

MML minor #11

Анализ зависимостей

Thanks

Рябенко Евгений
riabenko.e@gmail.com

2016 г.

Корреляция Пирсона

Коэффициент корреляции Пирсона — мера силы линейной взаимосвязи:

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

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

Выборочный коэффициент

Выборка пар $(X_{1i}, X_{2i}), i = 1, \dots, n$.

Выборочный коэффициент корреляции Пирсона:

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

Свойства

`https://yadi.sk/i/1idYe9YEzHgsk`

`https://yadi.sk/i/lfqo3s5ezHhQk`

`https://yadi.sk/i/xm4Jg_r9zHhVy`

`https://yadi.sk/i/JBoSC5IIzHhTP`

Корреляция Спирмена

Коэффициент корреляции Спирмена — мера силы **монотонной** взаимосвязи;
равен коэффициенту корреляции Пирсона между рангами наблюдений.

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

Выборочный коэффициент

Выборка пар $(X_{1i}, X_{2i}), i = 1, \dots, n;$

$\text{rank}(X_{1i}), \text{rank}(X_{2i})$ — ранги.

Выборочный коэффициент корреляции Спирмена:

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

Свойства

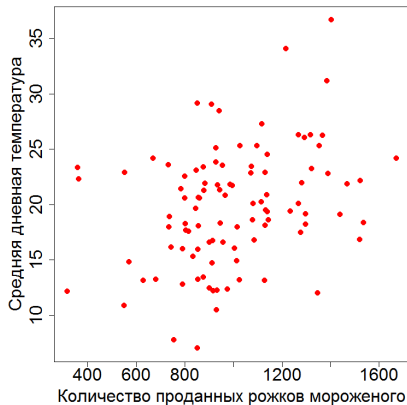
`https://yadi.sk/i/53Ct2IRdzHhbR`

`https://yadi.sk/i/jJYpVo9CzHhds`

`https://yadi.sk/i/yNaVMvkazHhiB`

`https://yadi.sk/i/O1_LATvyzHhfk`

Продажи мороженого



$$r_{X_1X_2} = 0.45, \quad \rho_{X_1X_2} = 0.44.$$

Связаны ли объём продаж мороженого и средняя дневная температура?

Критерий Стьюдента

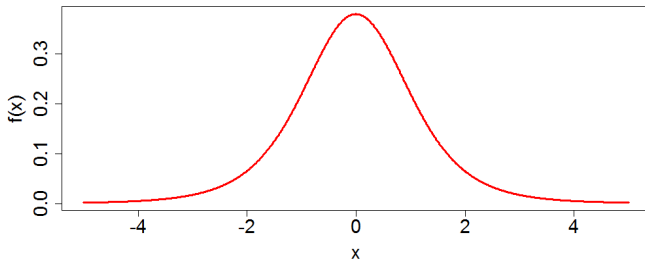
выборки: $(X_{1i}, X_{2i}), i = 1, \dots, n,$

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

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

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

нулевое распределение: $T \sim St(n-2).$



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

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

Для корреляции Спирмена

можно применять эти же критерии.

Продажи мороженого

H_0 : линейной связи нет, $r_{X_1X_2} = 0$.

H_1 : признаки линейно связаны, $r_{X_1X_2} \neq 0$.

Критерий Стьюдента: $p = 3 \times 10^{-6}$, признаки линейно связаны,
корреляция Пирсона — 0.45 (95% доверительный интервал — $[0.28, 0.59]$).

H_0 : монотонной связи нет, $\rho_{X_1X_2} = 0$.

H_1 : признаки монотонно связаны, $\rho_{X_1X_2} \neq 0$.

Критерий Стьюдента: $p = 3 \times 10^{-6}$, признаки монотонно связаны,
корреляция Спирмена — 0.44 (95% доверительный интервал — $[0.26, 0.60]$).

Корреляция Мэтьюса

Коэффициент корреляции Мэтьюса — мера силы взаимосвязи между двумя бинарными переменными.

Таблица сопряжённости:

$X_1 \backslash X_2$	0	1
0	a	b
1	c	d

$$MCC_{X_1 X_2} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}.$$

$MCC_{X_1 X_2} \in [-1, 1];$

0 — полное отсутствие взаимосвязи,

$1 \Leftrightarrow b = c = 0,$

$-1 \Leftrightarrow a = d = 0.$

Таблица сопряжённости $K_1 \times K_2$

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

V Крамера

Коэффициент V Крамера — мера силы взаимосвязи между двумя категориальными переменными.

