Меры связи: коэффициенты корреляции Пирсона и Спирмена

Алла Тамбовцева

Загрузка данных

Загрузим данные по результатам опроса студентов 1 курса (март 2020 года) по ссылке на CSV-файл:

```
dat <- read.csv("https://allatambov.github.io/psms/chip_n_dale_new.csv")
```

Опрос был посвящен мультсериалу «Чип и Дейл спешат на помощь», студенты указывали любимых героев и оценивали свои характеристики, измеряя их в героях этого мультфильма.

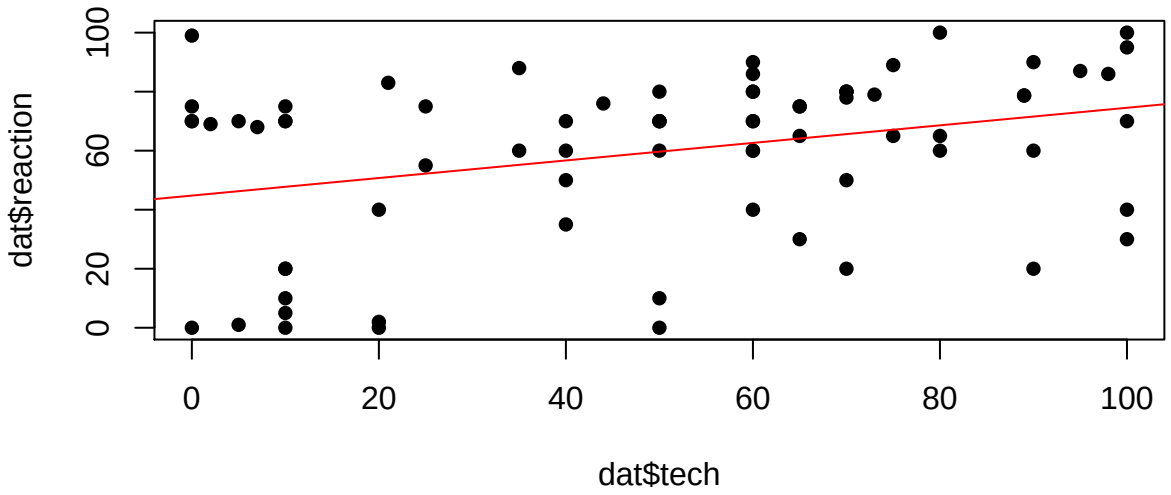
Описание переменных:

- `group`: группа;
- `female`: пол студента;
- `spec`: планируемый профиль обучения (политический анализ или управление);
- `fav`: любимый герой из команды спасателей в детстве, на кого хотелось быть похожим;
- `assoc`: с каким героем студент ассоциирует себя сейчас;
- `cheer`: показатель веселости в процентах от Дейла (Дейл = 100% веселость);
- `grump`: показатель ворчливости в процентах от Чипа (Чип = 100% ворчливость);
- `cheeze`: показатель любви к сыру в процентах от Рокфора (Рокфор = 100% любовь);
- `tech`: показатель любви к технике в процентах от Гаечки (Гаечка = 100% любовь);
- `reaction`: показатель скорости реакции в процентах от Вжика (Вжик = 100% скорость реакции).

Определение связей

Построим диаграмму рассеивания для двух показателей: любовь к технике и скорость реакции:

```
plot(dat$tech, dat$reaction, pch=16)
abline(reg = lm(dat$reaction ~ dat$tech), col = "red")
```



Пояснения к коду:

- В функции `plot()` указываем два показателя, для осих и для оси `y`. Опция `pch` определяет вид маркера для точки, здесь 16 — это закрашенная, довольно крупная, точка. Весь набор возможных маркеров можно получить, запросив помощь через строчку кода `?pch`.
- Функция `abline()` добавляет к предыдущему графику линию тренда, линию наклона облака рассеивания. Формально, это линия парной линейной регрессии, но о регрессиях мы будем говорить в следующем году.

Глядя на график, можно заметить, что какая-то положительная связь между показателями есть, но не очень сильная, средняя, ближе к слабой.

