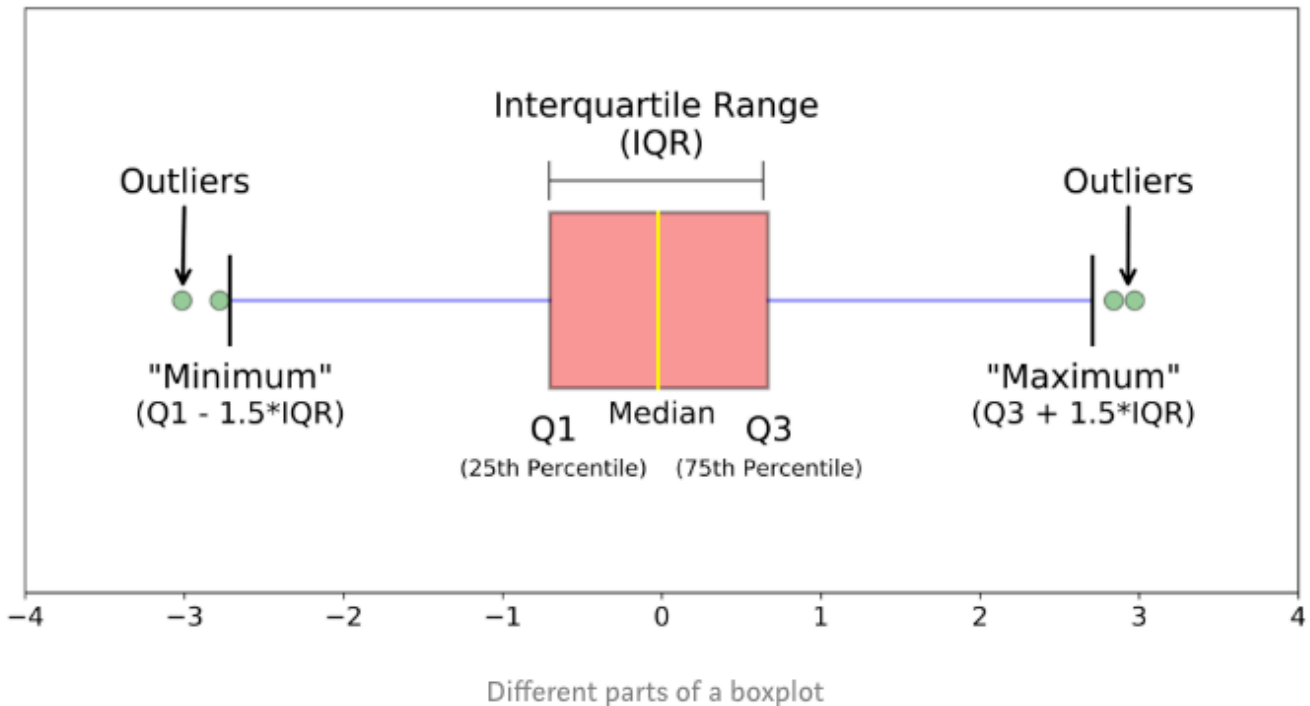


如何深刻理解箱线图 (boxplot)

如何深刻理解箱线图 (boxplot)



