Прикладная статистика в R

Лекция 1. Описательная статистика. Основы работы в R.

Елена Михайловна Парилина

д. ф.-м. н., проф.

2021

Почему R?

- Широкие возможности по визуализации данных и составления отчетов.
- Богатые библиотеки с разнообразием реализованных методов.
- Интеграция с другими программами анализа данных.
- Широкое распространение среди аналитиков и ученых.
- Продукт открытого доступа и бесплатный.
- Наличие бесплатных версий редакторов для работы в R (например, R Studio).

Литература

1. https://www.r-project.org/
2. https://www.rstudio.com
3. https://rstudio.cloud

4. Буре В.М., Парилина Е.М., Седаков А.А. Методы прикладной статистики в R и Excel (3-е изд., ст.). Изд-во "Лань" 2019. 152 стр. https://e.lanbook.com/book/112057
5. Буре В.М., Парилина Е.М. Теория вероятностей и математическая статистика. Изд-во "Лань" 2013. 416 стр. https://e.lanbook.com/book/10249

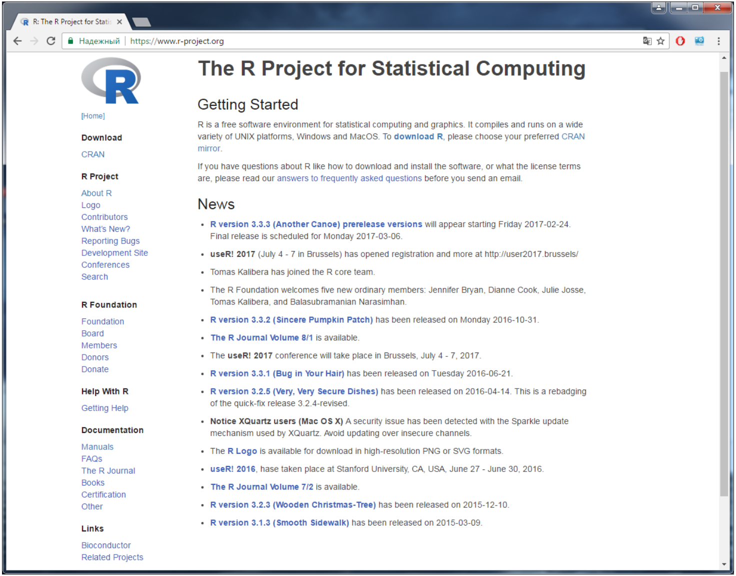
6. Adler J., “R in a Nutshell: A Desktop Quick Reference”, O’Reilly Media, 2010
7. Crawley M.J., “The R Book”, Willey, 2007, 950 p.

Модуль 1. Описательная статистика. Предварительная обработка данных.

- Знакомство с R.
- Описательная статистика.
- Корреляция и ковариация.
- Моделирование выборок из известных вероятностных распределений.
- Гистограммы, различные виды графического представления данных.
- Изображение временных рядов.
- Столбцовые и круговые диаграммы.
- «Ящик с усами».
- Графическое представление данных в R.

Знакомство с R

R Project [https://www.r-project.org/]



R: The R Project for Statis
Надежный | https://www.r-project.org

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Development Site
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

Help With R

Getting Help

Documentation

Manuals
FAQs
The R Journal
Books
Certification
Other

Links

Bioconductor
Related Projects

The R Project for Statistical Computing

Getting Started

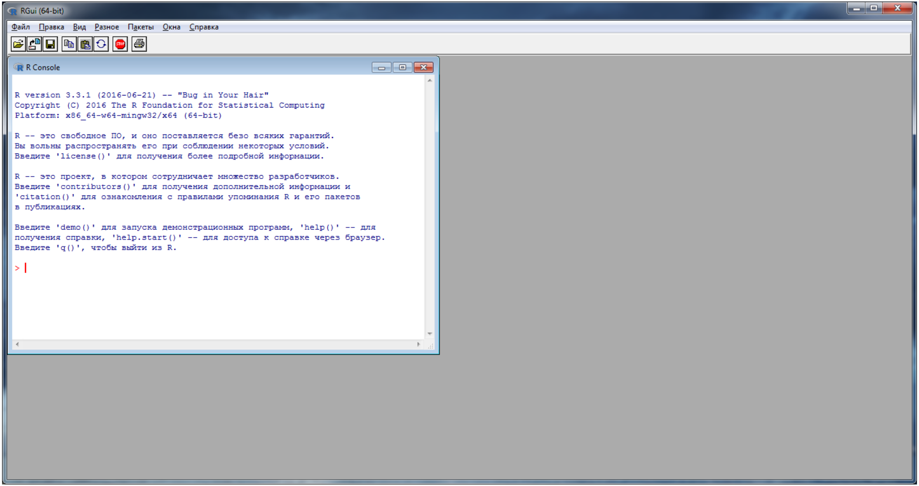
R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

