对信息学比赛中选手分数数据的分析

江苏省天一中学 邱天异

摘 要

本文建立了描述信息学比赛的数学模型，并基于该模型研究了过往比赛的选手分数数据。本文通过统计确定了同一名选手的得分波动所服从的分布，基于此从联赛分数推算出了选手整体水平的分布情况，并研究了竞赛选拔流程的效率、回答了有关比赛名次与得分的问题。上一句待完成后更新。本文中得到的结论对信息学竞赛赛制的优化、选手的日常训练和比赛策略制定具有参考意义。

1 引言

中国高中信息学竞赛的参赛人数和竞赛水平在最近十年中快速提高；这种迅猛的发展在让竞赛趋于繁盛的同时，也使得选手和教练对竞赛现状的认知难以跟上节拍。

这一情况引起了一些问题，例如：

- 选手对于自己所处的水平段认识不足，从而作出错误的学业规划。
- 出题人对于选手的水平认识不足，导致题目难度和部分分分配失当。
- 选手不了解对手的水平和发挥情况，导致选择了错误的考场策略。

本文将利用数学工具，基于过往比赛的选手分数数据来分析信息学竞赛的现状，以为上述问题的解决提供助力。

本文中用到的全部数据和计算程序可以在以下网址下载：

- https://files.cnblogs.com/files/turboboost/qty-thesis-statdata.zip
- https://github.com/TianyiQ/ioi2021-thesis/blob/main/qty-thesis-statdata.zip

正文分为五个部分：

第二节 建立用于描述信息学比赛的数学模型，作为后续分析的基础。
第三节 分析同一名选手的得分波动所服从的分布。
第四节 利用联赛初赛、复赛的得分数据推算出信息学竞赛选手整体水平的分布情况。
第五节 对于在比赛名次中出现的现象进行讨论。
第六节 待完成后更新。

由全文的目标决定，本文将不会对初中信息学竞赛进行研究，因此下文中在提到任何比赛时默认指面向高中生的比赛。

2 建立模型

2.1 赛程和赛制

在引入模型前，先对信息学竞赛的竞赛流程和比赛形式作简要介绍[1]。

信息学竞赛是一系列比赛的统称。这些比赛整体上呈现“逐级递进”的关系，即下一层比赛的优胜者晋级上一层比赛。这些比赛按照级别从低到高，大致排列为[2]：

a. 全国联赛 (NOIP/CSP) — 初赛
b. 全国联赛 (NOIP/CSP) — 复赛
c. 省队选拔赛
d. 清华/北大学科营 (THUWC/PKUWC/THUSC/PKUSC)
e. 亚太地区竞赛 (APIO)
f. 国家队选拔赛 (CTSC/CTS) — 非正式选手
g. 全国冬令营 (NOIWC) — 非正式选手
h. 全国决赛 (NOI)
i. 清华/北大集训 (CTT)
j. 全国冬令营 (NOIWC) — 正式选手
k. 国家队选拔赛 (CTSC/CTS) — 正式选手
l. 国际奥林匹克竞赛 (IOI)

图1展示了这些比赛间的关系。箭头从低级别比赛指向高级别比赛，表示该低级别比赛的优胜者可以晋级对应的高级别比赛，箭头上标记的数值表示大致晋级人数。

赛制即比赛的进行方式和比赛规则。信息学竞赛中采用笔试、COI 赛制（机试）、IOI 赛制（机试）这三种不同的赛制，表1给出了每种赛制的特点和先前提到的比赛所分别采用的赛制。

	时长	题数	题目类型	反馈机制	对应比赛
笔试	1~2h	数十	选择题、填空题	无反馈	a
COI 赛制	3~5h	3~4	编程题，有多档部分分	无反馈	bcfghk
IOI 赛制	3~5h	3~4	编程题，有多档部分分	多次提交、有反馈	deijl

表 1: 信息学比赛采用的赛制

[1]赛程和赛制在近几年有小幅变化，本小节中会尽量兼顾新旧两套机制
[2]后文将用下表中的字母标号来代指对应的比赛

$$
\begin{tikzcd}
{} \arrow[r, ">50000"] & a \arrow[d, "20000"] & \\
& b \arrow[dl, "500"'] \arrow[d, "1000"] \arrow[dr, "500"] & \\
d & c \arrow[l, "300"] \arrow[d, "300"] & efg \\
& h \arrow[d, "70"] & \\
& ijk \arrow[d, "4"] & \\
& l &
\end{tikzcd}
$$

图 1: 信息学比赛间的关系

2.2 数学模型

本小节中将建立用于描述一场信息学比赛的数学模型。

2.2.1 基本模型

为了更清晰地界定模型在现实中的适用范围，需要先明确：现实中怎样的对象能被称为一场“比赛”。

定义 2.1 (现实比赛). 一个**现实比赛**，即特定的人群在同样的规则下测试同一套题目的过程。一个现实比赛被参赛人群、规则和题目这三个要素所确定。

在这一定义下，每年中的 $a \sim l$ 这 12 个比赛，自然都是现实比赛。而且，不仅是包含两天考试的一场完整的比赛算作现实比赛，单独拿出其中一天也算现实比赛。

关于“参赛人群”这一概念需要注意两点：

• 参赛人群只是一个宽泛的范围，而不是具体的选手集合。例如我们可以规定参赛人群为“所有学习信息学的同学”，但这一规定并不关注张三、李四、王五是否是这个人群的成员。这样的规定不会给后续的分析带来不利影响，因为我们只关心关于比赛和人群的统计信息，而不关心每名选手的特点。

• 参赛人群不必囊括实际参赛的整个选手群体；例如在 NOIP 初赛中，“所有报名了初赛的女生”这一参赛群体依然能构成现实比赛。这一点对于后文中跨越不同比赛的分析大有帮助。

接下来定义从现实比赛抽象而来的数学模型。

定义 2.2 (理想比赛). **理想比赛** A 由二元函数 $\mathcal{H}_A : [0,1] \times \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ 确定，其中 $\mathcal{H}_A$ 连续且满足

$$\int_0^1 \int_{-\infty}^{+\infty} \mathcal{H}_A(x,\delta)\mathrm{d}\delta\mathrm{d}x = 1 \tag{1}$$

和

$$\int_{-\infty}^{+\infty} \delta \cdot \mathcal{H}_A(x_0,\delta)\mathrm{d}\delta = 0, \quad \forall x_0 \in [0,1] \tag{2}$$

此时我们把 $\mathcal{H}_A$ 称为 A 的**综合分布函数**。

接下来将定义：一个理想比赛何时被认为"描述"了一个现实比赛。这也将同时表明综合分布函数的实际含义。

首先约定一下记号：

- $\Pr[A]$ 表示事件 A 发生的概率。
- $\mathrm{E}[X]$ 表示随机变量 X 的期望值。

定义 2.3. 从现实比赛 B 可按如下方式确定一个理想比赛 A：

步骤 1. 记 B 的参赛选手集合为有限集 S_B，并在 B 的参赛人群（包括人群内部的具体构成）不变的情况下，假想参赛人数 $|S_B|$ 趋于无穷。我们之所以能够任意钦定 $|S_B|$，是因为——如先前所述——B 的定义并未指明具体的选手集合。

步骤 2. 每一名参赛选手 p 在比赛 B 中的实际得分 $score_p$ 是一个随机变量，它被各种偶然因素（如临场发挥）所支配，但是它的分布可以由选手 p 和现实比赛 B 的三个要素完全确定。假想对每一名选手 p 计算其期望得分 $exscore_p = \mathrm{E}[score_p]$，并取所有选手期望得分的最大值，记作 M_B。由于参赛人数趋于无穷，每一个个人的特征可以忽略，故 $M_B = \max_{p \in S_B} exscore_p$ 仅由 B 确定。

步骤 3. 从 S_B 中等概率随机选取一名选手 p，并：

- 定义 $[0,1]$ 上的随机变量 $X_B = \frac{exscore_p}{M_B}$。[3]易见随机变量 X_B 的实际取值与考场上的偶然因素无关，而是由选取 p 的方式确定。
- 定义 $\mathbb{R}$ 上的随机变量 $\Delta_B = \frac{score_p - exscore_p}{M_B}$。易见随机变量 Δ_B 的实际取值由选取 p 的方式和考场上的偶然因素（如选手临场发挥）共同确定。
  - 请注意，X_B 和 Δ_B 的定义中所用的 p 是同一名随机选择的选手，而不是独立的两次选择。

[3]"将每名选手的分数除以最高分数"这一操作，类似于信息学比赛中计算标准分的方式。另外注意到：虽然 $\frac{exscore_p}{M_B} \leqslant 1$，但 $\frac{score_p}{M_B}$ 可以大于 1

步骤 4. 取 A 的综合分布函数 $\mathcal{H}_A$ 为 X_B 与 Δ_B 的联合概率密度函数，从而确定 A 。换句话说，对所有 $x_0 \in [0,1], \delta_0 \in \mathbb{R}$ ，需要满足[4]

$$\int_0^{x_0}\int_{-\infty}^{\delta_0} \mathcal{H}_A(x,\delta)\mathrm{d}\delta\mathrm{d}x = \Pr\left[(X_B \leqslant x_0) \wedge (\Delta_B \leqslant \delta_0)\right] \tag{3}$$

由(3)知定义2.2的等式(1)满足；由 $\mathrm{E}\left[\Delta_B\right] = 0$ 知等式(2)满足。从而只要联合概率密度函数 $\mathcal{H}_A$ 存在且连续，A 就符合理想比赛的定义。

对于按上述方式得到的 A ，我们称 A 与 B **互相对应**。如果按上述过程得到的 $\mathcal{H}_A$ 不连续或根本不存在，则认为不存在与 B 对应的 A 。

冗长的定义可以用一句话来作直观的总结：$\mathcal{H}_A(x,\delta)$ 表示真实水平（即期望得分）约为 x（$x \in [0,1]$ 为按最高分折算后的标准分）、实际表现约为 $x+\delta$（同样表示标准分）的选手的期望人数占总人数的比例；之所以实际表现会偏离真实水平——以及这里之所以说“期望人数”——是因为考场上的各种偶然因素为比赛结果带来了随机性。

可以看到，理想比赛这一模型只考虑了哪些结果可能出现，而未考虑哪种结果实际出现。而在现实中，能够获知的却只有实际出现的结果——和它恰恰相反。下面定义的概念将处理这一问题。

定义 2.4 (分数分布函数)**.** 对理想比赛 A ，定义其**分数分布函数** $C_A : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ 满足

$$C_A(s) = \int_0^1 \mathcal{H}_A(x, s-x)\mathrm{d}x, \quad \forall s \in \mathbb{R}$$

命题 2.5 (分数分布函数的实际含义)**.** 对现实比赛 B 和与之对应的理想比赛 A ，假想比赛 B 的参赛人数 $|S_B|$ 趋于无穷，等概率随机选取选手 $p \in S_B$ ，则[5]：

$$\Pr\left[\frac{score_p}{M_B} \leqslant r\right] = \int_{-\infty}^{r} C_A(s)\mathrm{d}s, \quad \forall r \in \mathbb{R}$$

[4]也可以直观地理解为 $\mathcal{H}_A(x_0,\delta_0) = \Pr\left[(X_B \approx x_0) \wedge (\Delta_B \approx \delta_0)\right]$，不写作 $X_B = x_0$ 是因为取等概率为 0

[5]也就是说 C_A 为随机变量 $\frac{score_p}{M_B}$ 的概率密度函数。和先前类似，这里也可以直观理解为 $C_A(r) = \Pr\left[\frac{score_p}{M_B} \approx r\right]$

证明.

$$
\begin{aligned}
\Pr\left[\frac{score_p}{M_B} \leqslant r\right] &= \Pr\left[X_B + \Delta_B \leqslant r\right] \\
&= \iint\limits_{\{(x,\delta):x\in[0,1],\delta\in\mathbb{R},x+\delta\leqslant r\}} \mathcal{H}_A(x,\delta)\mathrm{d}(x,\delta) \\
&= \iint\limits_{\{(x,s):x\in[0,1],s\in(-\infty,r]\}} \mathcal{H}_A(x,s-x)\mathrm{d}(x,s) \\
&= \int\limits_{-\infty}^{r}\left(\int\limits_{0}^{1}\mathcal{H}_A(x,s-x)\mathrm{d}x\right)\mathrm{d}s \\
&= \int\limits_{-\infty}^{r} C_A(s)\mathrm{d}s
\end{aligned}
$$

□

在上面四个定义中，涉及到现实情况的部分难免有模糊之处；实际应用中对这几条定义的执行，也不可避免地需要作近似处理。但即便如此，作出这些规定依然能极大地帮助我们厘清思路并发现隐含的前提。

2.2.2 特殊情况下的模型

在一场现实比赛 B 中，每一个选手 $p \in S_B$ 的实际得分相比真实水平的“得分偏移量” $\frac{score_p - exscore_p}{M_B}$ 都是一个随机变量。如果所有选手的“得分偏移量”独立同分布，对我们的模型意味着什么？

容易想到，此时随机变量 Δ_B 的概率分布就和任何一个选手的“得分偏移量”的概率分布完全相同。换句话说，在定义2.3的步骤3中，不论我们钦定选取哪一个 p，Δ_B 取任何一个值的概率都是固定的，且恰好等于在不固定 p 的情况下 Δ_B 取这个值的概率。再换句话说[6]：

$$
\Pr\left[(\Delta_B \leqslant \delta)\middle|(X_B = x)\right] = \Pr\left[\Delta_B \leqslant \delta\right], \forall\left(\delta \in \mathbb{R}, x \in [0,1], \Pr\left[X_B = x\right] > 0\right)
$$

即随机变量 X_B, Δ_B 独立。在研究这件事之前，我们需要一对新的定义。

定义 2.6 (期望值分布函数和偏移量分布函数)**.** 对任意的理想比赛 A：

- 定义其**期望值分布函数** $\mathcal{X}_A : [0,1] \to \mathbb{R}_{\geqslant 0}$ 满足

$$
\mathcal{X}_A(x_0) = \int\limits_{-\infty}^{+\infty} \mathcal{H}_A(x_0,\delta)\mathrm{d}\delta, \quad \forall x_0 \in [0,1]
$$

[6] 和之前类似，这里之所以不写 $\Delta_B = \delta$，是因为取等概率为 0

- 定义其**偏移量分布函数** $\mathcal{D}_A : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ 满足

$$\mathcal{D}_A(\delta_0) = \int_0^1 \mathcal{H}_A(x, \delta_0)\mathrm{d}x, \quad \forall \delta_0 \in \mathbb{R}$$

命题 2.7 (期望值分布函数和偏移量分布函数的实际含义)**.** 对现实比赛 B 和与之对应的理想比赛 A：

- $\mathcal{X}_A$ 为 X_B 的概率密度函数。换句话说[7]：

$$\Pr[X_B \leqslant x_0] = \int_0^{x_0} \mathcal{X}_A(x)\mathrm{d}x, \quad \forall x_0 \in [0,1]$$

- $\mathcal{D}_A$ 为 Δ_B 的概率密度函数。换句话说[8]：

$$\Pr[\Delta_B \leqslant \delta_0] = \int_{-\infty}^{\delta_0} \mathcal{D}_A(\delta)\mathrm{d}\delta, \quad \forall \delta_0 \in \mathbb{R}$$

证明比较显然，这里略去。下面考虑 X_B, Δ_B 间的独立性带来的性质。

命题 2.8. 对现实比赛 B 和与之对应的理想比赛 A，如果 X_B 与 Δ_B 独立，则：

$$\mathcal{H}_A(x_0, \delta_0) = \mathcal{X}_A(x_0)\mathcal{D}_A(\delta_0), \quad \forall (x_0, \delta_0) \in [0,1] \times \mathbb{R} \tag{4}$$

更进一步，(4)是 X_B 与 Δ_B 独立的充要条件。

证明.

