2023 年教育数据统计与分析处理期末大作业 (A)

数据科学 2003　　吴名民　　2030090109

2023-6-18

1　第一题：

1.1　理论分析：

- 极差是最大值与最小值之间的差距，若分数用 X 表示，则其计算公式为：

$$Range = X_{max} - X_{min}$$

- 百分位数计算公式: 将分数数据从小到大排序，百分比值 p，及样本总量 n 有以下数学公式可以表示：

$$L = (n)(\frac{p}{100})$$

情况 1: 如果 L 是一个整数，则取第 L 和第 L+1 这两个位置数值的平均值。
情况 2: 如果 L 不是一个整数，则取下一个最近的整数。（比如 L=1.2，则取位置为第 2 个的数值）平均分计算公式为:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{N}$$

- 样本方差计算公式为:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N-1},$$

- 样本标准差计算公式为：

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N-1}}$$

1.2 计算结果：

表 1: 统计量计算结果

统计量	值
range	97.00
mean	625.27
var	715.99
std	26.76
25%	602.00
50%	623.00
80%	652.00

2 第二题：

2.1 理论分析：

- 由组距计算组数：

$$[\text{极差}/\text{组距}] + 1$$

- 频率分布直方图: 在频率分布直方图中横轴表示众多个连续变量离散化以后的区间，这个区间的大小称为组距，纵轴表示频率/组距。
- 密度曲线: 当长方形的宽度无限小，即组距无限小的时候，频率分布直方图就无限接近于一条光滑曲线，我们把这条曲线叫做概率密度曲线。

2.2 绘图结果：

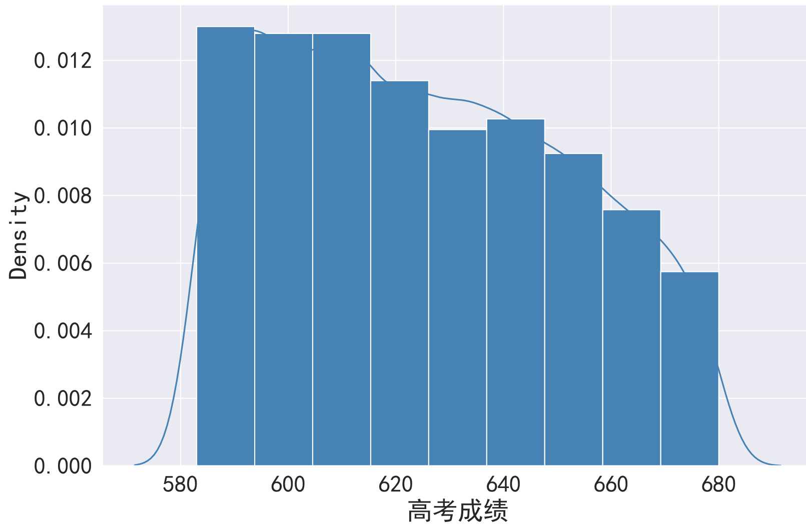

图 1: 频率分布直方图和概率密度曲线

3 第三题：

3.1 理论分析：

箱线图: 箱线图是用一组数据中的最小值、第一四分位数、中位数、第三四分位数和最大值来反映数据分布的中心位置和散布范围，可以粗略地看出数据是否具有对称性。通过将多组数据的箱线图画在同一坐标上，可以用于多组数据平均水平和变异程度的直观分析比较。四分位数（Quartile）是统计学中分位数的一种，即把所有数值由小到大排列并分成四等份，处于三个分割点位置的数值就是四分位数。

- 第一四分位数 (Q1)，等于该样本中所有数值由小到大排列第百分之 25 的数字。
- 第二四分位数 (Q2)，又称“中位数”，等于该样本中所有数值由小到大排列后第百分之 50 的数字。

