Визуализация данных и доверительные интервалы в R

Алла Тамбовцева

Загрузка данных

Загрузим данные по героям волшебного мира Дж.Роулинг по ссылке на CSV-файл и посмотрим на них:

```
hp <- read.csv("https://allatambov.github.io/twimc22/HP.csv",
               dec = ",", stringsAsFactors = TRUE)
View(hp)
```

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 `dec` сообщает R, что в качестве десятичного разделителя в дробях используется запятая (в R по умолчанию точка, записи с запятыми он не воспринимает как числа);
- функция `View()` открывает таблицу в новой вкладке внутри RStudio, тот же результат можно получить, кликнув на название датафрейма во вкладке *Environment*.

Описание переменных:

- `Name`: имя героя;
- `Gender`: пол героя;
- `Job`: должность/статус;
- `House`: факультет/школа;
- `Patronus`: патронус;
- `Species`: вид;
- `Blood.status`: статус крови;
- `Hair.colour`: цвет волос;
- `Eye.colour`: цвет глаз;
- `Loyalty`: кому герой предан;
- `Skills`: навыки;
- `Birth`: дата рождения;
- `Age`: возраст;
- `Wand.length`: длина волшебной палочки.

Описание данных

Запросим структуру загруженной таблицы:

```
str(hp)
```

```
## 'data.frame':    140 obs. of  14 variables:
##  $ Name        : Factor w/ 140 levels "(Bill) William Arthur Weasley",..: 65 118 68 4 121 97 53 57 6
```

```
##  $ Gender      : Factor w/ 3 levels "","Female","Male": 3 3 2 3 3 3 3 3 2 3 ...
##  $ Job         : Factor w/ 66 levels "","\nBlack family's house-elf (?-1996), \nHarry Potter's house-
##  $ House       : Factor w/ 7 levels "","Beauxbatons Academy of Magic",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Patronus    : Factor w/ 21 levels "","Boar","Cat",..: 17 10 14 15 13 12 19 19 8 19 ...
##  $ Species     : Factor w/ 10 levels "Centaur","Ghost",..: 5 5 5 5 3 5 5 5 5 5 ...
##  $ Blood.status: Factor w/ 16 levels "","Half-blood",..: 2 10 6 2 9 10 10 10 10 6 ...
##  $ Hair.colour : Factor w/ 37 levels "","Auburn","Bald",..: 4 20 7 27 4 5 20 20 20 4 ...
##  $ Eye.colour  : Factor w/ 26 levels "","Astonishingly blue",..: 7 4 8 4 3 1 8 8 6 8 ...
##  $ Loyalty     : Factor w/ 20 levels "","Albus Dumbledore | Dumbledore's Army | Order of the Phoenix
##  $ Skills      : Factor w/ 95 levels "","A highly accomplished Auror and an outstanding duellist, als
##  $ Birth       : Factor w/ 113 levels "","1 April, 1978 ",..: 74 26 51 92 77 71 25 2 43 16 ...
##  $ Age         : int  41 41 42 116 93 41 20 43 40 42 ...
##  $ Wand.length : num  11 12 10.8 15 16 ...
```

R сообщает нам, что в датафрейме 140 наблюдений (строк) и 14 переменных (столбцов), а также показывает, какого типа каждый столбец. В нашем случае очень много факторных (текстовых) столбцов, один целочисленный (`int` от *integer*) и один просто числовой (`num` от *numeric*).