$$\begin{aligned}
&\Pr\left[(\Delta_B \leqslant \delta_0) | (X_B = x_0)\right] = \Pr\left[\Delta_B \leqslant \delta_0\right], \ \forall \left(\delta_0 \in \mathbb{R}, \Pr\left[X_B = x_0\right] > 0\right) \\
\Leftrightarrow &\left(\int_{-\infty}^{\delta_0} \mathcal{H}_A(x_0, \delta)\mathrm{d}\delta\right) \Big/ \mathcal{X}_A(x_0) = \int_{-\infty}^{\delta_0} \mathcal{D}_A(\delta)\mathrm{d}\delta, \ \forall \left(\delta_0 \in \mathbb{R}, \mathcal{X}_A(x_0) > 0\right) \\
\Leftrightarrow &\frac{\mathcal{H}_A(x_0, \delta_0)}{\mathcal{X}_A(x_0)} = \mathcal{D}_A(\delta_0), \ \forall \left(\delta_0 \in \mathbb{R}, \mathcal{X}_A(x_0) > 0\right) \\
\Leftrightarrow &(4)
\end{aligned}$$

最后一步中还需要特别考虑 $\mathcal{X}_A(x_0) = 0$ 的情况，不难自行补全。 □

定义 2.9 (简单理想比赛)**.** 如果理想比赛 A 满足 (4) 式，则称它是**简单**的。

[7] 也可以直观地理解为：$\Pr[X_B \approx x_0] = \mathcal{X}_A(x_0)$

[8] 也可以直观地理解为：$\Pr[\Delta_B \approx \delta_0] = \mathcal{D}_A(\delta_0)$

由命题2.8，对于简单理想比赛 A ，从 $\mathcal{X}_A,\mathcal{D}_A$ 可唯一确定 $\mathcal{H}_A$ ，进而能够确定 C_A 。

命题 2.10 (简单理想比赛的分数分布函数)**.** 对简单理想比赛 A :

$$C_A(s)=\int_0^1 \mathcal{X}_A(x)\mathcal{D}_A(s-x)\mathrm{d}x,\quad \forall s\in\mathbb{R}$$

证明显然，这里略去。

2.3 几个关键的假设

为了使得后续分析成为可能，我们还需要对真实情况作一些近似处理。近似处理的具体方式由本小节的几个假设给出。

假设 2.11. 对任何一个信息学（现实）比赛 B ，都存在符合定义2.2的理想比赛 A 与其对应。

假设 2.12. 对任何一个现实比赛，如果它的<u>规则</u>基于 COI 或 IOI 赛制，则它对应的理想比赛是简单的。

假设 2.13. 考虑所有基于 COI 或 IOI 赛制的现实比赛，考察它们对应的理想比赛的偏移量分布，这些分布应该是<u>相似</u>的，即它们应该有相同的形式，即使其中的参数可能有不同的取值。

在给出下一个假设之前，还需要定义一个概念。

定义 2.14 (缩放等价)**.** 对简单理想比赛 A_1,A_2 ，当存在线性映射 $f(x)=\alpha x+\beta \quad (\alpha\in\mathbb{R}_{>0},\beta\in\mathbb{R})$ 同时满足以下条件时，称 A_1,A_2 **缩放等价**，称 f 为 A_1,A_2 间的**等价映射**：

1. $f(1)=1$
2. $\mathcal{D}_{A_2}(\alpha\delta)=\mathcal{D}_{A_1}(\delta)\cdot\alpha^{-1},\quad \forall\delta\in\mathbb{R}$
3. $\bar{\mathcal{X}}_{A_2}(f(x))=\bar{\mathcal{X}}_{A_1}(x)\cdot\alpha^{-1},\quad \forall x\in\mathbb{R}$ ，其中

$$\bar{\mathcal{X}}(x)=\begin{cases}0 & x\notin[0,1]\\ \mathcal{X}(x) & x\in[0,1]\end{cases}$$

如果上述映射 f 只满足条件1和3，则称 A_1,A_2 **弱缩放等价**，称 f 为 A_1,A_2 间的**弱等价映射**。

命题 2.15 (缩放等价的实际含义)**.** 对缩放等价的 A_1,A_2 及其等价映射 f ，有以下关系[9]：

[9]可以直观理解为：现实比赛 B_1（对应于 A_1 ）中的分数，经过 $f:x\mapsto\alpha x+\beta$ 的变换之后，变成了现实比赛 B_2（对应于 A_2 ）中的分数

1. $\int_{-\infty}^{\delta_0} \mathcal{D}_{A_1}(\delta)\mathrm{d}\delta = \int_{-\infty}^{\alpha\delta_0} \mathcal{D}_{A_2}(\delta)\mathrm{d}\delta, \quad \forall\delta \in \mathbb{R}$

2. $\int_{-\infty}^{x_0} \bar{X}_{A_1}(x)\mathrm{d}x = \int_{-\infty}^{f(x_0)} \bar{X}_{A_2}(x)\mathrm{d}x, \quad \forall x \in \mathbb{R}$

对于弱缩放等价类似。

证明. 先来看关于 $\mathcal{D}_{A_1}, \mathcal{D}_{A_2}$ 的部分：

$$\begin{aligned}
\int_{-\infty}^{\delta_0} \mathcal{D}_{A_1}(\delta)\mathrm{d}\delta &= \int_{-\infty}^{\delta_0} \mathcal{D}_{A_2}(\alpha\delta)\alpha\mathrm{d}\delta \\
&= \int_{-\infty}^{\alpha\delta_0} \mathcal{D}_{A_2}(\alpha\delta)\alpha\mathrm{d}(\alpha\delta)\cdot\alpha^{-1} \\
&= \int_{-\infty}^{\alpha\delta_0} \mathcal{D}_{A_2}(t)\mathrm{d}t
\end{aligned}$$

对于 $\bar{X}_{A_1}, \bar{X}_{A_2}$ 同理，这里不再重复。 □

假设 2.16. 对现实比赛 B_1 （对应理想比赛 A_1 ）和 B_2 （对应理想比赛 A_2 ），如果

1. B_1, B_2 的规则都基于 COI 或 IOI 赛制
2. B_1, B_2 的规则在除了赛制外的各方面均相同
3. B_1, B_2 的参赛人群相同

则 A_1, A_2 一定缩放等价。如果 B_1, B_2 满足条件1和3，则 A_1, A_2 一定弱缩放等价。

对这些假设无法予以严格的证明，但在此可以列举一些感性的理由，来说明它们大体上是可靠的。

1. 如果假设所有比赛在考查角度上没有差异（因为我们只关心普遍的统计特征，所以这种假设是合理的），那么一名选手的解题能力（即，能够在比赛中解出多大难度的题目）就一定是固定的。

2. 当组题人为一场比赛选择题目、出题人为命制的题目设置部分分时，他们会有意识地给较难的任务设置较高的分值、给较简单的任务设置较低的分值，而具体多高、多低，则取决于他们心中作的判断。虽然不同的人可能作出不一样的判断，但这些判断应该大体上是“成比例”的。例如：张三认为算法 2 应当获得三倍于算法 1 的得分、李四认为算法 2 应当获得 2.5 倍于算法 1 的得分，这两种判断在比例上是大致相符的。

3. 综合1和2，我们知道了：每个选手的能力可以看作是不变的；选手比赛中完成的任务难度与所获分数间的关系，这一关系在不同比赛之间应该是“成比例”的。所以只要选手集合不变，不同比赛的“选手期望得分构成的分布”也应该是“成比例”的（特别地，这两

个分布的最大值也应该是相对应的，所以在定义2.14中要求 $f(1) = 1$）。这就为假设2.16关于期望值分布的部分和对 $f(1) = 1$ 的要求提供了依据。

4. 根据经验，一名选手考场发挥的稳定与否与水平高低等因素没有明显的相关性；所以虽然不同选手的稳定性存在差异，但是在样本很大时，这种差异不会给统计结果带来较大的系统性的偏差，因此我们近似地认为所有选手水平发挥的稳定性是相同的。又因为得分与实际表现出的能力是“成比例”的，所以所有选手比赛得分的稳定性也是相同的。这为假设2.12和假设2.13提供了依据。

5. 不同的比赛因为比赛天数、试题数目等的不同，可能导致选手得分稳定性的不同（一般来说比赛天数越多，选手得分越稳定）。但如果两场比赛的天数、题数（算作比赛规则的一部分）等都相同，就可以用4中的论证，来为假设2.16关于偏移量分布的部分提供依据。

6. 真实的比赛中“离散”的特性——比如选择题三分一道——可以在理想化的模型中忽略。这样在人数趋于无穷时，我们很容易想到：其各种统计数据会是“连续”的。因此2.11是一个很自然的假设。

7. 根据经验，在 COI 赛制中表现好的选手，在 IOI 赛制往往表现也很好；反之亦然。因此 COI/IOI 赛制间的差异至多会对选手期望得分的分布起到缩放的作用，而不会带来本质的改变。类似地，选手在 COI/IOI 赛制中发挥稳定性的差异，也只有量的差别而无质的差别。所以，认为 COI/IOI 赛制的比赛有着本质相同（即在缩放后完全相同）的期望值分布、偏移量分布，是合理的。

8. 假设2.12会带来一个问题：如果一名选手的期望得分十分接近 0，但他的分数波动的幅度仍被认为与其他选手相同，就会使他可能考出“负分数”，并使得分数分布函数在负数处的点值非零。由于本文只研究近似的结果，且考虑到该现象并不会十分显著（因为一场比赛中只会有很少的选手期望得分接近 0），所以可以容忍这一不合理的现象。

3 偏移量分布的测量

由假设2.13，COI/IOI 赛制下偏移量分布有一定的形式。本节中，将利用过往比赛的分数数据得到偏移量分布的形式。

3.1 数据的获取

数据来自以下三场比赛：

- 2018 年北大集训（字母标号 i）
- 2019 年北大集训（字母标号 i）
- 2020 年北大集训（字母标号 i）

选用它们的原因是，北大集训包含连续进行的四场考试，更多的考试场数使得我们能够更精确地估计每一名选手的期望分数。

这些比赛的参赛情况见表2。

	参赛总人数	正式选手人数	非正式选手人数	选拔人数
北大集训 2018	约 60	50	约 10	15
北大集训 2019	约 70	50	约 20	15
北大集训 2020	约 90	50	约 40	30

表 2: 三场比赛的参赛情况

根据经验判断，这三场比赛中并非所有选手都全情投入。因此为了保证数据可靠性，对每场比赛只取总排名[10]中最靠前的 $1.5K \sim 2K$ 名选手的数据，其中 K 表示当场比赛的选拔人数。具体地说：北大集训 2018 取前 30 名、北大集训 2019 取前 30 名、北大集训 2020 取前 50 名。另外为保证比赛之间的统一性，后文中在计算考试分数标准差时，每场比赛只取总排名中前 30 名的分数。

3.2 数据的加工处理

三场比赛的参赛选手共计 110 人次，我们将他们视为 110 名不同的选手。三场比赛共计 12 场考试，我们将它们视为 12 个不同的现实比赛。参加这些现实比赛的共计 440 人次。

虽然这 12 个现实比赛的参赛人群是相同的（国家集训队选手和精英培训选手），但它们在题目难度等方面并不相同，如果直接将它们的数据汇总起来的话，会使得数据失去意义。为解决这一问题，我们需对比赛得分进行变换。

命题 3.1. 对缩放等价的理想比赛 A_1, A_2 及其等价映射 $f(x) = \alpha x + \beta$，有

$$\alpha = \frac{\mathrm{Stddev}\left[C_{A_2}\right]}{\mathrm{Stddev}\left[C_{A_1}\right]}$$

其中 $\mathrm{Stddev}\left[F\right]$ 表示以 F 为概率密度函数的随机变量[11]的标准差。

另外注意到由 $f(1) = 1$ 可得 $f(x) = 1 - \alpha(1 - x)$，所以不必再考虑 β 的取值。

[10] 总排名中按每天标准分总和降序排列

[11] 换句话说，这样的随机变量 Y 满足 $\Pr\left[Y \leqslant t\right] = \int_{-\infty}^{t} F(s)\mathrm{d}s, \quad \forall t \in \mathbb{R}$

证明.

$$\begin{aligned}
C_{A_1}(s) &= \int_0^1 \mathcal{X}_{A_1}(x)\mathcal{D}_{A_1}(s-x)\mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \bar{\mathcal{X}}_{A_1}(x)\mathcal{D}_{A_1}(s-x)\mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \left(\bar{\mathcal{X}}_{A_2}(\alpha x+\beta)\cdot\alpha\right)\left(\mathcal{D}_{A_2}(\alpha(s-x))\cdot\alpha\right)\mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \alpha^2\bar{\mathcal{X}}_{A_2}(\alpha x+\beta)\,\mathcal{D}_{A_2}(\alpha s+\beta-(\alpha x+\beta))\,\mathrm{d}(\alpha x+\beta)\cdot\alpha^{-1} \\
&= \alpha\int_{-\infty}^{+\infty} \bar{\mathcal{X}}_{A_2}(t)\mathcal{D}_{A_2}((\alpha s+\beta)-t)\mathrm{d}t \\
&= \alpha C_{A_2}(\alpha s+\beta), \quad \forall s\in\mathbb{R}
\end{aligned}$$

设连续型随机变量 Y_1 满足其概率密度函数为 C_{A_1} ，Y_2 满足其概率密度函数为 C_{A_1} ，则 $\alpha Y_1+\beta$ 与 Y_2 同分布。从而[12] $\mathrm{Var}\left[Y_2\right]=\mathrm{Var}\left[\alpha Y_1\right]=\alpha^2\cdot\mathrm{Var}\left[Y_1\right]$ ，于是 $\mathrm{Stddev}\left[Y_2\right]=\alpha\cdot\mathrm{Stddev}\left[Y_1\right]$ 。 □

结合等价映射的实际含义和命题3.1，可以得到对前述 12 个现实比赛 $B_{1\ldots12}$ 的分数做变换的方法：

步骤 1. 记 $B_{1\ldots12}$ 对应的理想比赛为 $A_{1\ldots12}$ 。

步骤 2. 构造 $A'_{1\ldots12}$ 满足 A_i 与 A'_i 缩放等价，且等价映射为 $f_i(x)=1-\frac{1-x}{c\cdot\mathrm{Stddev}\left[C_{A_i}\right]}$ 。这里 $c=4$ 为根据实际数据所选取的固定常数，用来避免产生负分数。

步骤 3. 则 $A'_{1\ldots12}$ 这 12 个理想比赛完全相同（即它们的综合分布函数相同），且与 $A_{1\ldots12}$ 中的每一个缩放等价。

另外须注意，根据定义2.3的步骤2，我们需要对每个现实比赛 B_i 确定选手期望分数的最大值 M_{B_i} 。这里可以用实际分数的最大值来近似地代替期望分数的最大值。

因为 $A'_{1\ldots12}$ 与 $A_{1\ldots12}$ 中的每一个缩放等价，所以我们只需测量 $A'_{1\ldots12}$ 的偏移量分布，即可得到结论。现在开始目标将转为测量 $A'_{1\ldots12}$ 的偏移量分布；为便于表述，记 $B'_{1\ldots12}$ 表示 $A'_{1\ldots12}$ 对应的现实比赛。

现在我们得到了 12 个完全相同的理想比赛 $A'_{1\ldots12}$ ，和每个理想比赛对应的现实比赛的分数数据；而因为 $A'_{1\ldots12}$ 完全相同，所以所有这些分数数据可以直接合并。现在我们有了一

