基于简单python脚本的SRTP三组公平性分析

A.I.B.

Qiushi Honor Class, College of Life Science, Zhejiang University 310000;

Abstract: 本次SRTP立项采用三组分开答辩的形式，为了探究本次立项过程中三组的评分标准是否一致。我们利用简单的python脚本进行平均值、方差和分布分析，并通过平均值和方差的关系得到了最终的结果，随后借用统计物理的基本假设；随后将本次的平均分组看成为热力学的正则系综，继续利用脚本判断公平性。

Key Words: SRTP；公平性分析；Python；正则系综

1 背景

1.1 SRTP

SRTP即浙江大学学生科研训练计划（Student Research Training Project）。简言之，即学生可以自行组建团队，寻找导师，进入实验室参与科研项目。项目申请成功可获得一定经费支持，成功结题可获得二课分。通常而言，SRTP会在每年春学期开学初进行项目申报，分为国家级、省级、校级、院级四级项目，面向对象一般为大二、大三学生，一个项目通常持续约一年时间，项目成果（如论文、专利等）属于项目团队。[1]

1.2 正则系综

1.2.1 系综理论

在林宗涵的《热力学与统计物理学》中对系综做出描述：**系综**是假想的、和所研究的系统性质完全相同的、彼此独立、**各自处于某一微观状态**的大量系统的集合.。这些微观状态会形成一个分布，哈密顿量 $H(\mathbf{p},\mathbf{q})$常用于描述这些微观状态

常见的系综有：

- 微正则系综- (N,V,E)系综：保持粒子数、体积、能量不变，下同
- 正则系综- (N,V,T)系综：可用于**描述社会现象**
- 巨正则系综- (μ,V,T)系综：用于描述开放体系，比如表面吸附过程等

1.2.2 系综平均

刘维尔定理：将系综在相空间中的运动看成代表点组成流体的运动，这个流体是**不可压缩的**，数学形式：

$$\frac{\partial \rho}{\partial t}=\{\rho,H\}\quad \frac{d\rho}{dt}=0 \tag{1}$$

上式中大括号表示泊松括号。因此， $\rho(\mathbf{p},\mathbf{q})$的表达式中不显含时间，对于平衡态，分布函数 $\rho(\mathbf{p},\mathbf{q})$是相空间中的位形函数。因此，对于系综平均可以用经典概率论中的 pdf函数计算，物理量 $u(\mathbf{p},\mathbf{q},t)$在 t时刻的系综平均：

$$<u(\mathbf{p},\mathbf{q},t)>=\iint u(\mathbf{p},\mathbf{q},t)\rho(\mathbf{p},\mathbf{q})d\mathbf{p}d\mathbf{q} \tag{2}$$

对于经典正则系综满足麦克斯韦分布：

$$\rho(\mathbf{p},\mathbf{q})d\mathbf{p}d\mathbf{q}=\frac{e^{-\beta H(\mathbf{p},\mathbf{q})}d\mathbf{p}d\mathbf{q}}{\iint e^{-\beta H(\mathbf{p},\mathbf{q})}d\mathbf{p}d\mathbf{q}} \tag{3}$$

期中分母正比于配分函数 Z：

$$Z = \frac{1}{h^r} \iint e^{-\beta H(\mathbf{p},\mathbf{q})} d\mathbf{p} d\mathbf{q} \tag{4}$$

因此：

$$< u(\mathbf{p},\mathbf{q},t) >= \iint u(\mathbf{p},\mathbf{q},t) \frac{e^{-\beta H(\mathbf{p},\mathbf{q})} d\mathbf{p} d\mathbf{q}}{\iint e^{-\beta H(\mathbf{p},\mathbf{q})} d\mathbf{p} d\mathbf{q}} d\mathbf{p} d\mathbf{q} \overset{h=1}{=} \frac{1}{Z} \iint u(\mathbf{p},\mathbf{q},t) e^{-\beta H(\mathbf{p},\mathbf{q})} d\mathbf{p} d\mathbf{q} \tag{5}$$

2 材料与方法

2.1 材料

来源于生科院官网的初次公示名单中，对各组的排名进行人为统计，讲三组同学的排名分别统计到表格中，详见附录1.[2]