Теперь запросим более интересную инофрмацию — описательные статистики для всех столбцов в датафрейме:

```
summary(hp)
```

```
##                                     Name        Gender
##  (Bill) William Arthur Weasley        :  1         : 1
##  Aberforth Dumbledore                 :  1   Female:49
##  Alastor Moody                        :  1   Male  :90
##  Albus Percival Wulfric Brian Dumbledore:  1
##  Albus Severus Potter                 :  1
##  Alecto Carrow                        :  1
##  (Other)                              :134
##                                                                                  Job
##  Student                                                                          :52
##                                                                                   :19
##  Advance Guard                                                                    : 3
##  Auror                                                                            : 3
##  Professor of Divination                                                          : 2
##  \nBlack family's house-elf (?-1996), \nHarry Potter's house-elf, \nHogwarts kitchen worker : 1
##  (Other)                                                                          :60
##                            House             Patronus
##                               :39   Unknown      :75
##  Beauxbatons Academy of Magic: 3   Non-corporeal:28
##  Durmstrang Institute        : 1                 :10
##  Gryffindor                  :38   None         : 7
##  Hufflepuff                  :13   Cat          : 2
##  Ravenclaw                   :18   Doe          : 2
##  Slytherin                   :28   (Other)      :16
##                  Species                       Blood.status  Hair.colour
##  Human               :105   Pure-blood or half-blood:38    Black  :25
##  Human               : 19   Pure-blood              :34           :17
##  Ghost               :  6   Half-blood              :23    Red    :14
##  Half-Human/Half-Giant:  2                           :17    Brown  :12
##  House elf           :  2   Muggle-born             : 7    Grey   :11
##  Werewolf            :  2   Pure-blood or Half-blood: 5    Blonde :10
##  (Other)             :  4   (Other)                 :16    (Other):51
##    Eye.colour
```

```
##         :54
##  Brown  :16
##  Blue   :13
##  Grey   :12
##  Dark   :11
##  Black  : 7
##  (Other):27
##                                                                                                  Loyalty
##                                                                                                      :51
##  Order of the Phoenix                                                                                :16
##  Dumbledore's Army |Hogwarts School of Witchcraft and Wizardry                                       :14
##  Lord Voldemort  | Death Eaters                                                                     :12
##  Dumbledore's Army | Order of the Phoenix | Hogwarts School of Witchcraft and Wizardry: 8
##  Hogwarts School of Witchcraft and Wizardry                                                          : 8
##  (Other)                                                                                             :31
##       Skills                                    Birth          Age
##          :27                                       : 13   Min.   : 15.00
##  Chaser  : 7   Pre 976                            :  4   1st Qu.: 39.50
##  Beater  : 4    1 September 1979- 31 August 1980:  3   Median : 43.00
##  Prefect : 4   Pre 1964                           :  3   Mean   : 48.91
##  Duelling: 3    1 September 1975- 31 August 1976:  2   3rd Qu.: 57.00
##  Auror   : 2    1 September 1978- 31 August 1979:  2   Max.   :141.00
##  (Other) :93   (Other)                           :113   NA's   :25
##   Wand.length
##  Min.   : 8.00
##  1st Qu.: 9.50
##  Median :10.75
##  Mean   :11.28
##  3rd Qu.:12.75
##  Max.   :16.00
##  NA's   :115
```

Для факторных столбцов функция `summary()` выдает количество различных значений, для числовых — набор базовых описательных статистик. В него входят: минимальное и максимальное значения (`Min` и `Max`), среднее (`Mean`), медиана (`Median`) и нижний и верхний квартили (`1st Qu` и `3rd Qu`). Кроме того, если в столбце встречаются пропущенные значения, R тоже об этом сообщает (`NA` и их количество).

При желании можно запросить описательные статистики отдельно для какого-нибудь столбца. Выберем столбец со значениями возраста героев:

```
summary(hp$Age)
```

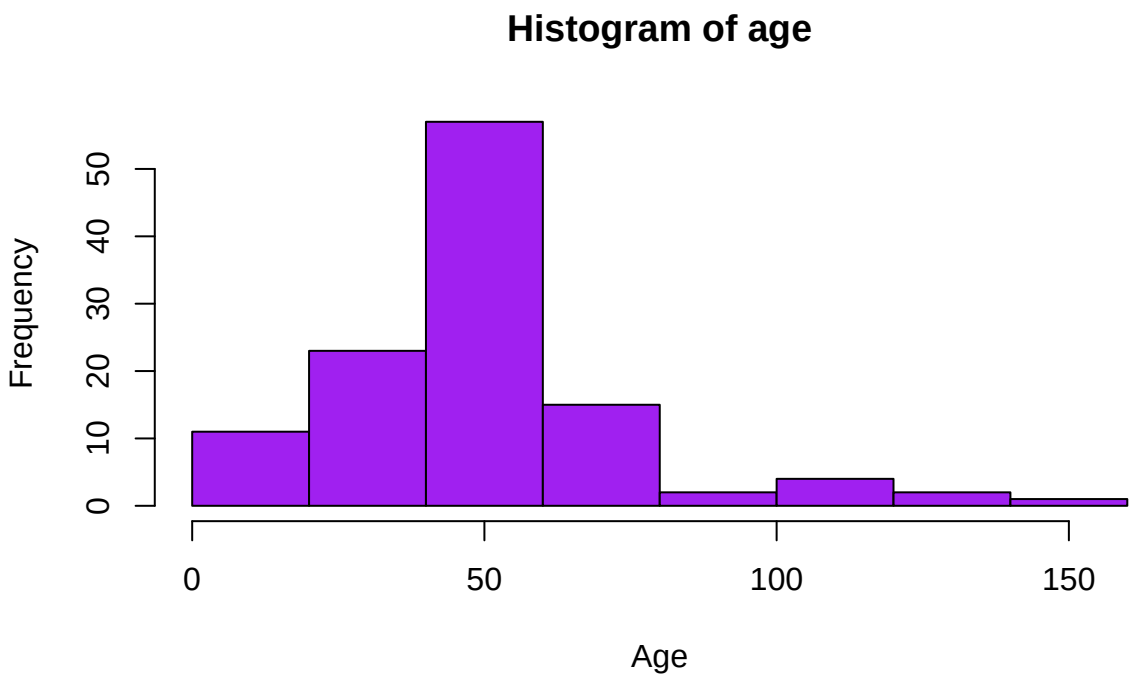
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   15.00   39.50   43.00   48.91   57.00  141.00      25
```

Проинтерпретируем результаты. Самому молодому герою 15 лет, самому старому — 141 год. В среднем, героям из волшебного мира примерно 49 лет, причём возраст половины героев не превышает 43 года. У 25% героев возраст не более 39.5 лет, а у 75% героев — не более 57 лет. По 25 героям никакой информации о возрасте у нас нет.

Визуализация данных

Построим гистограмму для возраста героев:

