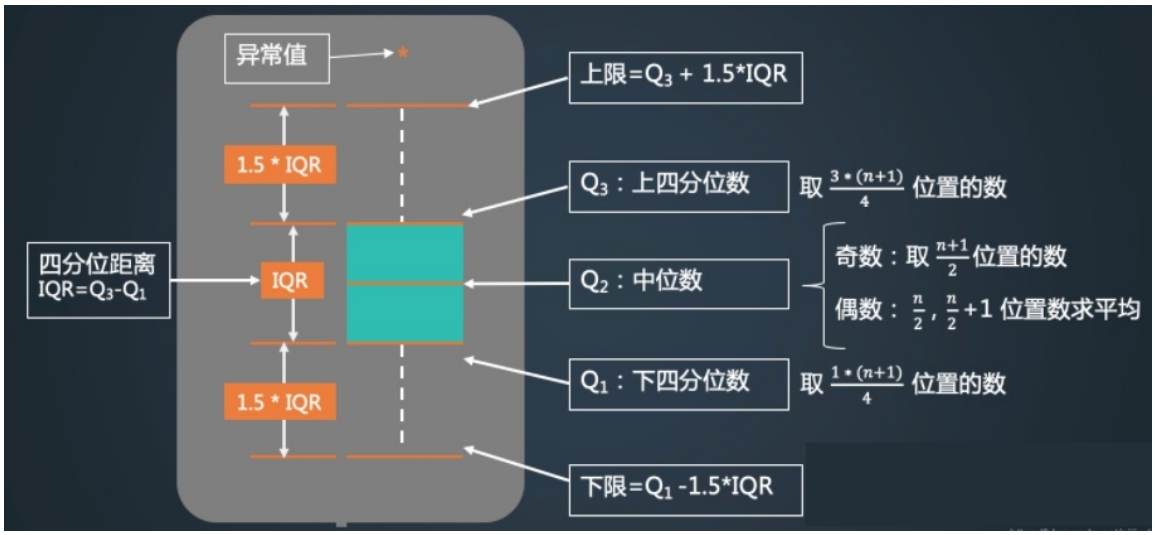
箱图

概述

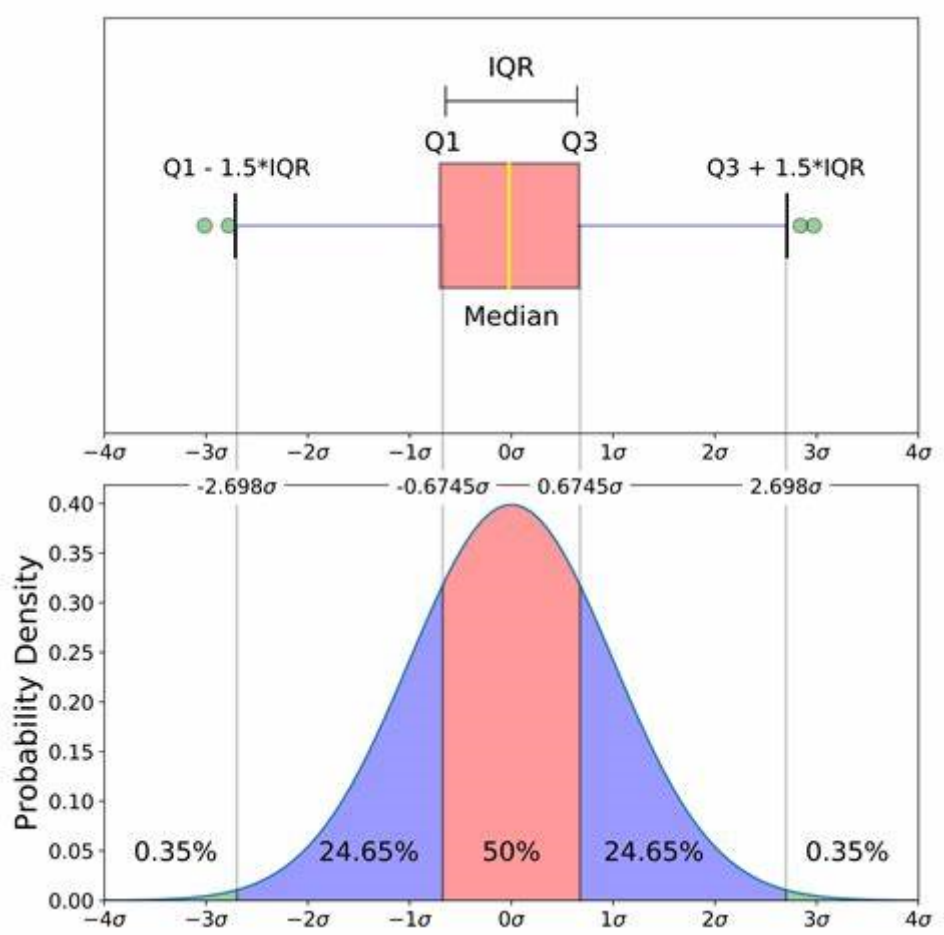
箱图，全称叫箱形图（Box-plot），是一种用作显示一组数据分散情况资料的统计图，因形状如箱子而得名。箱图主要用于反映原始数据分布的特征，还可以进行多组数据分布特征的比较。