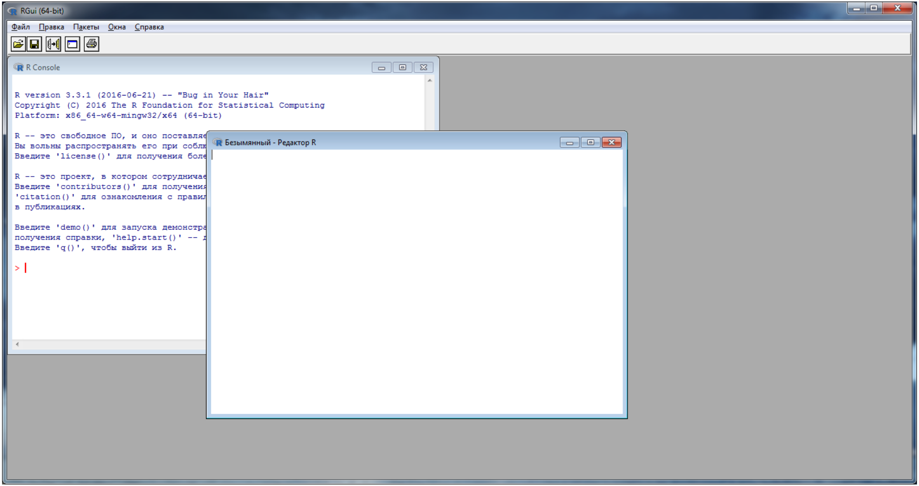
News

- **R version 3.3.3 (Another Canoe) prerelease versions** will appear starting Friday 2017-02-24. Final release is scheduled for Monday 2017-03-06.
- **useR! 2017** (July 4 - 7 in Brussels) has opened registration and more at http://user2017.brussels/
- Tomas Kalibera has joined the R core team.
- The R Foundation welcomes five new ordinary members: Jennifer Bryan, Dianne Cook, Julie Josse, Tomas Kalibera, and Balasubramanian Narasimhan.
- **R version 3.3.2 (Sincere Pumpkin Patch)** has been released on Monday 2016-10-31.
- **The R Journal Volume 8/1** is available.
- The **useR! 2017** conference will take place in Brussels, July 4 - 7, 2017.
- **R version 3.3.1 (Bug in Your Hair)** has been released on Tuesday 2016-06-21.
- **R version 3.2.5 (Very, Very Secure Dishes)** has been released on 2016-04-14. This is a rebadging of the quick-fix release 3.2.4-revised.
- **Notice XQuartz users (Mac OS X)** A security issue has been detected with the Sparkle update mechanism used by XQuartz. Avoid updating over insecure channels.
- The **R Logo** is available for download in high-resolution PNG or SVG formats.
- **useR! 2016**, hase taken place at Stanford University, CA, USA, June 27 - June 30, 2016.
- **The R Journal Volume 7/2** is available.
- **R version 3.2.3 (Wooden Christmas-Tree)** has been released on 2015-12-10.
- **R version 3.1.3 (Smooth Sidewalk)** has been released on 2015-03-09.

RGui

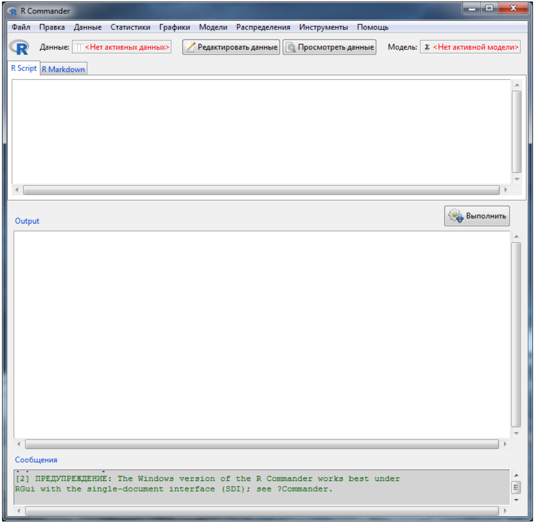

RGui (64-bit)
Файл Правка Вид Разное Пакеты Окна Справка
R Console
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R -- это свободное ПО, и оно поставляется безо всяких гарантий.
Вы вольны распространять его при соблюдении некоторых условий.
Введите 'license()' для получения более подробной информации.
R -- это проект, в котором сотрудничает множество разработчиков.
Введите 'contributors()' для получения дополнительной информации и
'citation()' для ознакомления с правилами упоминания R и его пакетов
в публикациях.
Введите 'demo()' для запуска демонстрационных программ, 'help()' -- для
получения справки, 'help.start()' -- для доступа к справке через браузер.
Введите 'q()', чтобы выйти из R.
>

R script

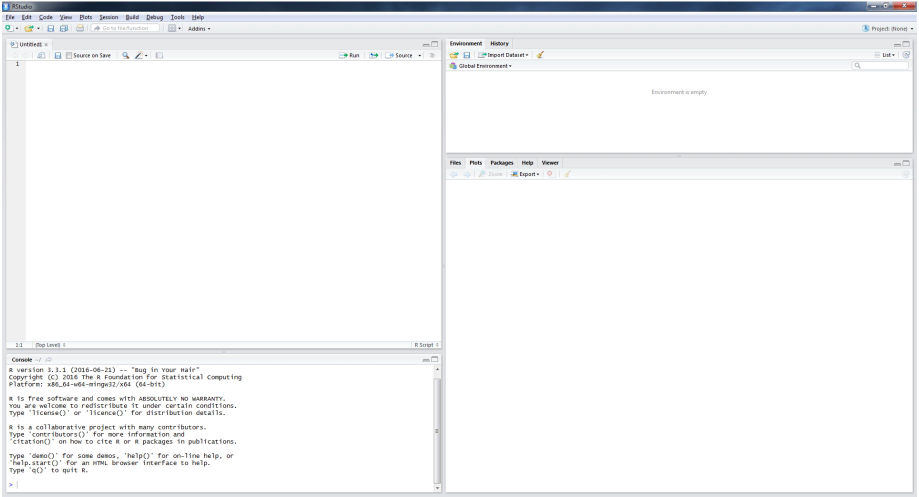

RGui (64-bit)
Файл Правка Пакеты Окна Справка
R Console
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
Безымянный - Редактор R