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

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

0 — полное отсутствие взаимосвязи,

1 — совпадение переменных (с точностью до переименования уровней).

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 $\mathbb{E}(X_1 | X_2 = 1) - \mathbb{E}(X_1 | X_2 = 0)$.

Эффективность лечения

Исследуется влияние сахарного диабета на эффективность тромболитической терапии.

	Выздоровели	Не выздоровели
Диабет	48	30
Нет	92	36

$MCC = -0.1074$.

Связано ли наличие диабета с эффективностью лечения?

Критерий хи-квадрат

выборки: $(X_{1i}, X_{2i}), i = 1, \dots, n,$

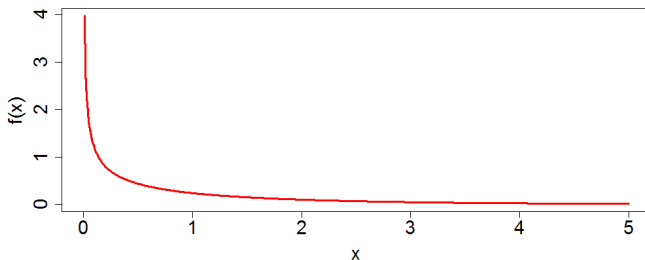
$X_1, X_2 \in \{0, 1\};$

нулевая гипотеза: $H_0: \text{MCC}_{X_1 X_2} = 0;$

альтернатива: $H_1: \text{MCC}_{X_1 X_2} \neq 0;$

статистика: $\chi^2 = n \text{MCC}_{X_1 X_2}^2;$

нулевое распределение: $\chi^2 \sim \chi_1^2.$



Условия применимости:

- $n \geq 40$
- $\frac{(a+c)(a+b)}{n}, \frac{(a+c)(c+d)}{n}, \frac{(b+d)(a+b)}{n}, \frac{(b+d)(c+d)}{n} > 5$

Эффективность лечения

	Выздоровели	Не выздоровели
Диабет	48	30
Нет	92	36

$$MCC = -0.1074.$$

H_0 : эффективность лечения не зависит от наличия диабета.

H_1 : эффективность лечения зависит от наличия диабета.

Критерий хи-квадрат: $p = 0.1651$, нельзя утверждать, что связь есть.

Пол и диета

Для 26 опрошенных известен пол и сидят ли они на диете.

	М	Ж
На диете	1	9
Не на диете	13	3

$$MCC = -0.6953.$$

Есть ли связь между этими признаками?

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 сопряжённости 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:

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

Пол и диета

Для 26 опрошенных известен пол и сидят ли они на диете.

	М	Ж
На диете	1	9
Не на диете	13	3

$$MCC = -0.6953.$$

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: $p = 0.0014$.

Категориальные признаки

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

Критерий хи-квадрат

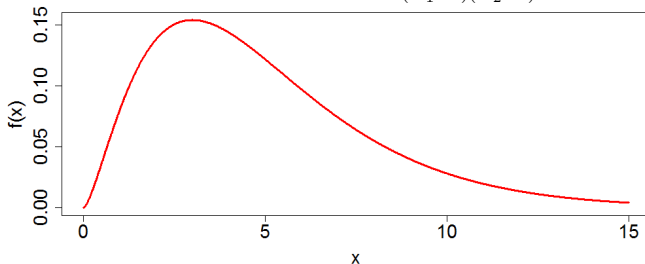


выборки: $(X_{1i}, X_{2i}), i = 1, \dots, n,$
 $X_1 \in \{1, \dots, K_1\}, X_2 \in \{1, \dots, K_2\},$

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

альтернатива: $H_1: H_0$ неверна;

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



Критерий хи-квадрат

Условия применимости:

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

Проверяет значимость отличия от нуля коэффициента V Крамера:

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

Другие критерии

Точный критерий Фишера и перестановочный критерий существуют и строятся по аналогии.

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)

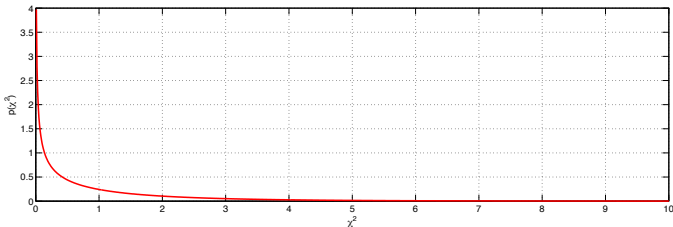
Пример (Bikel at el., 1975): в 1973 году на университет Беркли, Калифорния, подали в суд: доля поступивших абитуриентов мужского пола была выше, чем доля поступивших женского пола.

