

方差分析实验报告

1. 回顾并且写下单因素方差分析所基于的假设。

第一，满足独立性假设的分布：数据是随机选取的。

第二，满足方差齐性假设，这里有两种情况：第一种是方差完全相等；第二种是指方差符合经验法则——最大的均方差与最小的均方差的比不超过 2: 1。

第三，满足正态性分布的假设。

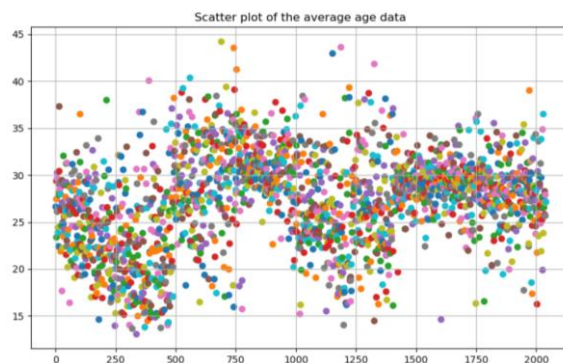
2. 我们想衡量一下平均年龄（第七列）对于类别（第二列）而言是否有着显著的差异。请为这一任务清晰地陈述原假设和备择假设。

答：原假设 (H_0)：第七列的平均年龄对于第二列的类别而言**没有显著差异**，即所有类别的平均年龄是**一致的**。

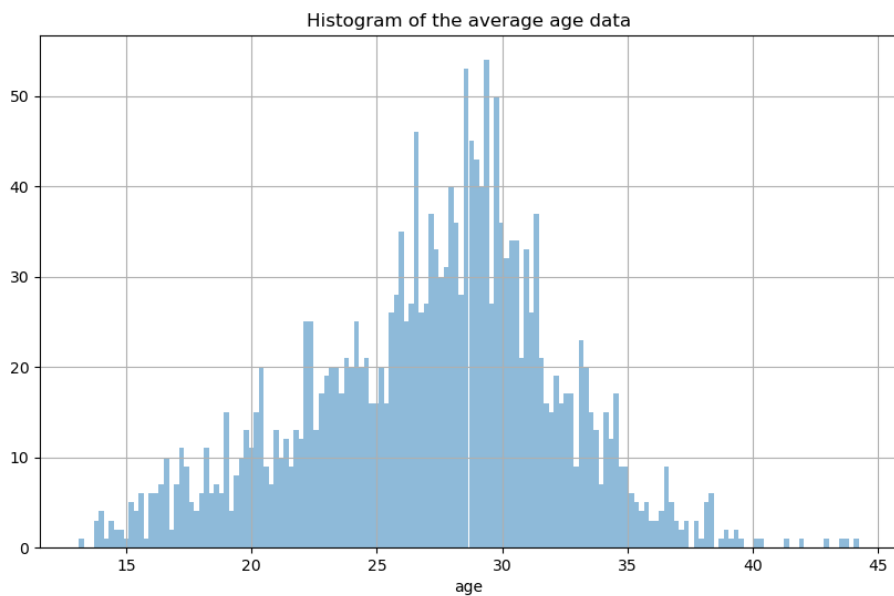
备择假设 (H_1)：第七列的平均年龄对于第二列的类别而言是**有显著差异**，即所有类别的平均年龄是**不一致的**。

