Типы данных. Простые графики

Корпусные методы исследований языковых процессов

Даша Попова

16.11.2022

R: Basic Data Types

- **logical**

```
> TRUE
[1]
TRUE
> class(TRUE)
[1] "logical"
> FALSE
[1] FALSE
> class(NA)
[1] "logical"
> T
[1] TRUE
> F
[1] FALSE
```

- **class()** to reveal type

R: Basic Data Types

- **numeric**

```
> 2
[1] 2
> 2.5
[1] 2.5
> class(2)
[1] "numeric“
> is.numeric(2)
[1] TRUE
```

R: Basic Data Types

character

```
> "I love data science!"
[1] "I love data science!"
> class("I love data science!")
[1] "character"
```

R: Basic Data Types

Coercion

```
> as.numeric(TRUE)
[1] 1
> as.numeric(FALSE)
[1] 0
> as.character(4)
[1] "4"
> as.numeric("4.5")
[1] 4.5
> as.integer("4.5")
[1] 4
> as.numeric("Hello")
[1] NA
Warning message:
NAs introduced by coercion
```

R: Basic Data Types

Try out:

>height = 1.67

>weight = 52

>ls()

What does ls() do?

>rm(weight)

>ls()

>weight

What does rm() do?

Workspace

DataCamp handouts

1. Create and name vectors
2. Create and name matrices
3. Create and name lists
4. Factors
5. Data frame
6. Vector arithmetic
7. Subsetting vectors
8. Matrix arithmetic
9. Subsetting matrices
10. Subset and extend lists
11. Subset, extend and sort data frames

Простые графики: линейный график

- ***A line chart*** is a graph that connects a series of points by drawing line segments between them. These points are ordered in one of their coordinate (usually the x-coordinate) value.

- Line charts are usually used in identifying the ***trends*** in data.

- The **plot()** function in R is used to create the line graph.

Простые графики: линейный график

- ***Syntax***
- The basic syntax to create a line chart in R is

  plot(v, type, col, xlab, ylab)

- **v** is a vector containing the numeric values;
- **type** takes the value "p" to draw only the points, “l” to draw only the lines and "o" to draw both points and lines;
- **xlab** is the label for x axis;
- **ylab** is the label for y axis;
- **main** is the Title of the chart;
- **col** is used to give colors to both the points and lines.

Простые графики: линейный график

- *# create the data for the chart*

```
>v <- c(1,1,0,1,1,0,2,2,0,0,2,2)
# plot the line chart
>plot(v, type = "o")
```

В каком месяце Вы родились?

```
>plot(v,type = "o", col = "red", xlab = "Месяц", ylab = "Number of
    birthdays", main = "Birthday distribution")

> plot(v, xaxt = "n", type = "o", col = "red", xlab = "Месяц", ylab =
    "Number of birthdays", main = "Birthday distribution")
> axis(1, at=1:5, labels=c("jan","feb","mar","apr","may"))
```

https://www.tutorialspoint.com/r/r_line_graphs.htm

Простые графики: диаграмма

- ***A pie-chart*** is a representation of values as slices of a circle with different colors. The slices are labeled and the numbers corresponding to each slice is also represented in the chart.

- In R the pie chart is created using the **pie()** function which takes positive numbers as a vector input. The additional parameters are used to control labels, color, title etc.

Простые графики: диаграмма

- **Syntax**
- The basic syntax for creating a pie-chart using the R is −

  pie(x, labels, radius, main, col, clockwise)
- **x** is a vector containing the numeric values used in the pie chart;
- **labels** is used to give description to the slices;
- **radius** indicates the radius of the circle of the pie chart.(value between −1 and +1);
- **main** indicates the title of the chart;
- **col** indicates the color palette;
- **clockwise** is a logical value indicating if the slices are drawn clockwise or counterclockwise.