2.2 方法

采用最为简单的Python脚本编写，代码由两块组成，第一部分进行简单的平衡值、方差、分布的分析；第二部分讲分组假设为正则系综，进行正态分布分析。脚本如下：

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

plt.rcParams['font.size']=21
plt.rcParams['font.family']='Times New Roman'

def analyze_and_plot(file_path):
    # 读取Excel文件
    df = pd.read_excel(file_path)

    # 计算每组的平均数和方差
    means = df.mean()
    variances = df.var()

    # 绘制平均数图表
    plt.figure(figsize=(10, 6))
    means.plot(kind='bar', color='skyblue')
    plt.title('Average Rank of Each Group')
    plt.xlabel('Group')
    plt.ylabel('Average Rank')
    plt.xticks(ticks=range(3), labels=['Group 1', 'Group 2', 'Group 3'], rotation=0)
    plt.tight_layout()
    plt.show()

    # 绘制方差图表
    plt.figure(figsize=(10, 6))
    variances.plot(kind='bar', color='orange')
    plt.title('Variance of Ranks in Each Group')
    plt.xlabel('Group')
    plt.ylabel('Variance')
    plt.xticks(ticks=range(3), labels=['Group 1', 'Group 2', 'Group 3'], rotation=0)
    plt.tight_layout()
    plt.show()

    # 绘制每组排名的分布图
    for i in range(3):
```

```
        plt.figure(figsize=(10, 6))
        plt.hist(df.iloc[:, i], bins=range(1, 58), alpha=0.7, label=f'Group {i+1}')
        plt.axvline(x=np.mean(df.iloc[:, i]), color='k', linestyle='dashed', linewidth=1)
        plt.text(np.mean(df.iloc[:, i]) * 1.1, 5, 'Mean: {:.2f}'.format(np.mean(df.iloc[:,
i])))
        plt.xlabel('Rank')
        plt.ylabel('Frequency')
        plt.title(f'Rank Distribution for Group {i+1}')
        plt.legend()
        plt.tight_layout()
        plt.show()

    # 公平性判断逻辑
    mean_diff = np.max(means) - np.min(means)
    var_diff = np.max(variances) - np.min(variances)

    if mean_diff < 5 and var_diff < 50:
        print("Based on the analysis, the competition appears to be FAIR.")
    else:
        print("Based on the analysis, the competition may NOT be fair.")

# 调用函数
file_path = r'C:\Users\Administrator\Desktop\ranking.xlsx'
analyze_and_plot(file_path)


# 正则系综分析部分
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

def analyze_and_plot(file_path):
    df = pd.read_excel(file_path)

    # 计算每组的均值和标准差
    means = df.mean()
    std_devs = df.std()

    # 绘制均值和标准差的图表
    plt.figure(figsize=(12, 6))
    means.plot(kind='bar', yerr=std_devs, capsize=4, color='skyblue', position=0,
label='Mean')
    plt.title('Mean and Standard Deviation of Each Group')
    plt.xlabel('Group')
    plt.ylabel('Value')
    plt.xticks(ticks=range(3), labels=['Group 1', 'Group 2', 'Group 3'], rotation=0)
    plt.legend()
    plt.tight_layout()
    plt.show()

    # 判断公平性的新逻辑
    mean_diff = np.max(means) - np.min(means)
    std_diff = np.max(std_devs) - np.min(std_devs)

    if mean_diff < 10 and std_diff < 5:
        print("Based on the analysis, the competition appears to be FAIR.")
```

```
    else:
        print("Based on the analysis, the competition may NOT be fair.")

