Проверка статистических гипотез в R

Алла Тамбовцева

Проверка гипотезы о доле

Знакомство с данными

Загрузим данные из файла `anames_new.csv` по ссылке и сохраним их в датафрейм (таблицу) с названием `anames`:

```
anames <- read.csv("https://allatambov.github.io/twimc/anames_new.csv",
stringsAsFactors = TRUE,
encoding = "UTF-8")
```

Пояснения к коду:

- функция `read.csv()` принимает на вход название файла или ссылку на него;
- опция `stringsAsFactors = TRUE` нужна для того, чтобы текстовые значения считались факторными; факторная переменная в R — текст с закрепленным за ним числовой меткой (как кодирование ответов в опросах для удобства);
- опция `encoding` нужна для того, чтобы файл с текстом на кириллице считался корректно, без крокозябр, на любой системе (файлы на кириллице, созданные на Mac/Linux, плохо считываются на Windows и наоборот).

Посмотрим на эти данные (также можно кликнуть на название датафрейма во вкладке *Environment*), таблица откроется в новой вкладке.

```
View(anames)
```

Описание переменных:

- `ID`: id наблюдения;
- `Name`: имя ребенка;
- `NumberOfPersons`: число родившихся с таким именем;
- `Year`: год;
- `Month`: месяц;
- `Sex`: пол ребенка;
- `Name_en`: транслитерация имени;
- `Month_en`: транслитерация месяца.

Отбор данных

Так как данные представлены за разные месяцы за разные годы, причем как по мальчикам, так и по девочкам, давайте для определённости и возможности сравнений выберем сначала только те строки датафрейма, которые соответствуют девочкам, родившимся в марте 2020 года, а затем — мальчикам, родившимся в марте 2020 года.

```
girls <- anames[anames$Year == 2020 & anames$Month_en == "Mart" & anames$Sex == "female", ]

boys <- anames[anames$Year == 2020 & anames$Month_en == "Mart" & anames$Sex == "male", ]
```

Пояснения к коду:

- R видит столбцы `Year`, `Month_en` и `Sex` только внутри датафрейма `anames`, без обращения к нему он не поймет, откуда эти значения брать. Поэтому мы вызываем их через `$`.
- Условия для отбора наблюдений прописываются в квадратных скобках. На первом месте указываются фильтры для строк, на втором – для столбцов. Здесь нам нужны все столбцы, поэтому на втором месте после запятой в квадратных скобках ничего нет, ограничений не ставим.
- А на строки мы накладываем ограничения, мы хотим отобрать только те строки в таблице, где в столбцах `Year`, `Month_en` и `Sex` стоят определенные значения. Проверка равенства значению осуществляется с помощью оператора `==`, а одновременность выполнения условий обеспечивается оператором `&` (AND).

Проверка гипотезы о равенстве доли числу

Давайте проверим формально гипотезу о том, что доля девочек с именем Анна равна 0.05. Это будет доля девочек в генеральной совокупности (все родившиеся девочки в Москве в марте 2020), мы хотим получить вывод, касающийся не только тех девочек, которые есть нашей таблице.

Сформулируем гипотезу и двустороннюю альтернативу:

$$H_0 : p = 0.05$$

$$H_1 : p \neq 0.05$$

Для того, чтобы проверить гипотезу о равенстве числу с помощью z-теста в R, нам понадобятся два значения: число успехов `x` (число девочек с именем Анна) и общее число испытаний `n` (общее число девочек).

Число девочек с именем Анна можно посмотреть в таблице `girls`, это 170. Общее число родившихся девочек – это сумма значений столбца `NumberOfPersons`:

```
n1 <- sum(girls$NumberOfPersons)
n1
```

```
## [1] 3716
```

Итак, всё для теста у нас готово. Запустим его с помощью функции `prop.test()`:

```
prop.test(x = 170, n = n1, p = 0.05)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  170 out of n1, null probability 0.05
## X-squared = 1.3262, df = 1, p-value = 0.2495
## alternative hypothesis: true p is not equal to 0.05
## 95 percent confidence interval:
##  0.03936146 0.05309171
## sample estimates:
##          p
## 0.04574812
```

Пояснения к коду:

- В `x` записываем число успехов, в `n` — общее количество испытаний.
- Значение из нулевой гипотезы записываем в `p`. По умолчанию альтернативная гипотеза выбирается двусторонняя, это можно изменить, но мы сделаем это позже.

Перейдем к выдаче. Вариант теста, который использует R для проверки гипотезы, немного отличается от того, что мы использовали на занятии, но это нестрашно. Что мы видим в выдаче?

Во-первых, наблюдаемое значение статистики, `X-squared`. Это наблюдаемое значение z-статистики, возведённое в квадрат (вспомните про распределение хи-квадрат с одной степенью свободы). Во-вторых, p-value, то есть такая вероятность:

$$\text{p-value} = \text{P}(\text{X-squared} > 1.3262) = \text{P}(|Z| > \sqrt{1.3262}) = 2\text{P}(Z > 1.152) = 0.2495.$$

Если мы выберем уровень значимости 5% ($\alpha = 0.05$), то нулевую гипотезу не следует отвергать, так как p-value больше уровня значимости. Следовательно, на 5%-ном уровне значимости мы можем утверждать, что доля девочек с именем Анна равна 0.05.

Помимо значения статистики и p-value функция выдает значение выборочной доли $\hat{p}$ (значение 0.0457 в `sample estimates`) и рассчитывает 95%-ный доверительный интервал для доли (уровень доверия при желании можно изменить, опция `conf.level`).

