[EEG3235A]Python和数据科学-2019年秋

**作 业 二**

姓名 学号

汕头大学工学院电子信息工程系

1. 已知 。证明。求。

**加分题：求。**

。

。

因为(因为是一个奇函数，它的积分区间关于0对称），（概率密度函数的性质），所以我们可以得到。

。

因为

，

，

和，所以我们有。

其中，我们用到等式

，

因为是一个奇函数。

因为，和

，

所以，。

1. 异常值（Outliers）是指与数据集中的其他值存在明显差异的数据点（数值型数据：Numeric）。异常值的检测方法有哪些（查阅文献，给出参考文献）？编写Python代码，检测Adult 数据集中的数值型数据是否存在异常值。
2. 编写Python代码，应用蒙特卡罗方法（Monte Carlo Method）来验证中心极限定理（Central Limit Theorem）。
3. 由计算机从均匀分布 随机产生个数据这里，我们选取。我们可以得到。
4. 计算。
5. 重复 (1)-(2) 次，我们可以得到个数据。
6. 绘制Q-Q图（图1所示），其中横坐标为理论计算的分位数，纵坐标为从小到大排序后的观测值。由Q-Q图可以看到所有的散点都在直线附近。所以，我们验证了，即我们用数值计算验证了中心极限定理。

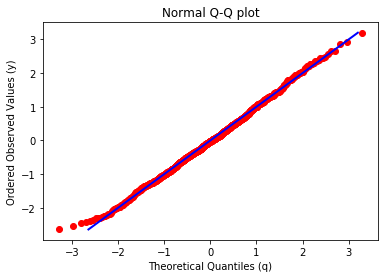


图1：Q-Q图

**Q-Q plot**

分位数（Quantile），亦称分位点。四分位数，Quartile。

分位数指的就是连续分布函数中的一个点，这个点对应概率。若概率，随机变量或它的概率分布的分位数，是指满足条件的实数

Q-Q图用于直观验证一组数据是否来自某个分布，或者验证某两组数据是否来自同一（族）分布。经常用于检验数据是否来自于正态分布。

检验一组数据（）是否来自分布函数为的某个分布的步骤如下：

1. 将数据从小到大排序：。的概率分别是：；
2. 由算出；
3. 画散点图；
4. 在图上添加一条直线做参考。看散点是否落在这条线的附近。

注：也有一些研究人员选择用来替换。

**箱形图**

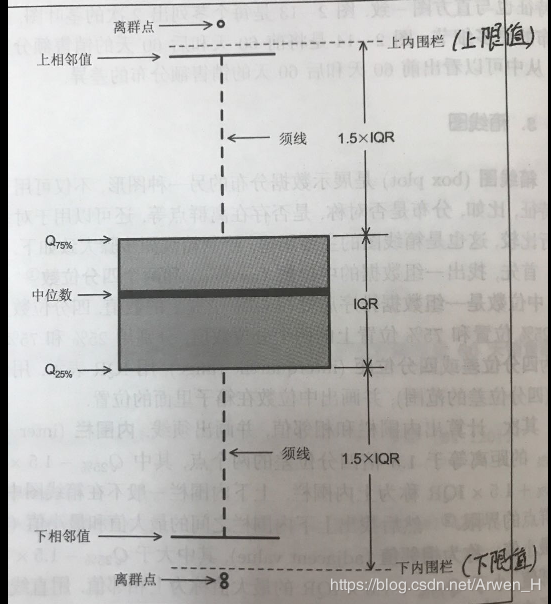
箱形图统计学知识

1）第一四分位数(Quartile，Q1)，又称“较小四分位数”，等于该样本中所有数值由小到大排列后第25%的数字；

2）第二四分位数(Q2)，又称“中位数”，等于该样本中所有数值由小到大排列后第50%的数字；

3）第三四分位数(Q3)，又称“较大四分位数”，等于该样本中所有数值由小到大排列后第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）。



上限值：Q1-1.5×IQR

上相邻值：距离上限值最近的值

须线：上下分位数各自与上下相邻值的距离

上四分位数（Q1)：一组数据按顺序排列，从小至大第25%位置的数值

中位数：一组数据按顺序排列，从小至大第50%位置的数值

四分位距（Interquartile Range, IQR）：Q3-Q1上四分位数至下四分位数的距离

下四分位数（Q3）：一组数据按顺序排列，从小至大第75%位置的数值

下相邻值：距离下限值最近的值

下限值：Q3+1.5×IQR

离群值（异常值）：一组数据中超过上下限的真实值