- 第三四分位数 (Q3)，等于该样本中所有数值由小到大排列后第百分之 75 的数字。

3.2 绘图结果：

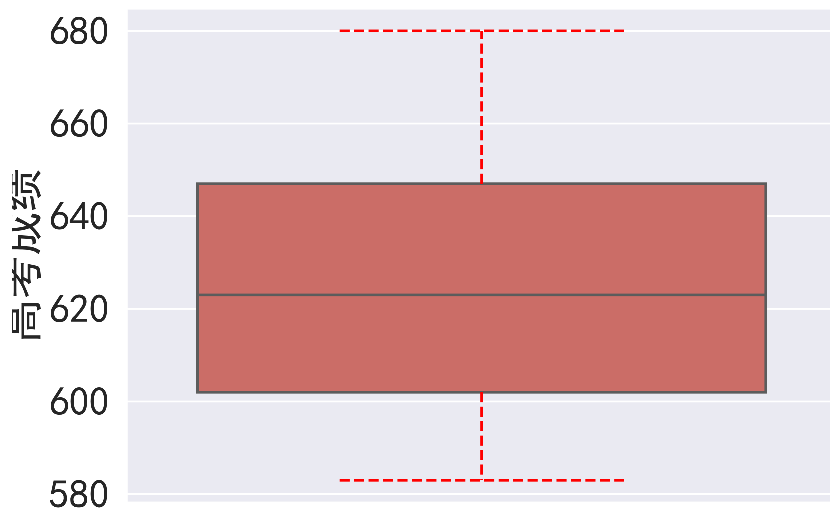

图 2: 箱线图

4 第四题：

4.1 理论分析：

4.1.1 第一问：

假设：H_0:$\mu \geq \mu_0 = 589$，即向小明能上天津师范大学。

H_1:$\mu < \mu_0$，即向小明不能上天津师范大学。

因为 H_0 中的全部 μ 都比 H_1 中的 μ 大，当 H_1 为真时 $\bar{X}$ 往往小，所

以拒绝域的形式为 $\bar{X} \leq k$。并且需要满足 $P\left\{当H_0为真且拒绝H_0\right\} \leq \alpha$，则

$$\begin{aligned} P\left\{当H_0为真且拒绝H_0\right\} &= P_{\mu \in H_0}\left\{\bar{X} \leq k\right\} \\ &= P_{\mu \geq \mu_0}\left\{\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \leq \frac{k-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right\} \\ &\leq P_{\mu \geq \mu_0}\left\{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{k-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right\} \end{aligned}$$

要控制 $P\left\{当H_0为真且拒绝H_0\right\} \leq \alpha$，需令 $P_{\mu \geq \mu_0}\left\{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{k-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right\} = \alpha$。因为 $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$，所以 $\frac{k-\mu_0}{\frac{\sigma}{\sqrt{n}}} = -z_\alpha \Rightarrow k = \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha$。因此拒绝域为 $\bar{x} \leq \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha$，即 $z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \leq -z_\alpha$，通过计算得 $\bar{X} = 573.667$。因为 $\sigma = 2, z_{0.05} = 1.645, \mu_0 = 589$，所以

$$z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = -13.279 \leq -1.645 = -z_\alpha$$

落在拒绝域内, 所以拒绝 H_0, 则预测向小明考不上天津师范大学。

4.1.2 第二问:

由题意假设：H_0:$\mu \geq \mu_0 = 589$，即向小明能上天津师范大学。

H_1:$\mu < \mu_0$，即向小明不能上天津师范大学。

是做 σ^2 关于 μ 的检验，由于 σ^2 未知不能利用 $\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ 来确定拒绝域，但 S^2 是 σ^2 的无偏估计，所以用 S 来替代 σ 采用：

$$t = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}}$$

来作为检验统计量，因为 $\frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}} \sim t(n-1)$, 由第一问同理可得该问题的拒绝域为:

$$t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} \leq -t_\alpha(n-1)$$

取 $\alpha = 0.05$, 则现在 $n = 3, t_{0.05}(2) = 2.9200$ 又算得 $\bar{x} = 573.667, s = 24.664$, 即有

$$t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = -1.077 > -2.9200 = -t_{0.05}(2)$$

t 未落在拒绝域中，故接受 H_0, 认为向小明考的上天津师范大学。

5 完整代码:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import pdfplumber
import seaborn as sns
import math
import warnings
warnings.filterwarnings("ignore")