Проверим такую же гипотезу, но уже для мальчиков с именем Александр:

```
n2 <- sum(boys$NumberOfPersons)
prop.test(x = 210, n = n2, p = 0.05)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  210 out of n2, null probability 0.05
## X-squared = 0.55504, df = 1, p-value = 0.4563
## alternative hypothesis: true p is not equal to 0.05
## 95 percent confidence interval:
##  0.04606376 0.06020857
## sample estimates:
##          p
## 0.05269762
```

Та же история, на 5%-ном уровне значимости не отвергаем гипотезу о равенстве доли мальчиков с именем Александр 0.05.

Проверка гипотезы о среднем

Загрузка данных

Загрузим данные с обработанными результатами психологического опросника:

```
cowles <- read.csv("https://allatambov.github.io/twimc/Cowles.csv")
View(cowles)
```

Описание переменных:

- `extraversion`: уровень экстраверсии, от 1 до 20;
- `neuroticism`: уровень невротичности (тревожности), от 1 до 20;
- `sex`: пол респондента;
- `volunteer`: участвует ли респондент в волонтёрской деятельности.

Проверим гипотезу о равенстве среднего уровня экстраверсии 15. Сформулируем эту гипотезу:

$$H_0 : \mu = 15$$

Сформулируем двустороннюю альтернативу:

$$H_1 : \mu \neq 15$$

Запустим функцию `t.test()`, чтобы воспользоваться критерием Стьюдента для одной выборки:

```
t.test(cowles$extraversion, mu = 15)
```

```
##
##  One Sample t-test
##
## data:  cowles$extraversion
## t = -25.433, df = 1420, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 15
## 95 percent confidence interval:
##  12.17036 12.57560
## sample estimates:
## mean of x
##  12.37298
```

Пояснения к коду:

- На первом месте указываем сам показатель, выбираем столбец через `$` из таблицы `cowles`.
- Затем формулируем нулевую гипотезу, задаём значение среднего генеральной совокупности `mu`.

Вернёмся к выдаче. Здесь у нас есть наблюдаемое значение статистики, число степеней свободы и p-value. Наблюдаемое значение статистики – то самое значение, которое мы умеем считать по формуле:

$$t = \frac{\bar{x} - a}{\frac{s}{\sqrt{n}}}.$$

Здесь это наблюдаемое значение равно -25.433.

Число степеней свободы равно 1420, это ожидаемо. Оно вычисляется как $\mathrm{df} = \mathrm{n} - 1$, а количество наблюдений (строк в таблице) здесь 1421.

Значение p-value в такого рода задачах мы не считали вручную в силу ограниченности имеющихся таблиц распределения, но в теории это следующая вероятность:

$$\text{p-value} = \mathrm{P}(|t| > t) = 2\mathrm{P}(t > t) = 2.2e - 16 = 2.2 \times 10^{-16} \approx 0.$$

В данном случае p-value примерно равняется 0, поэтому нулевая гипотеза о равенстве среднего уровня экстраверсии 15 отвергается на любом разумном уровне значимости.

В выдаче есть 95%-ный доверительный интервал для среднего и выборочное среднее $\bar{x} = 12.37$ (`mean` в `sample estimates`). Кроме того, R напоминает нам альтернативную гипотезу, чтобы мы не забыли, на что соглашаемся в случае, если нулевую гипотезу отвергаем.

Изменим тип альтернативной гипотезы на одностороннюю. Исходя из данных, альтернативу нужно выбирать левостороннюю, так как выборочное среднее меньше значения 15 из гипотезы:

$$H_1 : \mu < 15$$

Проверяем в R:

```
t.test(cowles$extraversion, mu = 15, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  cowles$extraversion
## t = -25.433, df = 1420, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 15
## 95 percent confidence interval:
##      -Inf 12.54299
## sample estimates:
## mean of x
##  12.37298
```

Итак, снова делаем вывод о том, что нулевую гипотезу надо отвергнуть. По идее, значение p-value здесь в 2 раза меньше, чем в предыдущем случае, так как это следующая вероятность:

$$\text{p-value} = \mathrm{P}(t > t)$$

Однако здесь из-за того, что p-value примерно 0, это заметить невозможно.

В завершение обзора статистических тестов посмотрим на то, как проверить гипотезу о равенстве средних в двух группах (двух генеральных совокупностях). Проверим, можно ли считать средний уровень экстраверсии одинаковым у волонтеров и не-волонтеров.

Сформулируем гипотезы:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Применим критерий Стьюдента для двух выборок:

```
t.test(cowles$extraversion ~ cowles$volunteer)
```

```
##
##  Welch Two Sample t-test
##
## data:  cowles$extraversion by cowles$volunteer
## t = -4.6907, df = 1270.1, p-value = 3.018e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3860765 -0.5685632
## sample estimates:
##  mean in group no mean in group yes
##          11.96238          12.93970
```

Пояснения к коду:

- Через ~ указываем показатель группировки — тот, что даёт деление на две группы.

Разберём полученную выдачу. Итак, у нас снова есть наблюдаемое значение t-статистики, равное -4.6907, число степеней свободы (считается сложно, можно посмотреть в задачнике, нормально, что оно дробное) и p-value. Значение p-value примерно 0, поэтому мы можем на любом разумном уровне значимости отвергнуть нулевую гипотезу о равенстве средних в двух группах. Следовательно, есть основания считать, что средний уровень экстраверсии отличается у волонтёров и не-волонтёров. Это

вполне объяснимо и, как можно определить по выборочным средним, средний уровень экстраверсии выше у волонтёров (`mean in group yes` равно 12.94, `mean in group no` равно 11.96).