利用R语言进行数据可视化

R for Data Visualization

李雄辉

21017年5月

Introduction

Scatter plot : plot()

Histogram : hist()

Introduction / Description

箱线图（Box Plot 或 Box-and-Whisker Plot）主要是从四分位数的角度出发描述数据的分布，它通过最大值（Q4）、上四分位数（Q3）、中位数（Q2）下四分位数（Q1）和最小值（Q0）五处位置来获取一维数据的分布概况。

我们知道，这五处位置之间依次包含了四段数据，每段数据量均为总数据量的1/4。通过每一段数据占据的长度，我们可以大致推断出数据的集中或离散趋势（长度越短，说明数据在该区间上越密集，反之则稀疏）。

Usage

```
boxplot(x, ...)

## S3 method for class 'formula'
boxplot(formula, data = NULL, ..., subset, na.action =

## Default S3 method:
boxplot(x, ..., range = 1.5, width = NULL, varwidth = F
        notch = FALSE, outline = TRUE, names, plot = TR
        border = par("fg"), col = NULL, log = "",
        pars = list(boxwex = 0.8, staplewex = 0.5, outw
        horizontal = FALSE, add = FALSE, at = NULL)
```

boxplot()是一个泛型函数，所以它可以适应不同的参数类型。目前它支持两种参数类型：公式（formula）和数据，后者对我们来说可能更容易理解（给一批数据、作相应的箱线图），而前者在某些情况下更为方便。

Arguments

x：为一个数值向量或者列表或者数据框，若为列表则对列表中每一个子对象依次作出箱线图

range：是一个延伸倍数，决定了箱线图的末端（须）延伸到什么位置，这主要是考虑到离群点的原因，在数据中存在离群点的情况下，将箱线图末端直接延伸到最大值和最小值对描述数据分布来说并不合适（图形缺乏稳健性），所以R中的箱线图默认只将图形延伸到离箱子两端range × (Q3 − Q1)处，即上下四分位数分别加/减内四分位距（Interquartile Range，简称IQR ≡ Q3 − Q1）的倍数，超过这个范围的数据点就被视作离群点，在图中直接以点的形式表示出来。

width：给定箱子的宽度。

varwidth：逻辑值，若为TRUE，那么箱子的宽度与样本量的平方根成比例，这在多批数据同时画多个箱线图时比较有用，能进一步反映出样本量的大小。

notch：是一个有用的逻辑参数，它决定了是否在箱子上画凹槽，凹槽所表示的实际上是中位数的一个区间估计，其计算式为Q2 + / − 1.58IQR/√n 区间置信水平为95%，在比较两组数据中位数差异时，我们只需要观察箱线图的凹槽是否有重叠部分，若两个凹槽互不交叠，那么说明这两组数据的中位数有显著差异（P值小于0.05）。

horizontal：逻辑值，设定箱线图是否水平放置。

Retrun / Value

List with the following components:

stats: a matrix, each column contains the extreme of the lower whisker, the lower hinge, the median, the upper hinge and the extreme of the upper whisker for one group/plot. If all the inputs have the same class attribute, so will this component.

n: a vector with the number of observations in each group.

conf: a matrix where each column contains the lower and upper extremes of the notch.

out : the values of any data points which lie beyond the extremes of the whiskers.

group: a vector of the same length as out whose elements indicate to which group the outlier belongs.

names: a vector of names for the groups.

`boxplot.stats()`

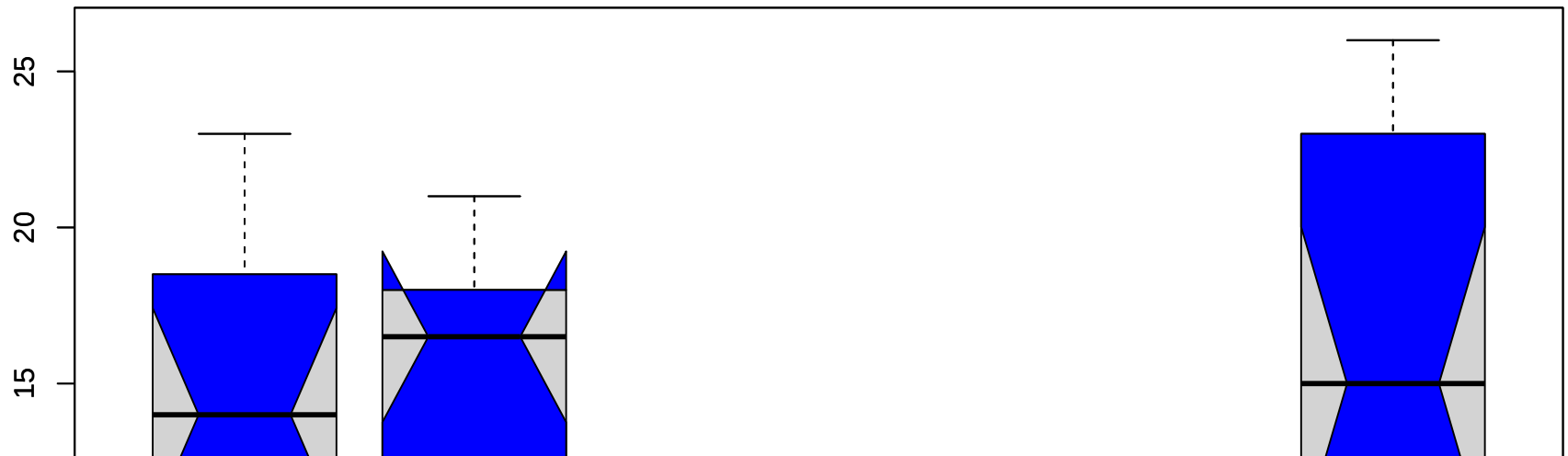
`five.stats()`

Examples

```
## boxplot on a formula:
boxplot(count ~ spray, data = InsectSprays, col = "lig

# *add* notches (somewhat funny here):
# □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □
boxplot(count ~ spray, data = InsectSprays,
        notch = TRUE, add = TRUE, col = "blue")
```

```
## Warning in bxp(structure(list(stats = structure(c(7,
## some notches went outside hinges ('box'): maybe set
```



示例1

我们使用数据集 “mtcars” 可在R环境中创建一个基本的箱线图。让我们来看看在mtcars 的 “mpg” 和 “cyl” 列。

```
input <- mtcars[,c('mpg','cyl')]
print(head(input))
```

```
##                    mpg cyl
## Mazda RX4         21.0   6
## Mazda RX4 Wag     21.0   6
## Datsun 710        22.8   4
## Hornet 4 Drive    21.4   6
## Hornet Sportabout 18.7   8
## Valiant           18.1   6
```

创建箱线图

下面的脚本将创建 mpg(英里每加仑)和cyl(气缸数)之间的关系的一个箱线图。