Урок 8

Корреляции

8.1. Корреляция Пирсона

Самый распространенный способ формализации корреляции — это коэффициент корреляции Пирсона. Корреляция Пирсона — это мера силы линейной взаимосвязи между двумя случайными величинами X_1 и X_2. Определяется она следующим образом:

$$r_{X_1X_2} = \frac{\mathbb{E}\left((X_1 - \mathbb{E}X_1)\left(X_2 - \mathbb{E}X_2\right)\right)}{\sqrt{\mathbb{D}X_1\mathbb{D}X_2}}, \quad r_{X_1X_2} \in [-1, 1],$$

где $r_{X_1X_2} = 1$ соответствует идеальной линейной взаимосвязи между случайными величинами, в которой при росте X_1 растет и X_2. $r_{X_1X_2} = -1$ — это идеальная линейная связь с отрицательным знаком, то есть, когда X_1 растет, X_2 падает. $r_{X_1X_2} = 0$ — это случай отсутствия корреляции; это значит, что две случайные величины меняются независимо друг от друга.

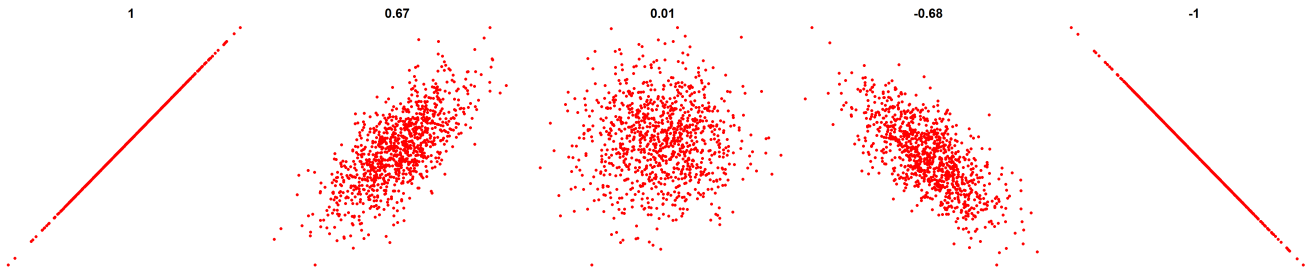
8.1.1. Выборочный коэффициент корреляции Пирсона

Если имеется выборка пар (X_{1i}, X_{2i}) объема n, по ней очень легко посчитать выборочный коэффициент корреляции Пирсона:

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

8.1.2. Примеры

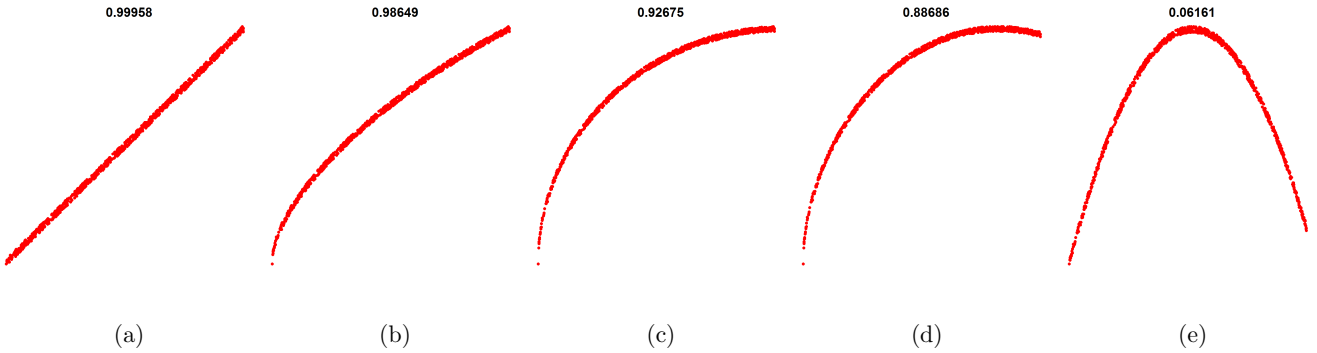
На рисунке 8.1a показаны диаграммы рассеяния — это графики, на одной оси которых отложены значения X_1, а на другой — X_2.



(a) (b) (c) (d) (e)

Рис. 8.1: Изменение корреляции Пирсона при размытии облака точке

Первый график (рисунок 8.1a) показывает облако точек с идеальной положительной корреляцией ($r_{X_1X_2} = 1$). Если начать это облако размывать (рисунки 8.1b, 8.1c), то коэффициент корреляции Пирсона постепенно уменьшится до 0. Если затем облако точек начать сжимать в обратном направлении, коэффициент растет по модулю и постепенно становится равным -1.



(a) (b) (c) (d) (e)

Рис. 8.2: Изменение корреляции Пирсона при увеличении изгиба облака точке

Следующий эксперимент показан на рисунке 8.2. На графике 8.2a показано облако с высокой положительной корреляцией между случайными величинами. Если начать его постепенно загибать, то коэффициент корреляции Пирсона будет уменьшаться (рисунки 8.2b, 8.2c, 8.2d). Когда форма облака становится похожей на параболу, значение выборочного коэффициента корреляции приближается к 0. Так происходит, потому что корреляция Пирсона — это мера силы линейной взаимосвязи между случайными величинами. То есть все нелинейные функциональные зависимости, даже если они очень хорошо выражены, коэффициент корреляции Пирсона не обнаруживают. Это демонстрируют примеры на рисуноке 8.3. Если между случайными величинами X_1 и X_2 наблюдаются сложные зависимости, далекие от линейных, коэффициент корреляции Пирсона будет всё равно близким к 0 (рисунок 8.3).

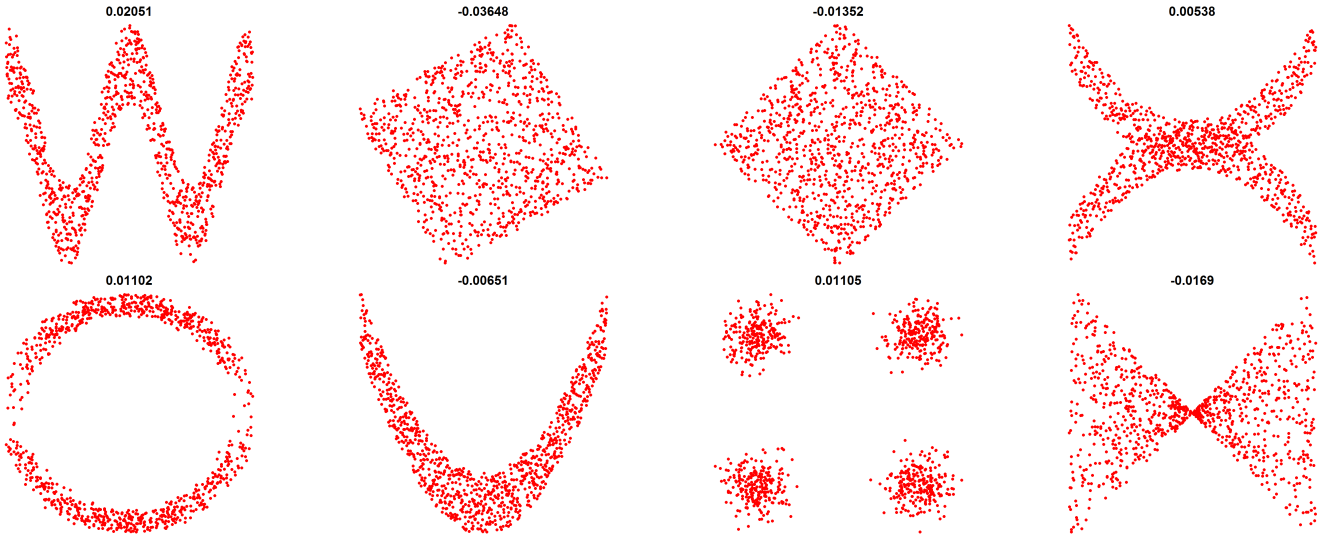
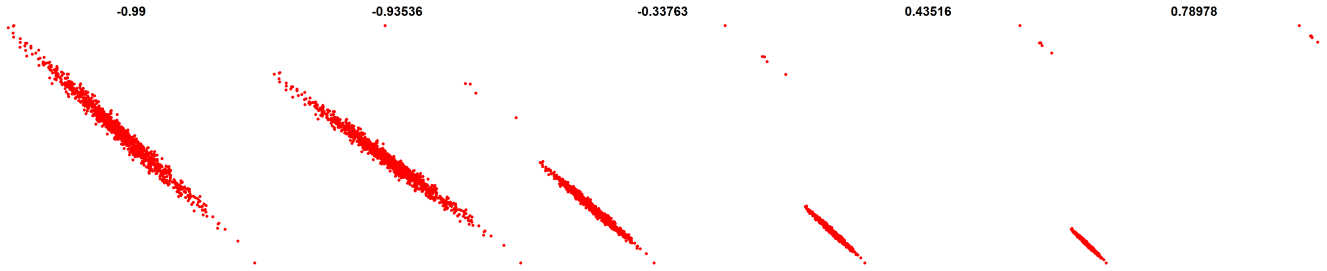