3. 编程解决问题

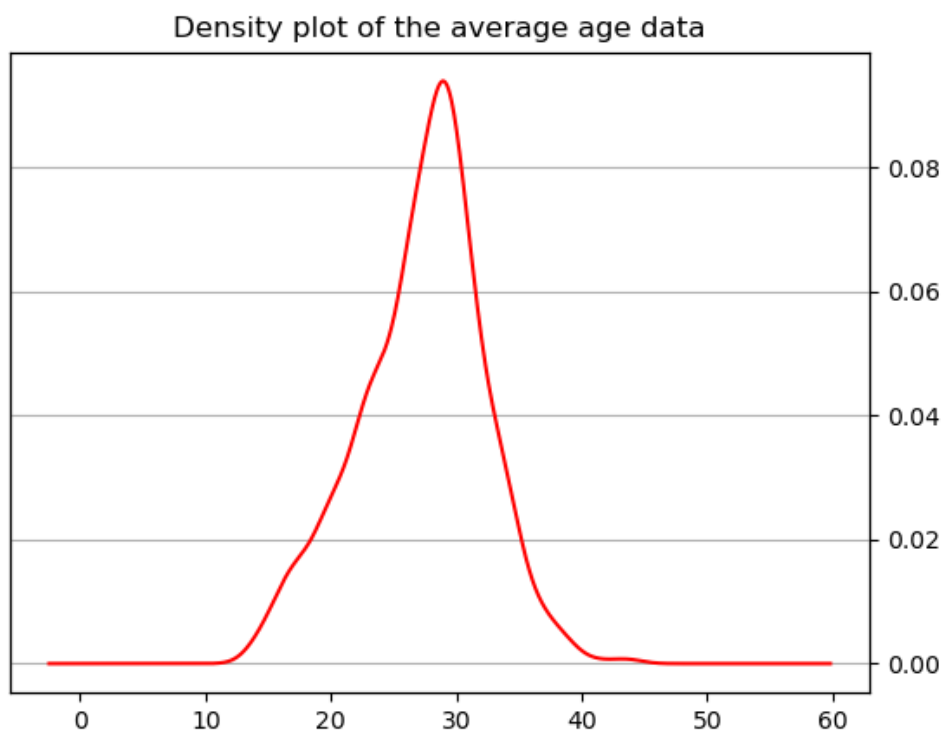
(a) 做出第七列的经验概率密度分布函数，探究这一维度的数据是否符合正态分布。



上图是平均年龄这一维度的散点图



上图是平均年龄这一维度的直方图。



上图是平均年龄这一维度的密度图。

从图中可以初步看出生成的数据与正态分布差距较大。为了得到更具说服力的结果，我们可以使用统计检验的方法。利用 `scipy.stats` 模块的 `kstest`、`normaltest`、`shapiro` 函数进行正态检验，得到的结果如下图所示。

```
scipy.stats.kstest统计检验结果: -----
KstestResult(statistic=0.05509327404455733, pvalue=8.012343213416506e-06, statistic_location=28.0181818182, statistic_sign=-1)

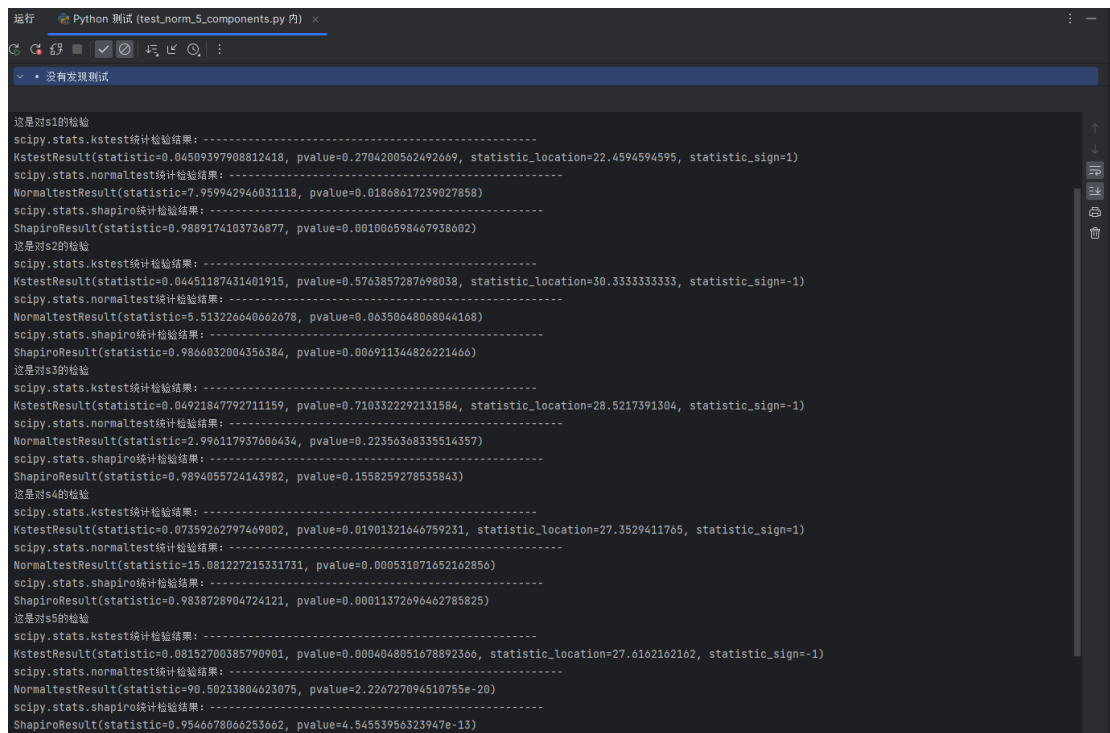
scipy.stats.normaltest统计检验结果: -----
NormaltestResult(statistic=24.479341544296894, pvalue=4.8348001102946654e-06)

scipy.stats.shapiro统计检验结果: -----
ShapiroResult(statistic=0.9883765578269958, pvalue=8.808685551808804e-12)
```