R Commander

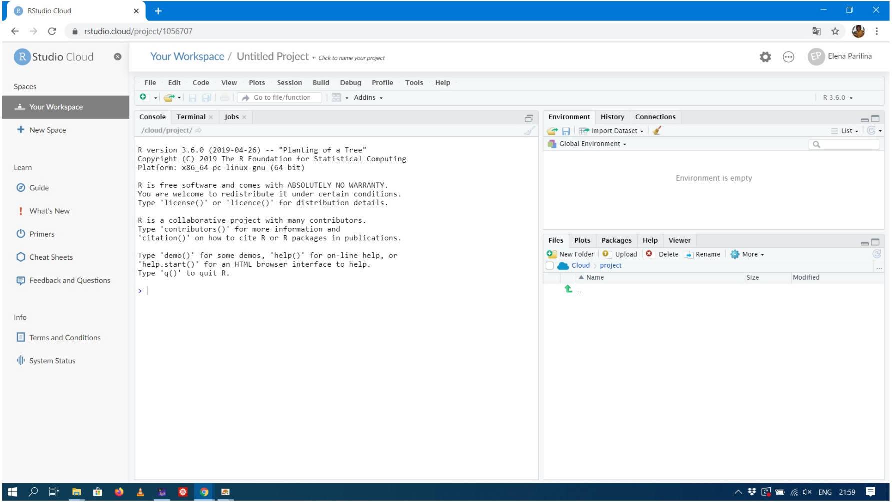
```
library(Rcmdr);
```



RStudio: [https://www.rstudio.com]



RStudio Cloud [https://rstudio.cloud]



R packages

- Обращение к библиотеке:

```
> library(name)
```

- Устновка пакета:

```
> install.packages("name")
```

- Установка нескольких пакетов:

```
> install.packages(c("tree" ,"maptree"))
```

Основные операции в R

- Составить вектор:

```
> c(1,2,3,4)
[1] 1 2 3 4
```

- Операции с векторами:

```
> c(1,2,3,4)+c(10,20,30,40)
[1] 11 22 33 44

> c(1,2,3,4)*c(10,20,30,40)
[1] 10 40 90 160

> 1/c(1,2,3,4)
[1] 1.000 0.500 0.333 0.250

> c(1,2,3,4)+c(10,100)
[1] 11 102 13 104
```

Переменные

- Присвоение значения переменной:

```
> x <- 1
> y <- 2
> z <- x+y
```

- Обращение к элементу вектора:

```
> b <- c(1,2,3,4,5,6,7,8,9,10)
> b[7]
[1] 7

> b[1:6]
[1] 1 2 3 4 5 6
```

Списки в R

- `list` — список, который содержит объекты. Например, следующая переменная x — список, содержащий объекты n, s, b и число 3:

```
> n = c(2, 3, 5)
> s = c("aa", "bb", "cc", "dd", "ee")
> b = c(TRUE, FALSE, TRUE, FALSE, FALSE)
> x = list(n, s, b, 3) # x содержит копии n, s, b
```

- Обращение к элементу списка:

```
> x[2]
[1] "aa" "bb" "cc" "dd" "ee"
```

Списки в R

- Для ссылки на элемент списка используйте оператор двойной квадратной скобки "[[ ]]" (ссылка на значение). Следующий объект $x[[2]]$ является вторым элементом списка x. Если $x[2]$ является копией второго элемента:

```
> x[[2]]
[1] "aa" "bb" "cc" "dd" "ee"
```

Список с названиями R

- ```
  > v = list(bob=c(2, 3, 5), john=c("aa" , "bb"))
  > v
  $bob
  [1] 2 3 5
  $john
  [1] "aa" "bb"
  ```

- Для обращения к элементу списка, можно использовать как двойную квадратную скобку "[[ ]]" или имя элемента:
  ```
  > v[["bob"]]
  [1] 2 3 5
  ```

- На именованный элемент списка можно также напрямую ссылаться с помощью оператора "$" вместо оператора двойной квадратной скобки:
  ```
  > v$bob
  [1] 2 3 5
  ```
  ```

Массив (фрейм) данных

Фрейм данных — это список, содержащий несколько именованных векторов одинаковой длины. Фрейм данных очень похож на таблицу базы данных. Например, рассмотрим массив данных с результатами побед и поражений в Национальной лиге (NL) East в 2008 году:

```
> teams <- c("PHI","NYM","FLA","ATL","WSN")
> w <- c(92, 89, 94, 72, 59)
> l <- c(70, 73, 77, 90, 102)
> nleast <- data.frame(teams,w,l)
> nleast
```

	teams	w	l
1	PHI	92	70
2	NYM	89	73
3	FLA	94	77
4	ATL	72	90
5	WSN	59	102

Массив данных

- Для обращения к компонентам массива данных используйте оператор "$":

```
> nleast$w
[1] 92 89 94 72 59
```

- Предположим, вы хотите узнать, где содержится количество проигрышей команды Florida Marlins (FLA). Вы можете рассчитать это так:

```
> nleast$teams=="FLA"
[1] FALSE FALSE TRUE FALSE FALSE
```

- Затем вы можете использовать следующую ссылку на правильный элемент в векторе проигрышей:

```
> nleast$l[nleast$teams=="FLA"]
[1] 77
```

Ввод данных

- XLS-file:

```
library(gdata)
data <- read.xls("C:/noname.xls", sheet = 1, header = TRUE)
```

- XLSX-file:

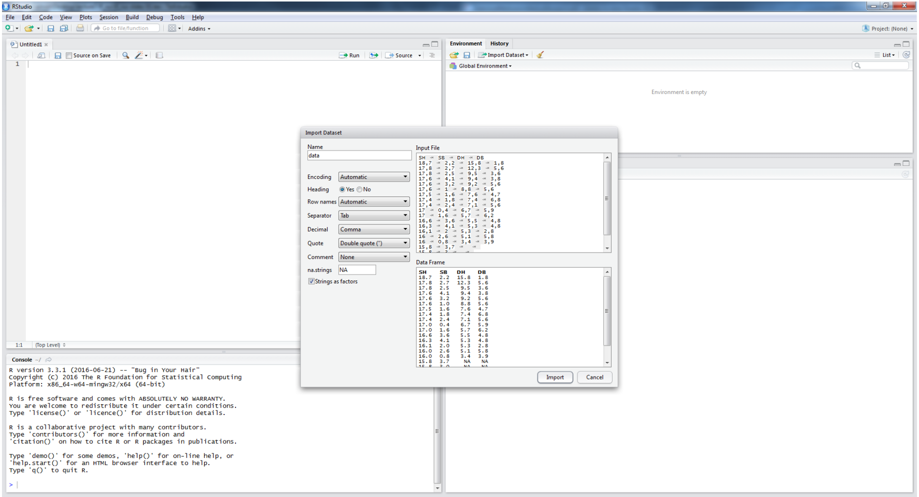
```
library(readxl)
data <- read_excel("C:/haemolytic.xlsx", sheet = 1, col_names
= TRUE)
```

- TXT-file:

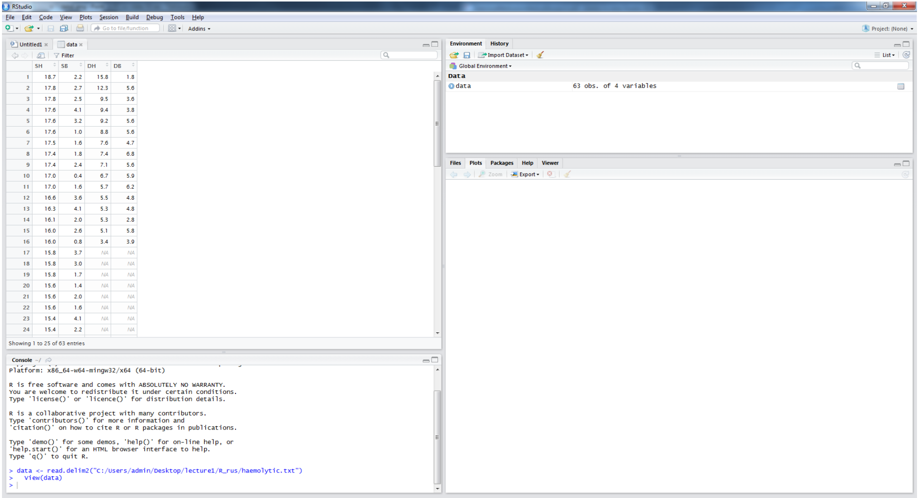
```
data <- read.table("C:/noname.txt", header=TRUE, sep="\t",
na.strings="NA", dec=",")
```

Ввод данных (RStudio)

File → Import Dataset → From ...



Ввод данных (RStudio)



Пропуски данных: NA

Чтобы проверить, есть ли в данных записи NA:

```
is.na(data$DH)
```

Чтобы создать вектор без записей NA:

```
data$DH[! is.na(data$DH)]
```

Вектор

```
x <- c(1,3,0,2,1,4,2,1,5,0)
x <- seq(-4,4,0.01)
```

```
length(x)
table(x)
summary(x)
```

```
x[4]
x[c(2,3,6)]
x[1:3]
x[-1]
x[-length(x)]
```

Работа с массивом "iris"

`require(fastR2)`: устанавливаем пакет `fastR2`;
`glimpse(iris)`: выводит краткое описане данных в базе `iris`;

```
> glimpse(iris)
Observations: 150
Variables: 5
$ Sepal.Length <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.8, 4.8, 4.3, 5.8, 5.7, 5.4,
 5.1, 5....
$ Sepal.Width  <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.4, 3.0, 3.0, 4.0, 4.4, 3.9,
 3.5, 3....
$ Petal.Length <dbl> 1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5, 1.5, 1.6, 1.4, 1.1, 1.2, 1.5, 1.3,
 1.4, 1....
$ Petal.Width  <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.2, 0.2, 0.1, 0.1, 0.2, 0.4, 0.4,
 0.3, 0....
$ Species      <fct> setosa, setosa, setosa, setosa, setosa, setosa, setosa, setosa, setosa, setosa, seto
sa, seto...
```

Работа с массивом "iris"

`head(iris, n=3)`: выводит первые 3 строки `iris`;
`tail(iris, n=4)`: выводит последние 3 строки `iris`;

```
> head(iris,n=3)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
> tail(iris,n=4)
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
147          6.3         2.5          5.0         1.9 virginica
148          6.5         3.0          5.2         2.0 virginica
149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
>
```

Работа с массивом "iris"

`iris[45:47,3:5]`: выводит определенные строки и столбцы массива `iris`;
`sample(iris, 5)`: выборка объема 5 из `iris`;

