

# 高考志愿填报建议系统

清华大学交叉信息院 赵梓硕

August 12, 2016

## 1 简介

### 1.1 作者信息

赵梓硕，本科生，就读于清华大学交叉信息研究院计科50班。E-mail: 741488923@qq.com

### 1.2 项目简介

每年高考后，填报志愿都是考生们的一大难题。由于每年高考的难易度和区分度都有所不同，每年各大高校及专业的录取分数线也有较大波动。因此，在估计高校的录取线时，一般人往往以历年录取线与批次线的线差作为重要依据。本项目正是以此思想为基础，试图建立一个更加合理的数学模型，对高校分数线给出一个相对更加准确的估计，并以此为依据，根据考生所在地区、高考分数和其偏好，为其推荐合适的高校与专业。

此项目为作者大一暑假实习内容。由于部分数据（如各地区历年考生分布）需要手工获取，以及实习时间、运算资源有限等原因，此项目暂只对北京、湖北、黑龙江、辽宁四个地区理科一本分数段的考生适用。

## 2 建模部分

### 2.1 原理

根据我国高招的规则，无论是顺序志愿或是平行志愿，高校的都是符合条件的考生中按分数从高到低排序录取的，录取的最后一名考生分数即是该院校（专业）的分数线。因此，当院校与专业的热门程度不变的前提下，相比录取线本身，其对应的位次相对更为稳定。

考试的最终目的是通过分数衡量考生水平，并选拔水平符合要求的考生。位次可以体现考生在总体中的相对水平，然而由于分布的非线性性，位次与水平成非线性关系。同时，正如熟知的“从95分到100分，远比从70分到75分要难”，即便排除了考试题的不确定性，考试的预期分数与实际水平也呈现非线性关系。此数学模型要得出的则是一个相对稳定的标尺，衡量考生水平与各院校的要求，从而给出合理的建议。

### 2.2 数学模型

每年高考中存在很多不确定的因素，在进行建模的过程中，作者进行了以下假设：

- (1) 考生实际水平 $L$ 呈正态分布，且每年保持稳定。
- (2) 考试成绩 $S$ 准确反映了考生的水平，即考分是水平的一个保序变换。
- (3) 考生总体意愿每年相同，各院校专业录取“水平线”在期望值附近作无规律波动。

参照四六级考试的“标准分”机制，以一分一段表为依据，若考生 $i$ 成绩 $S_i$ 对应的位次为 $k_i$ ，而当年该省考生总人数为 $n_i$ ，则定义相对位次

$$\alpha_i = \frac{n_i - k_i}{n_i}.$$

在忽略并列影响的前提下，对于任意考生 $j$ ，有

$$\Pr[S_i > S_j] = \alpha.$$

根据假设(2), 有

$$\Pr[L_i > L_j] = \alpha.$$

根据假设(1), 并参考人类智力分布特征, 不妨设(此步参数与最终结果无关, 但本系统及此文档中均使用此度量)

$$L \sim N(100, 15^2),$$

那么已知一分一段表与成绩 $S_i$ , 即可得出对应水平 $L_i$ 。

## 2.3 模型简化

在实际操作中, 要由 $S_i$ 得出 $L_i$ , 必须要该省当年完整的一分一段表, 然而各个省一分一段表的范围、格式往往不同, 难以在同一个网站上批量取得, 甚至许多还以图片形式展现, 因而难以使用程序批量处理, 需要手工获取。然而, 一张一分一段表往往有数百行, 再考虑数十个省的历年情况, 如果一一手工输入, 工作量会极为繁重。因此, 在实际操作中, 需要对一分一段表表现的分布状况进行拟合。

本项目中主要考虑一本线上的考生。对于较为优秀的学生, 其对高中知识应当已基本掌握, 因而高考对该群体而言“不是比谁得分多, 是比谁丢分少”。其能力与训练程度(这里统称为水平)的不同, 导致了其面对高考灵活多变的题型时, 失误的程度会有所不同, 从而最终得分有所不同。

