数据科学实验

实验目的：

通过对学生在某类题目上的掌握程度及各种题目的相似性分析，个性化推荐练习题目给同学。

实验方式：

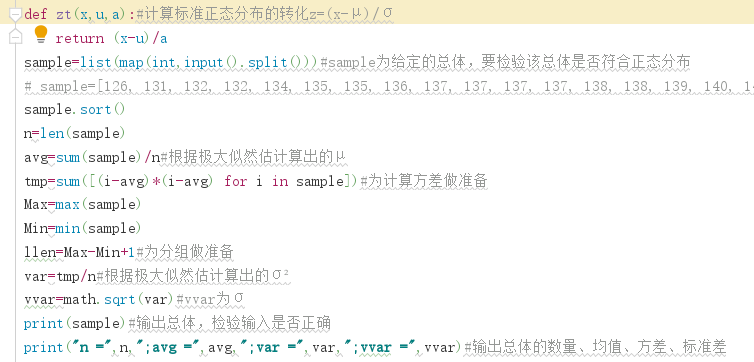
实验总体分为三步：

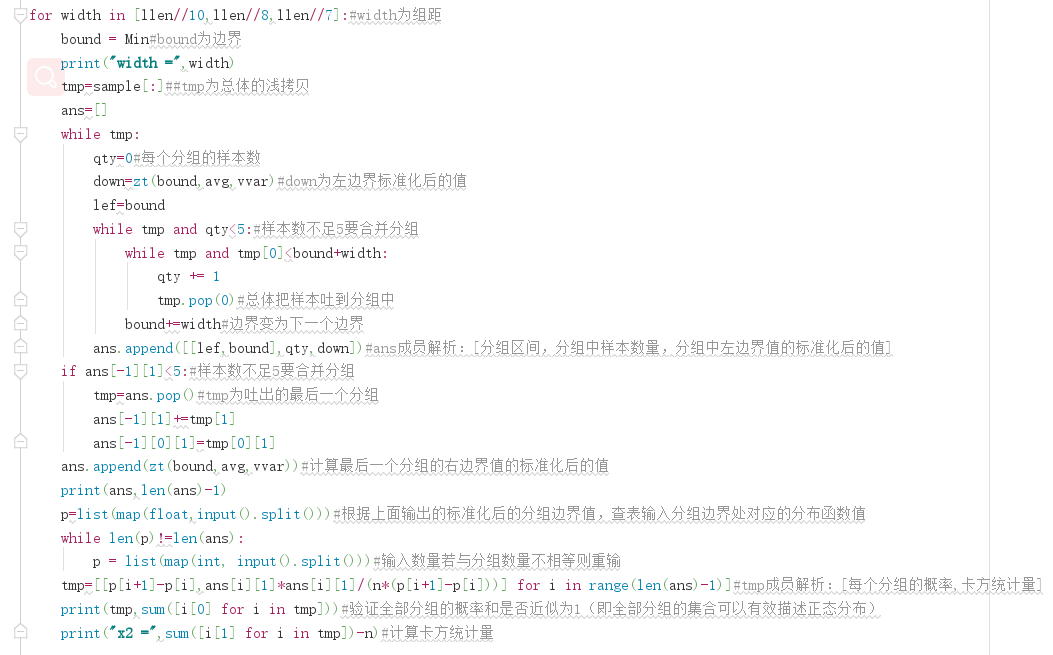
1. 我们想要探究对于给定题目学生的提交数据与学生的掌控程度之间的关系，并假设学生的做题水平分布满足“在x轴划分掌控程度区间而在y轴标定数量形成的直方图构成的图像近似正态分布曲线围成的图像”。于是对于从提交数据抽象出的某个指标（如最终分数、最长提交时间间隔的倒数等），对抽取的题目样本中每一题计算该指标，得到与做该题的学生人数相同数量的总体样本，在显著性水平α（如0.1）下用卡方拟合优度检验来检验假设“H0:总体服从正态分布”，若该题目样本中假设成立的比例超过p，则认为该项指标合理。而正态分布中σ越小，则分布越集中，我们认为该指标对掌控程度的描述越好。重复试验获取一系列描述掌控程度的有效指标，并用卡方独立性检验判断指标之间的相关性，确保所有指标都与掌控程度正相关。
2. 根据第一阶段的实验结果，根据各指标的σ决定权重，提出掌控值概念并计算。即通过统合第一阶段的有效指标给出一个描述学生对于代码题目的掌握能力较好的值。此时将样本范围由横向变为纵向，对于抽取的学生样本中的每个学生，计算学生对于给定的一套相似题目样本中的掌控值，得出与题目样本数量相同的掌控值总体，在显著性水平α（如0.1）下用卡方拟合优度检验来检验假设“H0:总体服从正态分布”，若该学生样本中假设成立的比例超过p，则认为该相似题目的难度分布合理，同时认为正态分布的σ越小，题目相似程度越高，由此验证相似题目分类的合理性。
3. 根据题目描述与分析题目答案代码结构为每个题目贴上若干特征标签，根据特征标签的重合数量判断题目的相似程度，从而由八个分类进一步精化评定相似程度的标准。对于学生做过的题目计算掌控值，对于掌控值在一定区间内的题目认定为“掌握不熟练但有一定基础”，在题库中依据相似程度计算推荐相似题目，即完成“对学生进行个性化推荐”的目标。

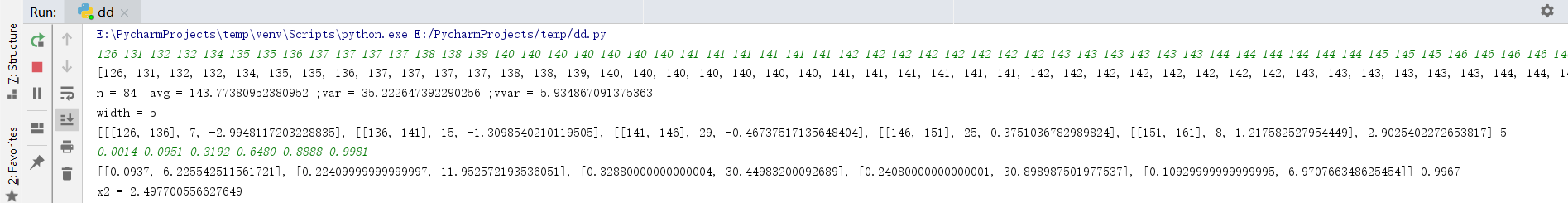
实验代码：

掌控值计算代码包括卡方拟合优度检验、卡方独立性检验，推荐功能代码包括相似代码分析与提取特征、根据标签搜索题库推荐相似题目。

其中卡方拟合优度检验代码示例如下：







卡方拟合优度检验知识如下：

