关于online judge的调研报告

1. 简介

Online Judge系统（简称OJ）是一个在线的判题系统。用户可以在线提交程序多种程序（如C、C++、Pascal）源代码，系统对源代码进行编译和执行，并通过预先设计的测试数据来检验程序源代码的正确性。

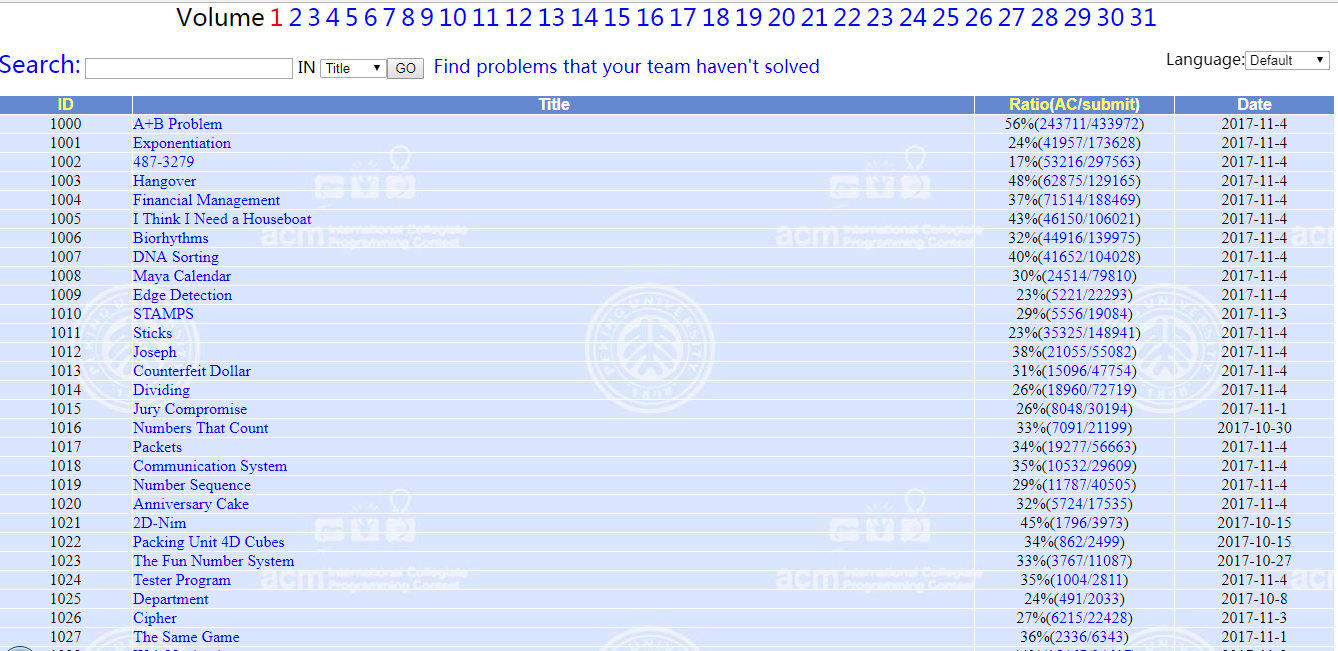
Online Judge系统最初使用于ACM-ICPC国际大学生程序设计竞赛和OI信息学奥林匹克竞赛中的自动判题和排名。现广泛应用于世界各地高校学生程序设计的训练、参赛队员的训练和选拔、各种程序设计竞赛以及数据结构和算法的学习和作业的自动提交判断中。

我们的数据结构课程也使用了交大的oj平台进行自动的作业评分。作为一名学弱，每次交作业都会看到其他同学一次通过，自己却要提交十数次，很是心碎。于是突发奇想，想要利用概率统计课程学习到的统计方法及指标，研究一下online judge上的正确率、提交数等的分布。