boxplot

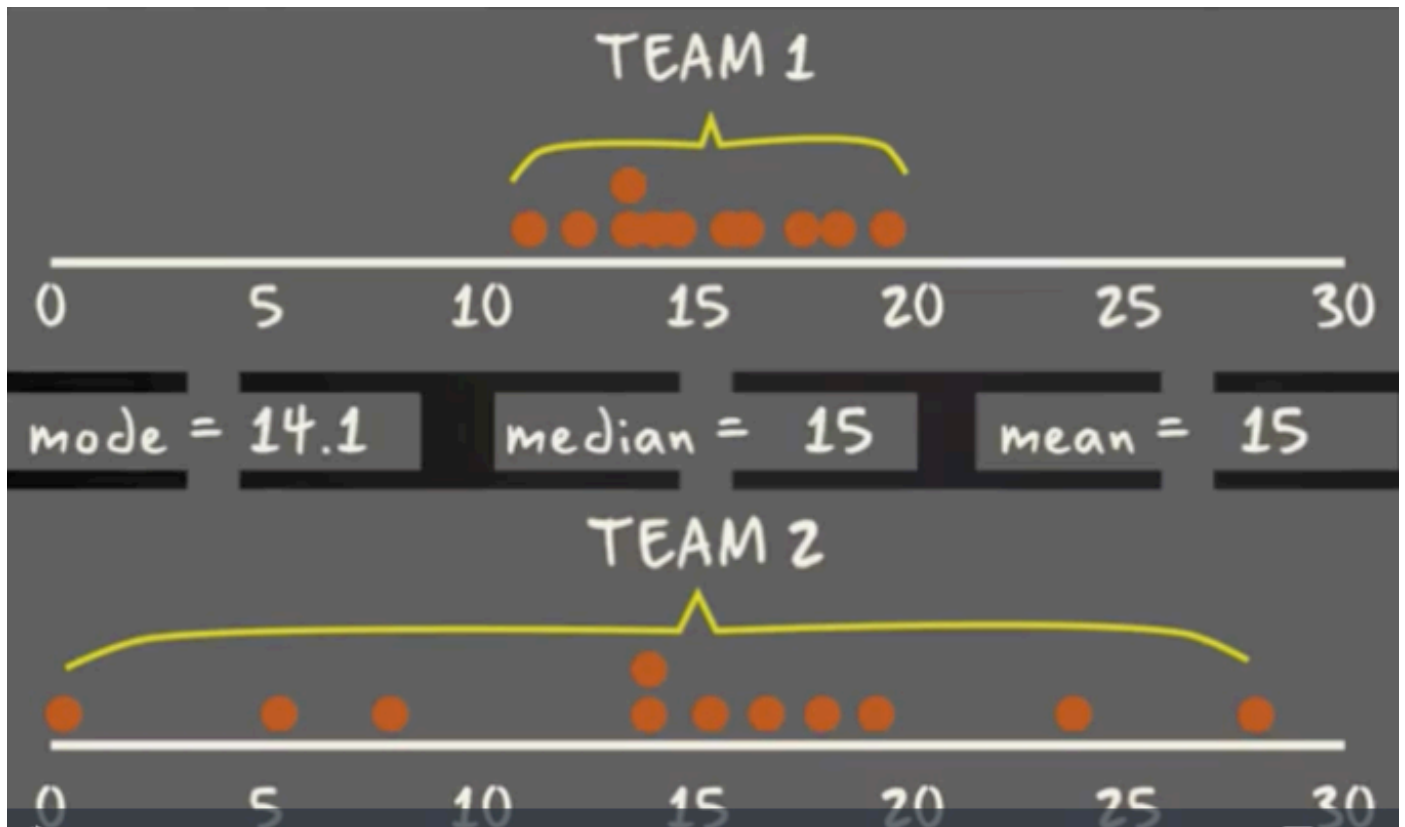
如上图箱线图，箱线图是一个能够通过5个数字来描述数据的分布的标准方式，这5个数字包括：最小值，第一分位，中位数，第三分位数，最大值，箱线图能够明确的展示离群点的信息，同时能够让我们了解数据是否对称，数据如何分组、数据的峰度；

