# R语言中的基本数据结构

Alex / 2023-07-05 / free_learner@163.com / AlexBrain.cn

总结一下R语言中的基本数据结构，包括向量和因子、矩阵和数组、列表和数据框，以及创建和索引方法。参考资料是*The Art of R Programming*的第2-6章的内容。我自己属于编程小白（虽然经常用R分析数据），理解上恐怕有很多错误，敬请指正。

## 一、向量和因子

在R语言中，向量（vector）是最基本的数据结构，比如，`x <- 1`是只包含一个元素的向量而不是标量，其他数据结构都和向量有所关联。可以使用`x <- c(1,5,8)`来创建一个向量。使用中括号索引向量中的元素`x[1]`或`x[c(1,3)]`，索引从1开始而不是0。向量要求每个元素的数据类型是一致的，元素数据类型包括整数（integer）、双精度浮点数（double/numeric）、布尔类型（Boolean/logical）、字符或字符串（character）。比如，`x <- c('abc','xyz')`就是一个包含字符元素的向量。除了使用数字进行索引，也可以用名字进行索引，比如，`x <- c(a=1, b=5, c=8)`可以用`x[1]`或`x['a']`索引第一个元素。另外，还需要注意的是，`NA`用来表示缺失值，比如，`x <- c(1, NA, 8)`或`x <- c('abc', NA, 'xyz')`均表示第二个元素缺失。

因子（factor）用来描述统计学上的类别数据（categorical data），比如，被试的性别。有时候类别数据会用数字来表示，比如用1表示男性、2表示女性，这个时候使用因子类型是重要的，否则使用统计模型可能会得到错误的结果（因为这些数字会当做连续变量）。使用`x <- factor(c(1,2,1,2))`创建因子，使用`levels(x)`查看类别水平。其他和向量相同。

## 二、矩阵和数组

矩阵（matrix）和向量一样，要求元素的数据类型一样，但是可以通过行和列来索引。比如，`x <- matrix(c(1:9), nrow=3, ncol=3)`就创建了一个3行3列的矩阵。可以使用`x[1,2]`来索引第一行第二列的元素，用`x[1,]`来索引第一行的元素。除了用行和列来索引，由于矩阵是按列存储的，所以也可以像向量一样用一个数字进行索引，比如，`x[4]`同样表示第一行第二列的元素。同样地，矩阵也可以用名字来索引，比如，`x <- matrix(c(1:9), nrow=3, ncol=3); rownames(x) <- c('R1', 'R2', 'R3'); colnames(x) <- c('C1', 'C2', 'C3')`，那么`x['R1','C2']`同样索引第一行第二列的元素。

矩阵是数组（array）的一个特例，即二维数组，数组可以包含三个或以上的维度，比如，`x <- array(c(1:12), dim = c(2,2,3))`创建了一个三维数组。可以使用三个数字来索引三维数组的元素`x[1,2,3]`。不过要注意，在提取子数组或子矩阵的时候，如果某些维度为1，可能数据结构会自动变成矩阵或向量，可以使用`drop=FALSE`选项避免这一点，比如`x[1,1,1,drop=FALSE]`。

## 三、列表

列表（list）形式上像向量，但是元素可以为不同的数据类型，比如，`x <- list(v1=123, v2='abc', v3=TRUE)`，对于列表元素（也叫作成分），有三种索引方式：`x[[1]]; x$v1; x[['v1']]`。注意如果是用`[]`索引，则返回的是子列表，比如`x[1]`表示就是只包含一个成分的列表，而不是成分本身的数据结构了（这里每个成分是向量）。列表的成分可以是其他类型的数据结构，比如，`x <- list(v1=matrix(1:9, nrow=3, ncol=3), v2=list(s1='abc', s2=TRUE))`，列表`x`包含两个成分，第一个是矩阵，第二个成分是包含两个成分的列表，比如`x$v2$s1`表示`'abc'`。因为这个原因，列表又称为递归向量（recursive vector），而第一节里的向量称为原子向量（atomic vector）。

## 四、数据框

数据框（data.frame）本质上是列表，只是每个成分是一个元素数量相同的向量，比如，`x <- data.frame(ID=c('S1','S2','S3'), Age=c(11, 33, 49), Sex=c('F','M','F'))`创建了包含三个成分的数据框。数据框在形式上像矩阵，因此数据框有四种索引方式，`x[[1]]; x[['ID']]; x$ID; x[,1]`，即除了列表的索引方式，还可以使用矩阵的索引方式。注意，在使用`data.frame()`时，字符向量会自动转换为因子类型，如果要避免这种转换，可以使用`stringAsFactors=FALSE`选项。