	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.

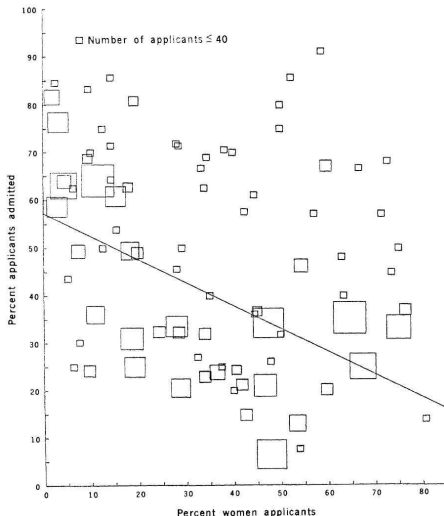
На 6 факультетов поступило значимо меньше мужчин, суммарный дефицит — 64.

Данные по 6 крупнейшим факультетам:

	Мужчины		Женщины	
	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 пытались поступить на факультеты с большим конкурсом.



Буллит и консервативность



RESEARCH ARTICLE

Misperceiving Bullshit as Profound Is Associated with Favorable Views of Cruz, Rubio, Trump and Conservatism

Stefan Pfattheicher^{1*}, Simon Schindler²

1 Ulm University, Ulm, Germany, **2** Kassel University, Kassel, Germany

<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0153419>

Определение

Буллит — бессодержательное, нелогичное или явно противоречащее элементарным научным знаниям утверждение.

Примеры:

- “Скрытый смысл трансформирует беспрецедентную абстрактную красоту”
- “Воображение лежит в основе экспоненциальных пространственно-временных событий”

Данные

Испытуемые: 196 граждан Америки (43.4% женщин, средний возраст 36.4 лет), набранные на Amazon Mechanical Turk.

Задание:

- 1 оценить глубокомысленность утверждений по шкале от 1 (“абсолютно не глубокое”) до 5 (“очень глубокое”)
- 2 оценить степень симпатии к трём кандидатам в президенты США от демократической и трём — от республиканской партий по шкале от 1 (“очень несимпатичен”) до 5 (“очень симпатичен”)
- 3 оценить степень консервативности собственных политических взглядов по семибалльной шкале Лайкерта.

Часть утверждений, оцениваемых респондентами, — буллит, часть — относительно редкие поговорки (“Промокший человек не боится дождя”).

Анализ

Для каждого испытуемого вычислялась средняя склонность считать буллит глубокомысленным.

Вычислялась корреляция Спирмена между итоговым признаком, консервативностью политических взглядов и степенями симпатии к кандидатам.

Для проверки значимости отличия корреляции от нуля использовался критерий Стьюдента.

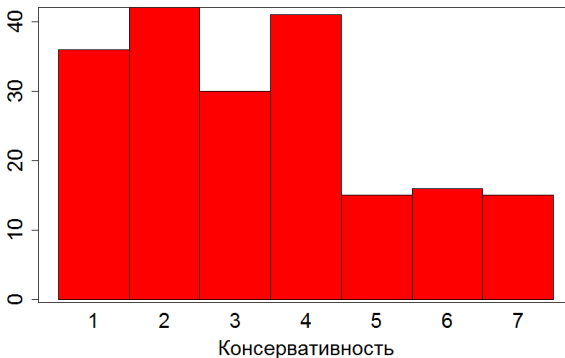
Результат

Обнаружена значимая положительная корреляция между тягой к буллитизму и:

- симпатией к Теду Крузу, Марку Рубио и Дональду Трампу
- степенью консерватизма

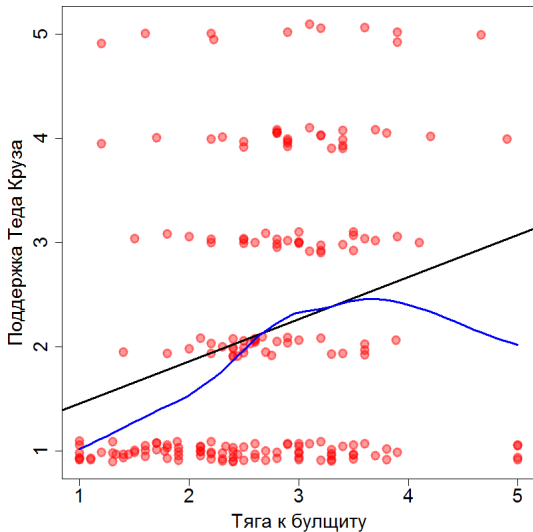
Проблемы

- репрезентативность выборки: аудитория Amazon Mechanical Turk — не случайная выборка граждан США
- несбалансированность выборки:



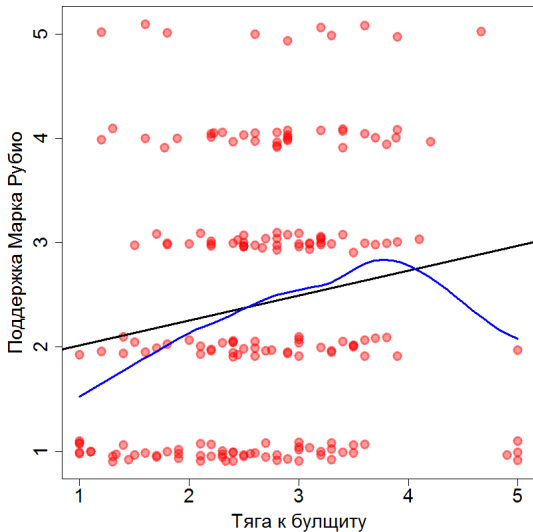
- сырые данные

Сырые данные



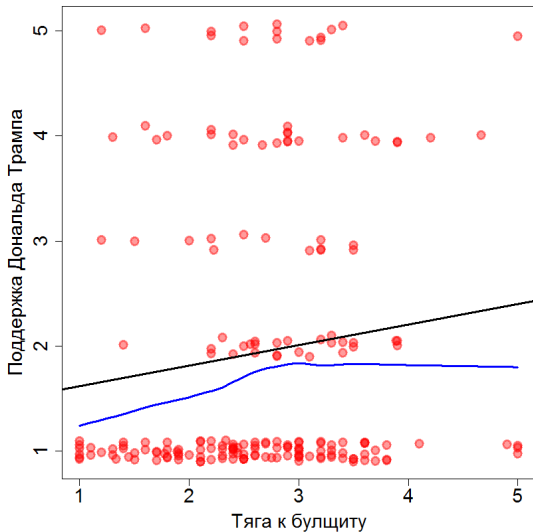
$$\rho_{XY} = 0.3, p = 2 \times 10^{-5}.$$

Сырые данные



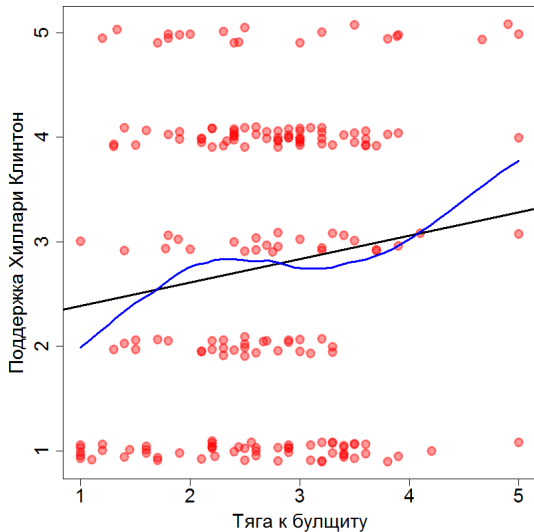
$$\rho_{XY} = 0.2, p = 0.0064.$$

Сырые данные



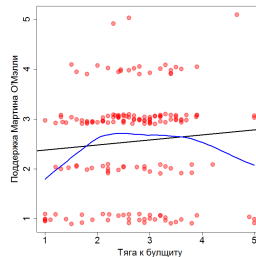
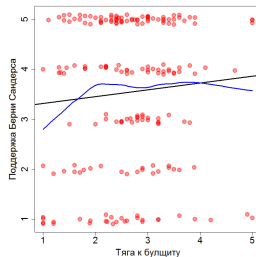
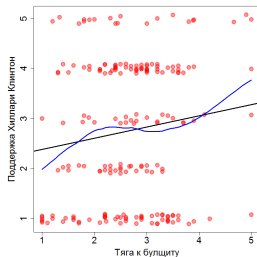
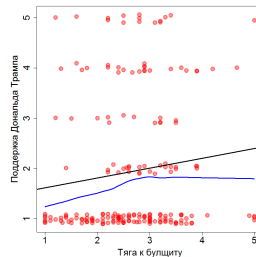
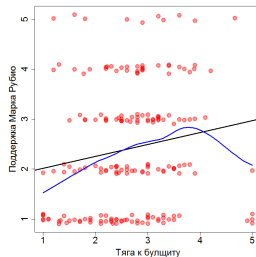
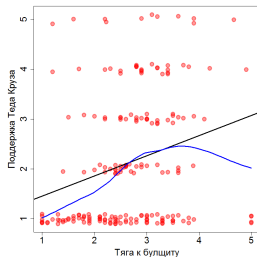
$$\rho_{XY} = 0.15, p = 0.0324.$$

