Фишера

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: $H_0\colon r_{X_1, X_2} = 0$;

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

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;

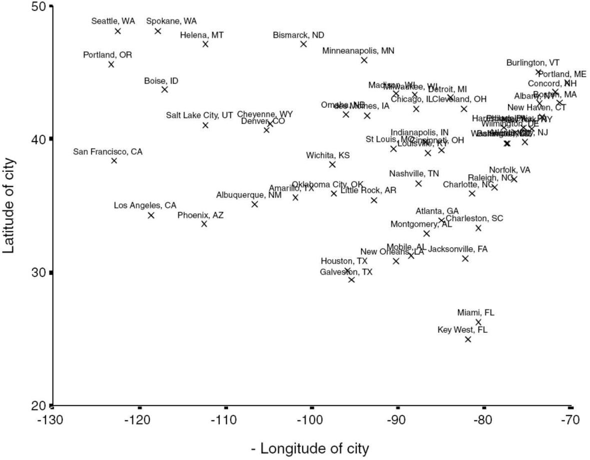


достигаемый уровень значимости:

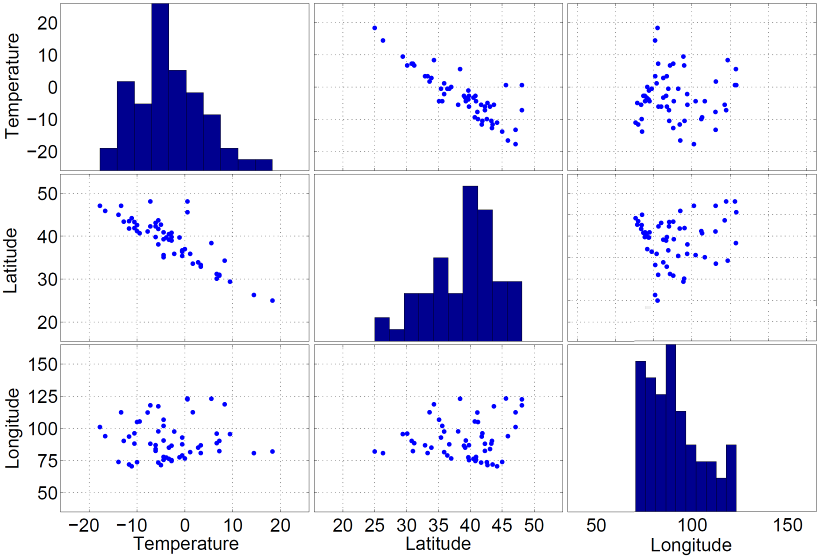
$$p(f) = 1 - fcdf(f, M-2, n-M-1).$$

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**	–

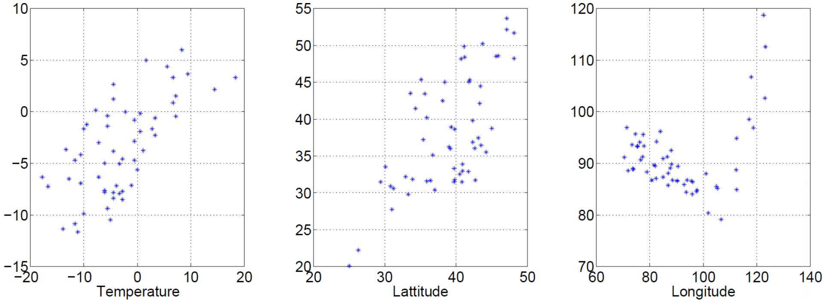
ρ	T	ϕ	λ
T	–	**0.000**	0.543
ϕ	**0.000**	–	0.552
λ	0.543	0.552	–

Температура воздуха и географическое положение

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

Коэффициенты множественной корреляции:

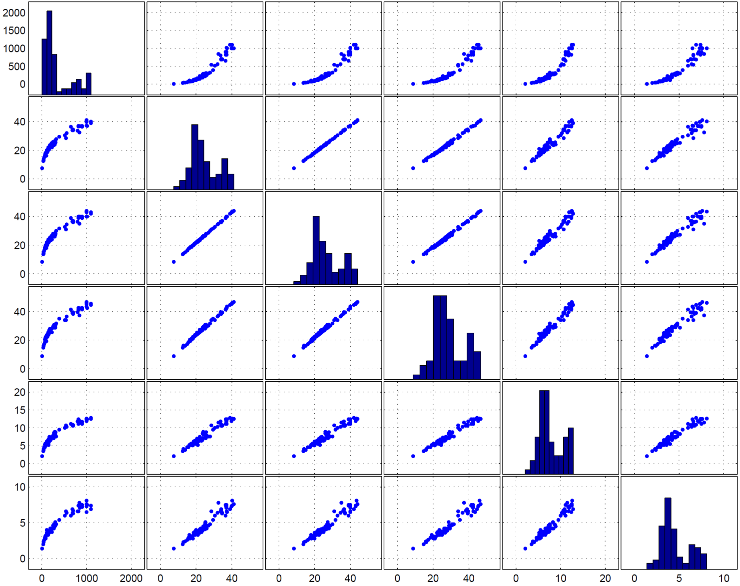
	T	ϕ	λ
R	0.659	0.667	0.312
p	6.0347×10^{-8}	3.6481×10^{-8}	0.0216
with	$0.235 \cdot \lambda - 0.638 \cdot \phi$	$0.397 \cdot \lambda - 0.678 \cdot T$	$1.542 \cdot T + 2.450 \cdot \phi$