Рис. 8.3: Нелинейные зависимости между случайными величинами

Следующий важный пример изображён на рисунке 8.4. На графике 8.4a показано облако из тысячи точек с сильной отрицательной корреляцией. Если взять 5 точек из этого облака и начать постепенно отодвигать в верхний правый угол диаграммы рассеяния, то, чем дальше отодвигаются эти 5 точек, тем меньше по модулю становится значение выборочного коэффициента корреляции (рисунки 8.4b, 8.4c). С какого-то момента оно переходит через 0 и начинает расти (рисунки 8.4d, 8.4e). Достаточно сильно отодвинув всего 5 точек из тысячи, можно получить большой положительный коэффициент корреляции. Это говорит о том, что коэффициент корреляции Пирсона неустойчив к выбросам: небольшое количество точек могут оказывать на него существенное влияние, если они находятся достаточно далеко от основного облака. Это существенная

особенность корреляции Пирсона, которую нужно иметь ввиду.



(a) (b) (c) (d) (e)

Рис. 8.4: Влияние выбросов на коэффициент корреляции Пирсона

8.2. Корреляция Спирмена

8.2.1. Определение, связь с корреляцией Пирсона

Ещё один способ формализации понятия корреляции — это корреляция Спирмена. Коэффициент корреляции Спирмена — это мера силы монотонной взаимосвязи между двумя случайными величинами, он равен коэффициенту корреляции Пирсона между рангами наблюдений. Для того, чтобы ее посчитать, нужно выборку пар $(X_{1i}, X_{2i}), i = 1, \ldots, n$ превратить наблюдение в каждой из подвыборок в ранги $\text{rank}(X_{1i})$, взять $\text{rank}(X_{2i})$, и уже на этих рангах посчитать значение коэффициента корреляции Пирсона. Именно за счет рангового преобразования получается, что корреляция Спирмена чувствительна к любой монотонной взаимосвязи между X_1 и X_2, поскольку ранговое преобразование превращает любую монотонную взаимосвязь в линейную.

Корреляция Спирмена наследует часть свойств корреляции Пирсона. Она точно так же меняется от -1 до 1, где крайние значения отрезка соответствуют идеальной, в данной случае монотонной, взаимосвязи между случайными величинами, а 0 — полному отсутствию монотонной взаимосвязи между ними.

8.2.2. Выборочный коэффициент корреляции Спирмена

Если имеется выборка пар $(X_{1i}, X_{2i}), i = 1, \ldots, n;$, то выборочный коэффициент корреляции Спирмена вычисляется следующим образом:

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

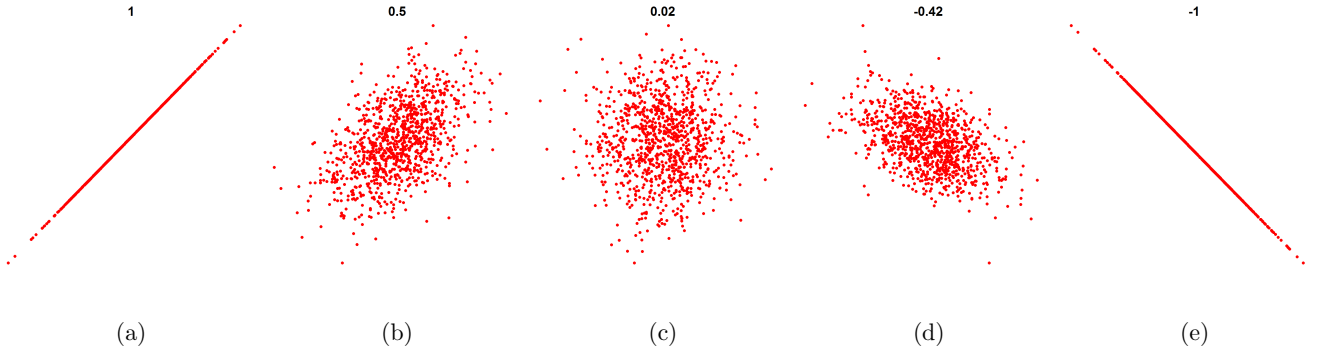
$$= 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}\left(\text{rank}(X_{1i}) - \text{rank}(X_{2i})\right)^2$$

В данном случае формулу выборочной корреляции Пирсона можно немного упростить, поскольку заранее известно, чему равны средние ранги в выборках и чему равны их дисперсии.

8.2.3. Примеры

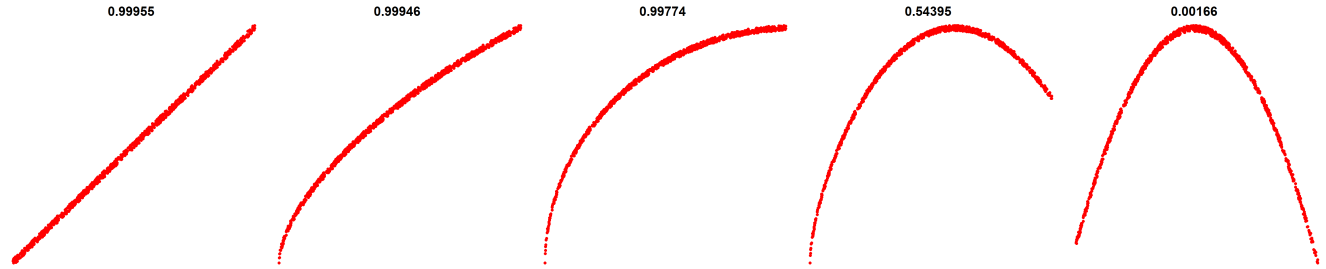
Чтобы посмотреть, какие из свойств корреляции Спирмена отличаются от свойств корреляции Пирсона, можно воспроизвести эксперименты с облаками точек.