file_path = r'C:\Users\Administrator\Desktop\ranking.xlsx'
analyze_and_plot(file_path)
```

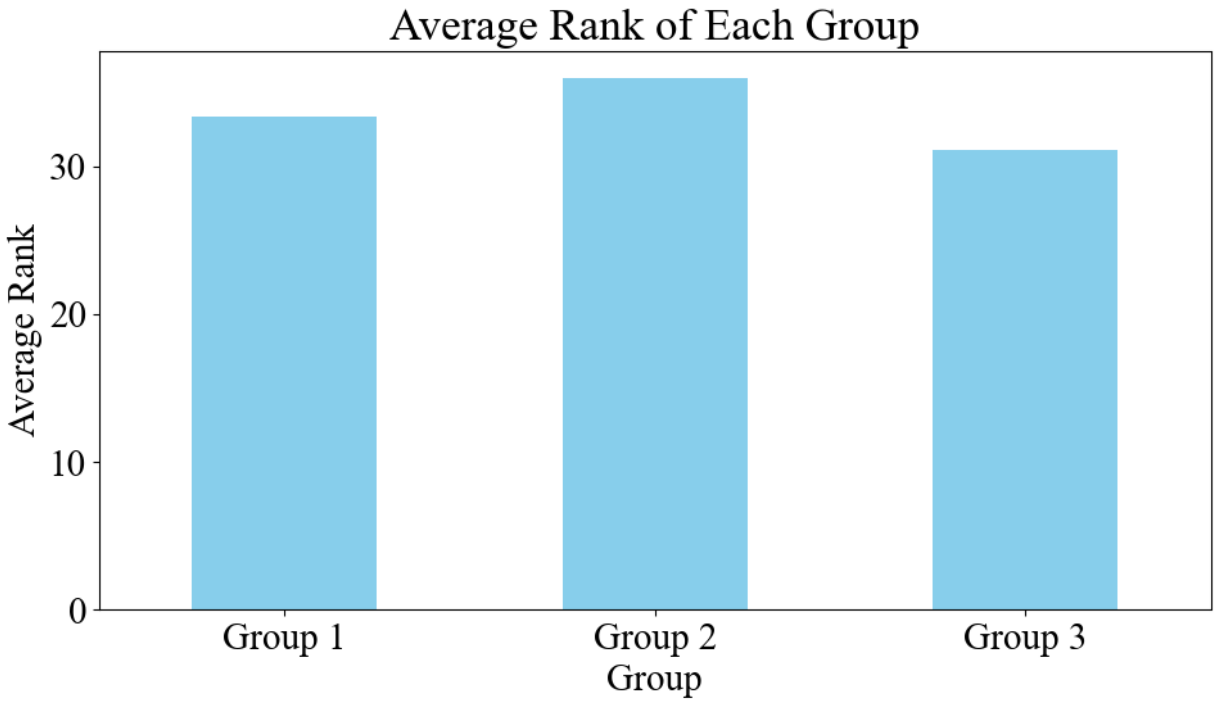
3 实验结果

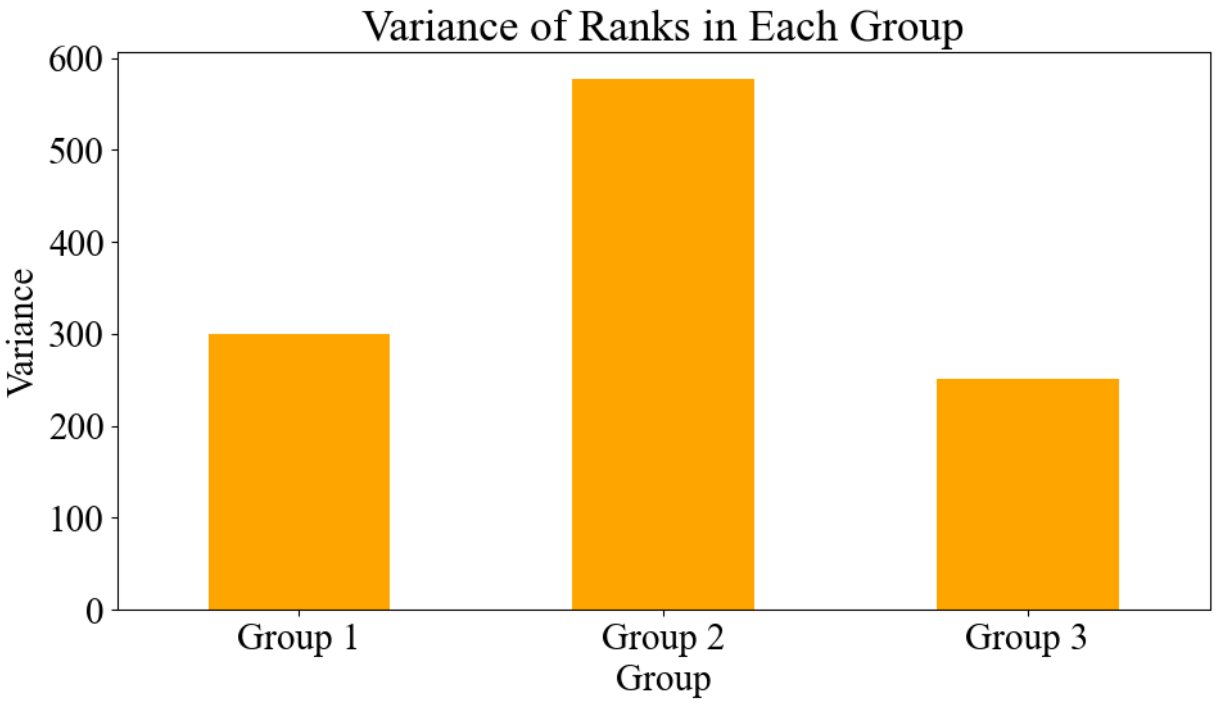
3.1 公平性判据

首先是运行判据结果(可以自行运行检测正确性)：

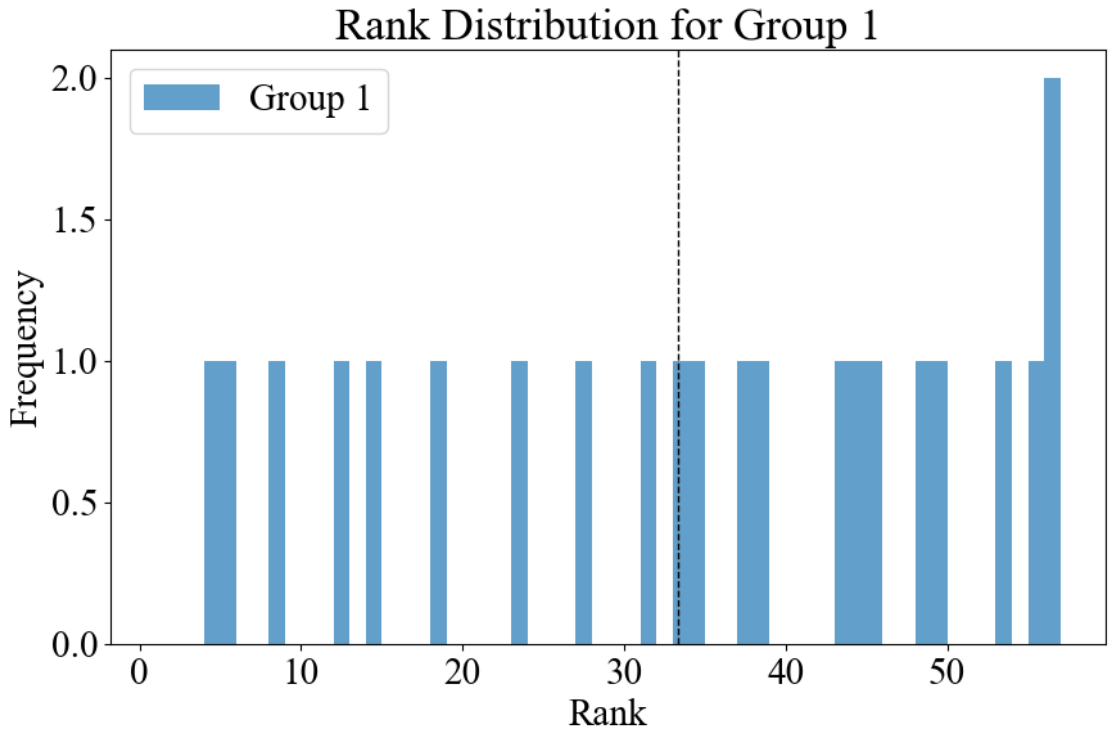
Based on the analysis, the competition may NOT be fair.