[12]这里 $\mathrm{Var}\left[Y\right]$ 表示随机变量 Y 的方差

个理想比赛（记为 A'，对应现实比赛 B'）和对应的 440 名选手的分数数据。原先的 110 名选手，每人对应着 B' 中的 4 名选手。

对于 110 名选手中的每一位，为了能够对比他在 B' 中的期望分数和他的四个“分身”的实际分数，我们还需要估算前者的值。这里可以用该名选手在他所参加的 4 场现实比赛 B'_i 中的平均分，来近似地代替在 B' 中的期望分数。

综上所述，我们会按如下的流程来加工分数数据：

步骤 1. 对 12 场考试中的每一场，将其中每一名选手的分数除以该场考试的最高分[13]，并以此代替原始分数。

步骤 2. 对 12 场考试中的每一场，计算总排名前 30 的选手的分数标准差 σ（这里的分数是指步骤1中得到的商），然后将其中每个选手的分数 x 施以变换[14] $x \mapsto 1 - \frac{1-x}{4\sigma}$ ，并以此代替原始分数。

步骤 3. 对 110 名选手中的每一位，计算他在 4 场考试中的平均分，然后计算他在每场考试中的得分与这一平均分的差。

这样可以对每名选手计算出 4 个差值，共计 440 个值，每个值都表示一名选手在一场比赛中实际得分与期望得分的差距。这 440 个值即对应着随机变量 $\Delta_{B'}$ 的取值，它们将会是下一小节的分析对象。

3.3 拟合的方法和结果

观察上一小节中获得的 440 个数值的分布情况，发现：

- 整个分布大体上对称，且以 0 为对称中心。
- 数值的分布中间稠密、两边稀疏，所有数值的绝对值都小于 1。
- 分布的形状类似钟形曲线。

受此启发，尝试用正态分布曲线来拟合这些数值。具体方法如下：

步骤 1. 对于 $t = -1.0, -0.9, \cdots, 1.0$ ，计算：落在 $[t-0.05, t+0.05)$ 中的数值个数与总个数 440 的比值。这个比值记作 $c(t)$ 。

步骤 2. 在平面直角坐标系中画出 $t - c(t)$ 散点图。

步骤 3. 选取合适的参数 $\sigma > 0$ ，以使得函数

$$f(t) = \int_{t-0.05}^{t+0.05} P_{\sigma^2}(x)\mathrm{d}x$$

的图像与这些 $t - c(t)$ 数据点尽可能贴近（即残差平方和最小）。

[13]即信息学比赛中计算标准分的过程

[14]除以标准差这一步的作用也可简单理解为，消除题目区分度不同所带来的影响

这里 P_{σ^2} 表示期望值为 0、方差为 σ^2 的正态分布（用 $N(0,\sigma^2)$ 表示）的概率密度函数，满足

$$P_{\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

再记

$$R_{\sigma^2}(t) = \int_{-\infty}^{t} P_{\sigma^2}(x)\mathrm{d}x$$

为正态分布 $N(0,\sigma^2)$ 的累积分布函数，则有[15] $R_{\sigma^2}(t) = \left(1+\mathrm{erf}\left(\frac{x}{\sqrt{2}\sigma}\right)\right)/2$，其中 erf 表示误差函数。

最后注意到

$$\int_{t-0.05}^{t+0.05} P_{\sigma^2}(x)\mathrm{d}x = R_{\sigma^2}(t+0.05) - R_{\sigma^2}(t-0.05)$$

于是在进行拟合的过程中我们可以方便地计算这一定积分。

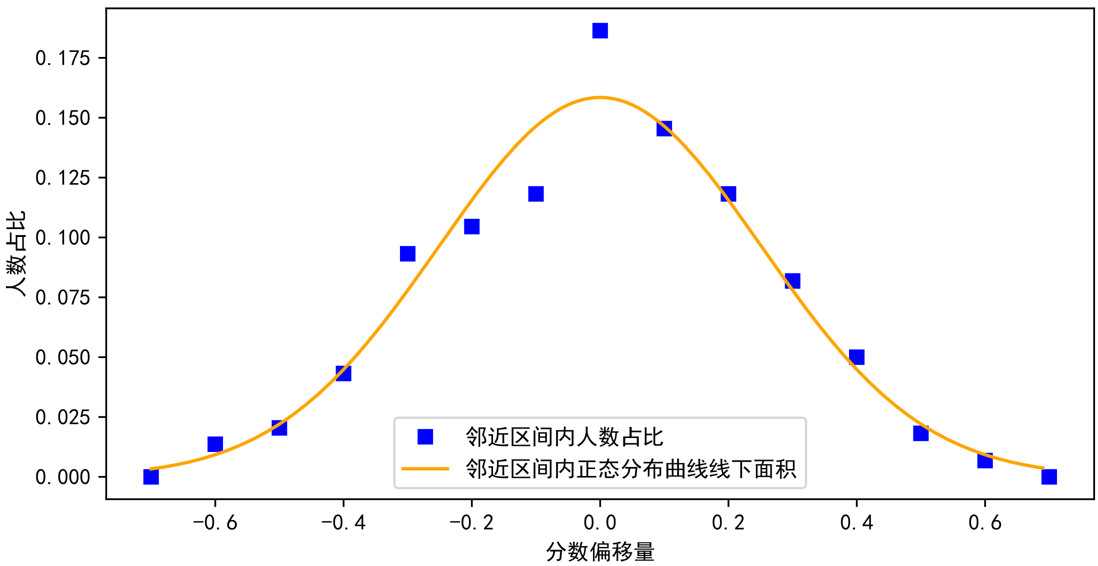

图 2: 散点图和拟合结果

图2展示了拟合的结果。可以看到，除了约 3 个数据点以外，其余数据点均与曲线贴合紧密。为了验证这些数据是否确实服从正态分布，还需绘制 Q-Q 图来进行检验。

图3展示了所绘制的 Q-Q 图。注意，该图的坐标轴经过缩放，故坐标轴上标注的数值仅能代表相对的比例关系。

[15] 误差函数 erf 没有闭合形式，这个式子可以视为 erf 函数的定义式

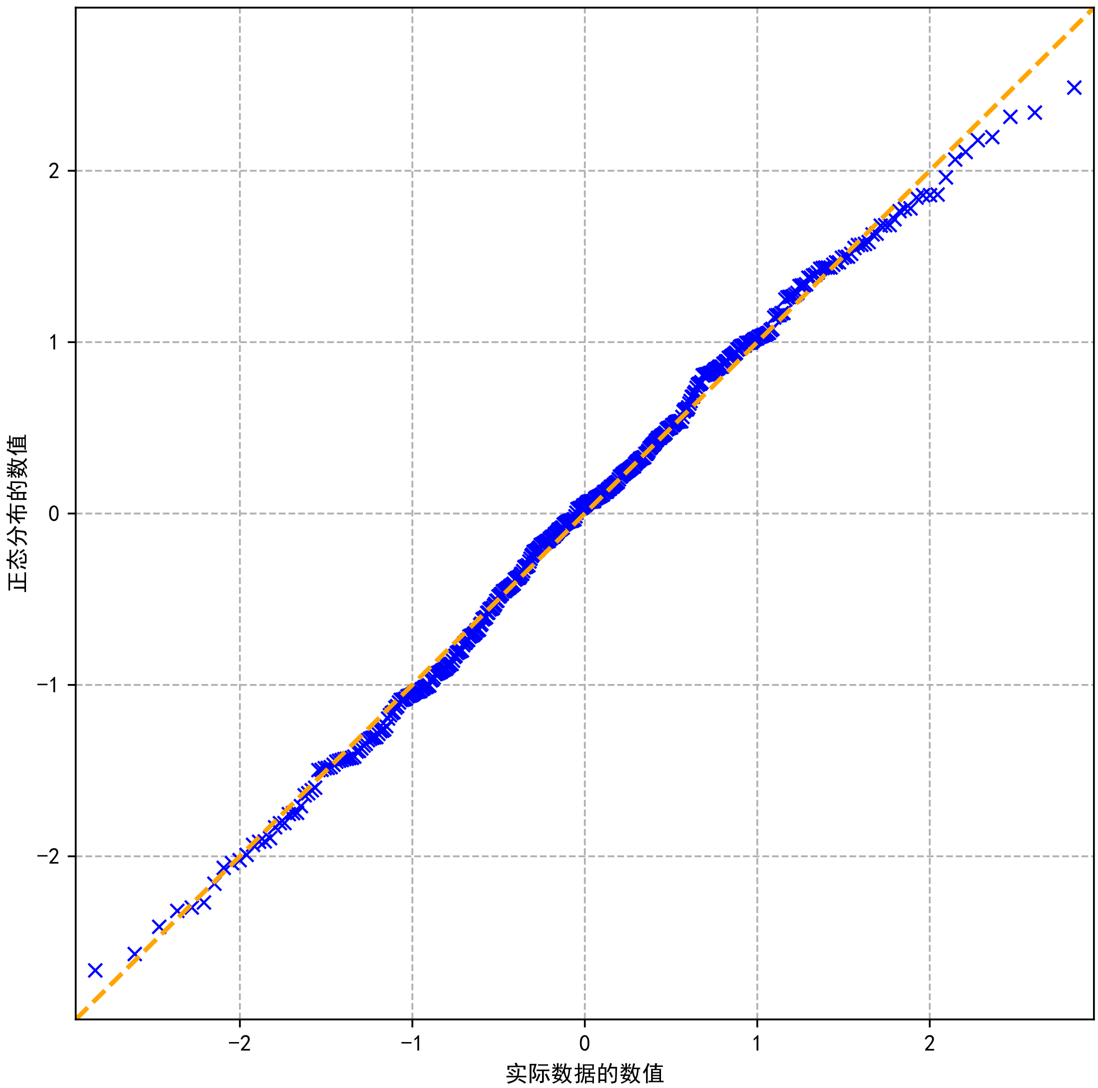

图 3: Q-Q 图，关于比赛分数差值数据和正态分布绘制

图3中有 440 个蓝色叉号，所有叉号的横坐标、纵坐标非严格递增。其中第 k 个叉号 $(1 \leqslant k \leqslant 440)$ 对应着 440 个数值中的第 k 小值 val_k ，叉号的横坐标 x_k 等于对应的数值 val_k ，而叉号的纵坐标 y_k 等于：440 个服从正态分布 $N(0,\sigma^2)$ 的数值中，第 k 小值的期望；其中的 σ 是待定的参数。可以证明 y_k 满足 $R_{\sigma^2}(y_k) = \frac{k}{440+1}$ ，于是由这一关系可以求出 y_k 。

如果这 440 个数值服从 $N(0,\sigma^2)$ 的话，容易想到应该有 $x_k \approx y_k$ ，也就是说所有叉号落在直线 $y = x$ 附近。我们通过选取合适的 $\sigma > 0$ 来让叉号尽可能贴近直线 $y = x$ ，最终的结果就是图3。可以看到，叉号与直线紧密贴合，所以这些数据确实服从正态分布[16]。又由假设2.13，这一规律对任何 COI/IOI 赛制的比赛都成立。

定理 3.2 (偏移量分布的形式). 对任何基于 *COI* 或 *IOI* 赛制的现实比赛，其对应的理想比赛的偏移量分布是期望值为 0 的正态分布。

4 选手整体水平的估计

本节将借助联赛初赛（字母标号 a）和联赛复赛（字母标号 b）的分数数据，来估计全国信息学竞赛选手整体水平的分布情况。

选手的“水平”是一个模糊的概念；为了将其量化，我们将用一名选手在联赛复赛中的期望分数来衡量这名选手的水平。

虽然由假设2.16，不同年份的联赛复赛（所对应的理想比赛）是缩放等价的；但它们毕竟不相同，因此“在联赛复赛中的期望分数”这一概念需要澄清。4.1小节将处理这一问题，并完成对复赛分数数据的初步分析。接着，4.2小节将从分数数据中，得到复赛在去除了初赛的筛选所带来的影响后，其（对应的理想比赛的）分数分布函数的表达式。最后，4.3小节将从分数分布计算出对应的期望值分布，这一分布即可体现全国选手整体水平的分布情况。

本节中会多次对现实情况作近似、作假设，于是也会不可避免地带来可观的误差。因此，本节的目标旨在估计而非精准计算，所得的结果仅能反映趋势而不保证精确。

4.1 复赛分数数据的获取、加工和拟合

数据来自以下 4 场比赛：

- NOIP2016 — 复赛（字母标号 b）
- NOIP2017 — 复赛（字母标号 b）
- NOIP2018 — 复赛（字母标号 b）
- CSP2019 — 复赛（字母标号 b）

之所以只采用 2016 年到 2019 年的比赛，是出于三个原因：

[16]注意到，缩放坐标轴和改变 σ 的取值，这两种操作对图像的改变其实是完全相同的，所以缩放坐标轴不会影响结论的可靠性

• 年代过于久远的比赛对当今的参考意义有限。

• 仅有的数据来源为NOI官网上的获奖名单公示，故只能取得获奖选手的分数信息。而自 2016 年起 CCF 更改了获奖规则，增加了获奖人数，使得可以获取的数据量大了许多。

• 2020 年的比赛规则有所更改（两天六题改为一天四题），所以假设2.16不再能保证 2020 年复赛与其他年份复赛缩放等价。

	参赛人数	获奖人数	满分	最高分	获奖分数线
NOIP2016 复赛	约 8300	约 5900	600	600	100
NOIP2017 复赛	约 10300	约 6600	600	600	80
NOIP2018 复赛	约 12900	约 8000	600	600	120
CSP2019 复赛	约 13900	约 8800	600	600	80
总计	约 45400	约 29300			

表 3: 4 场比赛的相关数据

表3展示了关于这 4 场比赛的几项统计数据。

由假设2.16，这 4 场现实比赛（所对应的理想比赛）是缩放等价的。进而由命题3.1，这 4 场现实比赛所对应的理想比赛，在对分数作变换（变换方式见3.2小节）后，将成为完全相同的理想比赛。

于是，与3.2小节类似，数据加工将按以下步骤进行：

步骤 1. 将所有比赛中所有选手的分数除以当场比赛的最高分（也就是计算标准分；注意到最高分等于满分），用以代替原始分数。

步骤 2. 去除所有 < 0.2 的分数。这是因为在这 4 场比赛中，获奖分数线与最高分的商的最大值为 0.2；这意味着分数低于 0.2 的选手中有一部分未能获奖，于是这些选手中其余部分的数据也失去意义，因此一并剔除。

步骤 3. 对每场比赛计算分数标准差 σ，然后对分数作变换 $R : x \mapsto 1 - \frac{1-x}{5.52\sigma}$，并用变换结果代替原分数。使用系数 5.52 的理由稍后说明。

步骤 4. 对每场比赛计算最低分，取所得的 4 个最低分的最大值 T，并剔除所有 $< T$ 的分数。这一步的理由与步骤2中的类似：分数低于 T 的部分选手未能获奖，故将这些选手连同已获奖的那些一并剔除。计算可得 $T \approx 0.200$，与步骤2中的阈值保持一致；这正是系数 5.52 的主要作用。

步骤 5. 现在所有这些分数数据属于同一理想比赛 A（满足 A 与原先 4 个现实比赛所对应的理想比赛缩放等价），将它们汇集起来即可。注意到我们所取得的并非完整的分数数据，而只是 $\geqslant 0.2$ 的那一部分分数。

本节中我们约定使用理想比赛 A 作为衡量选手水平的标尺，也就是说我们将用一名选手在 A 中的期望分数，来代表该名选手的水平。后文中如果作为一个现实比赛提到“联赛复赛”，则默认指 A 对应的现实比赛。

经过上述加工后，我们得到了 22093 个落在 $[0.2, 1]$ 之中的分数数据。由命题2.5，这些数据应当服从 C_A 所描述的概率分布。

接下来我们将确定函数 $C_A : (0, 1) \to \mathbb{R}_{\geqslant 0}$ ，满足在任何一个区间 (a, b) 上，C_A 的定积分在数值上约等于：分数落在 (a, b) 中的选手人数，与总人数 45400 的比值。