#提取pdf中的一分一段表, 转为dataframe后保存为csv文件
pdf=pdfplumber.open('./(分配+数据+要求)数据科学2003班期末大作业.pdf')
pages = pdf.pages[6:8]
table0=pages[0].extract_tables(table_settings={"explicit_horizontal_lines":[90]})
table1=pages[1].extract_tables(table_settings={"explicit_horizontal_lines":[90]})
tb0=pd.DataFrame(table0).T
tb1=pd.DataFrame(table1).T
data=pd.concat([tb0[0],tb0[1],tb0[2],tb0[3],tb0[4],tb0[5],tb1[0],
              tb1[1],tb1[2],tb1[3]]).dropna(how='all').reset_index()[0]
data = data.apply(pd.Series,index=['高考成绩','人数','累积人数'])
data.to_csv('一分一段表.csv',index=False)

#读取csv数据并获得分数在583-680的数据, 进行处理
df = pd.read_csv('一分一段表.csv')
df583_680 = df[(df['高考成绩']>=583) & (df['高考成绩']<=680)]
new_df583_680 = pd.DataFrame(np.repeat(df583_680.values,
                            df583_680['人数'],axis=0))
new_df583_680.columns=['高考成绩','人数','累积人数']

#极差计算
max_score=new_df583_680['高考成绩'].max()
min_score = new_df583_680['高考成绩'].min()
range_score = max_score-min_score
print("range:    %.6f"%range_score)
#计算样本方差
var = new_df583_680['高考成绩'].var()
print("var:     %.6f"%var)
```

```
    #计算 总体百分位数(25 百分位数、中位数、80 百分位数)、平均分、标准差等其它统计量
    display(new_df583_680['高考成绩'].describe(percentiles=[0.25,0.5,0.8]))

    #将一分一段表 分数段为 583-680 的数据按照组距为 10,
    #借助 python 绘制频率分布直方图和直方图的外廓曲线(概率密度曲线)
    plt.figure(figsize=(12, 8),dpi=200)
    plt.rcParams['font.sans-serif'] = ['SimHei'] # 黑体
    plt.rcParams['axes.unicode_minus'] = False #解决无法显示符号的问题
    sns.set(font='SimHei', font_scale=0.8) # 解决Seaborn中文显示问题
    sns.set_palette("hls")  #设置所有图的颜色，使用hls色彩空间
    sns.distplot(new_df583_680['高考成绩'],color="steelblue",bins=int(range_score/10))
    plt.savefig("hist.eps")
    plt.show()

    #将一分一段表提供的分数段为 583-680 的数据绘制箱线图。
    #图中体现出最值、第一四分位数、中位数、第三四分位数。
    plt.figure(figsize=(7, 5),dpi=100)
    sns.boxplot(y=new_df583_680['高考成绩'],
              capprops={'linestyle':'--','color':'red'},
              whiskerprops={'linestyle':'--','color':'red'})
    plt.savefig("box.eps")
    plt.show()

    #成绩样本分别为562 557 602 , 学校为平均分为589
    #计算小明分数的期望与标准差
    u0=589
    x = np.array([562,557,602])
    x_bar = x.mean()#样本均值
    score_S = x.std(ddof=1)#样本标准差
    print('向小明分数的数学期望: %.3f'%x_bar,', 标准差: %.3f'%score_S)
    #计算z和t
    z_alpha=1.645#z0.05
    t_alpha=2.9200#t0.05(2)
    z=(x_bar-u0)/(2/math.sqrt(3))#z统计量
    t=(x_bar-u0)/(score_S/math.sqrt(3))#t统计量
    print('z=%.3f'%z,'t=%.3f'%t)
```