

# Прикладной статистический анализ данных

## Анализ зависимостей

Олег Бахтеев  
psad@phystech.edu

2021

## Задача исследования взаимосвязи между признаками

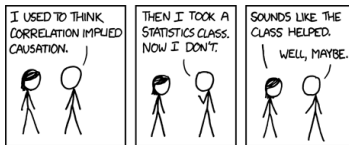
Дано: значения признаков  $X_1, X_2$  измерены на объектах  $1, \dots, n$ .

Эквивалентная формулировка: имеются связанные выборки  $X_1^n = (X_{11}, \dots, X_{1n})$  и  $X_2^n = (X_{21}, \dots, X_{2n})$ .

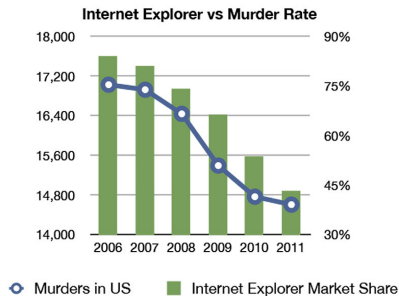
Насколько сильно признаки  $X_1, X_2$  связаны между собой?

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

# Корреляция и причинность



**Корреляция** — статистическая взаимосвязь между случайными величинами; не является достаточным условием причинно-следственной:



Другие примеры: <http://www.tylervigen.com/>

# Корреляция Пирсона

**Коэффициент корреляции Пирсона**  $r_{X_1 X_2}$  случайных величин  $X_1$  и  $X_2$  — мера силы **линейной** корреляции между ними:

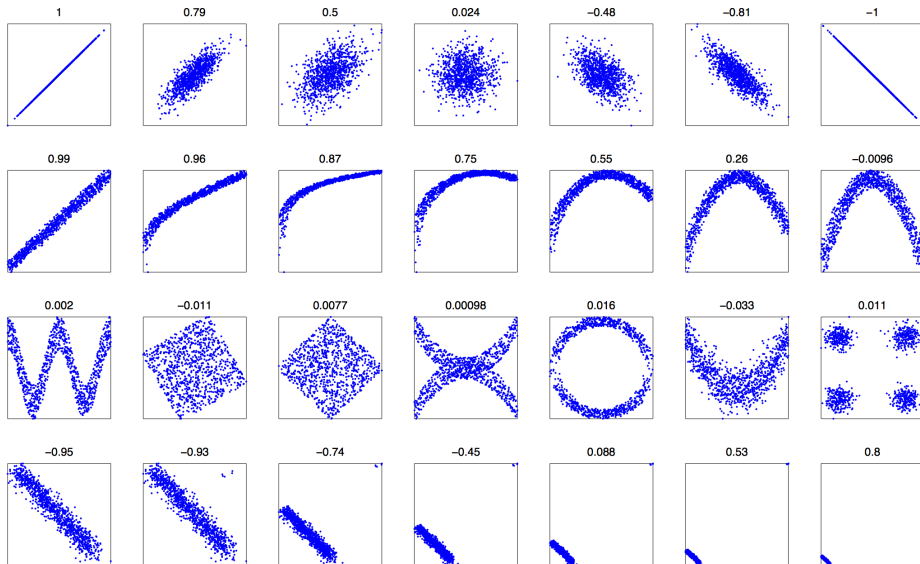
$$r_{X_1 X_2} = \frac{\mathbb{E}((X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2))}{\sqrt{\mathbb{D}X_1 \mathbb{D}X_2}}.$$

$$r_{X_1 X_2} \in [-1, 1].$$

Пусть имеется простая выборка пар  $(X_{1i}, X_{2i}), i = 1, \dots, n$ .  
Выборочный коэффициент корреляции Пирсона:

$$\hat{r}_{X_1 X_2} = \frac{\sum_{i=1}^n (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\sum_{i=1}^n (X_{1i} - \bar{X}_1)^2 \sum_{i=1}^n (X_{2i} - \bar{X}_2)^2}}.$$

# Корреляция Пирсона

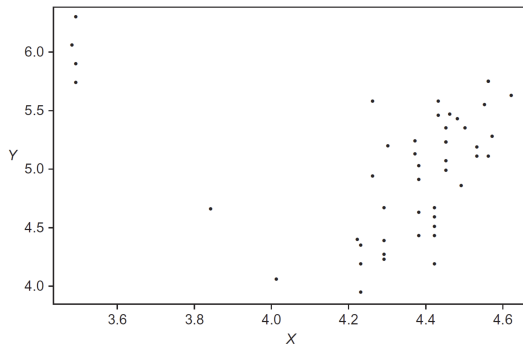


<http://guessthecorrelation.com/>

# Корреляция Пирсона

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 ( $\hat{r}_{XY} = -0.21$ ) из-за наличия в выборке красных гигантов.