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	-	0.958	0.958	0.955	0.953	0.958
Length1	0.958	-	0.999	0.997	0.975	0.972
Length2	0.958	0.999	-	0.997	0.975	0.972
Length3	0.955	0.997	0.997	-	0.982	0.971
Width	0.953	0.975	0.975	0.982	-	0.973
Thickness	0.958	0.972	0.972	0.971	0.973	-

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	-	0.913	0.916	0.918	0.881	0.883
Length1	0.913	-	0.983	0.961	0.847	0.856
Length2	0.916	0.983	-	0.952	0.844	0.863
Length3	0.918	0.961	0.952	-	0.872	0.854
Width	0.881	0.847	0.844	0.872	-	0.822
Thickness	0.883	0.856	0.863	0.854	0.822	-

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	-	0.987	0.986	0.988	0.974	0.972
Length1	0.987	-	0.998	0.996	0.963	0.964
Length2	0.986	0.998	-	0.993	0.959	0.966
Length3	0.988	0.996	0.993	-	0.972	0.964
Width	0.974	0.963	0.959	0.972	-	0.944
Thickness	0.972	0.964	0.966	0.964	0.944	-

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	-	0.046	0.088	−0.205	**0.240**	**0.237**
Length1	0.046	-	**0.888**	**0.281**	−0.148	0.032
Length2	0.088	**0.888**	-	0.174	−0.096	0.131
Length3	−0.205	**0.281**	0.174	-	**0.656**	**−0.253**
Width	**0.240**	−0.148	−0.096	**0.656**	-	**0.469**
Thickness	**0.237**	0.032	0.131	**−0.253**	**0.469**	-

Достигаемые уровни значимости:

p	Weight	Length1	Length2	Length3	Width	Thickness
Weight	-	0.697	0.447	0.073	**0.035**	**0.038**
Length1	0.693	-	**0.000**	**0.013**	0.198	0.784
Length2	0.447	**0.000**	-	0.129	0.407	0.255
Length3	0.073	**0.013**	0.129	-	**0.000**	**0.027**
Width	**0.035**	0.198	0.407	**0.000**	-	**0.000**
Thickness	**0.038**	0.784	0.255	**0.027**	**0.000**	-

Вес и линейные размеры рыб

Частные корреляции Спирмена:

ρ	Weight	Length1	Length2	Length3	Width	Thickness
Weight	-	0.063	0.104	0.096	**0.429**	**0.394**
Length1	0.063	-	**0.852**	**0.518**	−0.051	−0.131
Length2	0.104	**0.852**	-	−0.094	−0.143	0.189
Length3	0.096	**0.518**	−0.094	-	**0.405**	0.033
Width	**0.429**	−0.051	−0.143	**0.405**	-	−0.017
Thickness	**0.394**	−0.131	0.189	0.033	−0.017	-

Достигаемые уровни значимости:

p	Weight	Length1	Length2	Length3	Width	Thickness
Weight	-	0.589	0.367	0.408	**0.000**	**0.000**
Length1	0.589	-	**0.000**	**0.000**	0.660	0.256
Length2	0.367	**0.000**	-	0.416	0.214	0.100
Length3	0.408	**0.000**	0.416	-	**0.000**	0.777
Width	**0.000**	0.660	0.214	**0.000**	-	0.887
Thickness	**0.000**	0.256	0.100	0.777	0.887	-

Вес и линейные размеры рыб

Множественная корреляция всех признаков с весом: $R = 0.921$ ($p \approx 0$).
Максимизирующая корреляцию линейная комбинация: $292.25 \cdot Length1 - 151.16 \cdot Length2 - 151 \cdot Length3 + 148.89 \cdot Width + 83.43 \cdot Thickness$.