Based on the analysis, the competition may NOT be fair.

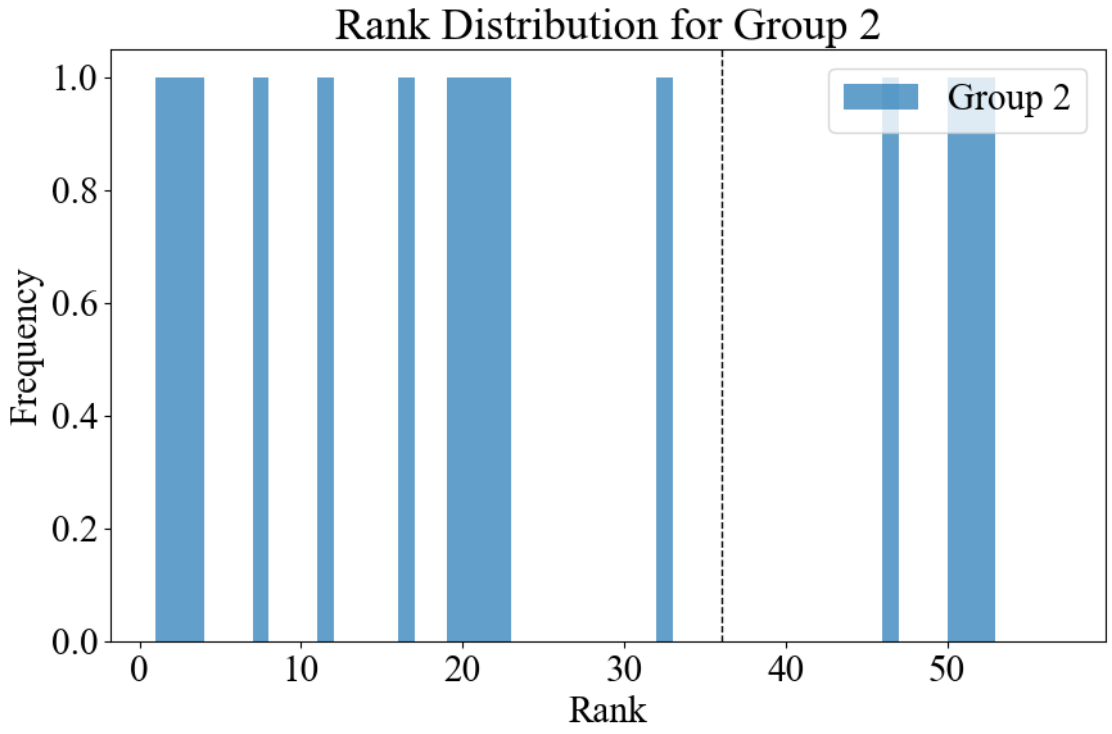
3.2 平衡值与方差

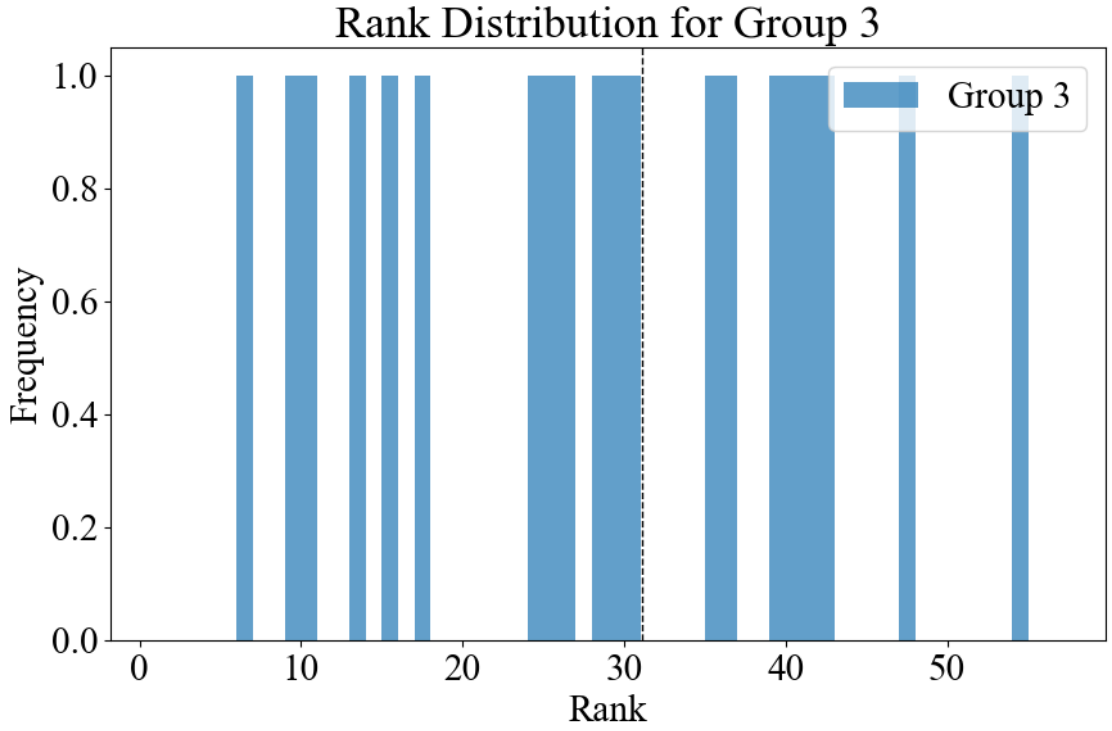



Variance of Ranks in Each Group
600
500
400
300
200
100
0
Variance
Group 1
Group 2
Group 3
Group

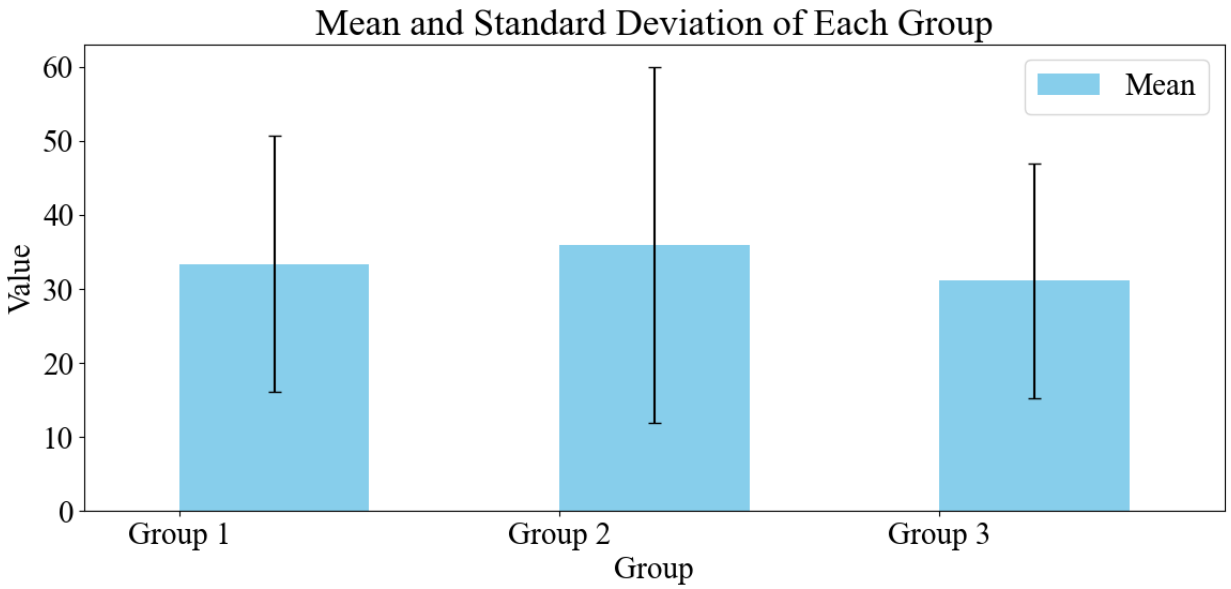
3.3 分布


Rank Distribution for Group 1
Group 1
2.0
1.5
1.0
0.5
0.0
Frequency
0
10
20
30
40
50
Rank


Rank Distribution for Group 2
Group 2
Frequency
Rank
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50


Rank Distribution for Group 3
Group 3
Frequency
Rank
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50

3.4 正态分布检验



结果图像如上图所示

4 分析与讨论

4.1 本方法的优劣势

- 相比与去年，三组确实可以提高效率
- 但是也出现了明显不公平现象
- 具体表现为三组间重点项目分配不均匀上
- 组间方差过大，平均值最大差值在5以上

4.2 本算法改进之处

- 过于简略
- 没有考虑干实验和湿实验的统计检验，相信结果肯定很震撼

参考文献：

[1]https://zjuers.com/welcome/learning/training_program/)

[2]http://www.cls.office.zju.edu.cn/2024/0401/c79352a2896740/page.htm