# Критерий Стьюдента

выборки:  $X_1^n = (X_{11}, \dots, X_{1n})$

$X_2^n = (X_{21}, \dots, X_{2n})$

выборки связанные

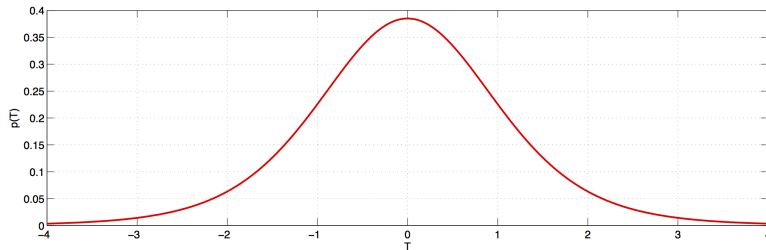
$(X_1, X_2) \sim N(\mu, \Sigma)$

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

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

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

нулевое распределение:  $St(n-2)$



# Преобразование Фишера

Стандартная ошибка:

$$SE = \frac{\sigma}{\sqrt{n}},$$

для t-критерия:

$$SE = \frac{1 - \hat{r}_{X_1 X_2}^2}{\sqrt{n - 2}}.$$

Преобразование Фишера:

$$z = \operatorname{arctanh}(\hat{r}_{X_1 X_2}), \quad z \sim \mathcal{N}$$

$$SE = \frac{1}{\sqrt{n - 3}}.$$



## Преобразование Фишера

Доверительный интервал для коэффициента корреляции Пирсона:

$$\left[ \hat{r}_{X_1 X_2} + \frac{t_{n-2, \alpha/2} (1 - \hat{r}_{X_1 X_2}^2)}{\sqrt{n}}, \hat{r}_{X_1 X_2} - \frac{t_{n-2, \alpha/2} (1 - \hat{r}_{X_1 X_2}^2)}{\sqrt{n}} \right].$$

С использованием преобразования Фишера:

$$\left[ \tanh \left( \operatorname{arctanh} \hat{r}_{X_1 X_2} + \frac{z_{\alpha/2}}{\sqrt{n-3}} \right), \tanh \left( \operatorname{arctanh} \hat{r}_{X_1 X_2} - \frac{z_{\alpha/2}}{\sqrt{n-3}} \right) \right].$$

# Преобразование Фишера

## Пример, Kanji, критерий 12

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?

$$H_0: r_{X_1X_2} = 0$$

$$H_1: r_{X_1X_2} \neq 0$$

Критерий Стьюдента:  $p = 0.0847$ .

Доверительный интервал:  $[-0.0157, 0.6557]$ .

С использованием преобразования Фишера:  $[-0.0455, 0.6100]$ .

## Перестановочный критерий

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

выборки связанные

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

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

статистика:  $T(X_1^n, X_2^n) = \hat{r}_{X_1 X_2}$

нулевое распределение: порождается перебором  $n!$  перестановок  
индексов одной из выборок

Достижимый уровень значимости — доля перестановок, на которых получилось такое же или ещё более экстремальное значение статистики.

## Перестановочный критерий

### Пример, Kanji, критерий 12

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?

$$\hat{r}_{X_1 X_2} = 0.32$$

$$H_0: r_{X_1 X_2} = 0$$

$$H_1: r_{X_1 X_2} \neq 0$$

Критерий Стьюдента:  $p = 0.0847$ .

Перестановочный критерий:  $p = 0.0564$ .