```
> iris[45:47, 3:5]
   Petal.Length Petal.Width Species
45          1.9         0.4  setosa
46          1.4         0.3  setosa
47          1.6         0.2  setosa
> sample(iris, 5)
   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species orig.id
57          6.3         3.3          4.7         1.6 versicolor      57
92          6.1         3.0          4.6         1.4 versicolor      92
53          6.9         3.1          4.9         1.5 versicolor      53
17          5.4         3.9          1.3         0.4     setosa      17
36          5.0         3.2          1.2         0.2     setosa      36
>
```

Работа с функциями в R

Определение зависимости:

$$\boxed{\texttt{goal}}\,(\,\boxed{\texttt{y}}\ \sim\ \boxed{\texttt{x}}\,,\ \texttt{data} = \boxed{\texttt{mydata}}\,)$$

"goal" — функция, "y $\sim$ x" — формула, "data=..." — название массива данных.

Описательная статистика

- Оценки числовых характеристик изучаемой случайной величины ξ, найденные по имеющейся у статистика выборке $X_{[n]} = (X_1, \ldots, X_n)$ объема n.
- Всевозможные функции от выборки.

Вариационный ряд Если элементы одномерной выборки упорядочить по возрастанию (построить *вариационный ряд* $X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}$) и отметить повторяемость наблюдений (подсчитать *частоту*), то получится *статистический ряд*, построенный по одномерной выборке $X_{[n]}$.

Размах — разность между максимальным и минимальным элементами выборки, $R = X_{\max} - X_{\min}$.

Полезные функции в R

- `max(x)`: максимальный элемент в `x`;
- `min(x)`: минимальный элемент в `x`;
- `sum(x)`: сумма элементов в `x`;
- `sort(x)`: вариационный ряд из элементов `x`.

Группированный статистический ряд При большом объеме выборки ее элементы иногда объединяются в группы, представляя результаты опытов в виде *группированного статистического ряда*. Для этого интервал, содержащий все элементы выборки, разбивается на k непересекающихся интервалов. Обычно разбиение производится на интервалы одинаковой длины $b = R/k$. После чего нетрудно определить частоты — количества n_i элементов выборки, попавших в i-ый интервал. Статистический ряд часто записывают в виде таблицы. В первой строке таблицы указывают середины интервалов группировки X_i, а во второй — частоты n_i. Подсчитываются также *накопленные частоты* $\sum_{j=1}^{i} n_j$, *относительные частоты* n_i/n, *накопленные относительные частоты* $\sum_{j=1}^{i} n_j/n$.

Полезные функции в R

- `colSums(x)`: сумма элементов по столбцам в `x`;
- `rowSums(x)`: сумма элементов по строкам в `x`.

Выборочное среднее — выборочный начальный момент 1-го порядка, который определяется равенством

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Выборочная дисперсия — выборочный центральный момент 2-го порядка, который определяется равенством

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2.$$

Выборочное среднее квадратическое отклонение равно s.

Полезные функции в R

- `mean(x)`: выборочное среднее `x`;
- `var(x)`: выборочная дисперсия `x`;
- `sd(x)`: выборочное с.к.о. `x`.

Выборочная квантиль x_p *порядка p* определяется как элемент вариационного ряда $X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}$ выборки $X_{[n]}$ с номером $[np]+1$, где $[a]$ — целая часть числа a.

Замечание

В описательной статистике используют ряд квантилей, имеющих специальные названия *персентили* (квантили порядков 0.01; 0.02;...;0.99), *децили* (квантили порядков 0.1; 0.2;...;0.9), *квартили* (квантили порядков 0.25; 0.5; 0.75).

Выборочная медиана — называется число, которое делит вариационный ряд на две части, содержащие равное количество элементов; если $n = 2k+1$, то медианой выборки является элемент вариационного ряда $X_{(k+1)}$, если $n = 2k$, то медианой выборки является число $(X_{(k)} + X_{(k+1)})/2$.

Выборочная мода — элемент выборки, имеющий наибольшую частоту.

Характеристики положения выборки

Наиболее распространенными характеристиками положения являются **выборочное среднее**, **выборочная медиана**, **выборочная мода**.

Характеристики рассеяния выборки

Наиболее распространенными мерами рассеяния являются **размах** (размах $R = X_{\max} - X_{\min}$), **средний межквартильный размах** (три квартили Q_1, Q_2, Q_3 делят вариационный ряд на четыре части с равным числом элементов, тогда средний межквартильный размах равен $(Q_3 - Q_1)/2$), **персентильный размах** (персентильный размах равен разности персентилей $P_{90} - P_{10}$), **выборочная дисперсия** $s^2 = a_2^{0*}$; исправленная дисперсия $\tilde{s}^2 = ns^2/(n-1)$ и **среднее квадратическое отклонение** $\tilde{s} = \sqrt{\tilde{s}^2}$).

Коэффициент вариации — мера относительного разброса выборки, вычисляется по формуле

$$v = s/\bar{X},$$

иногда коэффициент записывают в процентах $C_v = v \cdot 100\%$.

Коэффициент асимметрии вычисляется по формуле

$$S_{k1} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{s^3}$$

Коэффициент эксцесса вычисляется по формуле

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{s^4} - 3$$

Для нормального распределения теоретические коэффициенты асимметрии и эксцесса равны нулю.

Асимметрия и эксцесс

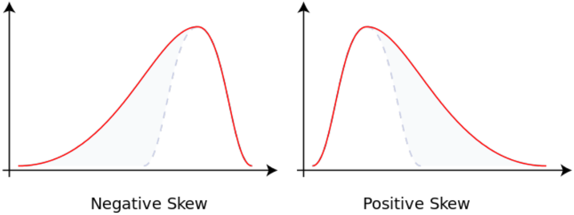


Рис.: Отрицательный и положительный коэффициенты асимметрии

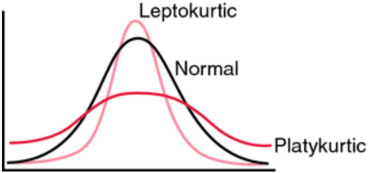


Рис.: Отрицательный и положительный коэффициенты эксцесса

- `mean(x)`: выборочное среднее `x`;
- `median(x)`: медиана `x`;
- `range(x)`: интервал (min(x); max(x));
- `var(x)`: выборочная дисперсия в `x`;
- `sd(x)`: выборочное с.к.о. `x`;
- `sort(x)`: вариационный ряд из элементов `x`;
- `rank(x)`: вектор рангов элементов из `x`;
- `quantile(x)`: вектор, содержащий минимум, квантиль уровня 0,25, медиану, квантиль уровня 0,75, максимум из `x`;
- `colMeans(x)`: выборочное среднее по столбцам в `x`;
- `rowMeans(x)`: выборочное среднее по строкам в `x`;
- `kurtosis(x)`: эксцесс `x` (в библиотеке `library(e1071)`);
- `skewness(x)`: асимметрия `x` (в библиотеке `library(e1071)`);
- `modeOf(x)`: мода `x` (в библиотеке `library(lsr)`);
- `summary(x)`: квартили выборки `x`.

Корреляция и ковариация

Коэффициент корреляции

Пусть имеется две выборки $X_{[n]} = \{X_1, \ldots, X_n\}$ и $Y_{[n]} = \{Y_1, \ldots, Y_n\}$, где пара элементов (X_i, Y_i) — две характеристики одного объекта. Тогда **коэффициент корреляции** вычисляется по формуле:

$$\hat{\rho}(X, Y) = \frac{\sum\limits_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2 \sum\limits_{i=1}^{n}(Y_i - \bar{Y})^2}},$$

а число $\sum\limits_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})/n$ называется **ковариацией** X и Y.

Свойства корреляции

Заметим, что $\hat{\rho}(X, Y) \in [-1, 1]$. Значения $\hat{\rho}$, близкие к -1 и +1, говорят о сильной линейной зависимости изучаемых случайных величин. О проверке гипотезы о значимой корреляции между величинами мы будем говорить в рамках Модуля 3.

Коэффициент корреляции

Функции в R

- `cor(x,y, method="pearson")`: коэффициент корреляции `x` и `y` (значение параметра `method` также может равняться `kendall` или `spearman`);
- `cov(x,y)`: ковариация `x` и `y`.