在对十余种常见函数和常见概率分布进行拟合之后，我们发现对数函数 $f(s) = -\log(s)$ 满足前述要求，且与已获得数据的贴合程度大幅好于所尝试的其他函数。此外，不难验证对数函数 $f(s) = -\log(s)$ 在 $(0, 1)$ 上非负且定积分等于 1 ，因此是一个合法的分数分布函数。以下将在实际数据和据对数函数计算出的数值之间进行比对，并展示结果。

步骤 1. 对于 $t = 0.25, 0.35, \cdots, 0.95$ ，计算：落在 $[t - 0.05, t + 0.05]$ 中的分数个数与总人数 45400 的比值。这个比值记为 $c(t)$ 。

步骤 2. 关于 < 0.2 的分数段，我们不了解其中具体的分数分布，只知道这一部分共有 $45400 - 22093 = 23307$ 人，占比 $\frac{23307}{45400} \approx 0.513$ 。为了与步骤4.1中的数据统一，我们将数值 0.513 乘以

$$\frac{\int_{0.05}^{0.15} -\log(s)\mathrm{d}s}{\int_{0}^{0.2} -\log(s)\mathrm{d}s}$$

以将其换算为分数段 $(0.05, 0.15)$ 上的人数占比，并记作 $c(0.1)$ 。

步骤 3. 在平面直角坐标系中画出 $t - c(t)$ 散点图。

步骤 4. 检查函数

$$f(t) = \int_{t-0.05}^{t+0.05} -\log(x)\mathrm{d}x$$

的图像是否与 $t - c(t)$ 散点图吻合。

图4展示了比对的结果。可以看到全部数据点与曲线贴合紧密，且算得残差平方和约为 $9.2 \cdot 10^{-5}$，显示出了较好的拟合效果。基于此，我们确定取 $C_A(s) = -\log(s)$ 。

命题 4.1. $C_A(s) = -\log(s)$ ，其中 $s \in (0, 1]$ 且 A 是根据*4.1*小节中描述的过程所确定的理想比赛。

4.2 结合初赛的分析

这一小节将对于复赛所对应的理想比赛 A ，在去除了初赛的筛选性所带来的影响后，计算所得的新的理想比赛（记为 A' ）的分数分布函数。其中，4.2.1小节将给出初赛分数数据的来源，4.2.2小节将结合这些数据给出关于初赛的几个假设；第4节的其余部分都将依赖于这些假设。4.2.3小节将计算初复赛（所对应理想比赛）的偏移量分布的标准差，以为4.2.4小节中 $C_{A'}$ 的计算做好准备。

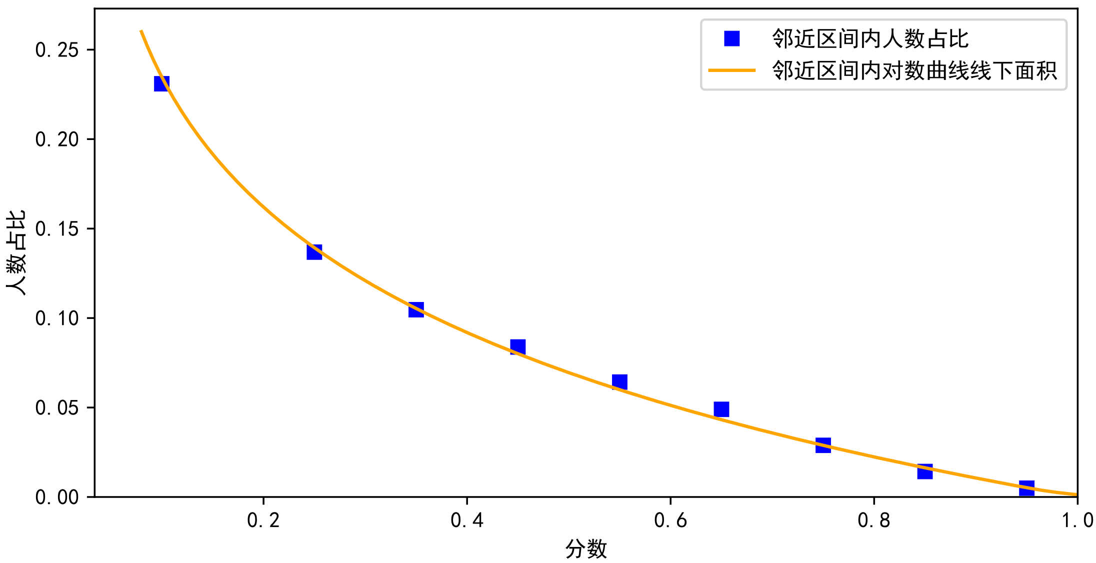

图 4: 散点图和拟合结果

4.2.1 初赛分数数据的获取

分数数据采用 NOIP2018 初赛北京赛区的成绩。该场比赛共 781 人获得非零分数（零分视为缺考），其中 536 人晋级复赛并获得非零分数；该场比赛满分 100 分，最高分 96 分，晋级分数线约为 35 分；全国最高分为 100 分。有关初赛的全部数据获取自官方网站上的成绩公示。

采用该场比赛的原因：后续分析需要分数表上包含选手姓名；而笔者所能找到的其他年份、其他省市的成绩公示，均未包含这一信息。

知道了选手姓名，我们就可以查询该名选手在 NOIP2018 复赛中的得分。通过这种方式，我们获得了 536 名晋级者的初赛和复赛分数。由于官方网站上的成绩公示仅包括获奖选手，这里所使用的北京选手复赛分数是按民间数据测试出的成绩。

除此以外，在4.2.4小节中，还将使用 CSP2019 初赛的全国分数数据，这些数据从官方网站上各省市发布的成绩公示汇总得到。CSP2019 初赛报名人数 48812 人，由于个别省份仅公示了晋级选手或未缺考选手的分数，最终收集到 47264 人的数据。

全国分数数据的来源之所以采用 CSP2019，是因为自 2019 年起才有完整的初赛分数公示。

4.2.2 关于初赛的几个假设

不同于复赛，信息学联赛的初赛是分省考试、分省排名的，这会给本文的分析带来很大困难。为了规避这一问题，我们作如下假设：

假设 4.2. 每一年的联赛初赛为全国统一考试、统一排名，全国范围内分数最高的若干名选手晋级复赛。

作出这一假设，意味着忽略不同省市间选手水平和竞争激烈程度上的差异，并用全国整体的选手水平和竞争激烈程度来代替之。即使如此，我们一般也并不能直接用一个地区的数据来“代表”全国的数据，而是只有在所研究的量与地域没有明显关联时（例如4.2.3小节中研究同一名选手的初赛得分与复赛得分间的关系）才能这样做。

在给出下一个假设前，先对 2018 北京初赛的分数做一点分析。

步骤 1. 将 536 名晋级选手按初赛分数分组：分数在 $[30, 40)$ 中的、在 $[40, 50)$ 中的、……、在 $[90, 100)$ 中的，分别分为一组，共计 7 组。

步骤 2. 对每一组计算初赛平均分和复赛平均分。

步骤 3. 对每一组，以复赛平均分为横坐标、初赛平均分为纵坐标，将数据点画在二维平面上，并将这些数据点连成折线图。

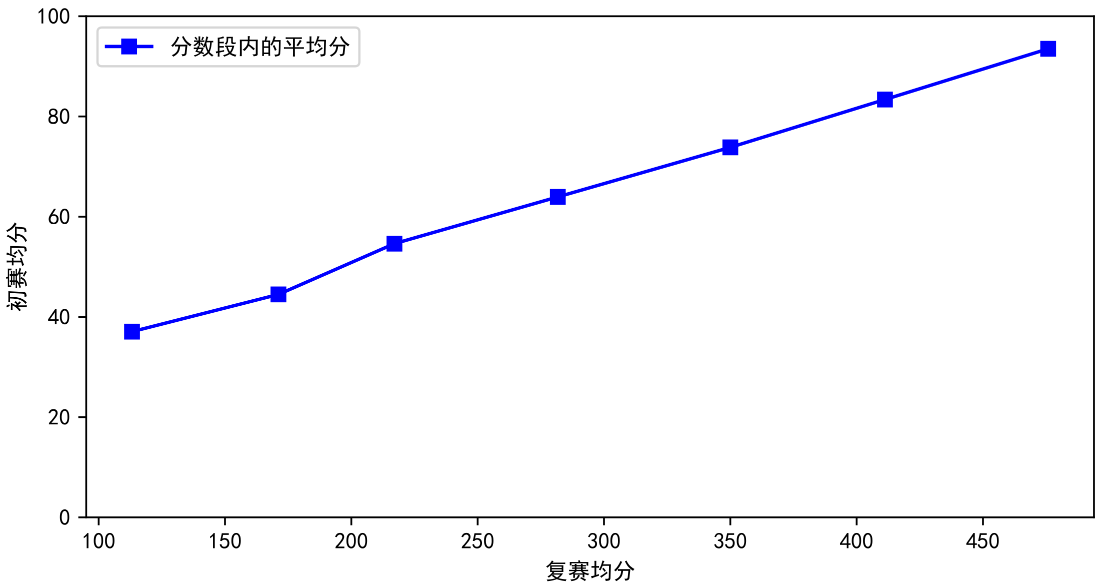

图 5: 平均分折线图

所得的折线图如图5所示。可以看到，这些数据点近似地连成一条直线；这提示我们，初赛分数与复赛分数之间存在一个线性的对应关系。

基于这一观察，我们作出如下假设：

假设 4.3. 记现实比赛 B_1 为联赛初赛，**取参赛人群为实际晋级复赛的全体选手**；记现实比赛 B_2 为联赛复赛；则 B_1, B_2 对应的理想比赛缩放等价。

注意：“取参赛人群为实际晋级复赛的全体选手”这一规定，只限制了参赛人群，而并未要求这些选手在理想比赛中的分数也一定得达到晋级的标准。也就是说，虽然我们只取

那些在现实中达到了晋级分数线的选手，但在构造对应的理想比赛时，我们忽略现实中发生了什么，仍然只考察每名选手分数波动的概率分布和他的期望分数。

关于这一假设需要作几点说明：

1. 2.3小节中，我们在为假设2.16予以辩护时，断言了“一名选手的水平是固定的，不会随比赛的改变而改变”。但是，由于考察内容的不同，一名选手在 B_1 和 B_2 中的能力差异可能较大，故上述断言在将初赛（即 B_1 ）加入考虑范围后似乎不再成立。

2. 为了使前述断言仍然成立，在4.2小节内，我们需暂时改变命题2.3中“期望得分”这一概念的所指，将其改为：一名选手在 B_1, B_2 中（在按最高分和标准差折算后）期望分数的平均值。这会改变从现实比赛构造理想比赛的方式，并使得 B_1, B_2 所对应理想比赛的期望值分布和偏移量分布发生变化，变化后 B_1, B_2 所对应的理想比赛分别记作 F_1, F_2 （所以 A 和 F_2 的区别，就是概念更改前和更改后的区别）。显然，此时 F_1, F_2 的期望值分布在经过缩放后是相同的。

3. 这样更改后，Δ_{B_2} 的值也发生了变化。原先 Δ_{B_2} 的取值等于选手实际表现与真实能力（即期望表现）的差；现在它的值还要在此基础上加上选手在复赛单项上的能力与初、复赛综合能力的差。但是，只要“单项能力减综合能力”这一随机变量服从正态分布，新的 Δ_{B_2} 也一定服从正态分布——因为服从正态分布的独立随机变量之和依然服从正态分布。另一方面，假如在原先定义下的随机变量 Δ_{B_1} 服从正态分布，则对新的 Δ_{B_1} 可做与刚才类似的论证。

4. 关于 F_1, F_2 间的缩放等价性，第2条对期望值分布的缩放等价予以了说明，第3条对偏移量分布的缩放等价予以了说明；这些说明都有一些感性的成分，它们仅用作对假设4.3含义的澄清，而并非尝试对其予以证明。需要注意，由于我们对概念的修改，关于 F_1, F_2 的期望值分布、偏移量分布所作的一切讨论，在4.2小节之外均没有意义。但是 F_1, F_2 的分数分布不会受这一修改的影响，故4.2小节计算出的分数分布函数会在后续分析中直接使用。

4.2.3 计算偏移量分布的参数

先前已经说明，理想比赛 F_2 的偏移量分布为正态分布 $N(0, \sigma^2)$ 。这一小节将基于4.2.1小节中获得的数据，来测量该分布的标准差 σ 。

引理 4.4. 独立随机变量 X_1, X_2 分别服从分布 $N(0, \sigma_1^2), N(0, \sigma_2^2)$，则 $X_1 + X_2$ 服从分布 $N(0, \sigma_1^2 + \sigma_2^2)$ 。

证明见维基百科相应条目 [1]，这里不再重复。

由假设4.3，在对 F_1 作线性的缩放变换 T 之后，可以使其与 F_2 相同；此时两者的偏移量分布均为 $N(0, \sigma^2)$ 。又注意到对现实比赛 B_1, B_2 ，其对应的随机变量 $\Delta_{B_1}, \Delta_{B_2}$ 应当是独立的（这里认为在定义2.3的步骤3中 B_1, B_2 共用同一个表示选手的随机变量 p ），所以由引

理4.4，$\Delta_{T[B_1]} + \Delta_{B_2}$ 服从正态分布 $N(0, 2\sigma^2)$ 。因此，选手在 $T[F_1]$ 和 F_2 中的分数之差，这一随机变量服从标准差为 $\sqrt{2}\sigma$ 的正态分布；只要测出它的标准差，即可得到 σ 的取值。

容易想到以下测量方式：

步骤 1. 对 F_1 的分数作线性变换，使得变换后它的期望值、偏移量分布与 F_2 相同。

步骤 2. 对先前提到的 536 名选手，计算每名选手在 F_2 中的分数和在变换后的 F_1 中的分数之差。

步骤 3. 这 536 个差值应该服从正态分布，那么计算这些值的标准差即可。

但一个问题是，这 536 个差值并非真正服从正态分布。如果一名选手考出了大幅低于自己期望分数的分数，那么他进入这 536 人之列的机会就会大大降低；换句话说，这 536 个数据的取样方法是有选择性的，而且选择的方式倾向于实际分数高于期望分数的选手，因此这些数据不能代表整体的分布。

假如我们召集那些没有晋级的选手，让他们也参加复赛考试并记录他们的分数，再把这些分数和原有的 536 个数据汇总，就能获得完整、有代表性的数据。但实际上，我们也可以“假装”已经获得了未晋级选手的数据，并对全体数据进行分析；如果这个过程中“碰巧”没有用到任何一个未晋级选手的数据，我们事实上就在只凭借已有的 536 个数据的情况下完成了测量。以下给出一个这样的测量方式。

步骤 1. 对 F_1 的分数作线性变换，使得变换后它的期望值、偏移量分布与 F_2 相同。

步骤 2. 对先前提到的 536 名选手，计算每名选手在变换后的 F_1 中的分数和在 F_2 中的分数之差。（前者减后者）

步骤 3. 取出这些差值中前 35 大的值，则这些值可以视为：某一组服从正态分布的 781 个数（781 即初赛参赛人数），其中前 35 大的值。通过测量这些值可以得到正态分布的标准差。

在第3步中，之所以说这 35 个值为 781 个数中的最大值，是因为：

• 计算发现第 35 大的差值约等于 0.26，高于初赛晋级分数线经过变换后的值。又因为复赛分数不可能小于 0，所以任何一个未达到晋级分数线的选手，其两试分数差值不可能达到 0.26。（计算发现 35 人中最低的初赛分数为 53 分，比晋级分数线高出近 20 分）