## Корреляция Спирмена

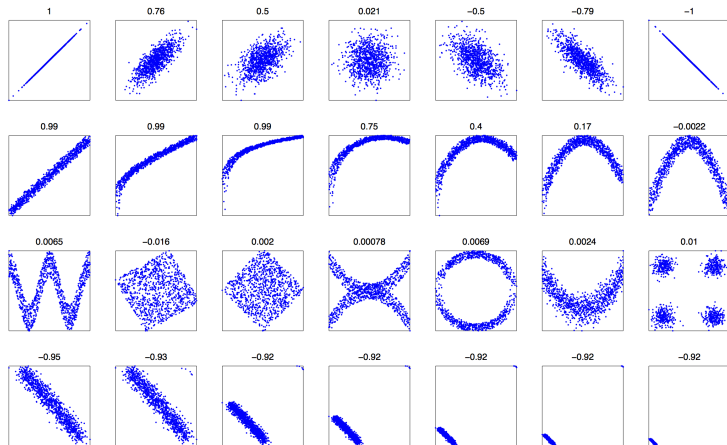
**Коэффициент корреляции Спирмена**  $\rho_{X_1 X_2}$  случайных величин  $X_1$  и  $X_2$  — мера силы **монотонной** корреляции между ними; равен коэффициенту корреляции Пирсона между рангами наблюдений.

Выборочный коэффициент корреляции Спирмена:

$$\begin{aligned}\hat{\rho}_{X_1 X_2} &= \frac{\sum_{i=1}^n \left( \text{rank}(X_{1i}) - \frac{n+1}{2} \right) \left( \text{rank}(X_{2i}) - \frac{n+1}{2} \right)}{\frac{1}{12} (n^3 - n)} = \\ &= 1 - \frac{6}{n^3 - n} \sum_{i=1}^n \left( \text{rank}(X_{1i}) - \text{rank}(X_{2i}) \right)^2,\end{aligned}$$

$\text{rank}(X_{1i}), \text{rank}(X_{2i})$  — ранги  $i$ -х наблюдений в соответствующих выборках.

# Корреляция Спирмена



# Критерий Стьюдента

выборки:  $X_1^n = (X_{11}, \dots, X_{1n})$

$X_2^n = (X_{21}, \dots, X_{2n})$

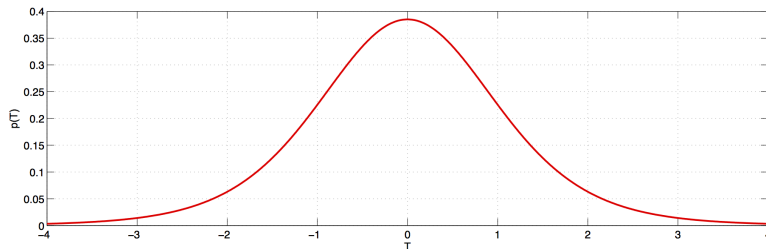
выборки связанные

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

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

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

нулевое распределение:  $St(n-2)$



# Критерий Стьюдента

## Пример, Kanji, критерий 58

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  $-0.854$

$$H_0: \rho_{X_1 X_2} = 0$$

$$H_1: \rho_{X_1 X_2} \neq 0$$

Критерий Стьюдента:  $p = 0.0024$ .



# Корреляция Кендалла

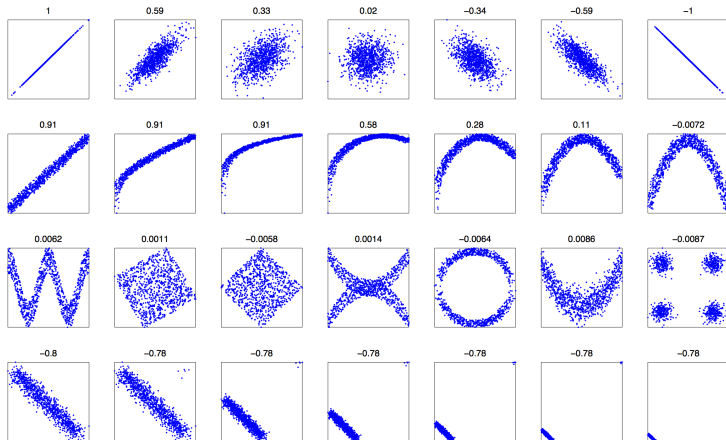
**Коэффициент корреляции Кендалла**  $\tau_{X_1 X_2}$  случайных величин  $X_1$  и  $X_2$  — мера их взаимной неупорядоченности; также оценивает силу **монотонной** корреляции между величинами.