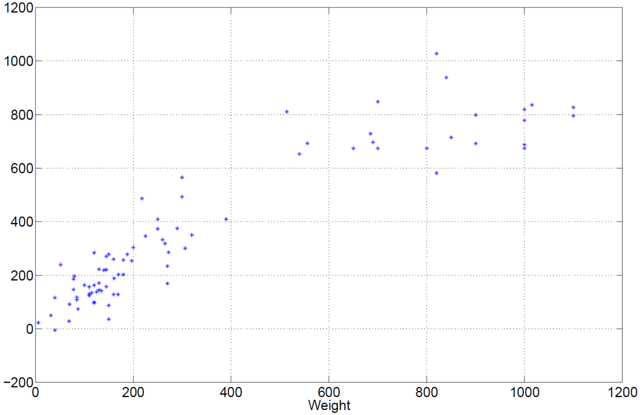


Таблица сопряжённости $K_1 \times K_2$

Имеются связные выборки $X_1^n = (X_{11}, \dots, X_{1n})$ и $X_2^n = (X_{21}, \dots, X_{2n})$, $X_{1k} \in \{1, \dots, K_1\}$, $X_{2k} \in \{1, \dots, K_2\}$, $k = 1, \dots, n$.

Таблица сопряжённости:

$X_1 \backslash X_2$	1	$\dots$	j	$\dots$	K_2	$\sum$
1						
$\vdots$						
i			n_{ij}			n_{i+}
$\vdots$						
K_1						
$\sum$			n_{+j}			n

Два случайных признака

Пусть π_{ij} — вероятность реализации пары (X_{1k}, X_{2k}) в ячейке (i, j).
$\{\pi_{ij}\}$ — совместное распределение (X_{1k}, X_{2k});
$\{\pi_{i+}\}, \{\pi_{+j}\}$ — маргинальные распределения.

X_1 и X_2 независимы, если

$$\pi_{ij} = \pi_{i+}\pi_{+j} \ \ \forall i = 1, \dots, K_1, j = 1, \dots, K_2.$$

Один случайный признак

Пусть X_{1k} — не случайная величина, а фиксированный признак. Тогда $\{\pi_{ij}\}$ не имеет смысла, вместо него рассматриваются $\left\{\pi_{1|i},\dots,\pi_{K_1|i}\right\}$ — условные распределения X_2 при $X_1=i$.

X_1 и X_2 независимы, если

$$\pi_{j|1}=\dots=\pi_{j|K_1}\quad\forall j=1,\dots,K_2.$$

Порождающие модели

1. Если все ячейки таблицы случайны, то распределение n_{ij} может быть, например, пуассоновским со средними μ_{ij}; совместная функция вероятности таблицы:
$$\prod_i \prod_j \frac{e^{-\mu_{ij}} \mu_{ij}^{n_{ij}}}{n_{ij}!}.$$
2. Если суммарный объём выборки n фиксирован, данные описываются мультиномиальной моделью:
$$\frac{n!}{n_{11}! \cdot \ldots \cdot n_{K_1 K_2}!} \prod_i \prod_j \pi_{ij}^{n_{ij}}.$$
3. Если X_1 не случайна, то фиксированы суммы по строкам n_{i+}, и каждая строка i порождается отдельной мультиномиальной моделью:
$$\frac{n_{i+}!}{\prod_j n_{ij}!} \prod_j \pi_{i|j}^{n_{ij}}.$$

Порождающие модели

Исследование: как исход автомобильной аварии на заданной магистрали X_1 (смертельный, несмертельный) зависит от использования ремня безопасности X_2 (был использован, не был использован)?

Исход \ Ремень	использован	не использован
смертельный		
несмертельный		

1. Исследователи собираются учесть все автомобильные аварии, которые произойдут на магистрали в течение года.
2. Исследователи запросят 200 случайных полицейских рапортов об авариях за последние годы.
3. Исследователи запросят 200 случайных полицейских рапортов об авариях за последние годы: 100 об авариях со смертельным исходом и 100 об авариях без смертельного исхода.

Таблица сопряжённости 2×2

Пусть X_{1k} и X_{2k} принимают значения 0 и 1.

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

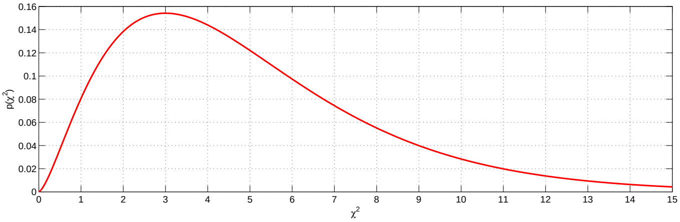
Критерий хи-квадрат

выборки: $X_1^n = (X_{11}, \dots, X_{1n})$, $X_{1k} \in \{1, \dots, K_1\}$,
$X_2^n = (X_{21}, \dots, X_{2n})$, $X_{2k} \in \{1, \dots, K_2\}$,
выборки связные;

нулевая гипотеза: H_0: X_1 и X_2 независимы;

альтернатива: H_1: H_0 неверна;

статистика: $$\chi^2\left(X_1^n, X_2^n\right) = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}} = n\left(\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\frac{n_{ij}^2}{n_{i+}n_{+j}} - 1\right);$$
$\chi^2\left(X_1^n, X_2^n\right) \sim \chi^2_{(K_1-1)(K_2-1)}$ при H_0;



достигаемый уровень значимости:
$$p\left(\chi^2\right) = 1 - chi2cdf\left(\chi^2, (K_1-1)(K_2-1)\right).$$

Критерий хи-квадрат

Условия применимости критерия:

- $n \geq 40$;
- $\frac{n_{i+} n_{+j}}{n} < 5$ не более, чем в 20% ячеек.