```
hist(hp$Age, col = "purple", xlab = "Age", main = "Histogram of age")
```

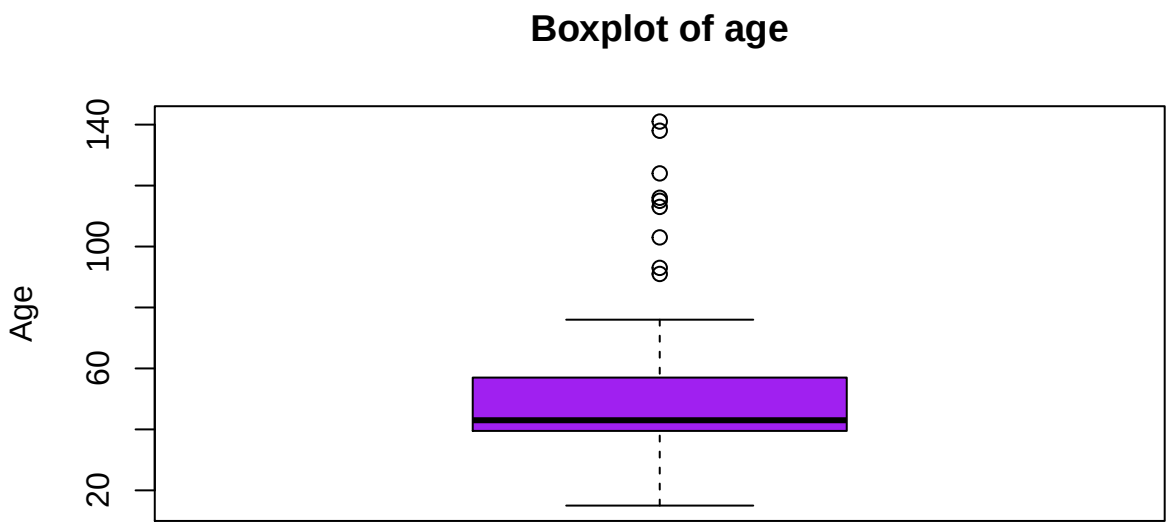


Пояснения к коду:

- Функция `hist()` строит гистограмму для числового показателя, в данном случае мы через `$` выбираем из датафрейма `hp` столбец `Age`.
- Опция `col` задаёт цвет заливки графика. Полный список цветов в R см. здесь.
- Опция `xlab` задаёт подпись по оси `x`.
- Опция `main` задаёт заголовок графика.

Как можно заметить, распределение возраста не похоже на симметричное, оно скошено, большинство значений сконцентрировано в районе 40-50 лет (да, выжившие герои уже давно не дети), при этом на графике явно видны нетипичные, слишком высокие, значения. Проверим это — построим ящик с усами:

```
boxplot(hp$Age, col = "purple", ylab = "Age", main = "Boxplot of age")
```

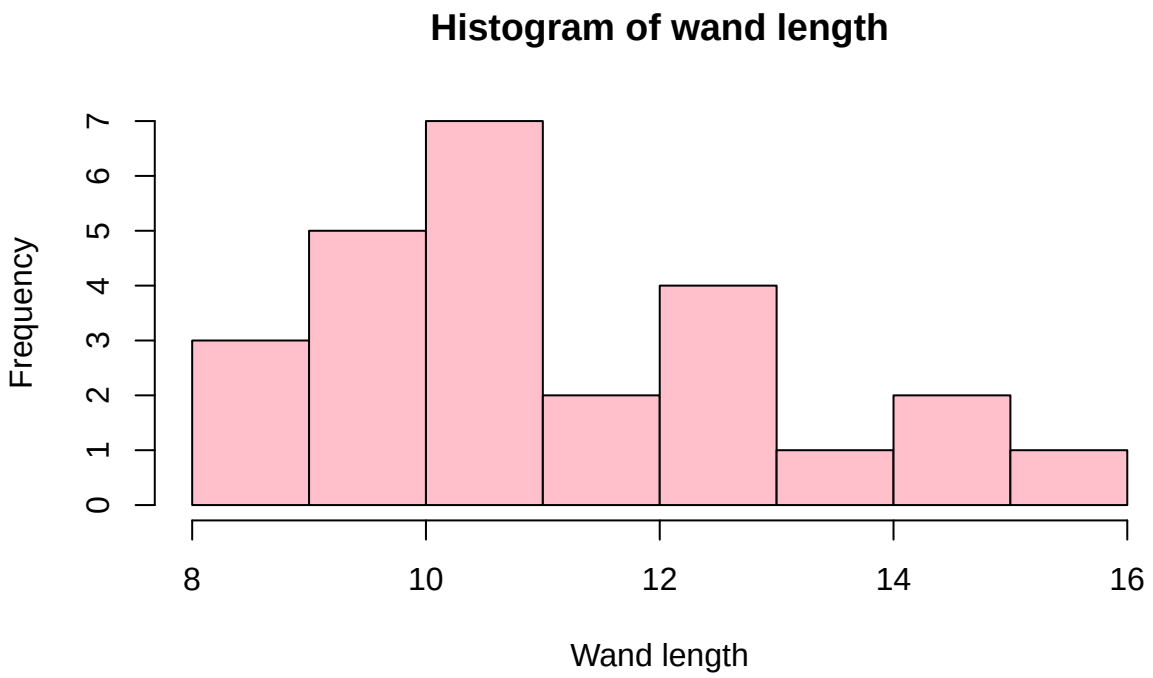


Действительно, судя по графику, выбросы есть, и находятся они в области нетипично больших значе-

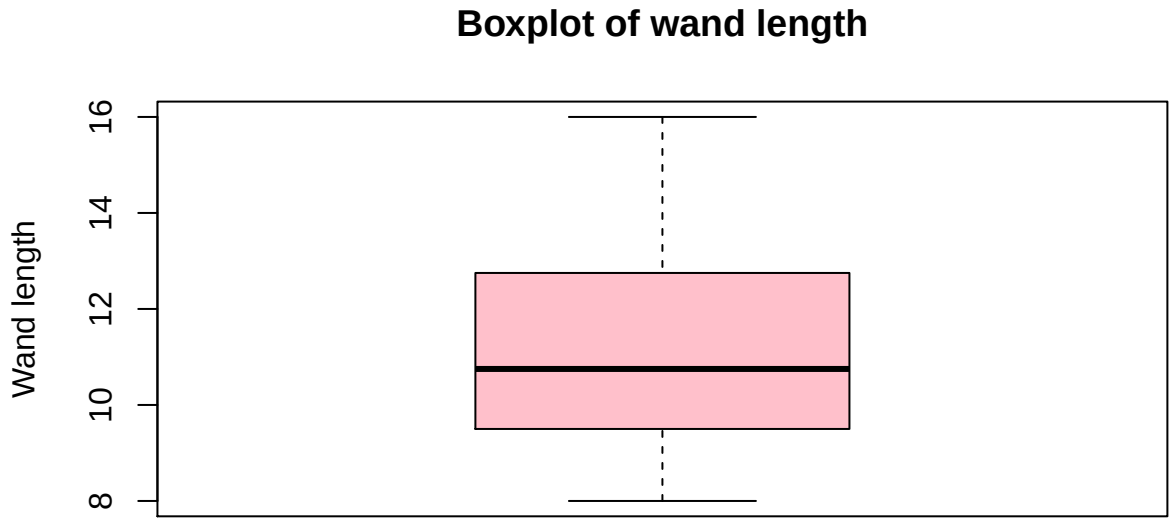
ний.

Построим гистограмму и ящик с усами для длины волшебных палочек:

```
hist(hp$Wand.length, col = "pink", xlab = "Wand length",
     main = "Histogram of wand length")
```