Корреляция Спирмена примерно так же, как и корреляция Пирсона, реагирует на сжатие и размывание облака точек на диаграмме рассеяния (рисунок 8.5). Видно, что крайние случаи идеальной линейной взаимосвязи (графики 8.5a, 8.5e) соответствуют -1 и 1, а в середине получаются значения коэффициента корреляции, близкие к 0 (график 8.5c).



(a) (b) (c) (d) (e)

Рис. 8.5: Изменение значения коэффициента корреляции Спирмена при размытии облака точке



(a) (b) (c) (d) (e)

Рис. 8.6: Изменение значения коэффициента корреляции Спирмена при увеличении изгиба облака точке

Более интересные результаты получаются в эксперименте с загибанием облака точек (рисунок 8.6). Видно, что пока зависимость между X_1 и X_2 остается монотонной (рисунки 8.6a, 8.6b), значение коэффициента корреляции Спирмена почти не убывает [1]. Однако потом, когда облако точек начинает превращаться в параболу (рисунки 8.6c, 8.6d, 8.6e), значение выборочного коэффициента корреляции Спирмена постепенно превращается в 0. Корреляция Спирмена не обнаруживает взаимосвязи между X_1 и X_2, отличные от монотонных. Это можно заметить и на следующих примерах. Когда между X_1 и X_2 есть какие-то сложные функциональные взаимосвязи (рисунок 8.7), корреляция Спирмена все равно остается близкой к 0, поскольку они далеки от монотонных.

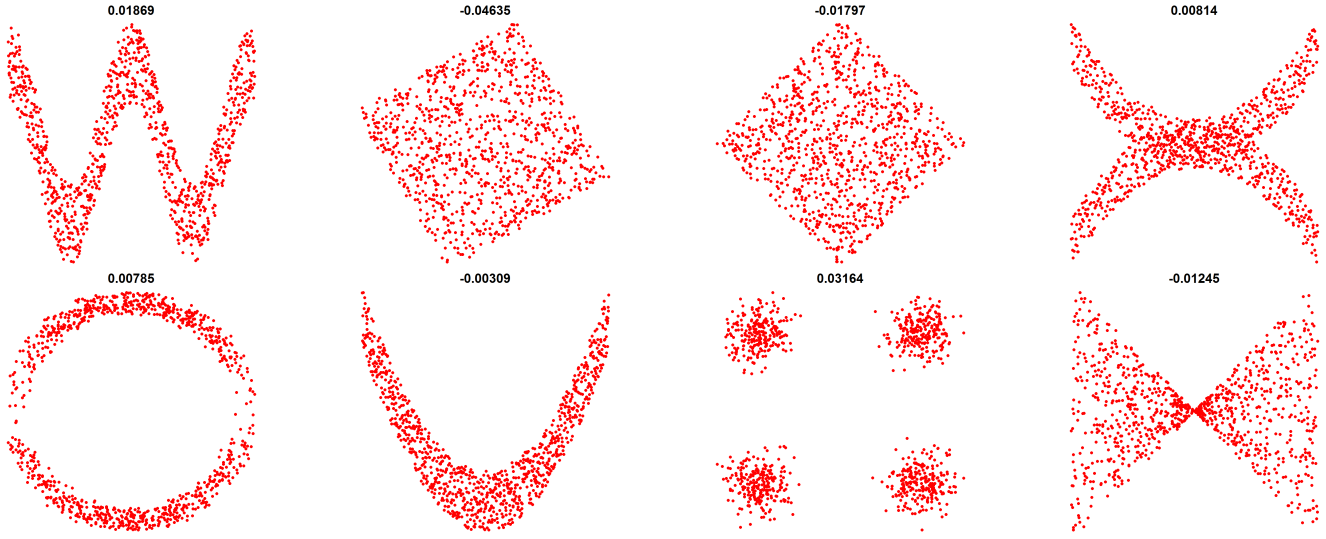
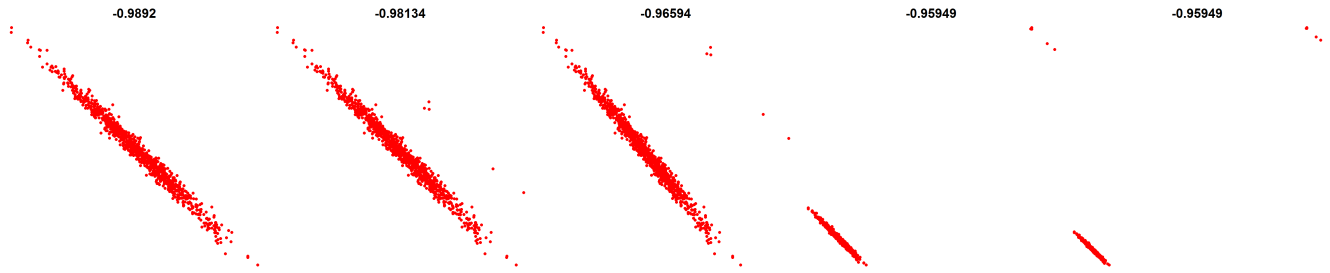


Рис. 8.7: Коэффициент корреляции Спирмена при нелинейных зависимостях между случайными величинами

[1] Небольшие изменения объясняются тем, что связь между признаками не в точности монотонная, а слегка зашумлена.

Гораздо более интересные результаты получаются в эксперименте с выбросами (рисунок 8.8). Когда из облака точек с сильной отрицательной корреляцией (рисунок 8.8a) пять точек начинают выдвигаться в правый верхний угол диаграммы рассеяния, значение коэффициента корреляции Спирмена сначала немного уменьшается (рисунки 8.8b, 8.8c). Однако как только эти пять точек оказываются за пределами диапазона изменений случайных величин в основном облаке, значение коэффициента корреляции Спирмена меняться перестает (рисунки 8.8d, 8.8e). Как бы далеко они ни отодвигались, не удаётся получить большую положительную корреляцию, как в случае с корреляцией Пирсона. Это говорит о том, что коэффициент корреляции Спирмена гораздо более устойчив к выбросам, то есть, небольшое количество точек с нетипичными значениями признаков очень слабо влияют на выборочное значение коэффициента корреляции.



(a) (b) (c) (d) (e)

Рис. 8.8: Влияние выбросов на коэффициент корреляции Спирмена

8.3. Корреляция Мэтьюса и коэффициент Крамера

8.3.1. Корреляция Мэтьюса

Коэффициент корреляции Мэтьюса — это мера силы взаимосвязи между двумя бинарными переменными. Для того чтобы его вычислить, необходимо использовать таблицу сопряженности (таблица 8.1).

X_1 \ X_2	0	1
0	a	b
1	c	d

Таблица 8.1: Таблица сопряжённости

В строках таблицы сопряжённости находятся значения одного признака, по столбцам — второго, в каждой ячейке — количество объектов, на которых реализовалась эта пара. Коэффициент корреляции Мэтьюса вычисляется по данным из таблицы сопряжённости следующим образом:

$$\mathrm{MCC}_{X_1X_2} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}.$$

Точно так же, как и коэффициенты Пирсона и Спирмена, корреляция Мэтьюса лежит в диапазоне от -1 до 1. $\mathrm{MCC}_{X_1X_2} = 0$ точно так же соответствует случаю полного отсутствия взаимосвязи между переменными. $\mathrm{MCC}_{X_1X_2} = 1$ соответствует ситуации, когда у X_1 и X_2 полностью совпадают, то есть $b = c = 0$, в выборке отсутствуют объекты, на которых значения X_1 и X_2 отличаются. $\mathrm{MCC}_{X_1X_2} = -1$ — это противоположная ситуация: в выборке нет ни одного объекта, на которых значения двух бинарных признаков совпадают.