主要特征点如下图:



一组数据按照从小到大顺序排列后，把该组数据四等分的数，称为四分位数。

- 第一四分位数 (Q1)
- 第二四分位数 (Q2，也叫“中位数”)
- 第三、四分位数 (Q3) 分别等于该样本中所有数值由小到大排列后第 25%、第 50% 和第 75% 的数字。第三四分位数与第一四分位数的差距又称四分位距（interquartile range, IQR）。



箱图的优点与特性

个人觉得，箱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。进一步分析来看，还有以下三个特性：

1、直观明了地识别异常数据

由于可以利用中位数、25% 分位数、75% 分位数、上边界、下边界等统计量的计算，可生成一个箱图，箱体区域包含的大部分为正常数据，而在箱体上边界和下边界之外的，就是异常数据。反之，箱形图可以用来直接观察数据整体的分布情况，凭借中位数、25/% 分位数、75/% 分位数等统计量，来描述数据的整体分布情况。

2、判断数据的偏态和尾重

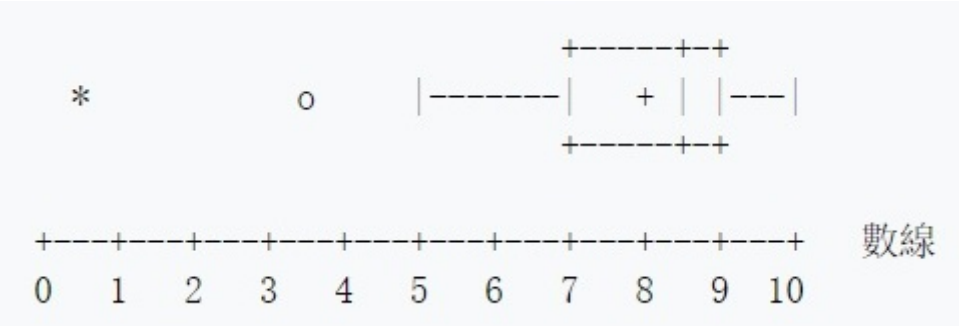
对于标准正态分布的大样本，中位数位于上下四分位数的中央，箱形图的方盒关于中位线对称。中位数越偏离上下四分位数的中心位置，分布偏态性越强。异常值集中在较大值一侧，则分布呈现右偏态；异常值集中在较小值一侧，则分布呈现左偏态。

3、多批数据通过形状来比较

箱子的上下限，分别是数据的上四分位数和下四分位数。这意味着箱子包含了 50% 的数据。因此，箱子的宽度在一定程度上反映了数据的波动程度。箱体越扁说明数据越集中，端线越短，也说明数据集中。

那么问题来了，究竟如何看?