Проверим наши догадки более формально — вычислим коэффициент корреляции Пирсона:

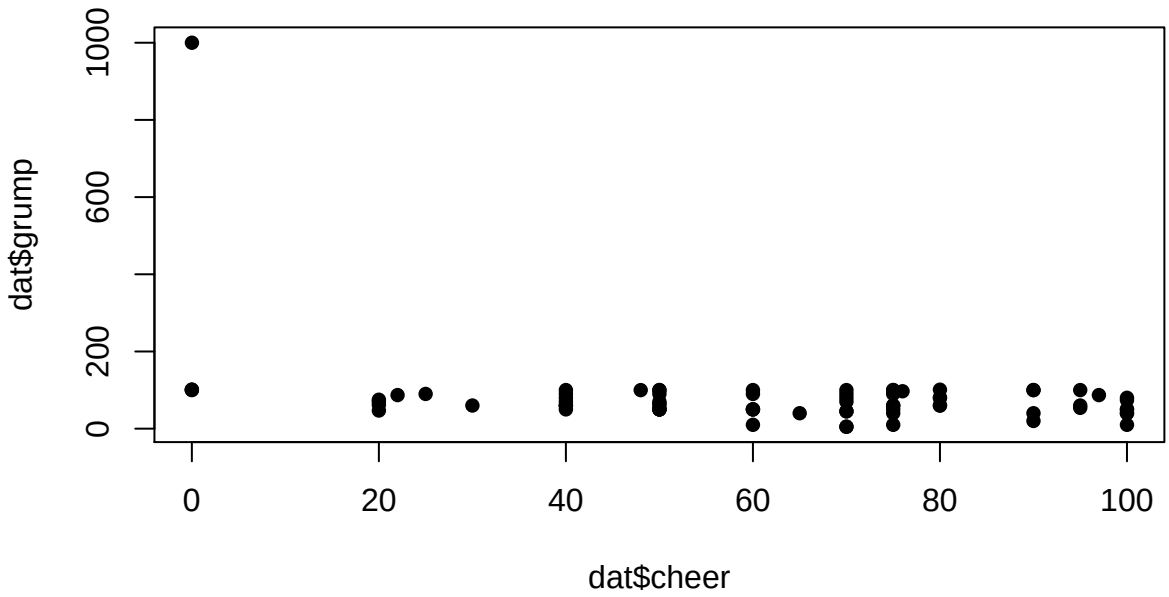
```
cor(dat$tech, dat$reaction)
```

```
## [1] 0.3325291
```

Действительно, коэффициент положительный, но не очень высокий.

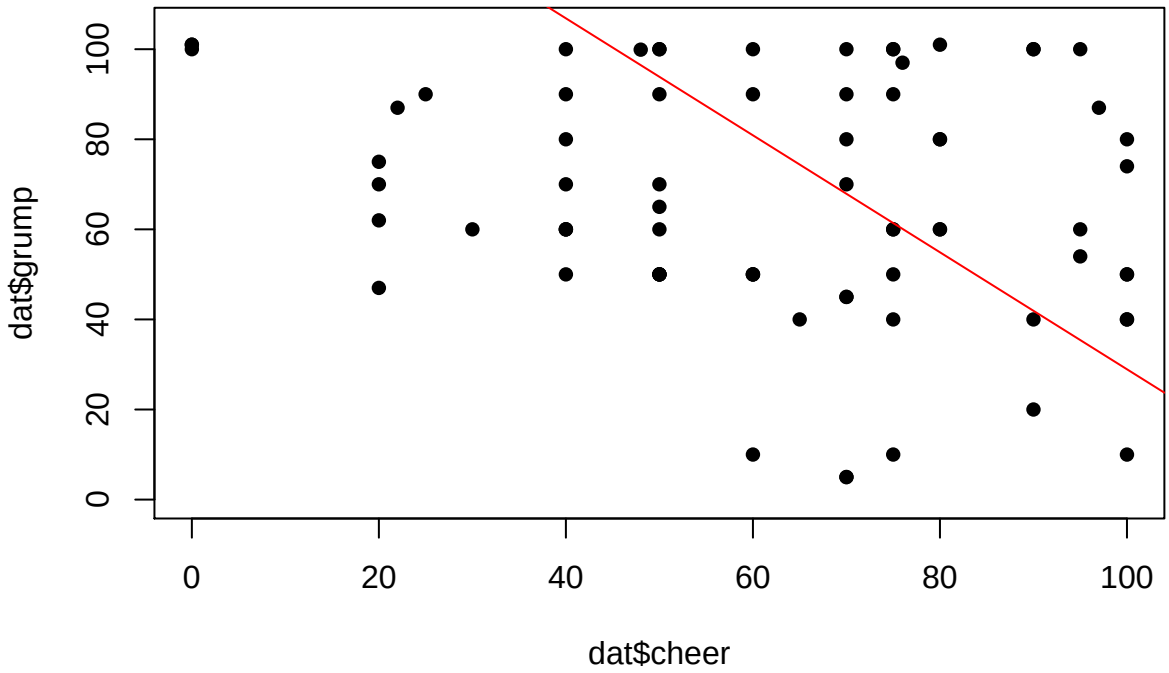
Посмотрим на другую пару показателей, более интересную. Проверим, есть ли связь между ворчливостью и любовью к сыру:

```
plot(dat$cheer, dat$grump, pch=16)
```