本文主要包括一下内容：

- 什么是箱线图
- 与概率密度图相比，箱线图的优劣
- 如何通过python 做箱线图

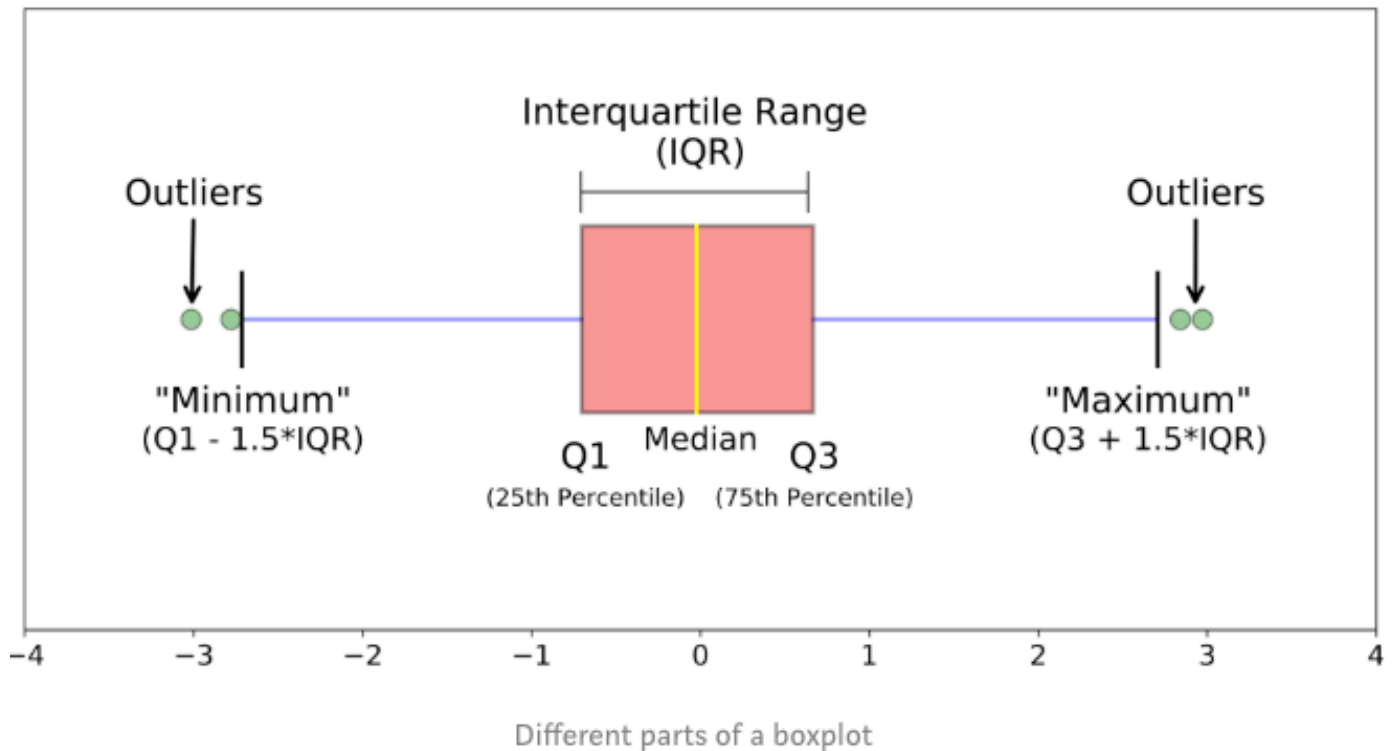
1. 什么是箱线图？

对于某些分布/数据集，您会发现除了集中趋势（中位数，均值和众数）的度量之外，您还需要更多信息。



boxplot2

您需要有关数据变异性或分散性的信息。箱形图是一张图表，它为您很好地指示数据中的值如何分布，尽管与直方图或密度图相比，箱线图似乎是原始的，但它们具有占用较少空间的优势，这在比较许多组或数据集之间的分布时非常有用。