Выборочный коэффициент корреляции Кендалла:

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

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

# Корреляция Кендалла



## Критерий без названия

выборки:  $X_1^n = (X_{11}, \dots, X_{1n})$

$X_2^n = (X_{21}, \dots, X_{2n})$

выборки связанные

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

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

статистика:  $\hat{\tau}_{X_1 X_2}$

нулевое распределение: табличное

При справедливости  $H_0$

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

Для  $n > 10$  справедлива аппроксимация нормальным распределением.

## Критерий без названия

### Пример, Kanji, критерий 59

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  $D = 5$ ,  $C = 38$ ,  $\hat{\tau}_{X_1 X_2} = 0.7821$ .

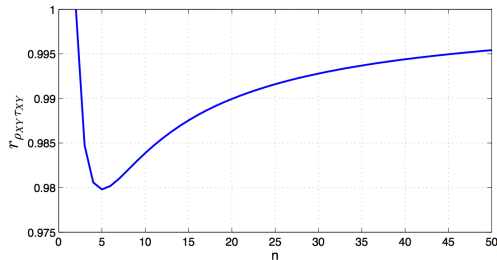
$$H_0: \tau_{X_1 X_2} = 0.$$

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

## Связь между коэффициентами корреляции

При справедливости  $H_0$  (отсутствии монотонной зависимости):

$$r_{\rho_{X_1 X_2} \tau_{X_1 X_2}} = \frac{2n+2}{\sqrt{4n^2+10n}}$$



<http://youtu.be/D56dvoVrBBE>: по сравнению с корреляцией Спирмена, корреляция Кендалла

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

$$(X_1, X_2) \sim N(\mu, \Sigma) \Rightarrow \lim_{n \rightarrow \infty} \mathbb{E} \tau_{X_1 X_2} = \lim_{n \rightarrow \infty} \mathbb{E} \rho_{X_1 X_2} = \frac{2}{\pi} \arcsin r_{X_1 X_2}.$$

## 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:

$$r_{X_1 X_2 | X_3} = \frac{r_{X_1 X_2} - r_{X_1 X_3} r_{X_2 X_3}}{\sqrt{(1 - r_{X_1 X_3}^2)(1 - r_{X_2 X_3}^2)}}.$$

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

$$r_{X_1 X_2 | X_3 X_4} = \frac{r_{X_1 X_2 | X_4} - r_{X_1 X_3 | X_4} r_{X_2 X_3 | X_4}}{\sqrt{(1 - r_{X_1 X_3 | X_4}^2)(1 - r_{X_2 X_3 | X_4}^2)}}.$$

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

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

# Критерий Стьюдента

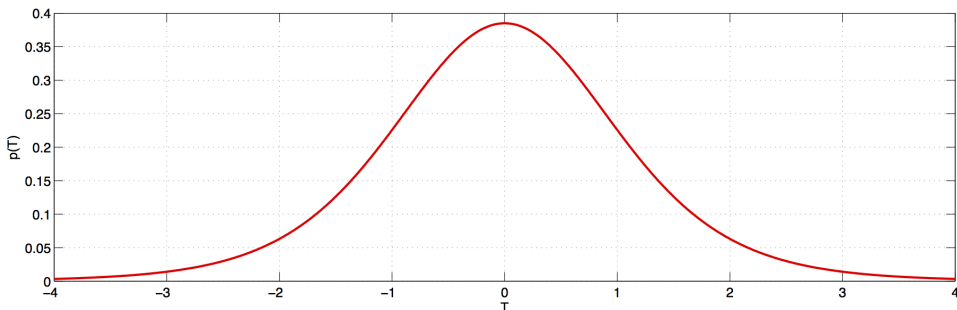
выборки:  $X_1^n = (X_{11}, \dots, X_{1n})$   
 $X_2^n = (X_{21}, \dots, X_{2n})$   
 $X_3^n = (X_{31}, \dots, X_{3n}), X_3 \in \mathbb{R}^M$   
 $(X_1, X_2, X_3) \sim N(\mu, \Sigma)$

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

нулевое распределение:  $St(n-M-2)$



## Множественная корреляция

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

$$r_{X_1, X_2, X_3} = \frac{r_{X_1 X_2}^2 + r_{X_1 X_3}^2 - 2r_{X_1 X_2}r_{X_1 X_3}r_{X_2 X_3}}{1 - r_{X_2 X_3}^2}.$$