由于校内oj最近不太稳定，且总提交量较少，此次调研使用了poj（PKU JudgeOnline）的数据。

1. 原始数据获取（更新于2017.11.21）

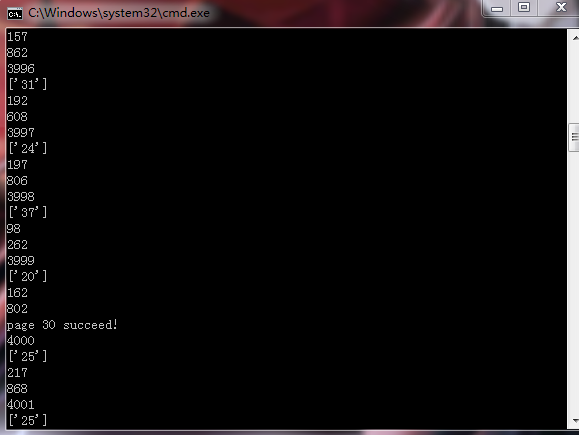
页面上可供获取的信息（如下图）



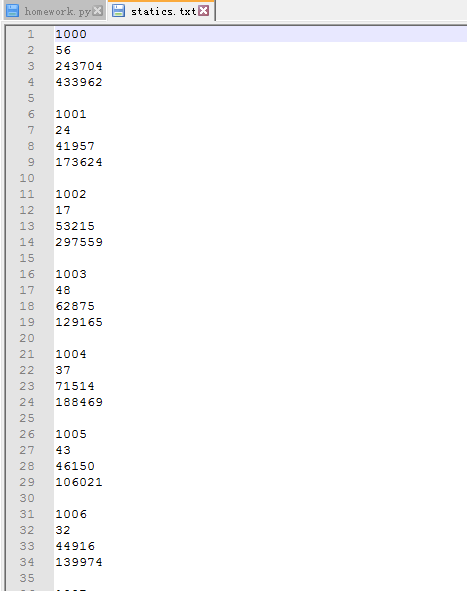
有题号、题目名称、通过比率、通过数量、总提交数、最后提交日期。其中需要的数据有题号、通过比率、通过数量、总提交数，于是我使用了python编写了一个小程序去获取这些数据（如下图，程序在附件中）。



运行情况（如下图）



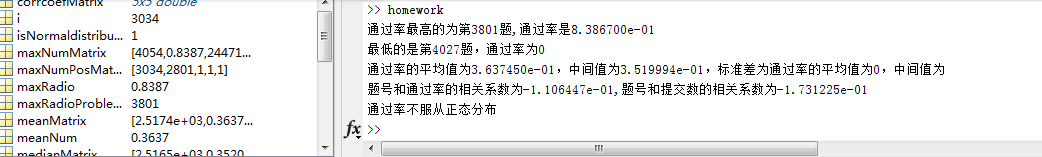
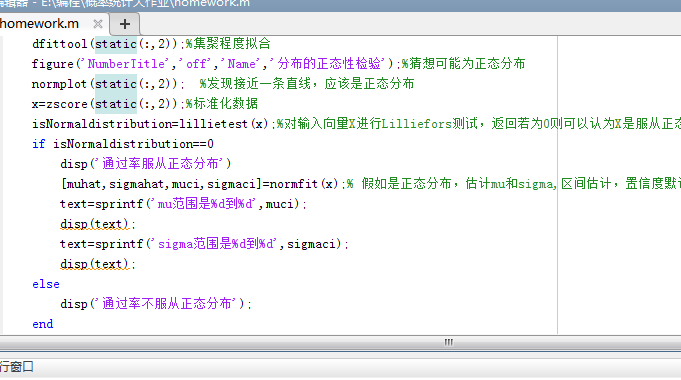
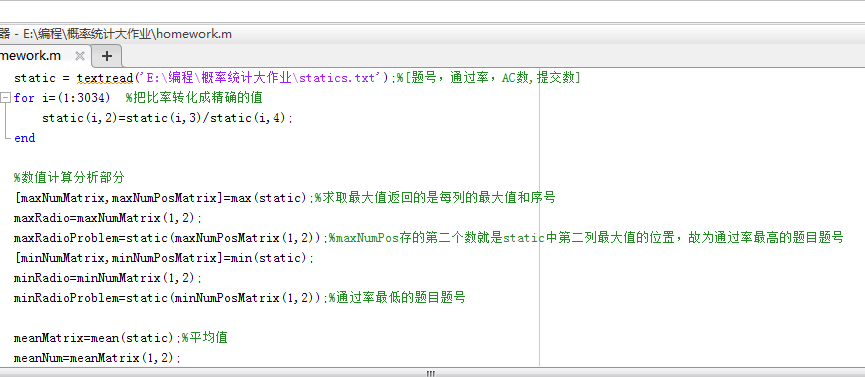
经过一段时间的运行，获取到了一系列数据，存储在了static.txt中（如下图）