```
boxplot(hp$Wand.length, col = "pink", ylab = "Wand length",
        main = "Boxplot of wand length")
```



Распределение показателя также не похоже на симметричное, но нетипичных значений здесь уже не наблюдается.

Доверительный интервал для доли

Допустим, мы хотим построить доверительный интервал для доли героев, которые когда-то учились на Гриффиндоре. Несложно заметить, что таких героев у нас 38 из 140:

```
summary(hp$House)
```

```
##                              Beauxbatons Academy of Magic
##                           39                            3
##         Durmstrang Institute                   Gryffindor
##                            1                           38
##                   Hufflepuff                    Ravenclaw
##                           13                           18
##                    Slytherin
##                           28
```

Для построения доверительных интервалов нам понадобится библиотека `DescTools`, её нужно установить:

```
install.packages("DescTools")
```

Установить библиотеку достаточно один раз, потом ее нужно будет вызывать через `library()`, чтобы R понимал, откуда брать те или иные специфические функции. Импортируем библиотеку:

```
library(DescTools)
```

Теперь воспользуемся функцией `BinomCI()` из этой библиотеки, она построит нам доверительный интервал для доли:

```
BinomCI(x = 38, n = 140, conf.level = 0.95)
```

```
##            est    lwr.ci    upr.ci
## [1,] 0.2714286 0.2046045 0.3504612
```

Пояснения к коду:

- Функция `BinomCI()` принимает на вход число успехов `x` и общее количество испытаний `n`.
- Опция `conf.level` задаёт уровень доверия. По умолчанию уровень доверия 0.95, здесь мы строим 95%-ный доверительный интервал для доли.

Выдача R для доверительного интервала довольно лаконичная. Это сама доля `est` ($\hat{p} = 0.27$), нижняя граница интервала `lwr.ci` (0.204) и верхняя граница интервала `upr.ci` (0.350). Можем проинтерпретировать полученный интервал следующим образом: с 95%-ной уверенностью можно утверждать, что доля героев-гриффиндорцев лежит в интервале от 0.20 до 0.35.

При желании можем построить аналогичные интервалы для других факультетов:

```
BinomCI(x = 28, n = 140, conf.level = 0.95)
```

```
##      est    lwr.ci    upr.ci
## [1,] 0.2 0.1421546 0.2738691
```

```
BinomCI(x = 13, n = 140, conf.level = 0.95)
```

```
##             est     lwr.ci    upr.ci
## [1,] 0.09285714 0.05507017 0.1523906
```

```
BinomCI(x = 18, n = 140, conf.level = 0.95)
```

```
##            est     lwr.ci    upr.ci
## [1,] 0.1285714 0.08289789 0.1940839
```

Проинтерпретировать полученные интервалы мы можете самостоятельно, а мы вернёмся к Гриффиндору и посмотрим, как выглядели бы расчеты доверительного интервала самостоятельно.

Зафиксируем число наблюдений в выборке `N = 140`. Вычислим выборочную долю успехов `phat` и долю неудач `qhat`

```
N <- 140
phat <- 38 / N
phat
```

```
## [1] 0.2714286
```

```
qhat <- 1 - phat
qhat
```

```
## [1] 0.7285714
```

Вычислим значение Z для уровня доверия 0.95 (квантиль уровня 0.975):

```
z <- qnorm(0.975)
z
```

```
## [1] 1.959964
```

Вспомним формулу для доверительного интервала для доли и подставим в нее компоненты, посчитанные ранее:

```
phat - z * sqrt(phat * qhat / N)
```

```
## [1] 0.1977658
```

```
phat + z * sqrt(phat * qhat / N)
```

```
## [1] 0.3450913
```

Получили примерно такой же доверительный интервал, как и ранее (небольшие отличия обусловлены округлением и корректировками, вшитыми в автоматический расчет интервалов).

Доверительный интервал для среднего

Построим 99%-ный доверительный интервал для средней длины волшебной палочки героев. Нам понадобится та же библиотека, только функция `MeanCI()`:

```
MeanCI(hp$Wand.length, na.rm = TRUE, conf.level = 0.99)
```

```
##     mean   lwr.ci   upr.ci
## 11.28000 10.10556 12.45444
```

Пояснения к коду:

- В функции `MeanCI()` мы просто указываем показатель, выборку, на основе которой мы строим доверительный интервал для среднего.
- Опция `na.rm=TRUE` добавлена для того, чтобы R при расчетах игнорировал пропущенные значения `NA`, иначе с ними он не сможет посчитать даже выборочное среднее.

Проинтерпретируем полученный доверительный интервал. С 99%-ной уверенностью можно утверждать, что средняя длина волшебных палочек лежит в интервале от 10.11 до 12.45 дюймов.