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

$$r_{X, M}^2 = c^T R c.$$

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



# Критерий Фишера

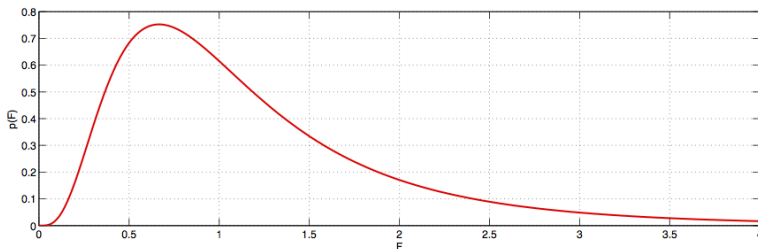
выборки:  $X_1^n = (X_{11}, \dots, X_{1n})$   
 $X_2^n = (X_{21}, \dots, X_{2n}), X_2 \in \mathbb{R}^M$   
 $(X_1, X_2) \sim N(\mu, \Sigma)$

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

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

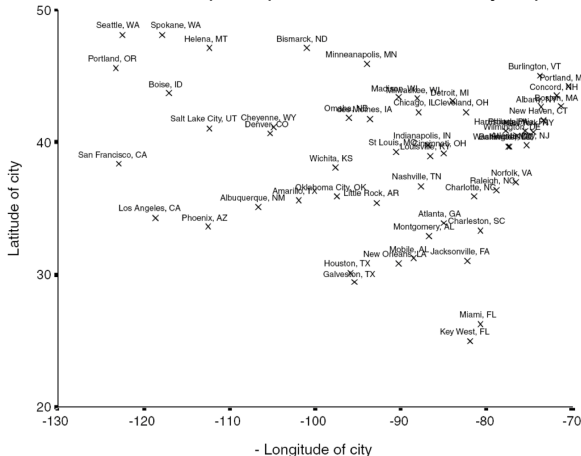
статистика:  $F(X_1^n, X_2^n) = \frac{\hat{r}_{X_1, X_2}^2}{1 - \hat{r}_{X_1, X_2}^2} \frac{n - M - 1}{M - 2}$

нулевое распределение:  $F(M - 2, n - M - 1)$

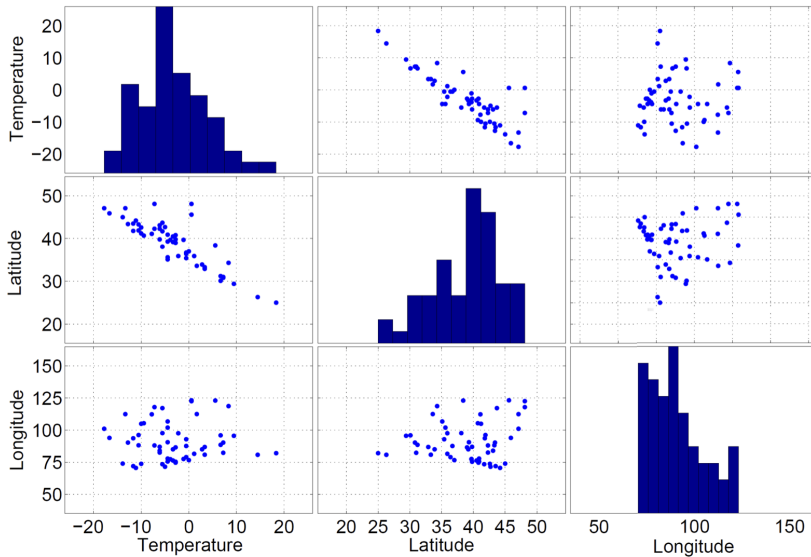


# Температура воздуха и географическое положение

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,  $\lambda$  — долгота,  $\phi$  — широта;  
 $r$  — корреляция Пирсона,  $\rho$  — Спирмена,  $\tau$  — Кендалла.