8.3.2. Коэффициент V Крамера

Этот подход можно обобщить на случай категориальных признаков. Пусть случайная величина X_1 принимает K_1 различных значений, а X_2 — K_2 разных значений. Можно составить большую таблицу сопряженности (таблица 8.2), у которой в строке i и столбце k будет стоять n_{ij} — количество объектов выборки, на которых $X_1 = i$, а $X_2 = j$.

X_1 \ X_2	1	…	j	…	K_2
1					
⋮					
i			n_{ij}		
⋮					
K_1					

Таблица 8.2: Таблица сопряжённости $K_1 \times K_2$

На основании этой таблицы сопряженности вычисляется мера взаимосвязи между X_1 и X_2. Эта мера называется коэффициентом V Крамера. Он обозначается ϕ_c, и равен корню из специальным образом нормированного значения статистики хи-квадрат:

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

Далее будет описано, как статистика хи-квадрат считается для таблицы сопряженности.

Коэффициент Крамера принимает значения исключительно в интервале от 0 до 1, то есть он не может быть отрицательным. 0, как и раньше, соответствует полному отсутствию взаимосвязи, а 1 — полному совпадению переменных X_1 и X_2 с точностью до переименования уровней. Корреляция между двумя категориальными переменными не может быть отрицательной, поскольку уровни категориальных переменных не связаны друг с другом отношениями порядков.

8.3.3. Пары переменных разных видов

Итак, в этом разделе было описано, как считать корреляцию между парами бинарных переменных, парами категориальных, а до этого — как считать корреляцию между парами непрерывных переменных. Однако до сих пор не сказано, что делать, если признаки в паре разных видов.

Например, пусть $X_1 \in \mathbb{R}$ — непрерывный признак, а $X_2 \in \{0, 1\}$ — бинарный. Чисто теоретически на этих данных можно посчитать корреляцию Пирсона или Спирмена. Никакая из них не сломается из-за того, что одна из выборок будет не непрерывной, а бинарной. Но так делать не стоит, это очень плохо. Корреляции Пирсона и Спирмена не рассчитаны на применение к бинарным или категориальным признакам. Полученная величина будет иметь мало смысла.

На самом деле для пар признаков, один из которых непрерывный, а другой — категориальный, вообще не нужно считать никакой коэффициент корреляции. $X_1 \in \mathbb{R}$ и $X_2 \in \{0, 1\}$ будут положительно коррелированы, если

$$\mathbb{E}\left(X_1 \mid X_2 = 1\right) > \mathbb{E}\left(X_1 \mid X_2 = 0\right).$$

Таким образом, мерой силы взаимосвязи между X_1 и X_2 может служить просто разность этих математических ожиданий:

$$\mathbb{E}\left(X_1 \mid X_2 = 1\right) - \mathbb{E}\left(X_1 \mid X_2 = 0\right)$$

Эта величина не нормированная, она может меняться в любом диапазоне, от $-\infty$ до $+\infty$. Однако её гораздо легче интерпретировать, чем коэффициент корреляции, который можно вычислить на такой паре выборок.

8.4. Значимость корреляции

В этой части пойдёт речь о том, как правильно интерпретировать значения коэффициентов корреляции. В частности, будет дан ответ на вопрос, можно ли по полученному выборочному значению коэффициента корреляции сказать, что он достаточно большой и отличается от 0.

8.4.1. Корреляция непрерывных величин

За 100 дней собраны данные о значениях средней дневной температуры и количестве проданных рожков мороженого. Значение коэффициента корреляции Пирсона, посчитанное по этой выборке: $r_{X_1 X_2} = 0.45$, Спирмена:

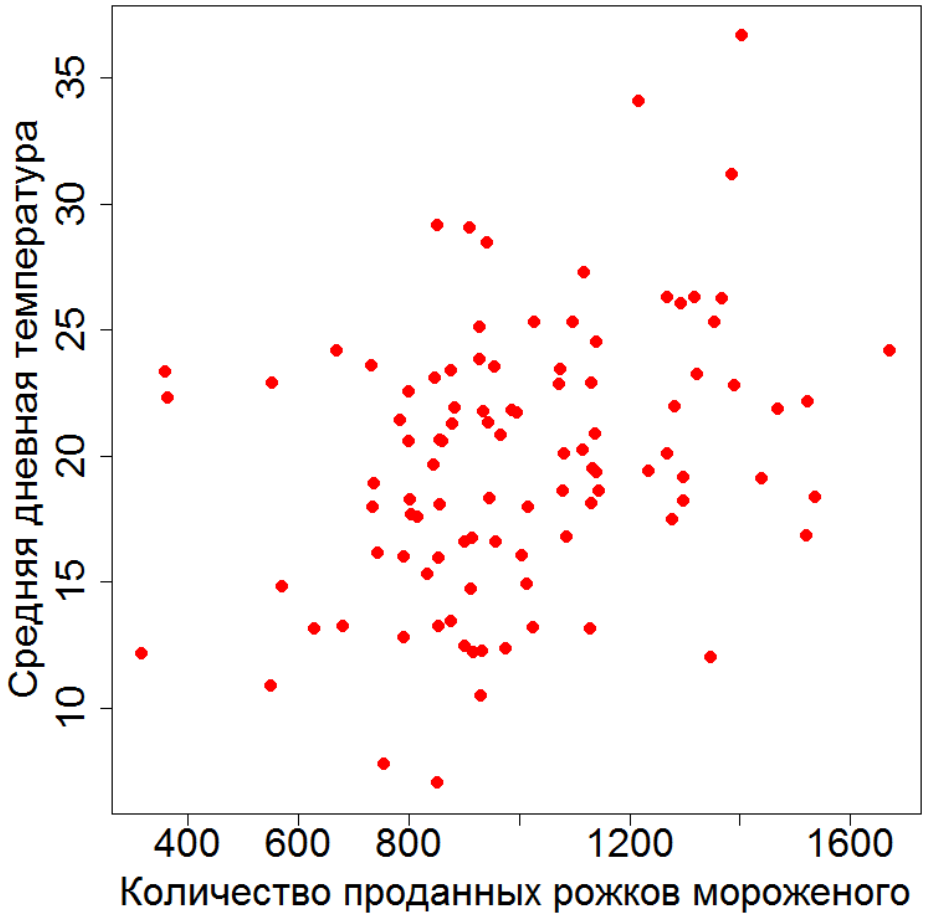


Рис. 8.9: Данные о продажах мороженого и средней дневной температуре

$\rho_{X_1X_2} = 0.44$. Можно ли по полученным значениям утверждать, что объем продаж мороженого и среднедневная температура статистически взаимосвязаны?

Ответить на этот вопрос позволяет статистический критерий Стьюдента (таблица 8.3).

выборки:	$(X_{1i}, X_{2i}), i = 1, \ldots, n,$
нулевая гипотеза:	$H_0 \colon r_{X_1X_2} = 0;$
альтернатива:	$H_1 \colon r_{X_1X_2} <\neq> 0;$
статистика:	$T = \frac{r_{X_1X_2}\sqrt{n-2}}{\sqrt{1-r^2_{X_1X_2}}};$
нулевое распределение:	$T \sim St(n-2).$

Таблица 8.3: Описание статистического критерия Стьюдента

Если нулевая гипотеза справедлива, то есть, корреляции нет, эта статистика имеет распределение Стьюдента с числом степеней свободы $n-2$ (рисунок 8.10).