Простые графики: диаграмма

- >x = c(?,?)
- > labels=c("апрель","январь")
- > pie(x,labels)

- x <- c(?, ?)
- >labels <- c("апрель", "январь")
- # plot the chart with title and rainbow color pallet
- >pie(x, labels, main = "Birthdays per month", col = rainbow(length(x)))

Простые графики: диаграмма

- >x = c(1,2)
- > labels=c("апрель","январь")
- piepercent<- round(100*x/sum(x), 1)
- # plot the chart with the legend
- > pie(x, labels = piepercent, main = "Birthdays per month",col = rainbow(length(x)))
- > legend("topright", c ("апрель","январь"), cex = 0.8, fill = rainbow(length(x)))

https://www.tutorialspoint.com/r/r_pie_charts.htm

Простые графики: столбчатая диаграмма

- ***A bar chart*** represents data in rectangular bars with length of the bar proportional to the value of the variable.
- R uses the function **barplot()** to create bar charts.
- R can draw both vertical and horizontal bars in the bar chart.
- In a bar chart each of the bars can be given different colors.

Простые графики: столбчатая диаграмма

- **Syntax**
- The basic syntax to create a bar-chart in R is −

  barplot(H, xlab, ylab, main, names.arg, col)
- **H** is a vector or matrix containing numeric values used in bar chart;
- **xlab** is the label for x axis;
- **ylab** is the label for y axis;
- **main** is the title of the bar chart;
- **names.arg** is a vector of names appearing under each bar;
- **col** is used to give colors to the bars in the graph.

Простые графики: столбчатая диаграмма

- *# create the data for the chart*
- > H <- c(7,12,28,3,41)
- *# plot the bar chart*
- > barplot(H)

- *# create the data for the chart*
- > H <- c(7,12,28,3,41)
- > M <- c("Mar","Apr","May","Jun","Jul")
- *# plot the bar chart*
- > barplot(H,names.arg=M,xlab="Month",ylab="Number of Birthdays",col="blue", main="B-day per month chart",border="red")

Простые графики: столбчатая диаграмма

- *# create the input vectors*
- > colors = c("green","orange","brown")
- > months <- c("Mar","Apr","May","Jun","Jul")
- > people <- c("J","S","P")
- *# create the matrix of the values*
- > Values <- matrix(c(2,9,3,11,9,4,8,7,3,12,5,2,8,10,11), nrow = 3, ncol = 5, byrow = TRUE)

- *# create the bar chart*
- > barplot(Values, main = "total number of B-days", names.arg = months, xlab = "month", ylab = "number of B-days", col = colors)
- *# add the legend to the chart*
- > legend("topleft", people, cex = 1.3, fill = colors)

https://www.tutorialspoint.com/r/r_bar_charts.htm

Выборка

- Задана совокупность наблюдений, объединенных некоторым общим признаком. Предположим, что эта совокупность бесконечна в том смысле, что в принципе наблюдения можно продолжить в любой момент времени, как, например, в серии бросаний монеты. Из этой совокупности "случайным образом" извлекается последовательность наблюдений. Если число этих наблюдений достаточно велико, то частота появления событий, обладающих указанным признаком, незначительно отклоняется от некоторой постоянной, называемой эмпирической вероятностью.

- На практике ответить на вопрос о том, может ли выбор из нашей совокупности рассматриваться как случайный, нелегко. Чаще всего этой несколько расплывчатой формулировкой о случайном выборе пользуются тогда, когда нет оснований предполагать наличие "привилегированных" наблюдений.

http://statistica.ru/theory/vyborochnyy-metod/

Выборка

- В этой связи часто говорят об "урновой" модели. Содержимое урны, например шары, неразличимые на ощупь, представляет совокупность, а извлечение шаров, которые мы предполагаем хорошо перемешанными, - случайный выбор.
- Целью такого случайного выбора из совокупности является выяснение ее структуры, в частности определение *эмпирической* вероятности. Здесь отчасти используется то эвристическое соображение, что при бесконечно большом числе наблюдений можно точно определить значение эмпирической вероятности.

http://statistica.ru/theory/vyborochnyy-metod/

Выборка

- Практически же проведение произвольно большого числа опытов или наблюдений связано с трудностями различного характера. Так, проведение большого числа опытов наталкивается на техническую невыполнимость или на экономические затруднения, что приводит к ограничению числа наблюдений. Приближение к идеальным условиям, которое имеет место в случае игр на разорение, в большинстве практически важных ситуаций не имеет места.