Коэффициенты корреляции:

$r$	$T$	$\phi$	$\lambda$
$T$	—	<b>-0.848</b>	0.024
$\phi$	<b>-0.848</b>	—	0.145
$\lambda$	0.024	0.145	—

$\tau$	$T$	$\phi$	$\lambda$
$T$	—	<b>-0.683</b>	0.030
$\phi$	<b>-0.683</b>	—	-0.011
$\lambda$	0.030	-0.011	—

$\rho$	$T$	$\phi$	$\lambda$
$T$	—	<b>-0.815</b>	0.030
$\phi$	<b>-0.815</b>	—	0.023
$\lambda$	0.030	0.023	—

Достигаемые уровни значимости:

$r$	$T$	$\phi$	$\lambda$
$T$	—	<b>0.000</b>	0.861
$\phi$	<b>0.000</b>	—	0.287
$\lambda$	0.861	0.287	—

$\tau$	$T$	$\phi$	$\lambda$
$T$	—	<b>0.000</b>	0.756
$\phi$	<b>0.000</b>	—	0.910
$\lambda$	0.756	0.910	—

$\rho$	$T$	$\phi$	$\lambda$
$T$	—	<b>0.000</b>	0.829
$\phi$	<b>0.000</b>	—	0.865
$\lambda$	0.829	0.865	—

## Температура воздуха и географическое положение

$T$  — температура,  $\lambda$  — долгота,  $\phi$  — широта;  
 $r$  — частная корреляция Пирсона,  $\rho$  — Спирмена.

Коэффициенты частной корреляции:

$r$	$T$	$\phi$	$\lambda$
$T$	—	<b>-0.861</b>	<b>0.280</b>
$\phi$	<b>-0.861</b>	—	<b>0.312</b>
$\lambda$	<b>0.280</b>	<b>0.312</b>	—

$\rho$	$T$	$\phi$	$\lambda$
$T$	—	<b>-0.817</b>	0.084
$\phi$	<b>-0.817</b>	—	0.082
$\lambda$	0.084	0.082	—

Достигаемые уровни значимости:

$r$	$T$	$\phi$	$\lambda$
$T$	—	<b>0.000</b>	<b>0.039</b>
$\phi$	<b>0.000</b>	—	<b>0.021</b>
$\lambda$	<b>0.039</b>	<b>0.021</b>	—

$\rho$	$T$	$\phi$	$\lambda$
$T$	—	<b>0.000</b>	0.543
$\phi$	<b>0.000</b>	—	0.552
$\lambda$	0.543	0.552	—

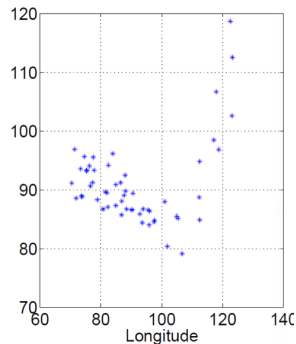
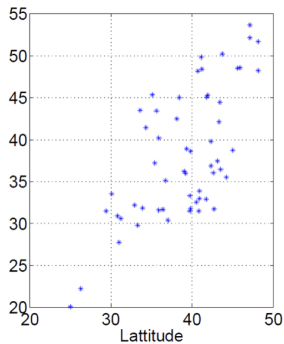
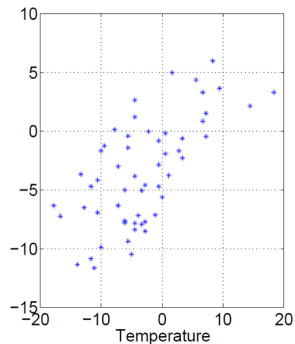
# Температура воздуха и географическое положение

$T$  — температура,  $\lambda$  — долгота,  $\phi$  — широта;

$R$  — множественная корреляция.

Коэффициенты множественной корреляции:

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



## Таблица сопряжённости $K_1 \times K_2$

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

Таблица сопряжённости:

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

## Два случайных признака

Пусть  $\pi_{ij}$  — вероятность реализации пары  $(X_1, X_2)$  в ячейке  $(i, j)$ .

$\{\pi_{ij}\}$  — совместное распределение  $(X_1, X_2)$ ;

$\{\pi_{i+}\}$ ,  $\{\pi_{+j}\}$  — маргинальные распределения:

$$\{\pi_{i+}\} = \sum_{j=1}^{K_2} \pi_{ij},$$

$$\{\pi_{+j}\} = \sum_{i=1}^{K_1} \pi_{ij}.$$

$X_1$  и  $X_2$  независимы, если

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



## Один случайный признак

Пусть  $X_1$  — не случайная величина, а фиксированный признак. Тогда  $\{\pi_{ij}\}$  не имеет смысла, вместо него рассматриваются  $\{\pi_{1|i}, \dots, \pi_{K_1|i}\}$  — 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  $\mu_{ij}$ ; 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:

$$\frac{n!}{n_{11}! \cdot \dots \cdot n_{K_1 K_2}!} \prod_i \prod_j \pi_{ij}^{n_{ij}}.$$

- 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
	использован	не использован
смертельный		
несмертельный		

- ❶ Исследователи собираются учесть все автомобильные аварии, которые произойдут на магистрали в течение года.
- ❷ Исследователи запросят 200 случайных полицейских рапортов об авариях за последние годы.
- ❸ 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 $2 \times 2$

Пусть  $X_1$  и  $X_2$  принимают значения 0 и 1.