Для проверки такой же точно гипотезы, но о корреляции Спирмена, а не Пирсона, можно использовать абсолютно тот же самый критерий Стьюдента.

В примере с мороженым нулевая гипотеза о том, что линейной связи нет против двусторонней альтернативы критерием Стьюдента уверенно отвергается. Признаки действительно линейно статистически взаимосвязаны. 95% доверительный интервал: $[0.28, 0.59]$. Такой доверительный интервал, кстати, можно построить, как на основе статистики критерия Стьюдента, так и с помощью бутстрепа.

Также можно использовать корреляцию Спирмена, чтобы проверить гипотезу об отсутствии монотонной взаимосвязи между двумя признаками:

$$H_0 \colon \rho_{X_1X_2} = 0,$$

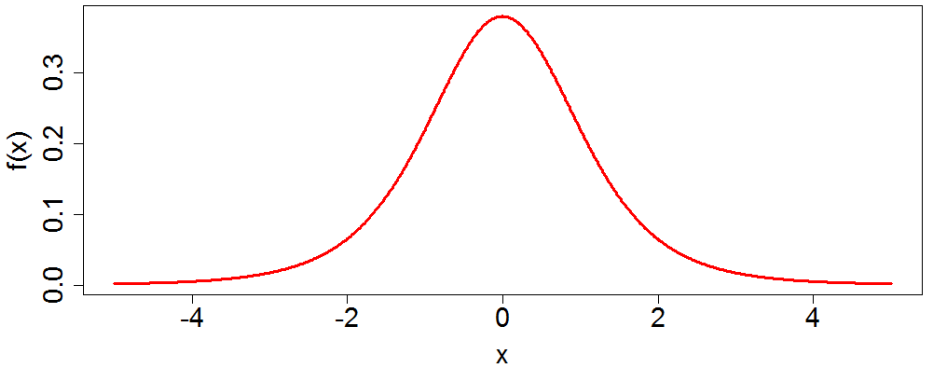


Рис. 8.10: Распределение Стьюдента

против двусторонней альтернативы

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

Критерием Стьюдента эта гипотеза также отвергается с очень похожим достигаемым уровнем значимости $p = 3 \times 10^{-6}$. Признаки действительно монотонно связаны. Это не удивительно, поскольку ранее было показано, что они связаны линейно, а линейная взаимосвязь — это частный случай монотонной. 95% доверительный интервал для корреляции Спирмена: $[0.26, 0.60]$.

8.4.2. Корреляция бинарных величин

В качестве примера работы с бинарными признаками можно рассмотреть задачу оценки эффективности тромболитической терапии по данным эксперимента, который проводился в Московской городской клинической больнице №25. В эксперименте участвовало 206 пациентов. Требуется понять, влияет ли наличие сахарного диабета у этих пациентов на эффективность тромболитической терапии.

	Выздоровели	Не выздоровели
Диабет	48	30
Нет	92	36

Таблица 8.4: Данные эксперимента по оценке эффективности тромболитической терапии

Данные эксперимента представляют собой таблицу 2×2 (таблица 8.4). Значение коэффициента корреляции Мэтьюса, подсчитанное по этой таблице: $\mathrm{MCC} = -0.1074$. Возможно, наличие сахарного диабета понижает шансы на выздоровление у пациентов. Эту гипотезу можно проверить формально с помощью критерия хи-квадрат (таблица 8.5).

выборки:	$(X_{1i}, X_{2i}), i = 1, \ldots, n,$ $X_1, X_2 \in \{0, 1\};$
нулевая гипотеза:	$H_0 \colon \mathrm{MCC}_{X_1 X_2} = 0;$
альтернатива:	$H_1 \colon \mathrm{MCC}_{X_1 X_2} \neq 0;$
статистика:	$\chi^2 = n\,\mathrm{MCC}^2_{X_1 X_2};$
нулевое распределение:	$\chi^2 \sim \chi^2_1.$

Таблица 8.5: Описание критерия хи-квадрат

Если нулевая гипотеза справедлива и значение коэффициента корреляции действительно равно 0, то статистика этого критерия имеет распределение хи-квадрат с одной степенью свободы (рисунок 8.11).

При рассмотрении задаче проверки нормальности уже шла речь о том, что критерий хи-квадрат достаточно капризный. Вот и в этом случае требуется, чтобы выборки были достаточно большими: $n \geq 40$. Кроме того, необходимо, чтобы каждая из следующих четырёх величин была больше 5:

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

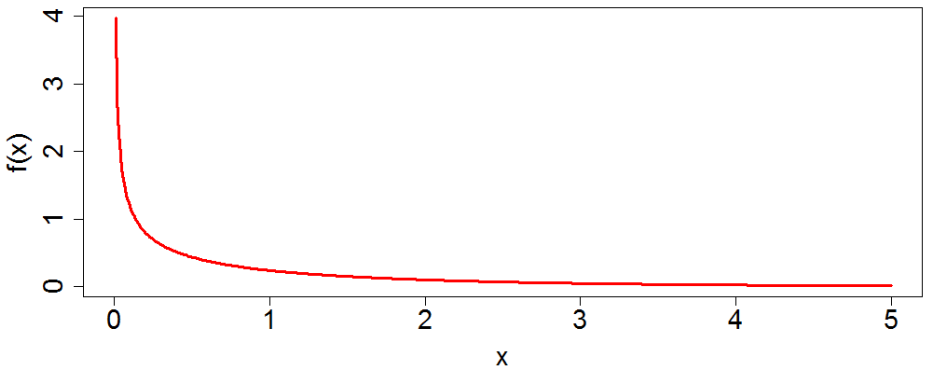


Рис. 8.11: Распределение хи-квадрат с 1 степенью свободы

Далее будет рассказано, откуда берутся эти четыре величины.

Итак, можно применить критерий хи-квадрат к данным эксперимента по оценке эффективности тромболитической терапии. Нулевая гипотеза H_0 : эффективность лечения не зависит от наличия диабета, против двусторонней альтернативы критерием хи-квадрат не отвергается. Достигаемый уровень значимости $p = 0.1651$, это больше, чем уровень значимости 0.05. Таким образом, нельзя утверждать, что между этими двумя признаками есть связь.

8.4.3. Корреляция категориальных величин

Критерий хи-квадрат можно обобщить на случай категориальных признаков (таблица 8.6). Таблица 8.7 — это таблица сопряженности для X_1 и X_2. Пусть X_1 принимает k_1 разных уровней, X_2 — k_2 разных уровней. В ячейке на пересечении строки i и столбца j стоит n_{ij} — количество объектов, на которых реализуется значение X_1 под номером i и значение X_2 под номером j. Дополнительно введены обозначения для сумм по строкам и столбцам: n_{i+} — это сумма по строке i, а n_{+j} — по столбцу j.