• 因此，除了 536 名晋级选手之外，其余选手不可能进入 35 人之列，故只考虑已有的 536 个数据是充分的。

在前述过程的步骤1中需要作分数变换，以下给出具体步骤。

步骤 1. 对于复赛分数 $s \in [120, 600]$ （120 为官方分数公示所覆盖的最低分数），计算该分数在全国范围内的排名 c ，并将 s 映射到 $t \in [0, 1]$ ，满足

$$\frac{\int_t^1 -\log(x)\mathrm{d}x}{\int_{0.186}^1 -\log(x)\mathrm{d}x} = \frac{c}{N}$$

其中 $N = 8044$ 为复赛不低于 120 分的选手总数，0.186 为最低分数 120 依4.1小节中的变换

R 映射到的值。

步骤 2. 对于低于 120 分的复赛分数，我们将 $[0,120)$ 均匀地映射到 $[0,0.186)$ 上去。以上两个步骤所描述的映射方式，保证了分数分布呈对数曲线，与命题4.1一致。

步骤 3. 计算北京初赛排名前 25% 选手（共 195 名）的分数标准差 σ_1 ，再对映射后的复赛分数计算北京选手前 195 名的分数标准差 σ_2 。然后对于初赛分数 $s \in [0,100]$ ，将其映射到 $1-(1-\frac{s}{100})\cdot\sigma_2/\frac{\sigma_1}{100}$ 。由命题3.1，这一映射方式保证了映射后两个理想比赛相同。这一步中只取前 25% 的理由：对于初赛期望分数离晋级分数线较近的选手，这些选手中有相当一部分未能进入复赛，故复赛在相应分数段的分布会比真实情况稀疏；只有把考察范围限制在分数足够高的选手，才能避免这一问题。

设得到的 35 个差值按降序排列为 $d_1,\cdots,d_{35}$ ，考虑如何由此推断全体差值的标准差 $\sqrt{2}\sigma$ 。

这里采用最大似然估计，即选取一个 σ 以最大化：在全体差值服从 $N(0,2\sigma^2)$ 的条件下，测量得到 $d_1,\cdots,d_{35}$ 的概率。注意到这个概率实际上必定等于 0，但可以通过取极限规避这一问题。下式给出 σ 的计算方式，其中 $u_1,\cdots,u_{781}$ 表示随意排列的 781 个差值，v_k 表示 $u_1,\cdots,u_{781}$ 中的第 k 大值。

$$
\begin{aligned}
&\lim_{\epsilon\to 0}\operatorname*{arg\,max}_{\sigma\in\mathbb{R}_{>0}} \Pr\left[v_k\in[d_k-\epsilon,d_k+\epsilon],\forall 1\leqslant k\leqslant 35\right]\\
=&\lim_{\epsilon\to 0}\operatorname*{arg\,max}_{\sigma\in\mathbb{R}_{>0}} \binom{781}{35}\cdot 35!\cdot\prod_{k=1}^{35}\left(R_{2\sigma^2}(d_k+\epsilon)-R_{2\sigma^2}(d_k-\epsilon)\right)\cdot R_{2\sigma^2}(d_{35})^{746}\\
=&\lim_{\epsilon\to 0}\operatorname*{arg\,max}_{\sigma\in\mathbb{R}_{>0}} (2\epsilon)^{-35}\cdot\prod_{k=1}^{35}\left(R_{2\sigma^2}(d_k+\epsilon)-R_{2\sigma^2}(d_k-\epsilon)\right)\cdot R_{2\sigma^2}(d_{35})^{746}\\
=&\lim_{\epsilon\to 0}\operatorname*{arg\,max}_{\sigma\in\mathbb{R}_{>0}} \prod_{k=1}^{35}\frac{R_{2\sigma^2}(d_k+\epsilon)-R_{2\sigma^2}(d_k-\epsilon)}{2\epsilon}\cdot R_{2\sigma^2}(d_{35})^{746}\\
=&\operatorname*{arg\,max}_{\sigma\in\mathbb{R}_{>0}}\left(\prod_{k=1}^{35}R'_{2\sigma^2}(d_k)\right)\cdot R_{2\sigma^2}(d_{35})^{746}\\
=&\operatorname*{arg\,max}_{\sigma\in\mathbb{R}_{>0}} 746\cdot\log\left(R_{2\sigma^2}(d_{35})\right)+\sum_{k=1}^{35}\log\left(P_{2\sigma^2}(d_k)\right)
\end{aligned}
$$

至此，问题完全转化为一个最优化问题。最优化方法采用 SciPy 提供的 BFGS 算法的实现 [2]，算得 $\sigma\approx 0.109$。

图6中的橙色曲线展示了差值的概率分布，35 条蓝色竖线表示实际测得最大的 35 个差值在其中的位置。

命题 4.5.

$$\mathcal{D}_{T[F_1]}(\delta)=\mathcal{D}_{F_2}(\delta)=P_{\sigma_F^2}(\delta),\quad\forall\delta\in\mathbb{R}$$

其中“缩放变换”T 满足 $T[F_1]=F_2$，常数 σ_F 约等于 0.109 。

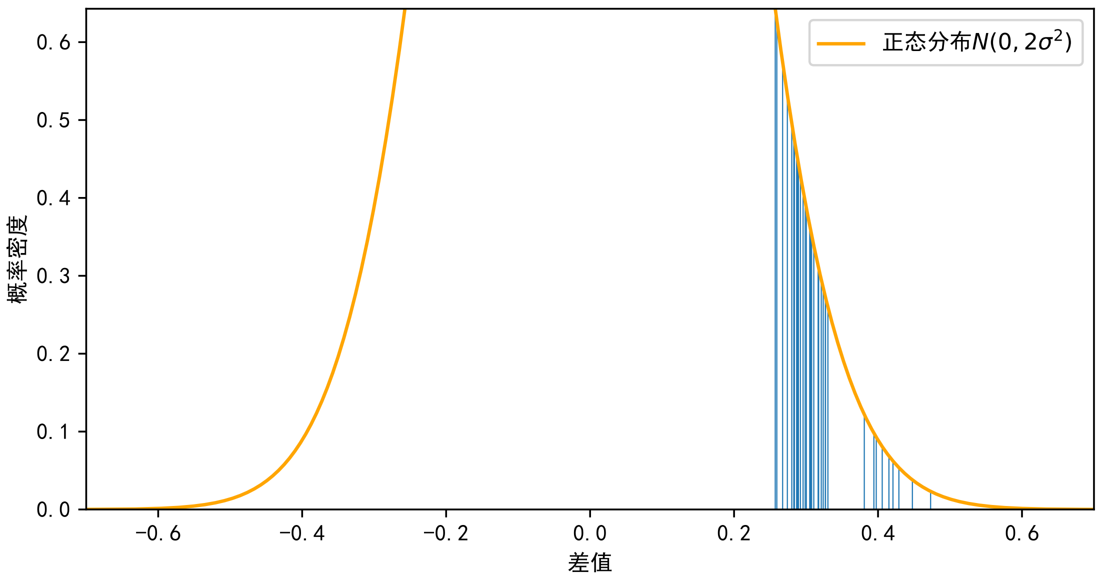

图 6: 差值的分布情况

4.2.4 消除初赛对分数分布的影响

这一小节将计算初赛分数线映射为复赛分数后的值，并借助该值计算出：消除初赛的筛选性带来的影响后（即假想初赛并未淘汰一人，所有选手都晋级复赛），复赛的分数分布函数。

在4.2.1小节中，已经得到了 CSP2019 初赛参赛选手的分数数据。由表3中的数据知，在 2016 到 2019 四年中，平均每年的初赛晋级人数约为 11350；因此我们选择 CSP2019 初赛全国第 11350 名的分数，作为假设4.2中的“全国统一晋级分数线”。

这里之所以对晋级人数而不是晋级率取平均数，是因为初赛的参赛选手总数受收费、政策等无关因素影响过大，而晋级复赛的人数与复赛获奖的人数呈固定比例，因而相对可靠。

最终算得分数线为 63 分。作为参照，CSP2019 初赛中，浙江、山东、江苏实际的分数线[17]分别为 72.5,60,53。

下面将这一分数线映射为复赛分数。为了与4.2.3小节保持一致，这里仍然使用同样的计算方式，并同样采用北京的数据。

我们将计算 CSP2019 初赛中，北京排名前 195 名的分数标准差 σ_1，再对（按4.2.3小节中的方式）映射后的 NOIP2018 复赛分数计算北京选手前 195 名的分数标准差 σ_2。然后对于 CSP2019 初赛分数 $s \in [0, 100]$，将其映射到 $1 - (1 - \frac{s}{100}) \cdot \sigma_2 / \frac{\sigma_1}{100}$。这一计算过程基于如下假设：2019 年北京选手整体水平，与 2018 年大体相同。

[17]这里给出的是全省分数线，省内各市的分数线可能高于全省分数线

对分数线 63 施加上述变换，得到其对应的复赛分数为 $h \approx 0.1035$。另一方面，由命题4.5可知，任何一名选手的初赛、复赛的（变换后）实际分数之差服从概率分布 $N(0, 2\sigma_F^2)$。从而，如果假想所有初赛选手都参加了复赛，则对于复赛实际分数为 t 的选手 p，其初赛分数达到 63 的概率为 $1 - R_{2\sigma_F^2}(h-t)$。

需要注意，这样得到的概率，是在获得具体的期望值分布前的先验概率。假如已知全体选手的期望分数分布情况，我们可以用贝叶斯公式得到前述选手 p 的期望分数取每一个值的概率，进而得到 p 的初赛分数取每一个值的概率，也就是后验概率。简便起见这里采用先验概率，即使它相比后验概率略失精确。

记 F_2' 为现实比赛 B_2 在消除初赛的筛选性带来的影响后所对应的理想比赛，则由以上讨论可得：

$$C_{F_2'}(s) \propto \frac{C_{F_2}(s)}{1 - R_{2\sigma_F^2}(h-s)} = \frac{-\log(s)}{1 - R_{2\sigma_F^2}(h-s)}, \quad \forall s \in (0,1]$$

上式中之所以使用“正比于”而不是“等于”，是因为分数分布函数表达的是分布“密度”，而不是样本“数量”。计算出对应的比例系数后得到：

$$C_{F_2'}(s) = \gamma \frac{-\log(s)}{1 - R_{2\sigma_F^2}(h-s)}, \quad \forall s \in (0,1]$$

其中常数 $\gamma \approx 0.549$，它使得 $C_{F_2'}$ 在 $[0,1]$ 上的定积分等于 1。

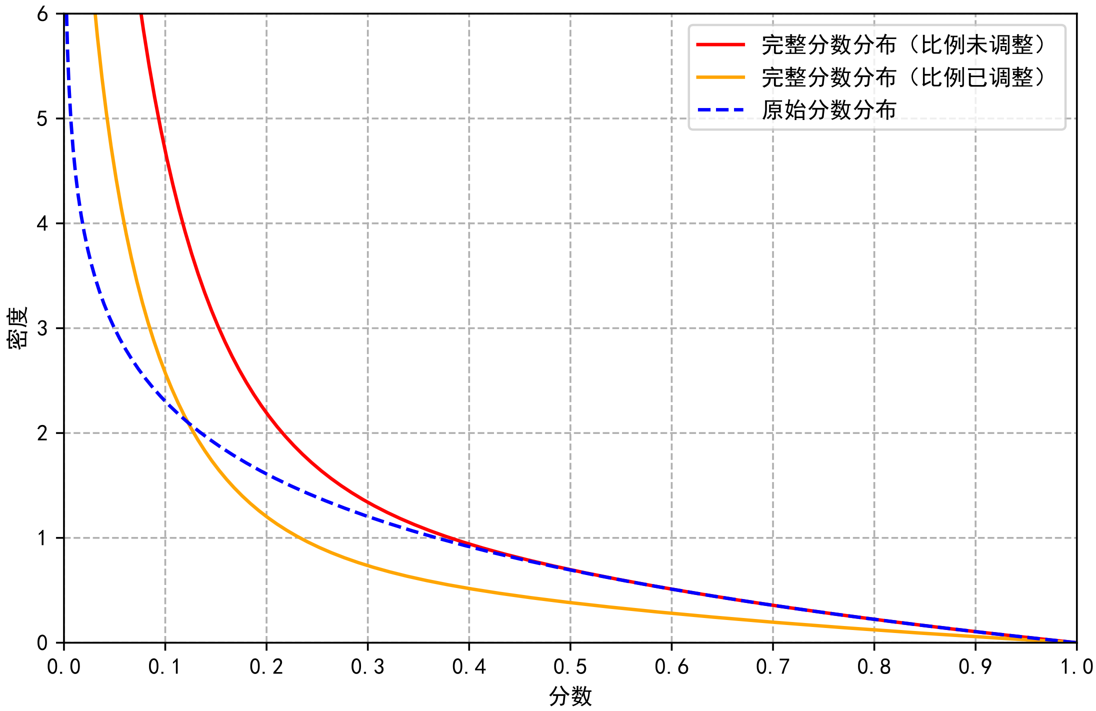

图 7: 消除初赛影响前后的分数分布函数

图7分别展示了以下三个函数的图像：

蓝色 $f(s) = -\log(s)$，即 $C_{F_2}(s)$ 或 $C_A(s)$。

红色 $f(s) = \frac{-\log(s)}{1-R_{2\sigma_F^2}(h-s)}$

橙色 $f(s) = \gamma\frac{-\log(s)}{1-R_{2\sigma_F^2}(h-s)}$，即 $C_{F_2'}(s)$。

由4.2.2小节中的讨论知，C_A 与 C_{F_2} 相同。同理，如果定义 A' 为：复赛在去除初赛影响后对应的理想比赛（这里采用**按原本的方式**解读的定义2.3），则 $C_{A'}$ 亦与 $C_{F_2'}$ 相同。因此有以下命题：

命题 4.6.

$$C_{A'}(s) = \gamma\frac{-\log(s)}{1-R_{2\sigma_F^2}(h-s)}, \quad \forall s \in (0,1]$$

其中常数 $\gamma \approx 0.549$ ，且 A' 为复赛在去除初赛影响后对应的理想比赛。

4.3 从分数分布还原期望值分布

在4.2.4小节中得到了 $C_{A'}(s)$ 的表达式；这一小节将由此计算 $\mathcal{X}_{A'}(x)$ 。

根据定理3.2，存在 $\sigma > 0$ 使得

$$\mathcal{D}_{A'}(\delta) = P_{\sigma^2}(\delta), \quad \forall \delta \in \mathbb{R}$$

进而由命题2.10得

$$C_{A'}(s) = \int_0^1 \mathcal{X}_{A'}(x) P_{\sigma^2}(s-x)\mathrm{d}x, \quad \forall s \in \mathbb{R} \tag{5}$$

于是，我们可以根据(5)，由 $C_{A'}$ 逆推出 $\mathcal{X}_{A'}$ 。

但是在这个问题中，期望分数接近 0 的选手占了总数中相当的比重，因此若单纯按(5)计算，则“负分数”问题（在2.3小节末尾有所提及）会变得十分显著。为解决这一问题，我们对于每一名选手，将其分数的概率分布中负的一段截掉，再将剩余部分的概率密度函数乘以某个系数，以使得乘完后剩余部分的总概率为 1。若一名选手期望分数为 x ，则容易证明：对该名选手而言，在上述过程中使用的系数 $c(x)$ 等于

$$\left(\int_0^1 P_{\sigma^2}(t-x)\mathrm{d}t\right)^{-1} = \left(R_{\sigma^2}(1-x) - R_{\sigma^2}(0-x)\right)^{-1}$$

于是 $C_{A'}$ 与 $\mathcal{X}_{A'}$ 间的关系被更新为

$$C_{A'}(s) = \int_0^1 \mathcal{X}_{A'}(x) \cdot \frac{P_{\sigma^2}(s-x)}{R_{\sigma^2}(1-x) - R_{\sigma^2}(0-x)}\mathrm{d}x, \quad \forall s \in \mathbb{R} \tag{6}$$