```
> d<-iris[,c(-5)]
> cor(d)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
> x<-c(2,6,4,5)
> y<-c(1,5,5,3.7)
> cor(x,y)
[1] 0.830634
```

Упражнение

Посчитать элементы описательной статистики для столбцов `Sepal.Width`, `Petal.Length`, `Petal.Width` в массиве `iris`.

Моделирование выборок

Биномиальное распределение

Случайная величина принимает значение, равное числу успехов в серии из n независимых испытаний, в каждом из которых происходит либо успех, либо неудача:

$$P\{\xi = m\} = C_n^m p^m q^{n-m}, \quad m = 0, 1, \ldots, n.$$

Математическое ожидание, дисперсия и среднеквадратическое отклонение случайной величины ξ равны соответственно:

$$E\xi = np, \quad D\xi = npq, \quad \sigma_\xi = \sqrt{npq}.$$

Функции в R

`choose(n,k)` $= C_n^k$,
`factorial(x)` $= x!$

Биномиальное распределение в R

Случайная величина ξ подчиняется биномиальному распределению с параметрами $(size, prob)$, где

- *size* — количество испытаний в серии,
- *prob* — вероятность успеха в одном испытании.

Функции	Значение функции
`dbinom(x,size,prob)`	$P\{\xi = x\}$
`pbinom(q,size,prob)`	$P\{\xi \leqslant q\}$
`qbinom(r,size,prob)`	наименьшее x такое, что $P\{\xi \leqslant x\} \geqslant r$
`rbinom(n,size,prob)`	генерирует n элем. из заданного распределения и возвращает их в виде вектора

Биномиальное распределение в R

```
> randomData<- rbinom(n=30,size=4,prob=0.5)
> tally(~randomData)
randomData
 0  1  2  3  4
 1  8 11  6  4
> randomData
 [1] 1 0 4 3 1 2 4 2 2 2 2 1 1 2 4 1 1 1 3 2 3 2 3 2 2 3 3 4 2 1
> dbinom(randomData,4,0.5)
 [1] 0.2500 0.0625 0.0625 0.2500 0.2500 0.3750 0.0625 0.3750 0.3750 0.3750 0.3750 0.2500 0.2500
[14] 0.3750 0.0625 0.2500 0.2500 0.2500 0.2500 0.3750 0.2500 0.3750 0.2500 0.3750 0.3750 0.2500
[27] 0.2500 0.0625 0.3750 0.2500
> dbinom(1,4,0.5)
[1] 0.25
> pbinom(1,4,0.5)
[1] 0.3125
> qbinom(0.7,4,0.5)
[1] 3
```

Упражнение:

Прокомментируйте вывод каждой функции.

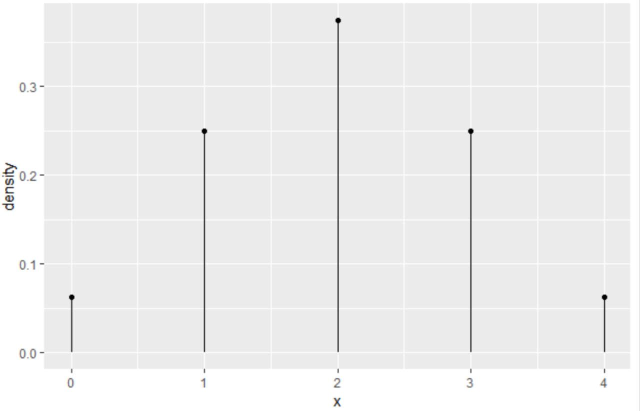
Биномиальное распределение в R

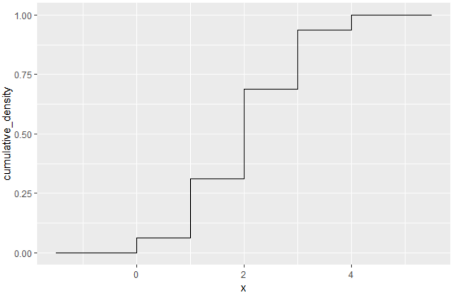
Замечания:

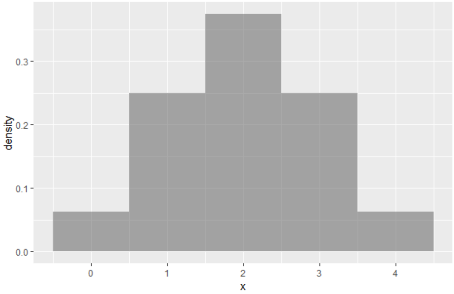
- В R параметр `size` используется для обозначения объема выборки, а параметр `n` — для обозначения числа элементов в случайной выборке.
- Схема задания других распределений такова, что неизменной остается первая буква функции, определяющая ее значение, а далее пишется название нужного распределения (например, в `dbinom(x,size,prob)` поменяется `binom` на другое распределение).

Графики (биномиальное распределение)