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

# Критерий хи-квадрат

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

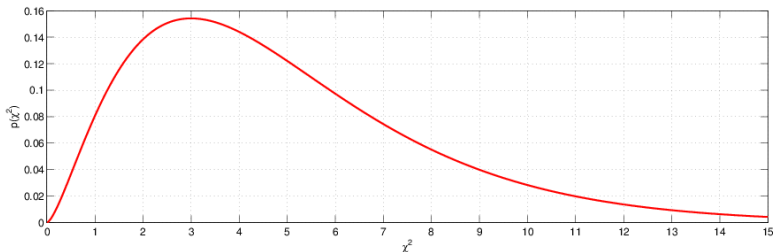
выборки связанные

нулевая гипотеза:  $H_0: X_1$  и  $X_2$  независимы

альтернатива:  $H_1: H_0$  неверна

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

нулевое распределение:  $\chi^2_{(K_1-1)(K_2-1)}$



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  $\Rightarrow p = 0.5398$ .

# G-критерий

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

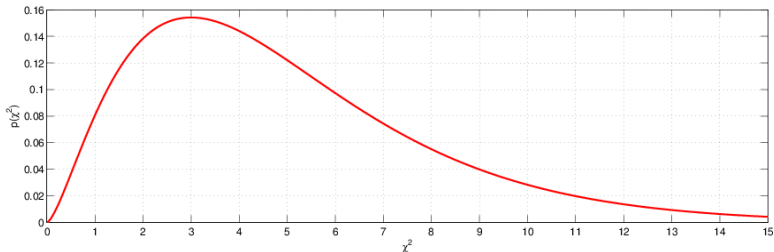
выборки связанные

нулевая гипотеза:  $H_0: X_1$  и  $X_2$  независимы

альтернатива:  $H_1: H_0$  неверна

статистика:  $G^2(X_1^n, X_2^n) = 2 \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} n_{ij} \ln \frac{n_{ij}n}{n_{i+}n_{+j}}$

нулевое распределение:  $\chi^2_{(K_1-1)(K_2-1)}$



## Точный критерий Фишера

выборки:  $X_1^n = (X_{11}, \dots, X_{1n}), X_1 \in \{0, 1\}$

$X_2^n = (X_{21}, \dots, X_{2n}), X_2 \in \{0, 1\}$

выборки связанные

нулевая гипотеза:  $H_0: X_1$  и  $X_2$  независимы

альтернатива:  $H_1: H_0$  неверна

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

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

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

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 26 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.

Точный критерий Фишера:  $p = 0.0008$ .

## Перестановочный критерий

Представим выборку в виде таблицы  $n \times 2$ :

	М	Ж
На диете	1	9
Не на диете	13	3

$\Rightarrow$

Строка	Столбец
1	1
1	2
...	
1	2
2	1
...	
2	1
2	2
2	2
2	2

Используем статистику критерия хи-квадрат, но её нулевое распределение будем оценивать по  $n!$  перестановок второй колонки.

$H_0$ : связи нет.

$H_1$ : признаки связаны.

Точный критерий Фишера:  $p = 0.0008$ .

Критерий хи-квадрат:  $p = 0.0004$ .

Перестановочный критерий со статистикой хи-квадрат:  $p = 0.0014$ .

## Коэффициент V Крамера

Мера взаимосвязи между двумя категориальными переменными — коэффициент V Крамера:

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

$$\phi_c(X_1^n, X_2^n) \in [0, 1];$$

0 соответствует полному отсутствию взаимосвязи, 1 — совпадению переменных.

## Корреляция между порядковыми переменными

Мера взаимосвязи между двумя порядковыми переменными — коэффициент  $\gamma$ :

$$\hat{\gamma} = \frac{p_C - p_D}{n^2 - p_t},$$

где  $p_C = \frac{C}{n}$  — частота появления согласованных пар элементов выборки, т. е., таких, что  $i_1 > i_2, j_1 > j_2$  или  $i_1 < i_2, j_1 < j_2$ ;

$p_D = \frac{D}{n}$  — частота несогласованных пар;

$p_t = \frac{T}{n}$  — частота таких пар, что  $i_1 = i_2$  или  $j_1 = j_2$ .