在进行逆推之前，先测量 σ 的值。我们获取了 CSP2019 复赛全体选手的民间分数（零分选手被去除，共计 12108 人获得非零分数），并按以下步骤进行测量：

步骤 1. 将每名选手每一天的分数除以当天最高分（两天最高分均为满分 300 分），再将每一天的所有分数做变换，以使得两天的分数分布分别呈对数曲线状。具体变换方式与4.2.3小节中相同；经过变换后，两天分别对应的理想比赛应当与 A' 相同。

步骤 2. 对每名选手计算两天分数之差，计算所有这些差值的标准差 σ_0 。

与4.2.3小节中类似，同一名选手的单日分数（变换后的分数），应该服从标准差为 $\sigma_1 = \frac{\sigma_0}{\sqrt{2}}$ 的正态分布。记 CSP2019 复赛第一天、第二天，这两个现实比赛分别为 D_1, D_2 ，则 $\Delta_{D_1}, \Delta_{D_2}$ 服从正态分布 $N(0, \sigma_1^2) = N(0, \frac{\sigma_0^2}{2})$ 。记现实比赛 D 为 CSP2019 复赛（两天综合），则有 $\Delta_D = \frac{\Delta_{D_1} + \Delta_{D_2}}{2}$ 。进而由引理4.4：

$$\begin{aligned}
\mathrm{Stddev}\left[\Delta_D\right] &= \frac{\sqrt{\mathrm{Stddev}\left[\Delta_{D_1}\right]^2 + \mathrm{Stddev}\left[\Delta_{D_2}\right]^2}}{2} \\
&= \frac{\sqrt{2}\sigma_1}{2} \\
&= \frac{\sigma_0}{2}
\end{aligned}$$

得到 $\sigma = \frac{\sigma_0}{2}$ 。换句话说：同一名选手在 CSP2019 复赛中（变换后）的分数波动，服从标准差为 $\sigma = \frac{\sigma_0}{2}$ 的正态分布。

最终算得 $\sigma \approx 0.078$ 。

从 $\mathcal{C}_{A'}$ 和 $\mathcal{D}_{A'}$ 逆推出 $\mathcal{X}_{A'}$ 难以精确地实现，因此这里只能近似地计算 $\mathcal{X}_{A'}$ 在许多个离散的点处的点值。

我们将区间 $(0, 1]$ 作 500 等分，并设立 500 个未知数 $x_{1\cdots500}$ ，分别表示在 500 个分点处 $\mathcal{X}_{A'}$ 的取值。另一方面，我们在(6)中将 s 取遍每一个分点，由此得到 500 个等式限制；注意到仅凭 $x_{1\cdots500}$ 无法表示出(6)中的定积分，因此定积分被换成离散的求和。在作了这样的"离散化"之后，原先的等式显然不再成立，因此改为最小化所有每一个等式两端之差的平方和。为了避免无意义的结果，我们额外加入了关于序列 $x_{1\cdots500}$ 非负性和"光滑性"的限制；后者通过序列 $x_{1\cdots500}$ 的高阶差分来表示。

上述问题最终归结到了一个二次规划模型的求解；可以证明其为凸二次规划，因此任何一个极值点都是最值点。最优化方法采用 SciPy 提供的信赖域算法的实现 [2]。用于计算的程序和最终算得的点值 $x_{1\cdots500}$ ，可以在本文开头的链接中找到。

观察所得的 500 个点值，发现：

1. 在与 0 紧邻的位置处，点值明显大于其他位置。

2. 在其余位置处，点值构成一条平滑的曲线。计算发现这些点值近似地符合二次函数 $f(x) = ax^2 + bx + c$ ，其中 $a \approx 1.697, b \approx -3.352, c \approx 1.655$ 。图8展示了该函数的图像。

至此，我们得到了函数 $\mathcal{X}_{A'}$ 的表达式。

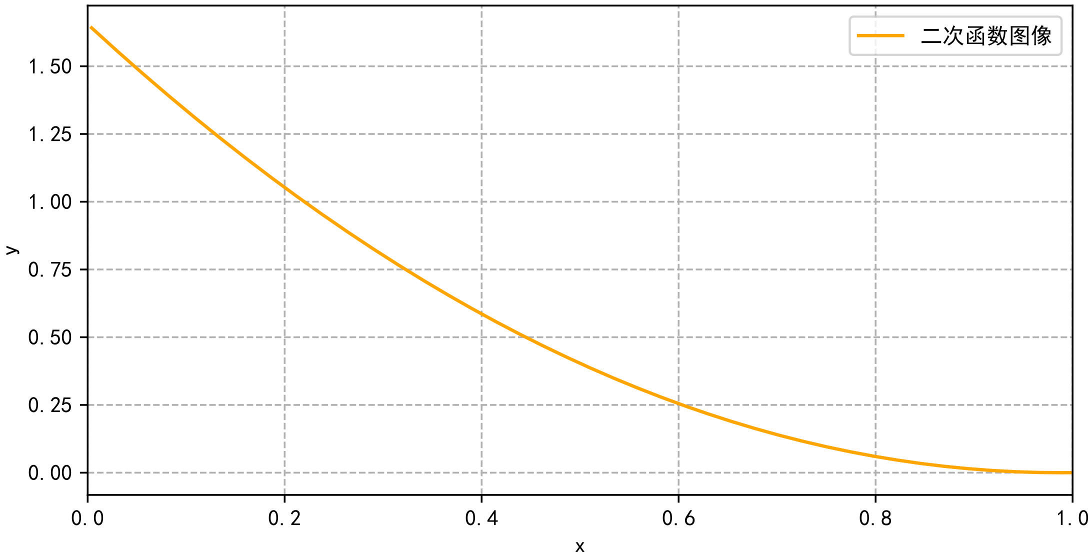

图 8: 二次函数 $f(x) = ax^2 + bx + c$

定理 4.7. 对于 $x \in (\epsilon, 1]$ 有 $\mathcal{X}_{A'}(x) = ax^2 + bx + c$，其中 $a \approx 1.697, b \approx -3.352, c \approx 1.655$，$\epsilon$ 为某个小常数，A' 为复赛在去除初赛影响后对应的理想比赛。

最后，如本节开头所说，本节的目标旨在估计而非精准计算，所得的结果仅能反映趋势而不保证精确；这对上述定理也同样成立。

5 关于比赛名次的讨论

本节中研究关于比赛名次的性质，这些性质可以为选手的比赛策略制订和日常训练导向提供参考。

5.1 比赛名次与实际水平的关系

一名选手在比赛得分上的波动，导致了他在比赛名次上的波动。考虑到各种奖项的颁发都是以名次而非得分为主要依据，研究选手比赛名次的概率分布就显得格外重要。

具体地说，对于 A, B 为一对互相对应的理想比赛和现实比赛，在这一小节中我们考虑：在该场比赛中额外加入一名选手 p（该选手得分的概率分布给定），则他的名次的概率分布为何。特别地，我们研究两个自变量（选手 p 的期望分数和分数波动幅度）对两个因变量（选手 p 的期望名次和中位名次）的影响。

本节中考虑的 B 均为 COI/IOI 赛制的信息学比赛。于是由定理3.2，Δ_B 和选手 p 的得分都服从正态分布。

5.1.1 期望分数与比赛名次的关系

记随机变量 C_p 为 p 的（与 B 中的选手分数一起，按定义2.3中的方式变换过后的）分数，设其服从期望值为 μ 、标准差为 σ 的正态分布，并用 $\mathcal{D}_p$ 来表示 $C_p-\mu$ 的概率密度函数。

令随机变量 U_p 表示：B 中有多大比例的选手实际得分高于 p 的实际得分（例如 $U_p=0.5$ 表示 B 中恰一半的选手实际得分高于 p）。

令实数 V_p 表示：B 中有多大比例的选手期望得分高于 p 的期望得分（例如 $V_p=0.5$ 表示 B 中恰一半的选手期望得分高于 p）。

此外设 Δ_B 服从正态分布 $N(0,\sigma_B^2)$ ，从而对任意 δ 有 $\mathcal{D}_A(\delta)=P_{\sigma_B^2}(\delta)$ 。

这一小节中将对变化的 μ ，考察以下两个量的变化趋势：

- $D_1=\mathrm{E}\left[U_p\right]-V_p$
- $D_2=\mathrm{Med}\left[U_p\right]-V_p$，其中**中位名次** $\mathrm{Med}\left[U_p\right]$ 满足 $\Pr\left[U_p>\mathrm{Med}\left[U_p\right]\right]=0.5$

对于 $\mathrm{E}\left[U_p\right]$ 不难发现：

$$
\begin{aligned}
\mathrm{E}\left[U_p\right] &= \int_0^1 \mathcal{X}_A(x)\left(\iint\limits_{\{(\delta_1,\delta_2):\mu+\delta_1\leqslant x+\delta_2\}} \mathcal{D}_p(\delta_1)\mathcal{D}_A(\delta_2)\mathrm{d}(\delta_1,\delta_2)\right)\mathrm{d}x \\
&= \int_0^1 \mathcal{X}_A(x)\left(\iint\limits_{\left\{(\delta_1,\delta_2^-):\delta_1+\delta_2^-\leqslant x-\mu\right\}} P_{\sigma^2}(\delta_1)P_{\sigma_B^2}(-\delta_2^-)\mathrm{d}(\delta_1,\delta_2^-)\right)\mathrm{d}x \\
&= \int_0^1 \mathcal{X}_A(x)\left(\int_{-\infty}^{x-\mu} P_{\sigma^2+\sigma_B^2}(\delta)\mathrm{d}\delta\right)\mathrm{d}x \\
&= \int_0^1 \mathcal{X}_A(x)R_{\sigma^2+\sigma_B^2}(x-\mu)\mathrm{d}x
\end{aligned}
$$

而对于 $\mathrm{Med}\left[U_p\right]$ 有

$$
\begin{aligned}
\mathrm{Med}\left[U_p\right] &= \int_\mu^{+\infty} C_A(s)\mathrm{d}s \\
&= \int_0^1 \mathcal{X}_A(x)R_{\sigma_B^2}(x-\mu)\mathrm{d}x
\end{aligned} \tag{7}
$$

其中等式(7)成立的理由是

$$\begin{aligned}&\Pr\left[U_p > \int_{\mu}^{+\infty} C_A(s)\mathrm{d}s\right]\\ =&\Pr\left[C_p > \mu\right]\\ =&0.5\end{aligned}$$

由此可以对给定的 $\mu, \sigma, \mathcal{X}_A$ 计算 D_1, D_2 的值。

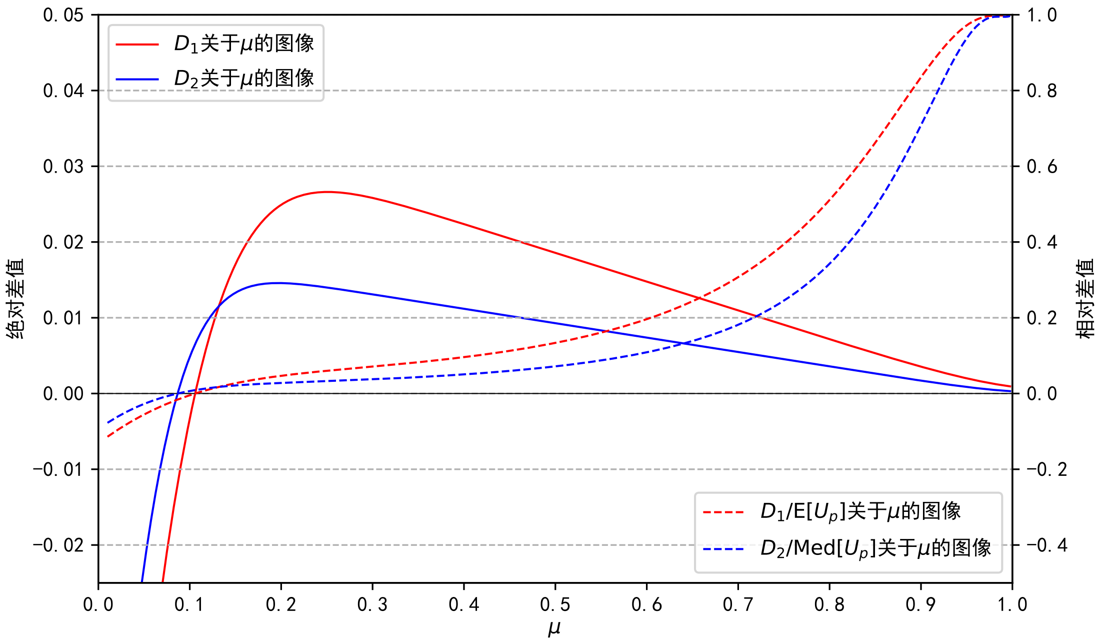

图 9: D_1, D_2 的图像，当 A 为定理4.7中的 A' 时

取理想比赛 A 为定理4.7中“复赛在去除初赛影响后对应的理想比赛” A'、取 $\sigma = \sigma_B \approx 0.078$（该数值来自于4.3小节中的测量），则 D_1, D_2 关于 μ 的图像如图9。这里对于期望分数小于（定理4.7中的）ϵ 的选手，将他们从 $\mathcal{X}_A$ 中剔除，并将 $\mathcal{X}_A$ 剩余部分的点值乘上某个系数，以使得其在 $[0, 1]$ 上定积分仍为 1。这样得到的 $\mathcal{X}_A$ 为二次函数。此外，D_1, D_2 本身是“绝对差值”，图中还用虚线绘制了“相对差值” $\frac{D_1}{\mathrm{E}[U_p]}, \frac{D_2}{\mathrm{Med}[U_p]}$ 的图像。

图9所考虑的情况是：现实比赛 B 的参赛人群为期望分数达到 ϵ 的**全体**信息学选手。但是对于实际中的比赛，其参赛选手往往是经过选拔的；以下将研究此类比赛。

简便起见，我们作如下假设：现实比赛 B 的参赛人群，是期望分数达到 ϵ 的**全体**信息学选手，经过一场现实比赛 C 的选拔后达到分数线的那些；其中 C 所对应的理想比赛与（定理4.7中的）A' 完全相同，且分数线为 $thres \in [0, 1]$ 。

由此易知，对于一名期望分数为 $\mu_q \in [0,1]$ 的选手 q，他达到分数线的概率为 $1 - R_{(0.078)^2}(thres - \mu_q)$。进而得到：

$$\mathcal{X}_A(x) \propto (ax^2+bx+c)\left(1 - R_{(0.078)^2}(thres - x)\right), \quad \forall x \in [0,1]$$

其中 a, b, c 为定理4.7中的系数。

为模拟联赛复赛的情况，取 $thres = 0.1053$（该数值为4.2.4小节中的计算结果），此时 D_1, D_2 的图像如图10；为模拟 NOI 的情况，取 $thres = 0.7$（计算发现近几年联赛中得分高于该数值的选手人数与 NOI 的参赛人数相近，故选择该数值），此时 D_1, D_2 的图像如图11。

观察图9、图10和图11，可以得到以下结论：

> 在一场信息学比赛中，对任何一名期望得分“不太低”的选手，他的期望名次、中位名次，均差于他的实际水平（以期望得分来衡量）在全体参赛选手中的位次。这一现象随着该名选手期望得分的升高而越发明显，在高分段尤其显著。

5.1.2 分数波动幅度与比赛名次的关系

这一小节中将考察 σ 的变化对 $\mathrm{E}[U_p]$ 的影响。注意到 σ 的变化不影响 $\mathrm{Med}[U_p]$ 的值，所以这里不考虑 $\mathrm{Med}[U_p]$。