```
> gf_dist("binom",params=list(4,0.5))
> gf_dist("binom",params=list(4,0.5),kind="cdf")
> gf_dist("binom",params=list(4,0.5),kind="histogram",binwidth=1)
```




1.00
0.75
0.50
0.25
0.00
cumulative_density
0
2
4
x


0.3
0.2
0.1
0.0
density
0
1
2
3
4
x

Отрицательное биномиальное распределение (NBinom)

Определение

Отрицательное биномиальное распределение, также называемое распределением Паскаля, — это распределение случайной величины, равной количеству произошедших неудач в последовательности испытаний Бернулли с вероятностью успеха p, проводимых до r-го успеха.

Функции	Значение функции
`dnbinom(x,size,prob)`	$P\{\xi = x\}$
`pnbinom(q,size,prob)`	$P\{\xi \leqslant q\}$
`qnbinom(r,size,prob)`	наименьшее x такое, что $P\{\xi \leqslant x\} \geqslant r$
`rnbinom(n,size,prob)`	генерирует n элем. из заданного распределения и возвращает их в виде вектора

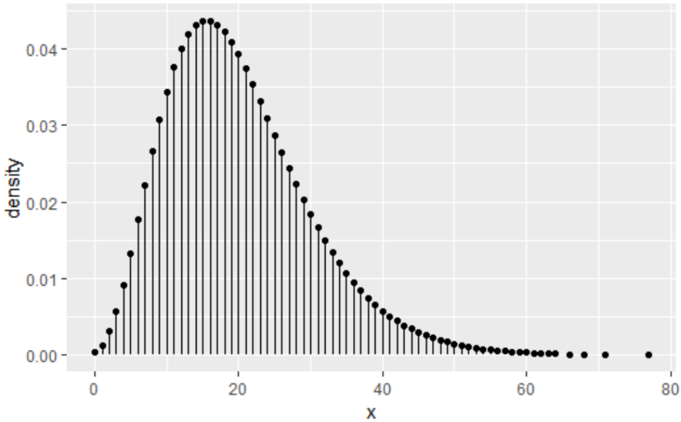
Отрицательное биномиальное распределение (NBinom)

```
> NrandomData<-rnbinom(50,10,0.3)
> NrandomData
 [1] 29 42  9 49 14 18 16 24 29 31 47 19 18 17 20 14 20 33 26 59 18  7 10 14
[25] 34 34 25 19 45 33 14 22 20 20 24 14 15 23 31 20 27 23 31 11 15 13 12 18
[49] 18 27
> tally(~NrandomData)
NrandomData
 7  9 10 11 12 13 14 15 16 17 18 19 20 22 23 24 25 26 27 29 31 33 34 42 45
 1  1  1  1  1  1  5  2  1  1  5  2  5  1  2  2  1  1  2  2  3  2  2  1  1
47 49 59
 1  1  1
```

```
> dnbinom(NrandomData,10,0.3)
 [1] 0.0309937642 0.0056044633 0.0115853688 0.0016155492 0.0327271473
 [6] 0.0450667256 0.0400907554 0.0436297266 0.0309937642 0.0254748746
[11] 0.0023456250 0.0464898854 0.0450667256 0.0429206911 0.0471872337
[16] 0.0327271473 0.0471872337 0.0203552932 0.0391392209 0.0002111875
[21] 0.0450667256 0.0055632023 0.0154085405 0.0327271473 0.0180204214
[26] 0.0180204214 0.0415354997 0.0464898854 0.0033602182 0.0203552932
[31] 0.0327271473 0.0465437714 0.0471872337 0.0471872337 0.0436297266
[36] 0.0327271473 0.0366544050 0.0453295861 0.0254748746 0.0471872337
[41] 0.0365299395 0.0453295861 0.0254748746 0.0196108697 0.0366544050
[46] 0.0284583889 0.0240233153 0.0450667256 0.0450667256 0.0365299395
> dnbinom(15,10,0.3)
[1] 0.0366544
> pnbinom(20,10,0.3)
[1] 0.4111913
> qnbinom(0.7,10,0.3)
[1] 27
```

Отрицательное биномиальное распределение (NBinom)

```
gf_dist("nbinom" , params=list(5,0.2))
```



Геометрическое распределение (Geom(prob))

Определение

Распределение случайной величины, равной количеству произошедших неудач в последовательности испытаний Бернулли с вероятностью успеха p, проводимых до **первого** успеха, называется **геометрическим распределением**.

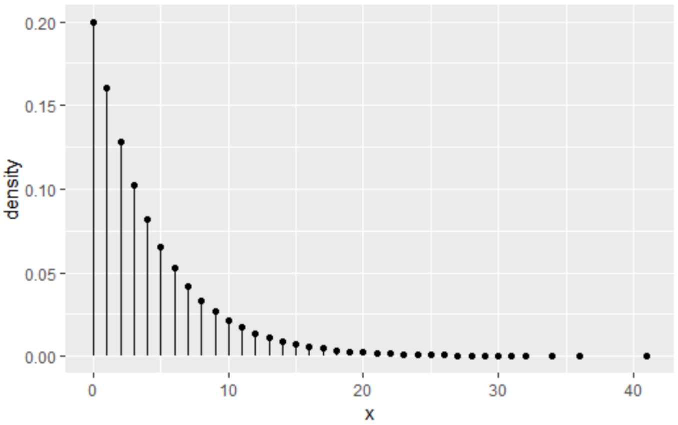
Функции	Значение функции
`dgeom(x,prob)`	$P\{\xi = x\}$
`pgeom(q,prob)`	$P\{\xi \leqslant q\}$
`qgeom(r,prob)`	наименьшее x такое, что $P\{\xi \leqslant x\} \geqslant r$
`rgeom(n,prob)`	генерирует n элем. из заданного распределения и возвращает их в виде вектора

Геометрическое распределение (Geom(prob))