$\gamma \in [-1, 1]$ ;  $-1$  соответствует полному отсутствию согласованных пар,  $1$  — отсутствию несогласованных.

### Пример

Политические взгляды	Счастье		
	Не слишком счастлив	Вполне счастлив	Очень счастлив
Либеральные	13	29	15
Умеренные	23	59	47
Консервативные	14	67	54

$\chi^2 = 7.07, p = 0.1322, \phi_c = 0.0742;$   
 $\hat{\gamma} = 0.185, 95\% \text{ доверительный интервал} — [0.032, 0.338].$

## 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.

## Пары переменных разных типов

Между категориальными и непрерывными признаками корреляции считать не нужно!

Пусть  $X_1 \in \mathbb{R}$ ,  $X_2 \in \{0, 1\}$ ;

$X_1$  и  $X_2$  положительно коррелированы, если  $\mathbb{E}(X_1 | X_2 = 1) > \mathbb{E}(X_1 | X_2 = 0)$ .

Мера взаимосвязи  $X_1$  и  $X_2$  — разность  $\mathbb{E}(X_1 | X_2 = 1) - \mathbb{E}(X_1 | X_2 = 0)$ .

## 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:  $\chi^2 = 17.555$ ,  $p = 2.8 \times 10^{-5}$ .

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)

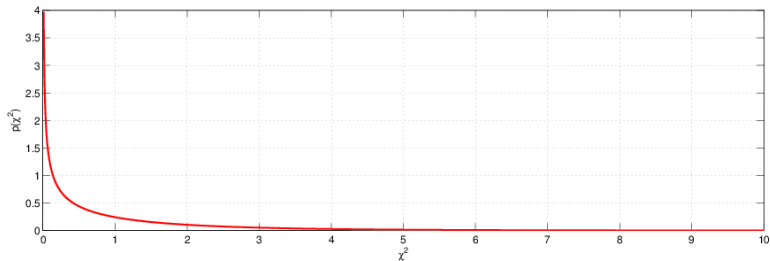
### Пример, Bikel et al., 1975

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.

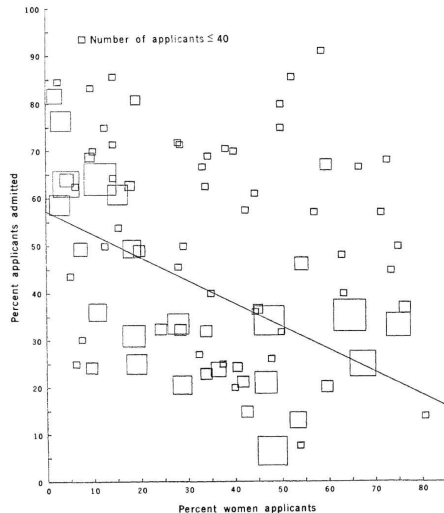
На 6 факультетов поступило значимо меньше мужчин, суммарный дефицит — 64.

Данные по 6 крупнейшим факультетам:

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

# Парадокс хи-квадрат (Симпсона)

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



## Парадокс хи-квадрат (Симпсона)

“Like fire, the chi-square statistic is an excellent servant and a bad master.” (Austin Bradford Hill)

# Литература

- непрерывные признаки — Лагутин, гл. 20;
- категориальные признаки — Agresti, гл. 2 и 3, Bilder, разделы 3.1, 3.2, 6.2.1, 6.2.2;
- значимость корреляции Пирсона — Kanji, №12, Good, 3.8;
- 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*, 2006.

Лагутин М.Б. *Наглядная математическая статистика*, 2007.

Agresti A. *Categorical Data Analysis*, 2013.

Bickel P.J., Hammel E.A., O'connell J.W. (1975). *Sex bias in graduate admissions: data from Berkeley*. Science, 187(4175), 398–404.

Bilder C.R., Loughin T.M. *Analysis of Categorical Data with R*, 2013.

Good P. *Permutation, Parametric and Bootstrap Tests of Hypotheses: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2005.

Kanji G.K. *100 statistical tests*, 2006.