выборки:	$(X_{1i}, X_{2i}), i = 1, \ldots, n,$ $X_1 \in \{1, \ldots, K_1\}, \quad X_2 \in \{1, \ldots, K_2\},$
нулевая гипотеза:	H_0: X_1 и X_2 независимы;
альтернатива:	H_1: H_0 неверна;
статистика:	$\chi^2\left(X_1^n, X_2^n\right) = \sum\limits_{i=1}^{K_1}\sum\limits_{j=1}^{K_2} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}} = n\left(\sum\limits_{i=1}^{K_1}\sum\limits_{j=1}^{K_2} \frac{n_{ij}^2}{n_{i+}n_{+j}} - 1\right);$ $\chi^2\left(X_1^n, X_2^n\right) \sim \chi^2_{(K_1-1)(K_2-1)}.$

Таблица 8.6: Описание критерия хи-квадрат для категориальных признаков

X_1 \ X_2	1	…	j	…	K_2	$\sum$
1						
⋮						
i			n_{ij}			n_{i+}
⋮						
K_1						
$\sum$			n_{+j}			n

Таблица 8.7: Таблица сопряжённости для категориальных признаков

В статистике этого критерия учитывается отклонение между n_{ij}, количеством объектов в каждой ячейке, и ожидаемым количеством объектов в этой ячейке при условии справедливости нулевой гипотезы.

При справедливости нулевой гипотезы статистика критерия имеет распределение хи-квадрат (рисунок 8.12).

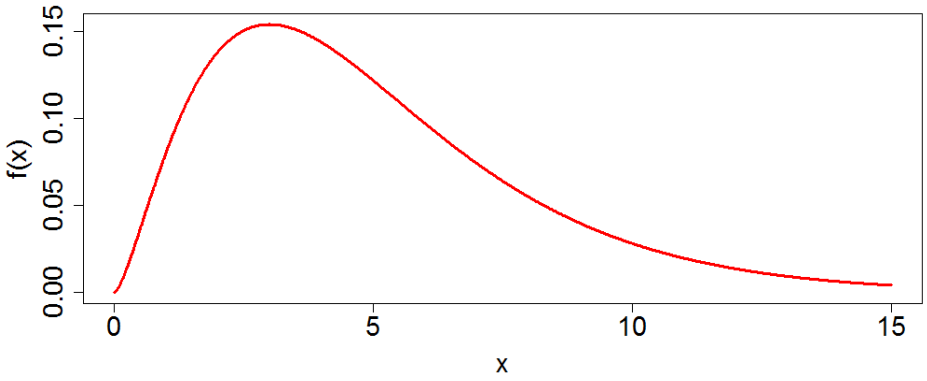


Рис. 8.12: Распределение хи-квадрат

Несложно показать, что критерий для таблиц 2×2, который рассматривался ранее, является частным случаем этого критерия.

Критерий хи-квадрат для таблиц сопряженности может применяться при выполнении следующих условий. Нужно, чтобы выборки были достаточно большими: $n \geq 40$. Кроме того, необходимо, чтобы ожидаемое количество элементов в каждой ячейке таблицы было меньше 5 ($\frac{n_{i+}n_{+j}}{n} < 5$), не более, чем в 20% ячеек.

Можно считать, что для категориальных признаков критерий хи-квадрат проверяет гипотезу о равенстве нулю коэффициента V Крамера против альтернативы, что он нулю не равен. Вообще говоря, коэффициент V Крамера определяется как раз через статистику критерия хи-квадрат:

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

8.5. Булщит и консервативность

8.5.1. Исследование и поставленный эксперимент

В конце апреля 2016 года в престижном журнале PLOS one вышла статья под названием «Восприятие «булщита» как глубокомысленного ассоциировано с поддержкой Круза, Рубио, Трампа и консерватизмом». Круз, Рубио и Трамп — это кандидаты в президенты США от республиканской партии. По определению авторов «булщит» — это бессодержательное, нелогичное или явно противоречащее элементарным научным знаниям утверждение. В качества примеров «булщита» они приводят такие фразы, как «скрытый смысл трансформирует беспрецедентную абстрактную красоту» или «воображение лежит в основе экспоненциальных пространственно-временных событий» (эти фразы получены специальным генератором «булщита»).

В статье анализируются данные эксперимента, в котором участвовали 196 граждан США, 43% из которых женщины, средний возраст испытуемых составляет 36 лет. Испытуемые (это достаточно необычно) набраны на платформе Amazon Mechanical Turk. Это платформа, на которой пользователям из Интернета можно заказать какое-то простое, достаточно механическое, не требующее высокой квалификации задание, и они его сделают за достаточно небольшие деньги. Например, типичное задание для этой платформы — это разметка картинок.

В эксперименте испытуемым нужно было ответить на ряд вопросов. Во-первых, им нужно было оценить глубокомысленность предъявляемых им утверждений по шкале от 1 (абсолютно не глубокое) до 5 (очень глубокое). Кроме того, каждый из них должен был оценить степень своей симпатии к трем кандидатам в президенты США от республиканской партии и трем кандидатам от демократической партии, также по шкале от 1 (очень не симпатичен) до 5 (очень симпатичен). Помимо этого, каждый из них должен был оценить степень консервативности собственных политических взглядов по семибалльной шкале Лайкерта, где 1 соответствует очень либеральным взглядам, а 7 — очень консервативным. Часть утверждений, предъявлен-

ных испытуемым, была «булщитом», а часть — относительно редкими поговорками (например, «промокший человек не боится дождя»).

Данные были проанализированы следующим образом. Для каждого испытуемого была вычислена средняя склонность читать «булщит» глубокомысленным. Для этого признака была посчитана корреляция Спирмена с консервативностью политических взглядов и степенями симпатии к шести кандидатам в президенты. Для проверки значимости отличия от нуля этой корреляции, использовался критерий Стьюдента.

Были получены следующие результаты. Обнаружена значимая положительная корреляция между тягой к «булщиту» и симпатией к Теду Крузу, Марку Рубио и Дональду Трампу (три кандидата от республиканской партии). Кроме того, тяга к «булщиту» положительно ассоциирована со степенью консервативности и эта корреляция тоже значима.

Это веселое исследование, но у него есть некоторые проблемы. Первая и самая важна проблема заключается в сомнительной репрезентативности выборки. Аудитория Amazon Mechanical Turk вовсе не является случайной выборкой из граждан США. В этой аудитории преобладают белые молодые мужчины с относительно невысоким доходом и достаточно хорошо образованные или получающие образование. Таким образом, непонятно, на какую генеральную совокупность можно обобщать результаты, полученные в исследовании: эта выборка крайне смещена для случая, когда хочется делать выводы о всех гражданах США. Связанная с этим проблема заключается в несбалансированности выборки по некоторым признакам.

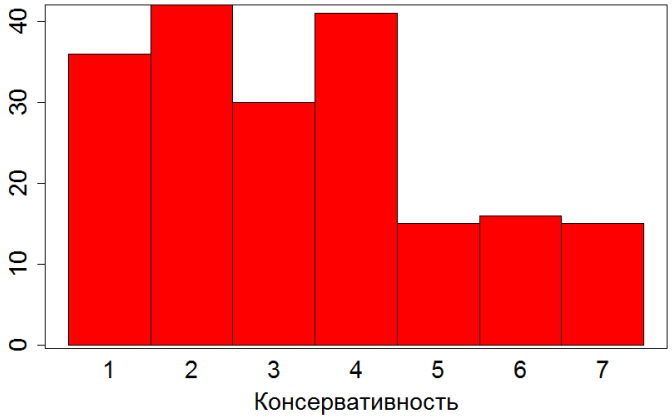


Рис. 8.13: Распределение консервативности испытуемых в выборке

На рисунке 8.13 показано распределение значения показателя консервативности в исследуемой выборке. В ней преобладают испытуемые с либеральными или умеренными взглядами, относительно небольшая часть испытуемых считает свои политические взгляды консервативными. Почему это проблема — станет понятно, если посмотреть на сырые данные.