Сырые данные



$$\rho_{XY} = 0.09, p = 0.212.$$

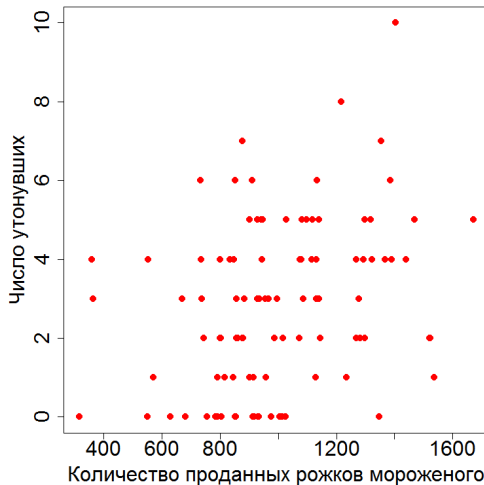
Сырые данные



Резюме

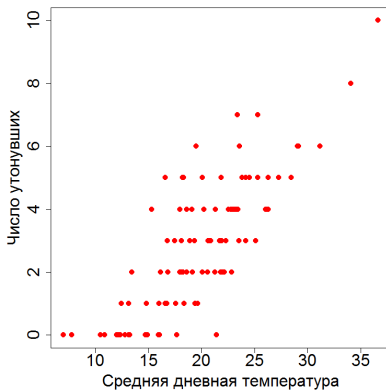
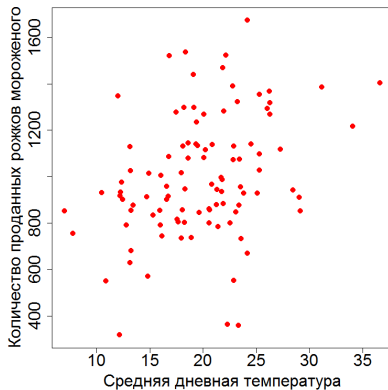
- всегда смотрите на сырые данные!
- корреляционный анализ — не сосисочная машина (справедливо для всех статистических методов)

Мороженое и смерть



$r_{X_1X_2} = 0.33, p = 0.0009$ (критерий Стьюдента), 95% доверительный интервал — $[0.138, 0.491]$.

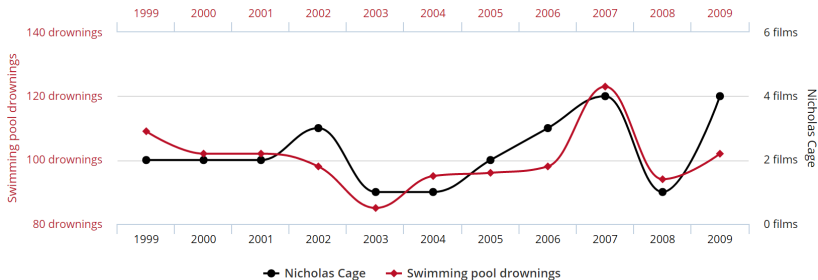
Мороженое, смерть и температура



Другие примеры

- количество самоубийств и радиоприёмников на душу населения (размер города)
- уровень CO_2 в атмосфере и распространённость ожирения (уровень жизни)
- рыночная доля Internet Explorer и количество убийств в США (время)

Смерть и Николас Кейдж



$r_{X_1X_2} = 0.67, p = 0.0253$ (критерий Стьюдента), 95% доверительный интервал — $[0.110, 0.905]$.

Больше: <http://www.tylervigen.com>

Резюме

- корреляция \nRightarrow причинно-следственная связь
- причинно-следственная связь $\xRightarrow{\text{возможно}}$ корреляция

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.