记 $\mathrm{E}[U_p](\mu_0, \sigma_0)$ 为 $\mu = \mu_0, \sigma = \sigma_0$ 时 $\mathrm{E}[U_p]$ 的取值，则我们将考察以下函数：

$$F(\mu_0, \sigma_0) = \frac{\mathrm{E}[U_p](\mu_0, \sigma_0) - \mathrm{E}[U_p](\mu_0, 0)}{\mathrm{E}[U_p](\mu_0, \sigma_0) + \mathrm{E}[U_p](\mu_0, 0)}$$

函数 F 的含义是 $\mathrm{E}[U_p]$ 随 σ 的变化而产生的增量；这里的增量用 $\mathrm{E}[U_p](\mu_0, \sigma_0)$ 与 $\mathrm{E}[U_p](\mu_0, 0)$ 的相对差值来描述，由于二者中任何一者都可能远小于对方，故分母上取二者之和而非二者之一。

上一小节中对三个比赛分别绘制了 D_1, D_2 的图像；这里我们对同样的三个比赛绘制 F 的图像，见图12、图13和图14。图中底部的线条为上方曲面的等高线。此外请注意图中坐标轴的方向。

观察三个图像，可以得到以下结论：

> 在一场信息学比赛中，对任何一名期望得分“不太低”的选手，若他的期望得分固定，则当他的分数分布的标准差提高时，他的期望名次变差。这一现象随着该名选手期望得分的升高而越发明显，在高分段尤其显著。

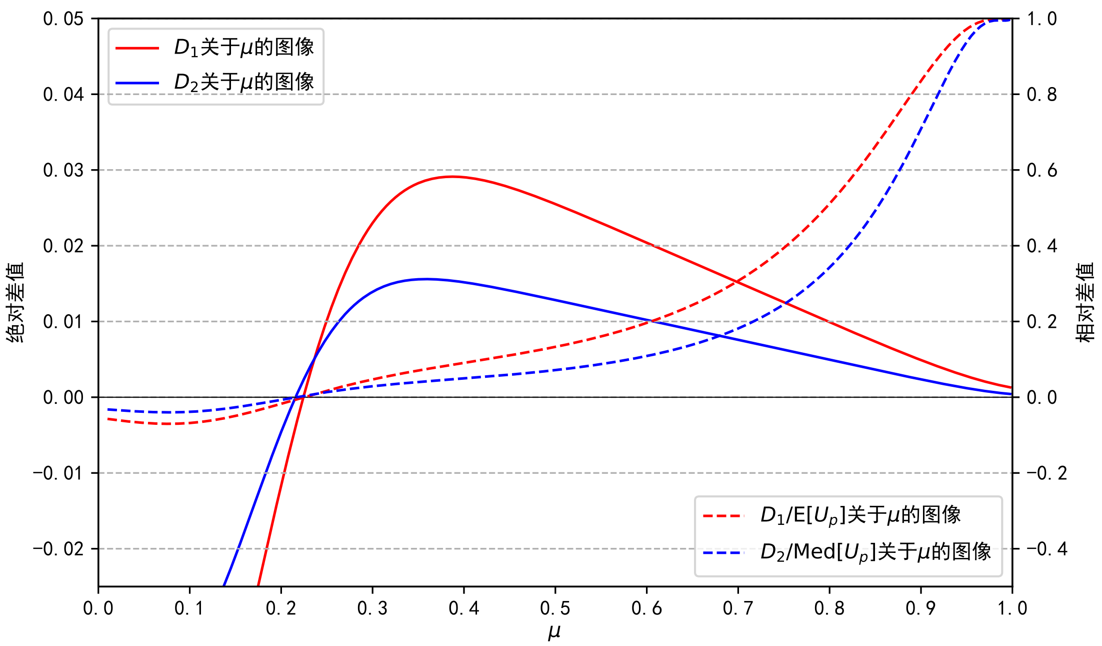

图 10: D_1, D_2 的图像，当 A 对应联赛复赛时

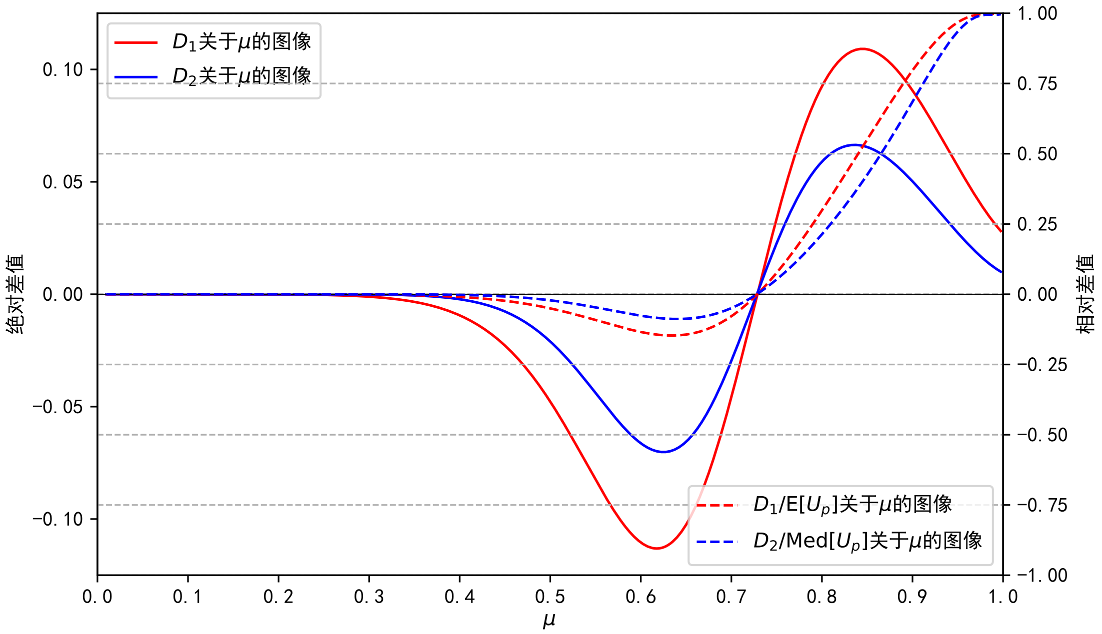

图 11: D_1, D_2 的图像，当 A 对应 NOI 时

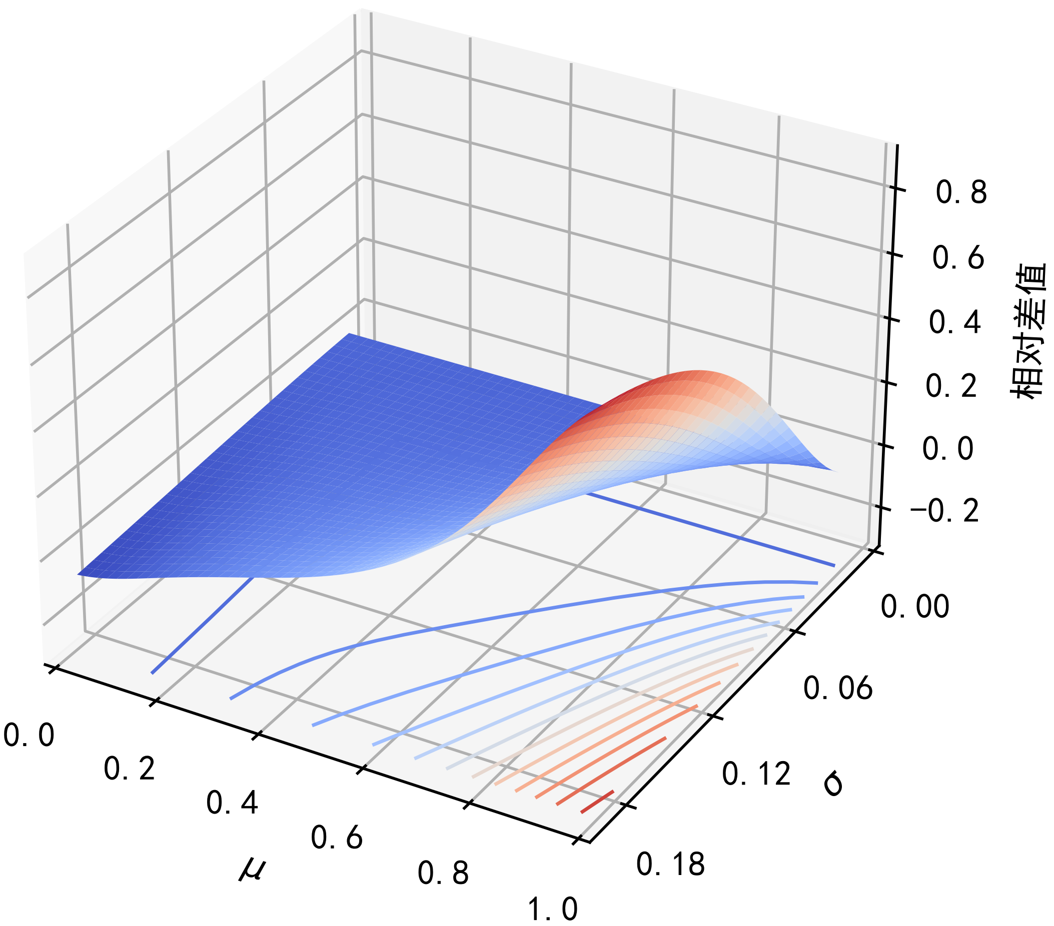

图 12: F 的图像，当 A 为定理4.7中的 A' 时

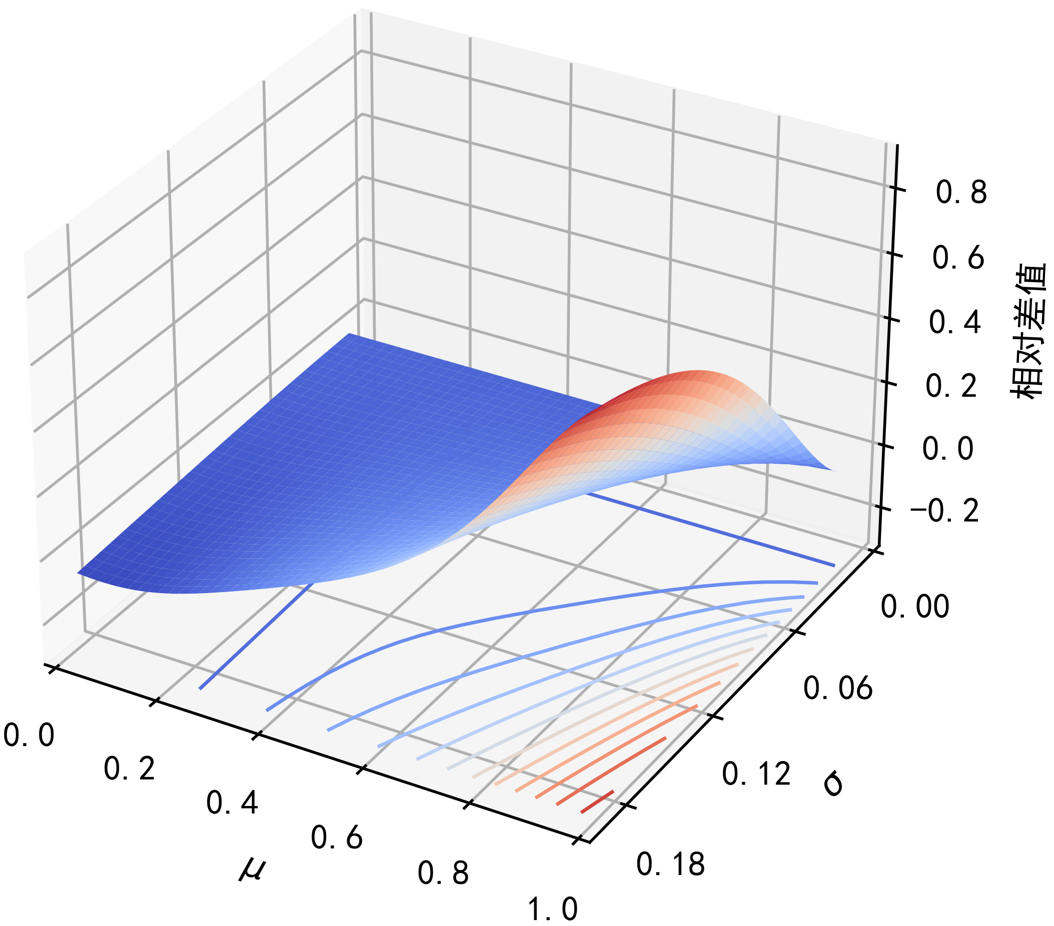

图 13: F 的图像，当 A 对应联赛复赛时

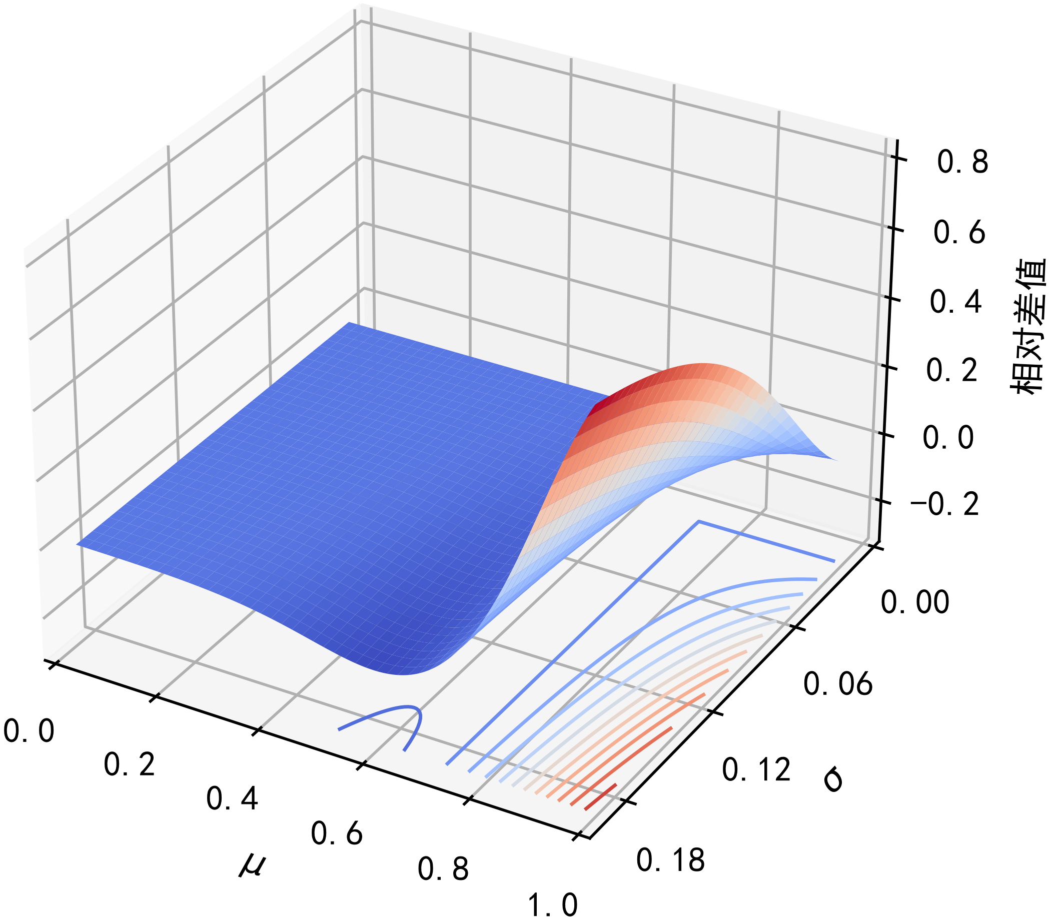

图 14: F 的图像，当 A 对应 NOI 时

5.2 多日比赛总名次与单日名次的关系

大多数信息学比赛都包括两天或更多的比赛日，而对于其中的大部分比赛，选手在每日考试后也都能直接或间接地知道自己当日的大致排名。这种情况下，一名选手在先前的考试日中的排名，对他在之后的考试日中的策略制定是一个重要信息。但要想有效利用这一信息，首先需要了解多日比赛总名次与单日名次之间的关系，而这正是本小节的目标。