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

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

Точный критерий Фишера

выборки: $X_1^n = (X_{11}, \dots, X_{1n})\,, X_{1k} \in \{0, 1\}$,
$X_2^n = (X_{21}, \dots, X_{2n})\,, X_{2k} \in \{0, 1\}$,
выборки связные;

нулевая гипотеза: $H_0 \colon X_1$ и X_2 независимы;

альтернатива: $H_1 \colon H_0$ неверна.

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

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

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

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

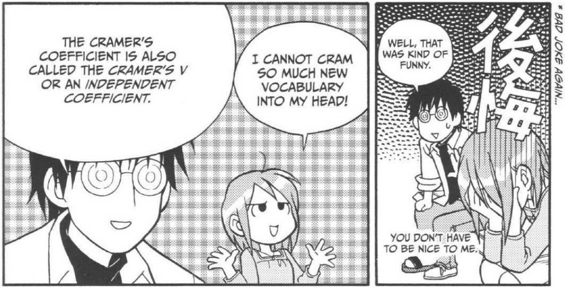
H_0: связи нет.
H_1: признаки связаны $\Rightarrow p = 0.0014$.

Коэффициент V Крамера

Мера взаимосвязи между двумя категориальными переменными — коэффициент V Крамера:

$$\phi_c\left(X_1^n, X_2^n\right) = \sqrt{\frac{\chi^2\left(X_1^n, X_2^n\right)}{n\left(\min\left(K_1, K_2\right) - 1\right)}}.$$

$\phi_c\left(X_1^n, X_2^n\right) \in [0, 1]$; 0 соответствует полному отсутствию взаимосвязи, 1 — совпадению переменных.



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.

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.

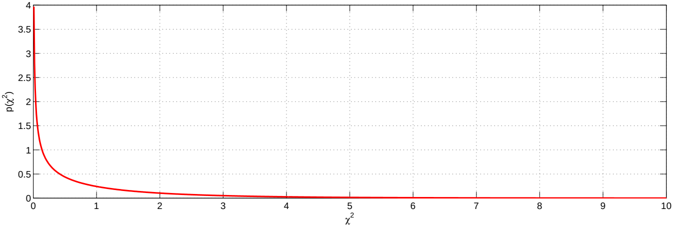
Парадокс хи-квадрат (Симпсона)

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.

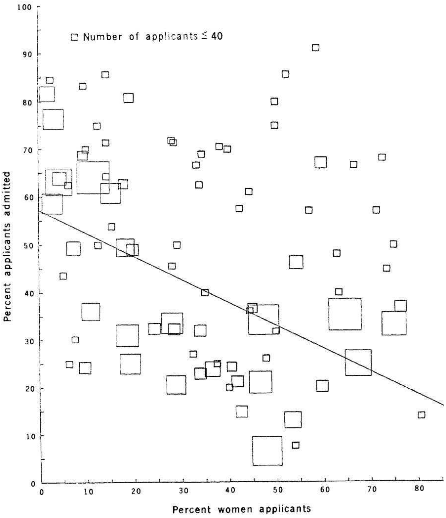
Значимо (при $\alpha = 0.05$) меньше женщин прошли отбор на 4 факультета, суммарный дефицит по ним — 26.
На 6 факультетов поступило значимо меньше мужчин, суммарный дефицит — 64.

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.



Литература

- непрерывные признаки — Лагутин, глава 20;
- категориальные признаки — Agresti, главы 2 и 3;
- значимость корреляции Пирсона — Kanji, №12;
- значимость корреляции Кендалла и Спирмена — Кобзарь, 5.2.2.2.1, 5.2.2.2.2;
- значимость частной и множественной корреляций — Кобзарь, 5.2.1.3.

Кобзарь А.И. *Прикладная математическая статистика*. — М.: Физматлит, 2006.
Лагутин М.Б. *Наглядная математическая статистика*. — Москва: Бином, 2007.
Agresti A. *Categorical Data Analysis*. — Hoboken: Wiley, 2002.
Kanji G.K. *100 statistical tests*. — London: SAGE Publications, 2006.

Прикладная статистика
6. Анализ зависимостей.

Рябенко Евгений
riabenko.e@gmail.com