На графике 8.14a по горизонтальной оси отложена тяга к «булщиту», а по вертикальной — степень поддержки Теда Круза. Каждая точка — это один испытуемый. Корреляция Спирмена между этими двумя признаками: $\rho_{XY} = 0.3$. Достигаемый уровень значимости критерия Стьюдента против двусторонней альтернативы: $p = 2 \times 10^{-5}$, то есть нулевая гипотеза об отсутствии корреляции отвергается. Если через это облако точек провести регрессионную прямую, то видно, что у неё, действительно, положительный наклон, но он не так уж велик. Если вместо линейной регрессии произвести локальное сглаживание методом LOESS, то получится синяя кривая, по которой видно, что картина не так однозначна. Действительно, до какой-то степени повышение тяги к «булщиту» соответствует повышению уровня поддержки Теда Круза, но это работает не на всей области определения тяги к «булщиту», признака, отложенного по горизонтальной оси.

Для сравнения можно посмотреть на графики по другим кандидатам в президенты. На рисунке 8.14b показан график для Марка Рубио. Здесь корреляция Спирмена $\rho_{XY} = 0.2$. Достигаемый уровень значимости соответствующего критерию Стьюдента против двусторонней альтернативы: $p = 0.0064$. Различия между полученным коэффициентом корреляции и нулем значимы. На графике при этом видна приблизительно та же ситуация, что и да этого, хотя угол наклона регрессионной прямой немножко уменьшился.

График для Дональда Трампа показан на рисунке 8.14c. Корреляция Спирмена здесь:$\rho_{XY} = 0.15$. Достигаемый уровень значимости $p = 0.0324$. Отличие корреляции от нуля все еще значимо на уровне значимости 0.05. Угол наклона регрессионной прямой еще немного уменьшился.

Для сравнения на рисунке 8.14d представлен график для Хиллари Клинтон. Корреляция Спирмена здесь $\rho_{XY} = 0.09$. Соответствующий достигаемый уровень значимости $p = 0.212$., то есть данные не позволяют

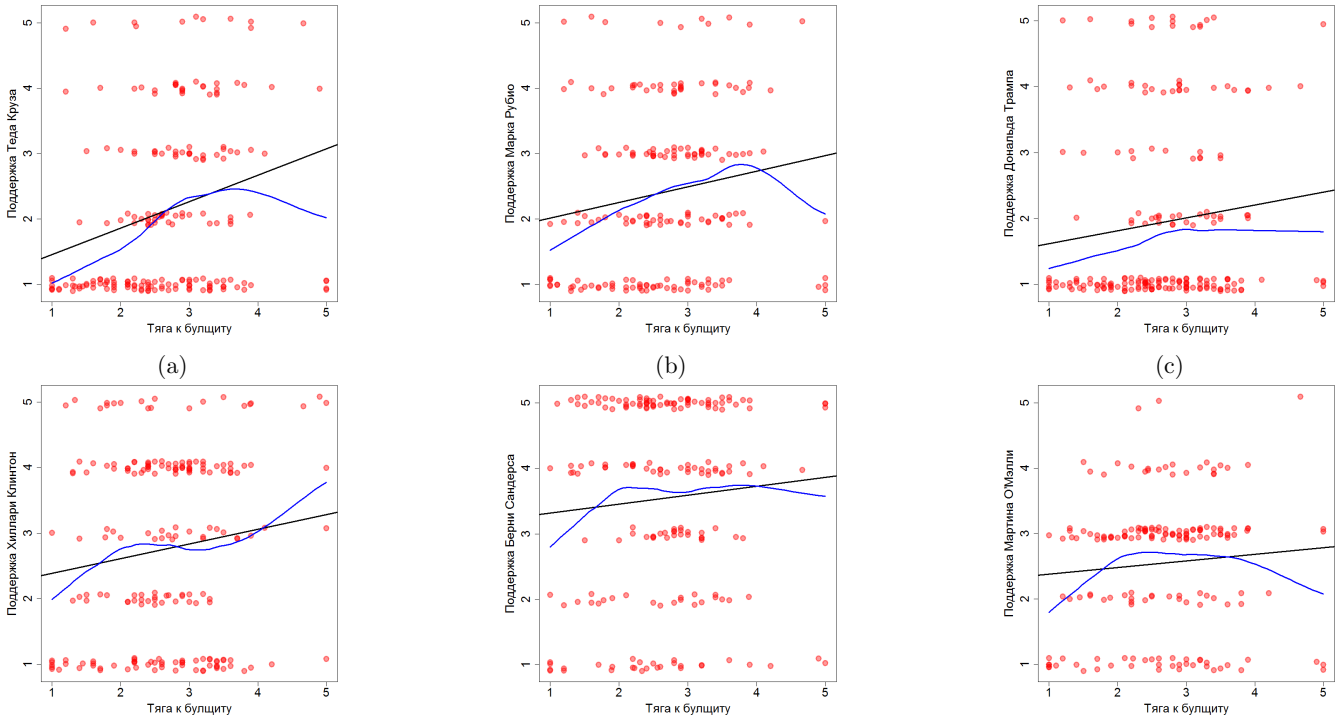


Рис. 8.14: Данные о тяге к «булщиту» и поддержке кандидатов в президенты США. В верхнем ряду — кандидаты республиканской партии, в нижнем — демократической.

отвергнуть нулевую гипотезу о том, что признаки не коррелированы. На графике, однако, происходит примерно то же самое: положительный угол наклона регрессионной прямой сохраняется, он только еще совсем немного уменьшился. Кривая, полученная методом LOESS, здесь имеет точно такой же повышающийся вид.

На рисунке 8.14 сгруппированы графики для всех шести кандидатов в президенты. В верхнем ряду — кандидаты от республиканской партии, где корреляция между признаками везде значима. В нижнем ряду — кандидаты от демократической партии, на которых корреляция везде незначима. При этом на всех графиках происходит примерно одно и то же. Основное отличие между ними, — это то, что в верхнем ряду большая часть точек сосредотачивается в облаке, соответствующем значению признака, отложенного по вертикальной оси, равного единице. Но поскольку в выборке большая часть испытуемых не придерживается консервативных взглядов, они кандидатов от республиканской партии и не поддерживают. Именно это множество точек и оказывает наибольшее влияние на коэффициенты корреляции.

Коэффициент корреляции Спирмена предназначен для работы с непрерывными признаками, а в данном случае рассматриваемые признаки существенным образом дискретны — они измерены в шкале от 1 до 5. Кроме того, лучше всего корреляции Спирмена и Пирсона и соответствующие им критерии Стьюдента работают в ситуациях, когда признаки, между которыми вычисляется корреляция, распределены нормально. В данном случае это совсем не так.

Главная мораль этой истории заключается в том, что всегда нужно смотреть на сырые данные. Никакая статистика, никакие цифры, вычисленные на этих данных, не могут полностью описать, что происходит в данных.

Корреляционный анализ — это не сосисочная машина, которой можно подать что угодно на вход, а на выходе получить идеально правильные выводы. Нужно всегда следить за качеством исходных данных и проверять, соответствует ли распределение признаков в выборках тому распределению, которое предполагается в применяемых методах.