Что здесь интересного? То, что у нас есть нетипичное наблюдение — кто-то совсем не любит сыр и ОЧЕНЬ ворчливый, на 1000%, даже не на 100%, карантин сделал свое дело. Из-за этого наблюдения нам не очень удобно оценивать наклон остального облака точек, масштаб не самый удачный. Уберем эту точку с графика, не удаляя ее из данных, выставим ограничения по оси `y` (опция `ylim`):

```
plot(dat$cheer, dat$grump, pch=16, ylim=c(0, 105))
abline(reg = lm(dat$grump ~ dat$cheer), col = "red")
```



Можно отметить, что связь обратная, хотя облако точек очень специфическое, довольно рассеянное. Скорее всего, связь тоже довольно слабая. Проверим:

```
cor(dat$cheer, dat$grump)
```

```
## [1] -0.314393
```

Связь слабая, однако, по факту, она должна быть еще слабее — эта нетипичная точка, которую мы оставили в данных, будет «оттягивать» на себя внимание и обеспечивать более сильный его наклон. Отфильтруем наблюдения и пересчитаем коэффициент Пирсона:

```
dat2 <- subset(dat, dat$grump < 105)
cor(dat2$cheer, dat2$grump)
```

```
## [1] -0.2423091
```

Действительно, связь между показателями еще более слабая. Коэффициент Пирсона неустойчив к наличию нетипичных наблюдений, поэтому перед его вычислением надо проверить, есть ли выбросы на диаграмме рассеяния.

Теперь проверим гипотезу о равенстве теоретического коэффициента корреляции нулю. Сформулируем гипотезы:

$$H_0 : r = 0$$
$$H_1 : r \neq 0$$

Воспользуемся функцией `cor.test()`, по умолчанию она проверяет гипотезу на основе коэффициента Пирсона:

```
cor.test(dat2$cheeze, dat2$cheer)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dat2$cheeze and dat2$cheer
```

```
## t = 0.58286, df = 71, p-value = 0.5618
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1636574  0.2944014
## sample estimates:
##       cor
## 0.06900817
```

Что представлено в этой выдаче? Во-первых, название коэффициента (`Pearson's correlation`). Во-вторых, сам коэффициент Пирсона, равный 0.069 (`cor` в `sample estimates`). В-третьих, наблюдаемое значение t-статистики, которое мы до этого считали вручную по формуле:

$$t_{\text{набл}} = R\sqrt{\frac{n-2}{1-R^2}}$$

Здесь наблюдаемое значение t примерно равно 0.58. Это наблюдаемое значение статистики принадлежит распределению Стьюдента с числом степеней свободы df $= 71$, что объяснимо — число степеней свободы для распределения статистики в таком тесте равно $n-2$, а n здесь 73 (всего 74 строки, но есть одно пропущенное значение).

И, наконец, в выдаче есть p-value:

$$\text{P}(|t| > t) = 2\text{P}(t > t) = 0.5618$$

Значение p-value довольно высокое, выше любого принятого уровня значимости (1%, 5%, 10%), поэтому нулевую гипотезу об отсутствии связи мы не отвергаем. Следовательно, нет связи между любовью к сыру и ворчливостью, что вполне логично.

Напоследок посчитаем коэффициент корреляции Спирмена для еще одной пары показателей: веселость и любовь к сыру. А также проверим следующую гипотезу:

$$H_0 : \text{признаки независимы}$$

Для этого понадобится та же функция `cor.test()`, только с опцией `method = "spearman"`:

```
# "spearman" can be abbreviated to "sp"
cor.test(dat2$cheeze, dat2$cheer, method = "spearman")
```

```
## Warning in cor.test.default(dat2$cheeze, dat2$cheer, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  dat2$cheeze and dat2$cheer
## S = 60082, p-value = 0.5385
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.07315055
```

Снова видим название коэффициента корреляции (`Spearman's rank correlation rho`) и его значение 0.07 (`rho` в `sample estimates`). Наблюдаемого значения статистики ($z_{\text{набл}} = R\sqrt{n-1}$) здесь нет, но есть сумма квадратов разностей рангов $S = 60082$. И, конечно, есть p-value:

$$\text{P}(|z| > z) = 2\text{P}(z > z) = 0.5385$$

Здесь p-value снова довольно большое, поэтому гипотеза о независимости признаков не отвергается, связи между любовью к сыру и веселостью нет.