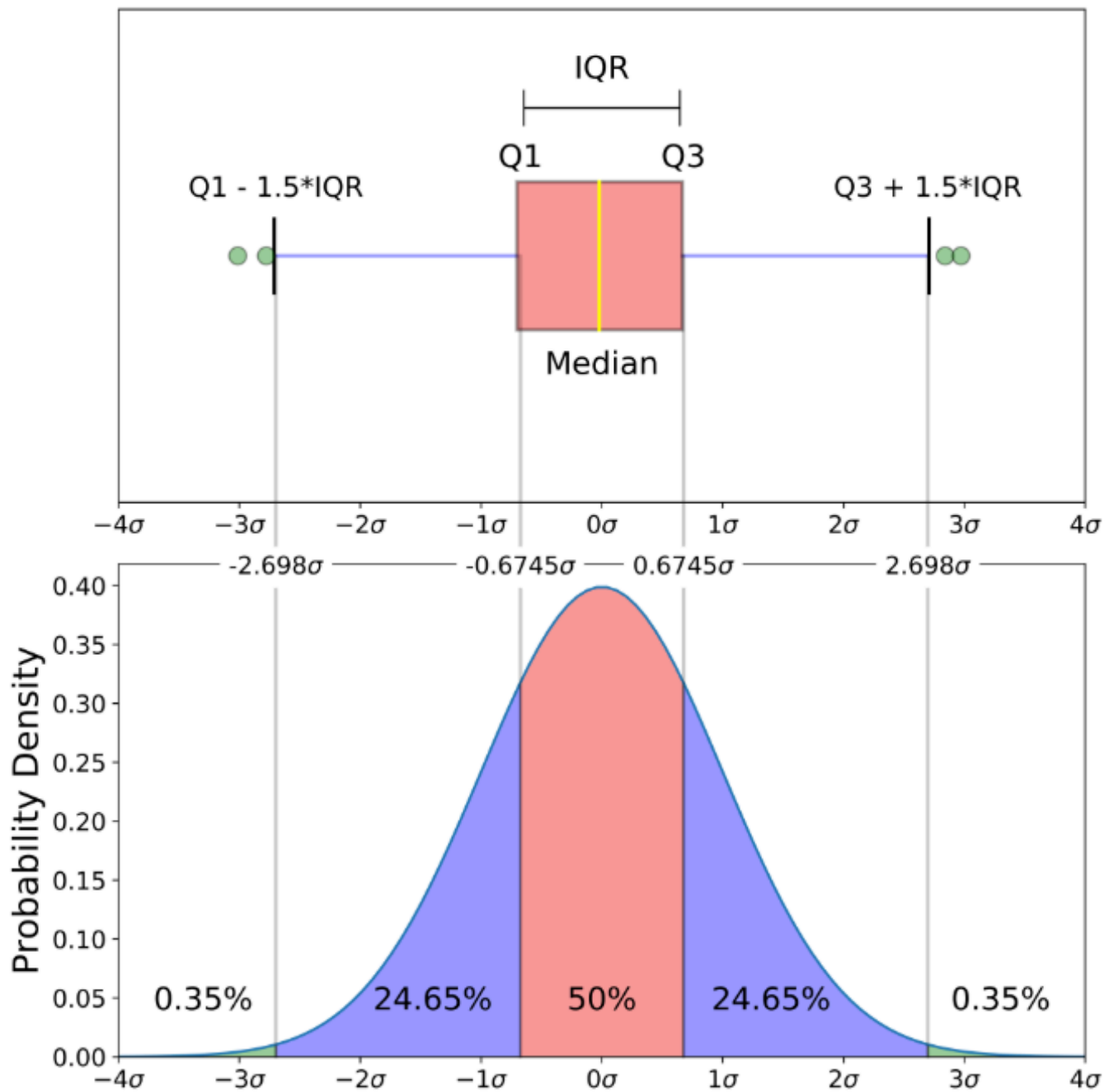
boxplot3

箱线图是一种基于五位数摘要（“最小”，第一四分位数（Q1），中位数，第三四分位数（Q3）和“最大”）显示数据分布的标准化方法。

1. 中位数（Q2 / 50th百分位数）：数据集的中间值；
2. 第一个四分位数（Q1 / 25百分位数）：最小数（不是“最小值”）和数据集的中位数之间的中间数；
3. 第三四分位数（Q3 / 75th Percentile）：数据集的中位数和最大值之间的中间值（不是“最大值”）；
4. 四分位间距（IQR）：第25至第75个百分点的距离；
5. 晶须（蓝色显示）
6. 离群值（显示为绿色圆圈）
7. “最大”： $Q3 + 1.5 \cdot IQR$
8. “最低”： $Q1 - 1.5 \cdot IQR$

离群值，“最小”或“最大”的内容可能尚不清楚。下一节将尽力为您解决问题；

Boxplot on a Normal Distribution



Comparison of a boxplot of a nearly normal distribution and a probability density function (pdf) for a normal distribution

boxplot4

上图是近似正态分布的箱线图与正态分布的概率密度函数 (pdf) 的比较,我

向您显示此图像的原因是，查看统计分布比查看箱形图更为普遍。换句话说，它可以帮助您理解箱线图；

本节将涵盖许多内容，包括：

1. 对于正态分布而言,0.7%的数据异常值;
2. 什么是“最小”和“最大”