（该图为初次获取的图片）每一段的第一行是题号，第二行是通过比率（百分比），第三行是通过数量，第四行是总提交数。后来为了方便导入matlab分析，我改变了数据的记录格式。

之后我使用了C++、matlab进行了数值分析。

1. 数据分析



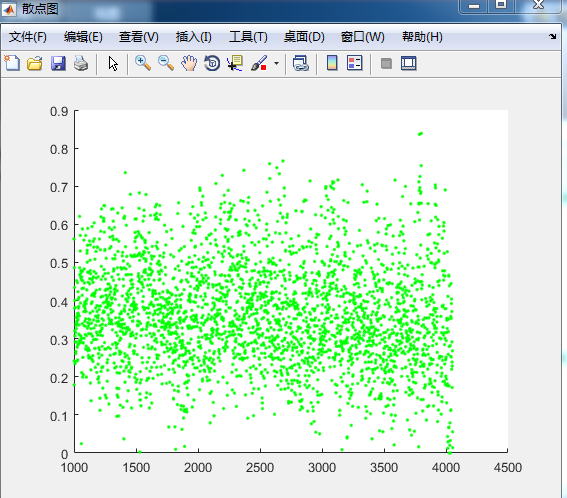
我们首先将static.txt中的数据读入到了matlab中。是一个3034×4的矩阵。

首先，我先是对比率进行了排序，找到了通过最低和最高的题目。由于在原始数据中的比率（从网站上爬取到的比率）只保留到了个位，故使用提交数和通过数求精确的比率。最后找到了这两道题。并对所有的题目的通过率求了平均值、中位数、标准差。

如图：平均值为0.3637，中位数为0.3520，标准差为0.1353。

提交通过率最大的是3801号题通过率为0.8387，最低的4027号题目竟然没有人通过。

我对所有的题目的通过率分布使用点状图进行绘制，分析它的集聚情况。



题号和题目的难度有必然的联系吗？题号和题目提交数量有必然的联系吗？

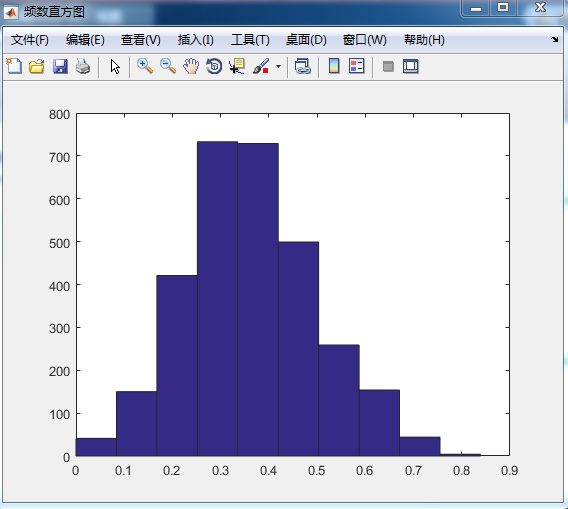
我求取了题号和题目通过率、提交数量的相关系数。如图，分别为-0.1106、-0.1731。可见其线性相关性较低。