8.6. Корреляция и причинно-следственная связь

Давайте проанализируем корреляцию между суммарными продажами мороженого за день и количеством людей, которое в этот день утонуло на всех пляжах города (рисунок 8.6e). Корреляция между этими двумя

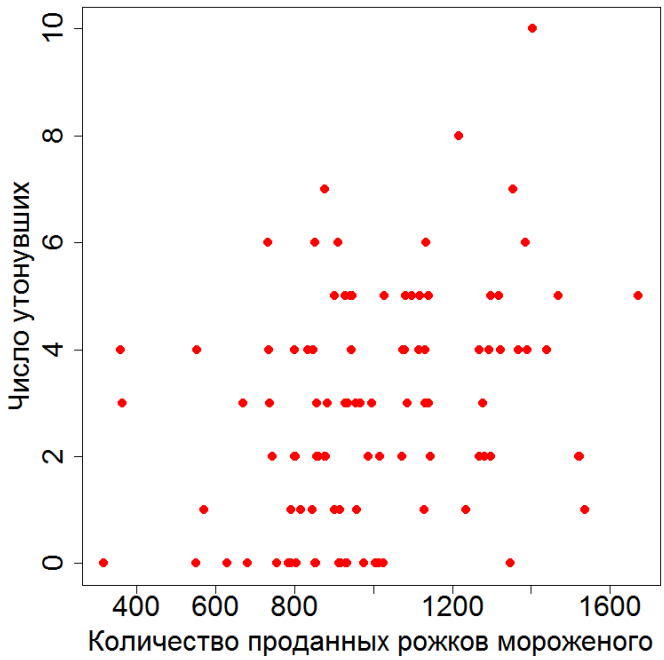
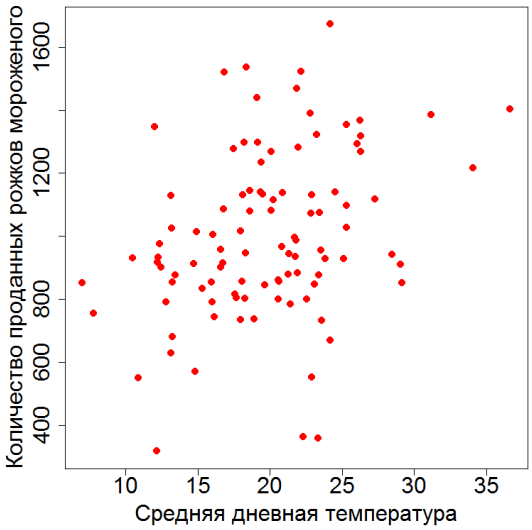
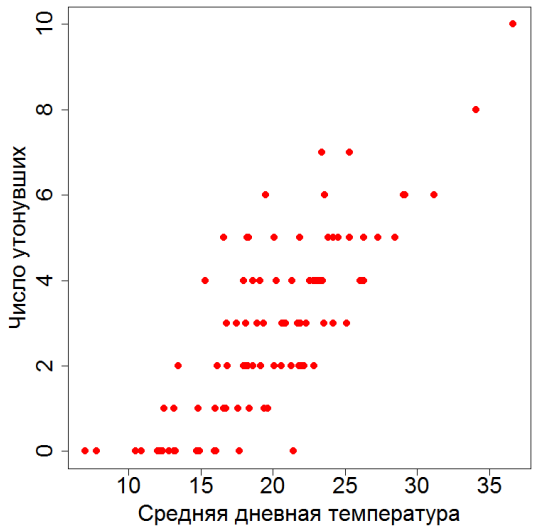


Рис. 8.15: Данные о продажах мороженого за день и количестве утонувших на пляжах города людей в этот день

признаками положительная: $r_{X_1X_2} = 0.33$. Достигаемый уровень значимости критерия Стьюдента: $p = 0.0009$. 95% доверительный интервал для корреляции Пирсона: $[0.138, 0.491]$.



(a) Данные о продажах мороженого за день и средней температуре за день



(b) Данные о продажах мороженого за день и средней температуре за день

Рис. 8.16

Из этих результатов можно сделать вывод, что чем больше люди едят мороженого, тем чаще они тонут. Или, например, что из-за того, что люди часто тонут, другие люди больше едят мороженого. Однако очевидно, что это не так. Ранее было показано, что продажи мороженого достаточно сильно скоррелированы со среднедневной температурой (рисунок 8.16a). Если посмотреть на корреляцию между среднедневной температурой и число утонувших людей (рисунок 8.16b), видно, что она еще больше, и это естественно. Таким образом, в данном примере значимость корреляции между продажами мороженого и числом утонувших

людей объясняется воздействием третьего признака — среднедневной температуры. Именно третий признак — единственный из трех, который оказывает причинно-следственное влияние на оставшиеся два. Никаких других причинно-следственных связей между этими тремя признаками быть просто не может.

В учебниках по статистике можно найти большое количество веселых примеров таких ложных корреляций, объясняющихся воздействием третьего скрытого признака. Например, количество самоубийств и радиоприемников на душу населения высоко положительно коррелировано, и это объясняется воздействием признака «размер города». Уровень углекислого газа в атмосфере планеты и распространенность ожирения также высоко положительно коррелированы — это объясняется ростом со временем уровня жизни. Рыночная доля браузера Internet Explorer и количество убийств в США тоже положительно коррелированы, и это объясняется в первую очередь фактором времени: во времени снижается и тот, и другой показатель.

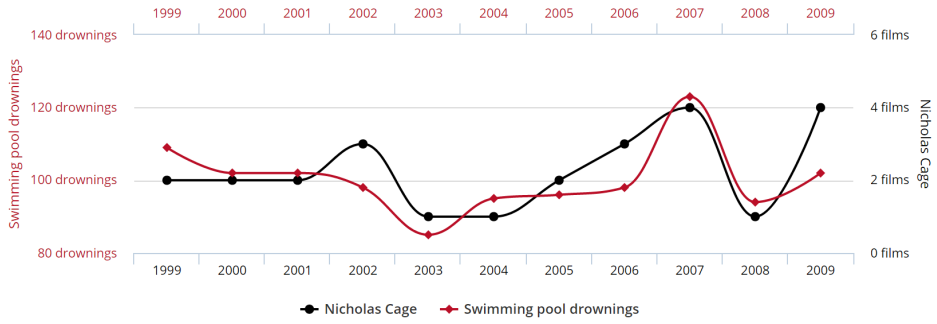


Рис. 8.17: Данные о количестве людей, утонувших в бассейне и количеством фильмов, в которых снялся Николас Кейдж

Иногда корреляцию между парой признаков нельзя объяснить даже влиянием никакого третьего другого, а эта корреляция просто случайна. Если взять достаточное количество величин и искать среди них все возможные попарные корреляции, найдётся очень много странного. Например, можно показать, что значима положительная корреляция между количеством людей, которые утонули при падении в бассейн, и количеством фильмов, в которых снялся за год Николас Кейдж. Корреляция Пирсона между этими двумя признаками $r_{X_1X_2} = 0.67$. Достигаемый уровень значимости критерия Стьюдента: $p = 0.0253$. 95% доверительный интервал для корреляции Пирсона: $[0.110, 0.905]$. Несмотря на то, что он довольно широкий, 0 в нём не содержится. Тем не менее, абсолютно очевидно, что связать эти два признака какой бы то ни было цепочкой причинно-следственных связей не представляется возможным. Этот эффект явно случайный, и то, что его нашли, — это следствие того, что его очень хорошо искали.

Главный вывод: из корреляции никогда не следует причинно-следственная связь, но из причинно«=следственной связи часто следуют корреляции. Причинно-следственная связь оставляет в данных какие-то следы, которые можно обнаружить в том числе и корреляционными методами. Однако для этого есть другие специальные методы, связанные с построением графов причинности, и лучше использовать именно их.