- Установилась следующая терминология. Бесконечная (гипотетическая) совокупность возможных наблюдений называется ***генеральной совокупностью***, и результаты наблюдений, из нее извлеченных, называются ***выборкой*** из этой совокупности. Число наблюдений в выборке называют ее ***объемом***.

- Понятие бесконечной совокупности представляет идеализацию действительного положения вещей, даже когда под этим понимается потенциальная возможность неограниченного повторения опытов. Практик рассматривает всякую совокупность, "достаточно большую" по сравнению с объемом имеющейся выборки, как бесконечную.

Выборка

Репрезентативная выборка

- Репрезентативная выборка (representative sample) - одно из ключевых понятий анализа данных. Репрезентативная выборка - это выборка из генеральной совокупности с распределением $F(x)$, представляющая основные особенности генеральной совокупности.

- Например, если в городе проживает 100 000 человек, половина из которых мужчины и половина женщины, то выборка 1000 человек из которых 10 мужчин и 990 женщин, конечно, не будет репрезентативной.

- Построенный на ее основе опрос общественного мнения, конечно, будет содержать смещение оценок и приводит к фальсификации результатов.
- Необходимым условием построения репрезентативной выборки является равная вероятность включения в нее каждого элемента генеральной совокупности.

http://statistica.ru/theory/vyborka-vyborochnoe-raspredelenie/

Случайная величина или переменная

- ***Случайная переменная*** — это величина, которая может принимать любое из набора взаимоисключающих значений с определенной вероятностью.

- Распределение вероятности показывает вероятности всех возможных значений случайной переменной. Это теоретическое распределение, которое выражено математически и имеет *среднее* и *дисперсию* — аналоги среднего и дисперсии в эмпирическом распределении.

http://statistica.ru/theory/raspredeleniya-veroyatnostey/

Нормальное распределение

- In a random collection of data from independent sources, it is generally observed that the distribution of data is normal.
- Which means, on plotting a graph with the value of the variable in the horizontal axis and the count of the values in the vertical axis we get a bell shaped curve.
- The center of the curve represents the mean of the data set. In the graph, fifty percent of values lie to the left of the mean and the other fifty percent lie to the right of the graph.
- This is referred as ***normal distribution*** in statistics.

https://www.tutorialspoint.com/r/r_normal_distribution.htm

Нормальное распределение

- dnorm(x, mean, sd)
- rnorm(n, mean, sd)
- **x** is a vector of numbers;
- **n** is number of observations(sample size);
- **mean** is the mean value of the sample data. It's default value is zero;
- **sd** is the standard deviation. It's default value is 1.

https://www.tutorialspoint.com/r/r_normal_distribution.htm

Нормальное распределение

- *# create a sequence of numbers between -10 and 10 incrementing by 0.1*
- > x <- seq(-10, 10, by = .1)
- *# choose the mean as 2.5 and standard deviation as 0.5*
- > y <- dnorm(x, mean = 2.5, sd = 0.5)
- > plot(x,y)

- rnorm is used to generate random numbers whose distribution is normal. It takes the sample size as input and generates that many random numbers. We draw a histogram to show the distribution of the generated numbers.
- # Create a sample of 50 numbers which are normally distributed
- > y <- rnorm(50, mean = 2.5, sd = 0.5)
- # Plot the histogram for this sample.
- >hist(y, main = "Normal Distribution")

https://www.tutorialspoint.com/r/r_normal_distribution.htm