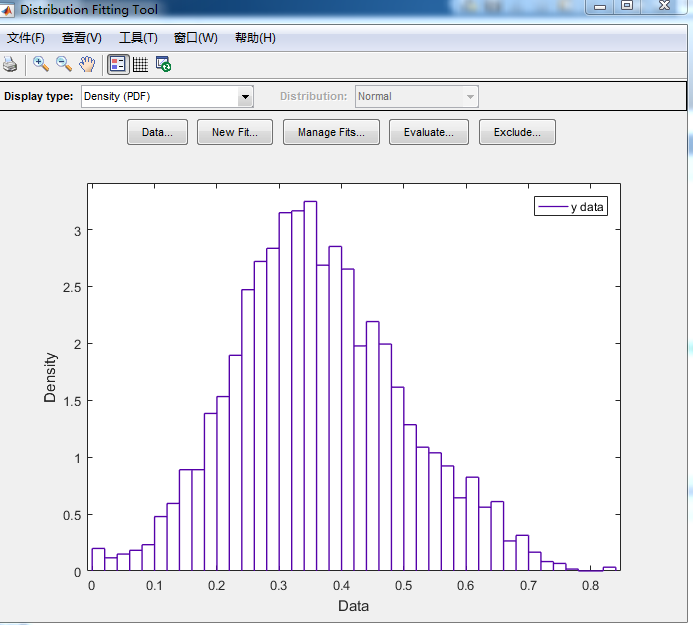
假使以平均的通过率作为我们单次提交通过的概率，那么我们提交完所有的题目各一次后，正确数目的期望值可以达到多少呢？

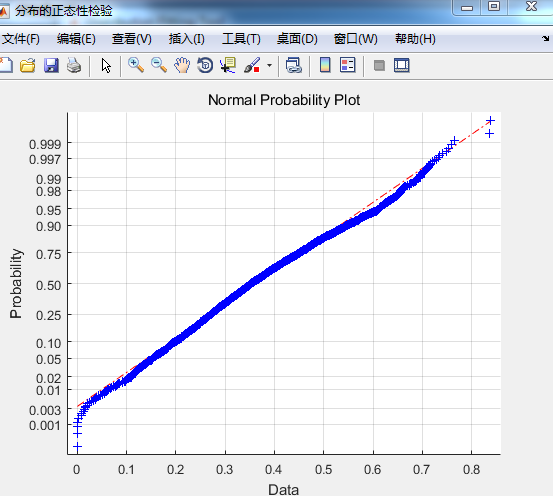
显然，由大数定律，可以使用频率估计概率，所以我们若以p0表示单次提交通过率，p0=0.3637，那么一共有3034道题目，设正确的题目数目为，则由数学期望值的定义可知E(x)=3034\*0.3637=1103。

使用相同的概率分析，像我一样提交15次还不能通过的可能性是多大呢？显然是一个二项分布。P(一道题目提交15次仍然不能通过)=p0^15=2.58×（10^-7）。事实是，这个概率低的令人发指，可见我是天选之子……

之后我想研究一下数据的分布，我做了频数分布直方图，发现类似正态分布，考虑到中心极限定理我猜测通过率的分布，应该的确符合正态分布。我又对通过率进行了拟合，之后又通过分布的正态性检验，发现是接近一条直线的，所以我进行了正态分布的lillietest检验，但是很不幸，对通过率数据标准化后，它并不符合正态分布。但是显然它是十分接近的。







综上分析结束。

1. 总结

当然，这样的分析是有缺陷、有问题的，因为题目的难度并不是可以使用通过率去直接量化的。有很多不确定的因素都会影响题目的通过率。比如有的题目是需要相关知识以后才能做，如果仅仅靠通过率去判断难度，可能有的题目一般人完全没有思路，在这种判断条件下却被当成了容易的题目。因此，想要通过找高通过率的方法找到比较容易的题去练习也是不现实的。但是，从一个宏观的层面上，通过率还是有一定的参考价值，我想网站给出每道题目的通过率，也有供使用者参考题目的难度的用意。

同时，通过此次分析，我在实际生活中运用到了统计的方法，学习了使用计算机获取数据和分析数据的方法。进一步来看，现实中的复杂情况和不确定的影响也激励着我们去探究更科学、更精确的方法去进行概率和统计的研究。