整个工作主要由两部分组成：

1.通过Gap统计方法确定各个聚类中的K值

2.依据k-medoids聚类中的中心点分析聚类结果

然后实验的课程对象还是选用了《大学计算机基础这门课》，按照之前的步骤先挑选出163个选课浏览者，剩下691个授课阶段浏览者和4442个高覆盖率学习者待聚类分析。

Part1：

Gap统计方法最初在《Estimating the number of clusters in a data set via

the gap statistic》一文中提出（<https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00293>）

原文定义了聚类结果中每一类别中的任意两点的距离之和为

然后对这个值进行均值化并求和：

其中需要注意的是在的计算中和是两个数值和实际意义上相同的值，故归一化中除以了2，而为类别中的元素数目。

由此聚类性能体现在上，值越小，证明每一类别中各点的间距越小，即聚类效果越好。这个目标函数的定义与肘部法则类似，但不一样的地方在于计算中，肘部法则没有计算任意两点间的距离求和，而是仅计算了每一类别中各点与质心点的距离进行求和。

然后文中定义了Gap统计方法的目标函数：

这里后面的为前面定义的取对数（考虑到有可能数值较大），前面的表征的期望值，涉及到生成一个参考数据集的工作。原文中提到这个数据集通常通过蒙特卡洛模拟产生，在样本里所在的矩形区域中（高维的话就是立方体区域）按照均匀分布随机地产生和原始样本数一样多的随机样本，并对这个随机样本做聚类分析，从而得到一个。如此往复多次（通常20次，为减少随机产生的误差），可以得到20个，对这20个数值求平均值，就得到了E(logDk)的近似值。