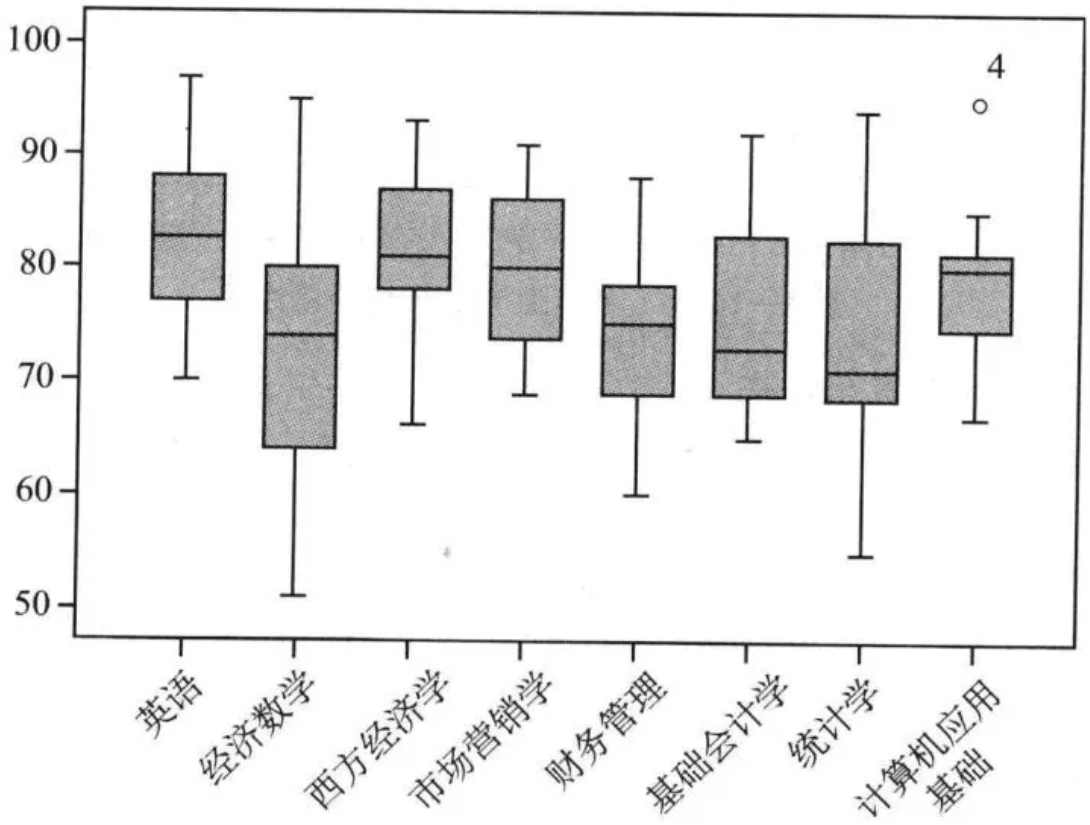
以一个箱形图具体例子，如下图:



这组数据显示出:

- 最小值 (*minimum*)=5
- 下四分位数 (*Q1*)=7
- 中位数 (*Med*-- 也就是 Q2)=8.5
- 上四分位数 (*Q3*)=9
- 最大值 (*maximum*)=10
- 平均值=8
- 四分位间距 (interquartile range)={\displaystyle Q3-Q1}=2 (即 ΔQ)

我们再以一个实际题目来分析，如下图:



- 各科成绩中，英语和西方经济学的平均成绩比较高，而统计学和基础会计学的平均成绩比较低。（用中位数来衡量整体情况比较稳定）
- 英语、市场营销学、西方经济学、计算机应用基础和财务管理成绩分布比较集中，因为箱子比较短。而经济数学、基础会计学和统计学成绩比较分散，我们可以对照考试成绩数据看看也可以证实。
- 从各个箱形图的中位数和上下四位数的间距也可以看出，英语和市场营销学的成绩分布是非常的对称，统计学则非常的不平衡——大部分数据都分布在 70 到 85(中位数到上四分位数) 分以上。同样，也可以从成绩单里的数据证实
- 在计算机应用基础对应的箱形图出现了个异常点，我们回去看看成绩单，计算机那一栏，出现了考 95 分的学霸，比第二名多了 10 分，其他同学的成绩整体在 80 分左右。

值得注意的是，**箱形图更多用于多组数据的比较**，相对直方图不仅节省了空间，还可以展示出许多直方图不能展示的信息；**单组数据则更适合采用直方图**，可视化效果更直观。