先给出这一小节用到的模型：现实比赛 $B_{1\cdots k}$ 分别表示每一天的比赛，它们对应同一个理想比赛 A 。一名选手的总分定义为：他在 $B_{1\cdots k}$ 中的分数 $score_{1\cdots k}$ （满足 $score_i \in [0,1]$ ）的加权平均数 $w_1 score_1 + \cdots + w_k score_k$ （满足 $w_1 + \cdots + w_k = 1$ ）。

在实际中，每天的比赛在难度、区分度上可能有差异，因此仅可以认为各天的比赛（所对应的理想比赛）缩放等价而不是完全相同；但通过引入权值 $w_{1\cdots k}$ ，可以将这些比赛“标准化”。也就是说, $w_{1\cdots k}$ 在此处相当于“缩放系数”，与等价映射中的一次项系数作用相同。

设 Δ_{B_i} 服从正态分布 $N(0,\sigma_B^2)$ 。以下我们关注比赛中的一名特定选手 p ，记他各天的分数为 $score_{1\cdots k}$ 、各天的名次为 $rank_{1\cdots k}$ （满足 $rank_i \in [0,1]$ ，定义方式与上一节中的 U_p 相同），我们将尝试从 $rank_{1\cdots k}$ 近似地计算 p 的总分名次 $rank_{total} \in [0,1]$ 。

这里我们还要求：

• $\mathcal{X}_A(x) = (a'x^2 + b'x + c')\left(1 - R_{\sigma_P^2}(thres - x)\right)$ ，其中 $(a',b',c') \propto (a,b,c)$ 为常系数、$thres, \sigma_P$ 为某两个不关心其具体取值的常数。

• 注意到对足够大的 x （不妨划定为 $x \geqslant T$ ，其中 T 为某个不关心其具体取值的常数）有 $\mathcal{X}_A(x) \approx a'x^2 + b'x + c'$ 。我们要求 p 的得分和名次足够高，以致于 $\mathcal{X}_A(x)$ 在 $x < T$ 的部分对答案几乎没有影响。

5.2.1 $\sigma_B = 0$ 的情况

在这一小节中，假设 $\sigma_B = 0$ 。严格地说，σ_B 必须是正数，否则 $N(0,\sigma^2)$ 就并非良定义；但这里 $\sigma_B = 0$ 仅用来表达 $C_A = \mathcal{X}_A$ ，故暂且忽略这一问题。

通过对系数 a',b',c' 的计算，发现 $a'x^2 + b'x + c' \approx a'(x-1)^2$ ，于是可以得到：

$$
\begin{aligned}
rank_i &= \int_{score_i}^{1} C_A(s)\mathrm{d}s \\
&\approx \int_{0}^{1-score_i} a's^2\mathrm{d}s \\
&= \frac{a'\left(1-score_i\right)^3}{3}
\end{aligned}
$$

所以有

$$
1 - score_i \approx \sqrt[3]{\frac{3rank_i}{a'}}
$$

进而 p 的总分 $score_{total}$ 满足

$$1 - score_{total} \approx \sum_{i=1}^{k} w_i \sqrt[3]{\frac{3rank_i}{a'}}$$

最后

$$\begin{aligned}
rank_{total} &\approx \frac{a'(1 - score_{total})^3}{3} \\
&\approx \frac{a'}{3} \cdot \left(\sum_{i=1}^{k} w_i \sqrt[3]{\frac{3rank_i}{a'}} \right)^3 \\
&= \left(\sum_{i=1}^{k} w_i \sqrt[3]{rank_i} \right)^3
\end{aligned}$$

注意到上式的左右两边是齐次的，所以，若将所有 $rank_{1\cdots k}$ 和 $rank_{total}$ 同时乘以常数，则正确性不会受影响。因此即使取 $rank_{1\cdots k}$ 和 $rank_{total}$ 为“原始排名”（即分数高于 p 的人数），上式仍然成立。

得到结论：

> 若忽略参赛选手分数的波动（也就是假设，除 p 以外的任何选手在任何一天中都恰会考出自己的期望得分），则对于得分足够高的选手 p ，其总名次约等于每日名次的<u>1/3 次加权幂平均</u>。

结合幂平均不等式 [3]，可以得到如下推论：

> 在上一结论的条件下，选手 p 的总名次一定不差于（即数值上不大于）每日名次的加权平均数，取等仅当每日名次相同。更进一步，在名次加权平均数固定的情况下，名次波动幅度越大，总名次一般越好（即数值上越小）；换句话说，对单场比赛而言，前进 n 名对总名次的正面影响，一般大于后退 n 名对总名次的负面影响。

5.2.2 $\sigma_B > 0$ 的情况

这一小节中考虑 $\sigma_B > 0$ 的情况。

首先，为了能够直接确定系数 a', b', c' ，我们不妨取 $thres$ 为 $-\infty$ ，也就是说对任何 $x \in [0, 1]$ 有 $\mathcal{X}_A(x) = a'x^2 + b'x + c'$ 。注意到对于 $\mathcal{X}_A(x)$ 在 $x \in [T, 1]$ 的部分而言，$thres$ 的变化只相当于对这一部分函数值作等比例缩放；而稍后会看到，对 $\mathcal{X}$ 和 C 的函数值的等比例缩放，不会影响后续推导的正确性。因此，钦定 $thres$ 的取值不会影响一般性。

为了使得后续推导成为可能，需要先对函数 C_A 作近似，以简化其形式。

对于确定的 σ_B ，可以计算得到实数 $r(\sigma_B) \geqslant 1$ 和 $a(\sigma_B) \geqslant 0$ ，以使得函数 $f_{\sigma_B}(s) = a(\sigma_B)(r(\sigma_B) - s)^2$ 的图像与 C_A 尽可能贴近（即残差平方和最小）。实验发现：

$$\max_{\sigma\in(0,0.4],s\in[0.6,1]} |f_{\sigma_B}(s) - C_A(s)| \approx 0.02$$

由此可见，这一近似的效果十分优秀。回忆到我们只关注“足够高”的那一部分分数分布，因此在上式中只考虑 $s \geqslant 0.6$ 。

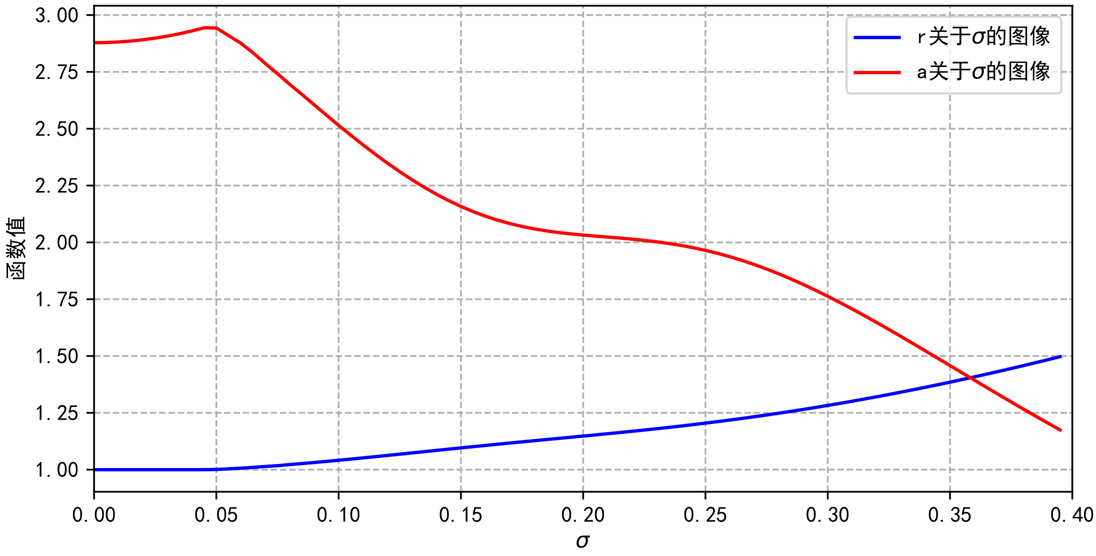

图 15: 函数 $a(\sigma)$ 和 $r(\sigma)$ 的图像

图15展示了函数 $a(\sigma)$ 和 $r(\sigma)$ 的图像。可以看到，前者大体上递降，而后者则递增。

将全部 k 个比赛合并视为一个现实比赛 B_{total} ，以下计算（服从正态分布的）随机变量 $\Delta_{B_{total}}$ 的标准差 σ_{total} 。

$$\begin{aligned}
\sigma_{total} &= \text{Stddev}\left[\sum_{i=1}^{k} w_i \Delta_{B_i}\right] \\
&= \sqrt{\sum_{i=1}^{k} \text{Stddev}\left[w_i \Delta_{B_i}\right]^2} \\
&= \sqrt{\sum_{i=1}^{k} w_i^2 \sigma_B^2} \\
&= \sqrt{\sum_{i=1}^{k} w_i^2} \cdot \sigma_B
\end{aligned}$$

特别地，当权重 w_i 全部相等时，$\sigma_{total} = k^{-1/2}\sigma_B$ 。

接下来类比上一小节进行推导。注意到 $C_A(s)$ 在 $s \in (1, r]$ 的部分也要纳入考虑。

$$\begin{aligned} rank_i &= \int_{score_i}^{r} (\sigma_B) C_A(s) \mathrm{d}s \\ &\approx \int_{0}^{r(\sigma_B)-score_i} a(\sigma_B) s^2 \mathrm{d}s \\ &= \frac{a(\sigma_B)\left(r(\sigma_B) - score_i\right)^3}{3} \end{aligned}$$

$$\begin{aligned} r(\sigma_B) - score_{total} &= \sum_{i=1}^{k} w_i \left(r(\sigma_B) - score_i\right) \\ &\approx \sum_{i=1}^{k} w_i \sqrt[3]{\frac{3rank_i}{a(\sigma_B)}} \end{aligned}$$

$$\begin{aligned} rank_{total} &\approx \frac{a(\sigma_{total})(r(\sigma_{total}) - score_{total})^3}{3} \\ &\approx \frac{a(\sigma_{total})}{3} \cdot \left(r(\sigma_{total}) - r(\sigma_B) + \sum_{i=1}^{k} w_i \sqrt[3]{\frac{3rank_i}{a(\sigma_B)}} \right)^3 \\ &= \frac{a(\sigma_{total})\left(r(\sigma_{total}) - r(\sigma_B)\right)}{3} + \frac{a(\sigma_{total})}{a(\sigma_B)} \left(\sum_{i=1}^{k} w_i \sqrt[3]{rank_i} \right)^3 \end{aligned}$$

发现总名次 $rank_{total}$ 是关于 $\sum_{i=1}^{k} w_i \sqrt[3]{rank_i}$ 的一次函数。接着注意到 $\sigma_{total} < \sigma_B$，故由 $a(\sigma), r(\sigma)$ 的增减性可知，该一次函数的一次项系数大于等于 1，而常数项小于等于 0；因此自变量和因变量的大小关系无法直接确定。

图16展示了以下 4 个量关于 σ_B 的图像：

蓝色 在典型情况下（指的是 $k = 2, w_1 = w_2$ 时的情况，此时 $\sigma_{total} = 2^{-1/2}\sigma_B$）的一次项系数 $K = a(\sigma_{total})/a(\sigma_B)$ 。

红色 在典型情况下的常数项 $B = a(\sigma_{total})\left(r(\sigma_{total}) - r(\sigma_B)\right)/3$ 。

黄色 最大的 $m \in [0, 1]$ 满足在典型情况下有 $Km + B < m$ 。

绿色 最大的 $m \in [0, 1]$ 满足在任何情况下（对 σ_{total} 只需符合 $\sigma_{total} < \sigma_B$ 即可）都有 $Km + B < m$ 。

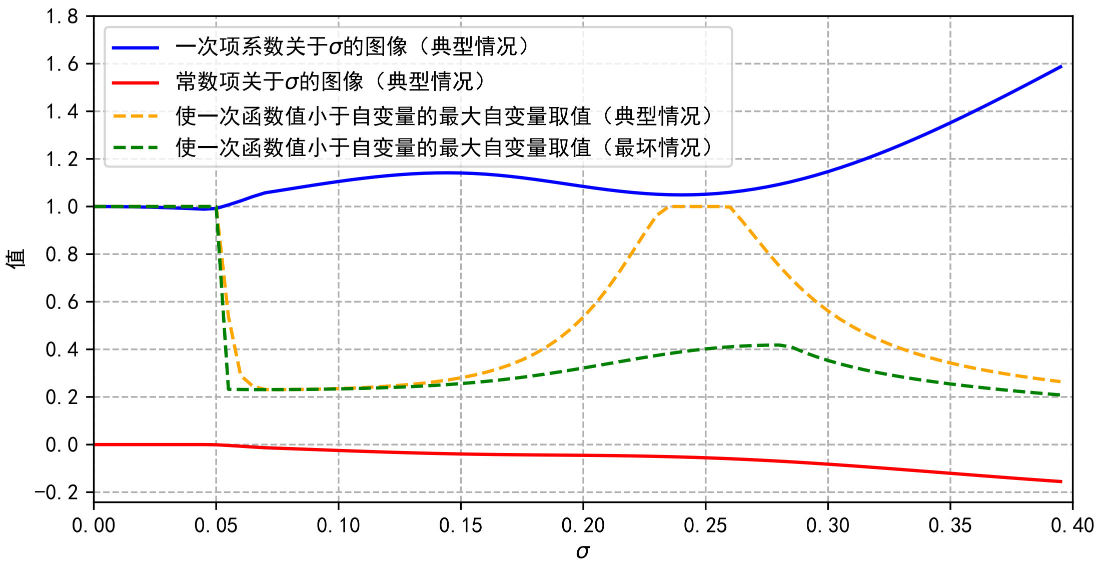

图 16: 一次函数的系数和大小关系改变的阈值

观察到绿色曲线的纵坐标始终没有低于 0.2，因此可以认为

$$\left(\sum_{i=1}^{k} w_i \sqrt[3]{rank_i} \leqslant 0.2\right) \Rightarrow \left(rank_{total} < \sum_{i=1}^{k} w_i \sqrt[3]{rank_i}\right)$$

回忆到先前要求 p 的得分“足够高”，因此我们暂不考虑 $\sum_{i=1}^{k} w_i \sqrt[3]{rank_i} > 0.2$ 的情况。

在本小节的开头，我们作了 $thres = -\infty$ 的假设。注意到随着 $thres$ 的升高，所有 $a(\sigma)$ 会等比例地变大，而 $r(\sigma)$ 不变，进而可以证明阈值 0.2 只会变大而不会变小。因此，我们所考虑的 $thres = -\infty$ 恰好是最坏情况。

最终得到结论：

> 对于得分足够高的选手 p，计算其每日名次的1/3 次加权幂平均 M，则 p 的总名次 m 近似地满足线性关系 $m = K \cdot M + B$，其中系数 K, B 由比赛本身确定。如果 p 的排名在前 20%，则进一步有 $m < M$ 。

从上述性质可以导出一个有趣的推论：如果 p 的每日名次都相等且在前 20%，则 p 的总名次一定好于他的每日名次。

6 致谢

感谢中国计算机学会提供交流和学习的平台；

感谢国家集训队高闻远教练的指导；
感谢教导过我的老师、教练们；
感谢清芷等同学与我讨论本文内容。

参考文献

[1] Wikipedia: Sum of normally distributed random variables,
https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables

[2] SciPy Documentation: scipy.optimize.minimize,
https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html

[3] Wikipedia: Inequality between any two power means,
https://en.wikipedia.org/wiki/Generalized_mean#Inequality_between_any_two_power_means