```
> geomrandomData<-rgeom(50,0.25)
> geomrandomData
 [1]  7  9  2  1 11  2  4  6  0  0  0  8  0  7  1  0  0  0  4  1  2  3  2  2
[25]  3 10  0  1  2  0  5  1  1  0  4  0  3  0  4 17  2  1  0 13  2 10 10  5
[49]  1  0
> tally(~geomrandomData)
geomrandomData
 0  1  2  3  4  5  6  7  8  9 10 11 13 17
14  8  8  3  4  2  1  2  1  1  3  1  1  1
> dgeom(c(0,1,2,3,4,5),0.3)
[1] 0.300000 0.210000 0.147000 0.102900 0.072030 0.050421
> pgeom(10,0.3)
[1] 0.9802267
> qgeom(0.5,0.3)
[1] 1
> qgeom(0.9,0.3)
[1] 6
```

Геометрическое распределение (Geom(prob))

```
gf_dist("geom" , params=list(0.2))
```



Мультиномиальное (полиномиальное) распределение (Multinom)

- Пусть производится n независимых испытаний, каждое из которых может закончится одним из r исходов из множества $\{1, \dots, r\}$.
- Исходу i соответствует вероятность p_i, $i = 1, \dots, r$.
- При этом, $\sum_{i=1}^{r} p_i = 1$.
- Пусть набор $(a_1, \dots, a_n)$ — упорядоченный набор чисел из множества $\{1, \dots, r\}$.

Вероятность того, что произойдет m_1 через $P_n(m_1, \dots, m_r)$, как легко видеть:

$$P_n(m_1, \dots, m_r) = \frac{n!}{m_1! m_2! \dots m_r!} p_1^{m_1} \dots p_r^{m_r}. \tag{1}$$

Мультиномиальное (полиномиальное) распределение

Определение

Распределение, определяемое формулой (1), называется **полиномиальным** распределением.

Функции	Значение функции
`dmultinom(vector1,size,prob_vector)`	$P\{\xi = vector1\}$
`rmultinom(x,size,prob_vector)`	генерирует n элем. из заданного распределения и возвращает матрицу

```
> multrandomData<-rmultinom(20,8,c(0.3,0.5,0.2))
> multrandomData
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
[1,]    2    2    2    2    0    3    2    0    4     1     3     3     1
[2,]    6    4    3    4    7    3    4    7    4     2     4     5     4
[3,]    0    2    3    2    1    2    2    1    0     5     1     0     3
     [,14] [,15] [,16] [,17] [,18] [,19] [,20]
[1,]     1     4     5     3     3     3     3
[2,]     5     4     3     4     2     3     4
[3,]     2     0     0     1     3     2     1
> dmultinom(c(5,5,0),10,c(0.3,0.5,0.2))
[1] 0.01913625
```

Другие важные распределения

1. *Нормальное распределение с параметрами* (a, σ^2)

   Плотность распределения: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$.

   Функция распределения: $F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{(y-a)^2}{2\sigma^2}} dy$.

   Математическое ожидание: a.
   Дисперсия: σ^2.

2. *Равномерное распределение на отрезке* $[a, b]$

   Плотность распределения: $f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0, & x \notin [a, b]. \end{cases}$

   Функция распределения: $F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \leqslant x < b, \\ 1, & x \geqslant b. \end{cases}$

   Математическое ожидание: $\frac{a+b}{2}$.

   Дисперсия: $\frac{(b-a)^2}{12}$.

Другие важные распределения

1. *Экспоненциальное распределение с параметром* $\lambda > 0$
   Плотность распределения: $f(x) = \begin{cases} 0, & x \leqslant 0, \\ \lambda e^{-\lambda x}, & x > 0. \end{cases}$
   Функция распределения: $F(x) = \begin{cases} 0, & x \leqslant 0, \\ 1 - e^{-\lambda x}, & x > 0. \end{cases}$
   Математическое ожидание: $\frac{1}{\lambda}$.
   Дисперсия: $\frac{1}{\lambda^2}$.
2. *Распределение Пуассона с параметром* $\alpha > 0$
   Вероятность: $\mathrm{P}\{\xi = k\} = \frac{\alpha^k}{k!} e^{-\alpha}, \quad k = 0, 1, \ldots$.
   Математическое ожидание: α.
   Дисперсия: α.

Распределения в R

- `rnorm(n,m,sd)` — нормальное распределение с параметрами `m`, `sd`^{2};
- `rweibull(n,shape,scale)` — распределение Вейбулла с параметрами `shape`, `scale`;
- `rpois(n,lambda)` — распределение Пуассона с параметром `lambda`;
- `rgamma(n,shape,scale)` — гамма-распределение с параметрами `shape`, `scale`;
- `rbinom(n,size,prob)` — биномиальное распределение с параметрами `size`, `prob`;
- `rchisq(n,df)` — χ^2-распределение с `df` степенями свободы;
- `rexp(n,rate)` — экспоненциальное распределение с параметром `rate`;
- `rf(n,df1,df2)` — распределение Фишера с `df1`, `df2` степенями свободы;
- `rt(n,df)` — распределение Стьюдента с `df` степенями свободы.

Итоги

Что мы узнали на Лекции 1?

- Что такое R, как его устанавливать и какое графическое приложение использовать при работе с R.
- Как пользоваться библиотеками, как искать нужные функции в R.
- Как вводить данные вручную и из имеющихся файлов.
- Как совершать простые операции с данными в R.
- Что такое описательная статистика и какие характеристики данных к ней относятся.
- Как интерпретировать элементы описательной статистики.
- Как посчитать корреляцию между переменными.
- Как создавать случайные выборки.

Что мы узнаем на Лекции 2?

Мы узнаем,

- какие возможности графического представления данных имеются в R.
- какие гипотезы о изучаемой случайной величине можно выдвинуть при составлении описательной статистики и анализе графиков.

Спасибо за внимание и до встречи на Лекции 2!