概率密度函数

帖子的这一部分与68-95-99.7规则文章非常相似，但适用于箱线图,为了了解百分比的来源，重要的是要了解概率密度函数（PDF），PDF用于指定随机变量落入特定值范围内的概率,而不是任何一个具体值，该概率由该变量在该范围内的PDF的积分得出，也就是说，它是由密度函数下方但水平轴上方以及范围的最小值和最大值之间的面积给出的，这个定义可能没有多大意义，因此让我们通过绘制正态分布的概率密度函数来理解它，下式是正态分布的概率密度函数：

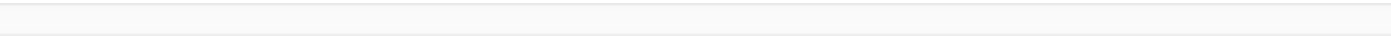
boxplot5

让我们简化一下，假设我们的平均值（ μ ）为0，标准偏差（ σ ）为1:

boxplot6

可以使用任何图形进行绘制，但是我选择使用Python进行图形绘制

```
# Import all libraries for this portion of the blog post
from scipy.integrate import quad
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
x = np.linspace(-4, 4, num = 100)
constant = 1.0 / np.sqrt(2*np.pi)
pdf_normal_distribution = constant * np.exp((-x**2) / 2.0)
fig, ax = plt.subplots(figsize=(10, 5));
ax.plot(x, pdf_normal_distribution);
ax.set_ylim(0);
ax.set_title('Normal Distribution', size = 20);
ax.set_ylabel('Probability Density', size = 20);
```



boxplot7

上图没有显示事件的概率，而是事件的概率密度,为了获得事件在给定范围内

的概率, 我们需要进行积分, 假设我们感兴趣的是寻找随机数据点落在四分位数范围内的概率。平均值的.6745标准偏差, 我们需要将-0.6745集成到0.6745。这可以用SciPy完成。

```
# Make PDF for the normal distribution a function
def normalProbabilityDensity(x):
    constant = 1.0 / np.sqrt(2*np.pi)
    return(constant * np.exp((-x**2) / 2.0) )
# Integrate PDF from -.6745 to .6745
result_50p, _ = quad(normalProbabilityDensity, -.6745, .6745, limit = 1000)
print(result_50p)
```

boxplot8

可以对“最小”和“最大”执行相同的操作

```
# Make a PDF for the normal distribution a function
def normalProbabilityDensity(x):
    constant = 1.0 / np.sqrt(2*np.pi)
    return(constant * np.exp((-x**2) / 2.0) )
# Integrate PDF from -2.698 to 2.698
result_99_3p, _ = quad(normalProbabilityDensity,
                        -2.698,
                        2.698,
                        limit = 1000)
print(result_99_3p)
```

boxplot9

如前所述，离群值是数据的剩余0.7%.

请务必注意，对于任何PDF，曲线下的面积必须为1（从函数范围内绘制任何数字的概率始终为1）

boxplot10

本部分主要基于我的Python for Data Visualization课程的免费预览视频。在上一节中，我们介绍了正态分布的箱线图,但是由于您显然并不总是具有基本的正态分布，因此让我们研究一下如何在真实数据集上利用箱形图。为此，我们将利用乳腺癌威斯康星州（诊断）数据集。

读取数据

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
# Put dataset on my github repo
```



```
df = pd.read_csv('https://raw.githubusercontent.com/mGalarnyk/Python_Tutorials
```

箱线图

下面使用箱线图来分析分类特征（恶性或良性肿瘤）和连续特征（area_mean）之间的关系。

seaborn

```
sns.boxplot(x='diagnosis', y='area_mean', data=df)
```

boxplot11

使用该图，我们可以比较Area_mean的范围和分布，以进行恶性和良性诊断。我们观察到，恶性肿瘤的Area_mean以及较大的异常值存在较大的变异性。

另外，由于箱线图上的凹口不重叠，因此可以得出结论，在95%的置信度下，真实中位数确实有所不同。

关于箱线图，还有一些其他注意事项：

1. 请记住，如果您想知道箱线图不同部分的数值是什么，可以随时从箱线图中提取数据。
2. Matplotlib不会首先估计正态分布，而是根据估计的分布参数计算四分

位数,中位数和四分位数直接从数据中计算得出。换句话说, 您的箱形图看起来可能会有所不同, 具体取决于数据的分布和样本的大小, 例如, 不对称且具有或多或少的异常值

原作者: Michael Galarnyk

原出处: medium (收费)

原链接: <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>