

CSDN

博客学院下载GitChatTinyMind论坛APP问答商城VIP活动招聘ITeye

6

搞懂箱形图

写评论

7年11月28日 09:47:19

收藏

微信

微博

QQ

箱形图，也叫盒须图，盒式图，boxplot。有95%的把握猜中你现在已经不太确定，这图中有几条线？每条线代表什么？是中位数，还是众数？

公司的深点，箱形图存在的意义为何？之于数据分析的实践意义在哪里？

下来，带你从概念开始，一步步剖析箱形图以及背后的故事。

什么是箱形图？

箱图的发明者John Tukey。Tukey先生1915年出生于美国麻省的新贝德福德。他22岁的时候在布朗大学获得了硕士学位，之后又在普林斯顿大学拿到了化学博士。有趣的是，他并没有直接去从事统计工作，而是在二战期间进入了火控研究室，在那里，大量武器相关的研究最终都转而需要先解决统计学问题。从此，Tukey改变了自己的人生方向，一个统计学的巨人即将出现。



联系我们

请扫描二维码联系客服



webmaster@csdn.net

400-660-0108

QQ客服 客服

关于 招聘 广告服务 网站地图

©2018 CSDN版权所有 京ICP证090024号

百度提供搜索支持

经营性网站备案信息

网络110报警服务

中国互联网举报中心

北京互联网违法和不良信息举报中心

箱形图最大的优点就是不受异常值的影响，可以以一种相对稳定的方式描述数据的离散分布情况。默念两遍，箱形图不受异常值的影响，这很重要。

为了更形象的说明，我们先画个图，看图说话。使用工具RStudio，假设有数据集num = c(1,6,2,7,4,2,3,3,8,25,30)，直接通过boxplot(num)画图，如下：



首先从外观上感知这是个什么东东。噢，中间是个矩形块，可以把它想象成一个盒子。盒子里面有一条线，外面有两个形似T的东西。哦，最外面还有圈，这个可不是所有的箱形图都会有。接下来一一解释这些玩意儿。

## 2.箱形图五要素

有一件重要的点，要交代一下，不然可能要被大多数人给忽略掉了。画箱形图，首先要把数据从大到小排序，没错，是从大到小。

(1) 中位数  
中位数，即二分之一分位数。所以计算的方法就是将一组数据（此处中位数，特别指的是从大到小排列的有序序列，平时求中位数并不要求是有序序列哦）分成两份，取中间这个数。

如果原始序列长度n是奇数，那么中位数所在位置是(n+1)/2；如果原始序列长度n是偶数，那么中位数所在位置是n/2，n/2+1，中位数的值等于这两个位置上的数的平均数。

(2) 上四分位数Q1  
强调一下，四分位数的求法，是将序列平均分成四份。具体的计算目前有(n+1)/4与(n-1)/4两种，一般使用(n+1)/4。

好吧，这部分我已经说不太清楚了，需要借助R语言这个强大的工具来举例说明。举个例子，有有序序列一个test = c(1,2,3,4,5,6,7,8)，通过summary(test)这个序列的中位数，上四分位数，下四分位数以及算数平均值。

这个Q1=2.75是怎么计算出来的呢？首先序列长度n=8，(1+n)/4=2.25，这是什么意思呢？说明上四分位数在第2.25个位置数，实际上这个数是不存在的，知道这个位置是在第2个数与第3个数之间的。

只能假想从第2个数到第3个数之间是均匀分布的。那么第2.25个数就是第二个数\*0.25+第三个数\*0.75，即2\*0.25+3\*0.75=2.5

summary(test)

6

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	2.75	4.50	4.50	6.25	8.00

写评论

下不是很酷~~

收藏

微信

微博

QQ

num

[1]

1	6	2	7	4	2	3	3	8	25	30
---	---	---	---	---	---	---	---	---	----	----

> summary(num)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	2.500	4.000	8.273	7.500	30.000

还是以开篇使用的栗子，来说明。  
IQR=Q3-Q1=7.5-2.5=5  
上内限=Q3+1.5\*IQR=7.5+1.5\*5=15，与剔除两个异常址30，25后的极大值8，两者取最小值，所以上内限就是8  
下内限=Q1-1.5\*IQR=2.5-1.5\*5=-5，与剔除两个异常址30，25后的极小值1，两者取最大值，所以下内限就是1

(5) 外限  
外限与内限的计算方法相同，唯一的区别就在与：上面的T形线段所延伸到的极远处，是Q3+3IQR(其中，IQR=Q3-Q1)与剔除异常值后的极大值两者取T形线段所延伸到的极远处，是Q1-3IQR与剔除异常值后的极小值两者取最大。

3.箱形图之与异常址清洗

箱形图最重要的用途就是识别异常值。数据清洗中，作用很大哦。今天先到这里～

联系我们

请扫描二维码联系客I  
✉webmaster@csd  
☎400-660-0108  
👤QQ客服 🗣客服

关于 招聘 广告服务 网站地

©2018 CSDN版权所有 京ICP证0900246

🔍 百度提供搜索支持

经营性网站备案信息

网络110报警服务

中国互联网举报中心

北京互联网违法和不良信息举报中心

欢迎大家围观，微信公众号：“数据分析师手记”，支持原创，记得打赏哦💎💎



文章标签： 统计学-箱形图    ▼查看关于本篇文章更多信息

上一篇

csv文件打开是乱码，怎么办？管用的方法，一个就够

下一篇

数据分析师面试准备

- 想对作者说点什么
- Jon\_www    2018-05-31 00:35:28    #6楼

excel里面试一下也是我回复的那个算法，可能祁那个时候流行就Q1是（n+1）/4是那么算的，现在EXCEI是这边这么算的，不过总之对于盒须图来说，两种算法不精确伤大雅。
- homehehe2014    2018-05-27 21:37:16    #5楼

作者，你在逗我？？异常值还没找到，内限值的计算就要用到异常值了？？这逻辑也太渣了吧...
- 大號子    2018-03-25 17:18:25    #4楼

按照楼主举的例子，此例的四分位数算法是这样的。Q1的位置=1+（n-1）x 0.25 Q2的位置=1+（n-1）x 0.5 Q3的位置=1+（n-1）x 0.75
- 数据分析师手记    2018-03-14 11:25:44    #3楼

查看 13 条热评