这里假设满分为 $\Omega$ , 定义失误率

$$\varepsilon = \frac{\Omega - S}{\Omega}.$$

显然 $\varepsilon$ 是关于 $L$ 的减函数。然而 $\varepsilon$ 与 $L$ 并不成线性关系。由常识可知, 当 $\varepsilon$ 已经足够小后, 进一步减小它的值将极为困难。由附件Excel表格Sheet1(数据来自北京市近三年一分一段表)可见, 其斜率绝对值随 $L$ 的增加而减小, 与预期相吻合。

根据四省近3~4年的数据,  $\varepsilon$ 与 $L$ 的关系大致为

$$\varepsilon^\lambda \approx kL + b.$$

取 $\lambda = -0.35$ (如Sheet2图表可见), 对北京市各年数据都有较好的拟合, 但由于各地区试卷风格不同, 加上同一地区各年亦有微小差异, 根据具体分布来确定 $\lambda$ 值, 拟合效果会更好。具体操作只需取若干个 $L$ 值, 调整不同的 $\lambda$ 值作线性回归, 使相关系数 $r$ 最大即可。(事实上 $\lambda$ 值一定的偏差对最终结果影响并不大, 精确至0.05即足够)

在此四省中, 最优 $\lambda$ 值在-0.05至-0.45之间不等。

经实际测试, 在 $L$ 不小于106~110, 不大于150左右时, 此拟合较为精确。因此该模型对排名前30%左右的考生有效( $L > 150$ 的考生往往对应清华、北大或至少复旦、交大的档次, 并不需要太多建议), 可用于一本院校分数线预测。

## 3 具体操作

首先, 需要得出各省各年高考分数转换对应表。只须在Sheet3中的“总人数”和“满分”中填入该次高考情况, 表格中“名次”一栏就会自动显示 $L = 150, 140, 130, 120, 110$ 对应的名次, 然后在“分数”栏对应位置填写该名次对应分数, 并调整“参数”栏数据(精确至0.05即可)使右侧图表中的 $R^2$ 最大, 再将Sheet4的内容复制至univ/std/xx-yyyy文件下(xx为省代号, yyyy为年份), 即录入了该省当年分数与水平值的对应表。

运行univ/data/getU.py, 即可在网站<http://gkcx.eol.cn/>上抓取各高校在各地区的录取平均分。(网站上无最低分, 由于自主招生等原因, 最低分也无太大意义, 因此结果会偏高, 但事实上踩线录取一般会进入冷门专业, 因此以平均分为基准有合理性)数据保存在univ/data/下, 格式为res-xx(xx为省代号)。

运行univ/make.sh, 可以将此项目的C++程序进行编译。

运行univ/c-bin/list-generator, 可生成各省份各高校平均分预测, 数据保存在univ/pred/下, 格式为xx-\*\*\*\*\* (xx为省代号, \*\*\*\*\*为省名拼音, 如00-BeiJing)。

运行univ/spec-xml/getS.py, 可抓取各专业xml信息并保存在spec-xml/data下, 再运行univ/spec-xml/proc.py, 可处理其信息并保存在spec-xml/processed下, 文件名为xx(xx为省代号)。

运行univ/tags/tags, 可由univ/data/res-xx生成各高校标签信息, 保存在univ/tags/data/下。  
将website/univ/header.php中的root\_path改为univ目录绝对路径, 即可通过website/univ/index.html运行该项目。

## 4 源代码路径

univ/data/getU.py, univ/spec-xml/getS.py, univ/spec-xml/proc.py: 获得数据  
univ/c-bin/list-generator.cpp: 生成院校分数预测  
univ/c-bin/search.cpp: 查询符合条件的院校  
univ/tags/tags.cpp: 生成学校标签数据  
univ/spec-xml/pred.cpp: 查询符合条件的专业 (-u为按学校查询, -s为按专业查询)