可以看出三种检验方式得到的 P-value 均小于 0.05, 这说明需要拒绝原假设, 原数据不服从正态分布。

注: 本题目的代码请见 code 文件夹的 test_normality.py 文件。

(b) 在第七列中, 根据第二列的类别标签可以分为五个部分。请测试每一个部分的正态性, 并且测试方差的齐次性。



```
运行 Python 测试 (test_norm_5_components.py 内) ×
• 没有发现测试

这是对s1的检验
scipy.stats.kstest统计检验结果: -----
KstestResult(statistic=0.04509397908812418, pvalue=0.2704200562492669, statistic_location=22.4594594595, statistic_sign=1)
scipy.stats.normaltest统计检验结果: -----
NormaltestResult(statistic=7.959942946031118, pvalue=0.01868617239027858)
scipy.stats.shapiro统计检验结果: -----
ShapiroResult(statistic=0.9889174103736877, pvalue=0.001006598467938602)
这是对s2的检验
scipy.stats.kstest统计检验结果: -----
KstestResult(statistic=0.04451187431401915, pvalue=0.5763857287698038, statistic_location=30.3333333333, statistic_sign=-1)
scipy.stats.normaltest统计检验结果: -----
NormaltestResult(statistic=5.513226648662678, pvalue=0.06358048068044168)
scipy.stats.shapiro统计检验结果: -----
ShapiroResult(statistic=0.9866032004356384, pvalue=0.006911344826221466)
这是对s3的检验
scipy.stats.kstest统计检验结果: -----
KstestResult(statistic=0.04921847792711159, pvalue=0.7103322292131584, statistic_location=28.5217391304, statistic_sign=-1)
scipy.stats.normaltest统计检验结果: -----
NormaltestResult(statistic=2.996117937606434, pvalue=0.22356368335514357)
scipy.stats.shapiro统计检验结果: -----
ShapiroResult(statistic=0.9894055724143982, pvalue=0.1558259278535843)
这是对s4的检验
scipy.stats.kstest统计检验结果: -----
KstestResult(statistic=0.07359262797469002, pvalue=0.01901321646759231, statistic_location=27.3529411765, statistic_sign=1)
scipy.stats.normaltest统计检验结果: -----
NormaltestResult(statistic=15.081227215331731, pvalue=0.000531071652162856)
scipy.stats.shapiro统计检验结果: -----
ShapiroResult(statistic=0.9838728904724121, pvalue=0.00011372696462785825)
这是对s5的检验
scipy.stats.kstest统计检验结果: -----
KstestResult(statistic=0.08152708385790901, pvalue=0.0004048051678892366, statistic_location=27.6162162162, statistic_sign=-1)
scipy.stats.normaltest统计检验结果: -----
NormaltestResult(statistic=90.50233804623075, pvalue=2.226727094510755e-20)
scipy.stats.shapiro统计检验结果: -----
ShapiroResult(statistic=0.9546678066253662, pvalue=4.54553956323947e-13)
```

在经过正态检验之后, 我们可以看到, 这五个部分中, 前三个部分的数据由于其 kstest 的 P-value 大于 0.05, 因此可以近似看作成符合正态分布, 后面两个部分的数据由于其 kstest 的 P-value 小于 0.05, 故认为不服从正态分布。

本题目的方差齐次性分析如下图所示:

```
第一部分的方差为24.2433843476337
第二部分的方差为27.22095889226124
第三部分的方差为6.517304980081701
第四部分的方差为25.992212914769162
第五部分的方差为9.114233025498347
方差的最大值为27.22095889226124，方差的最小值为6.517304980081701
比值为4.17672012825154

进程已结束，退出代码为 0
```

方差齐次性假设要求最大的方差与最小的方差的比不大于(2: 1)的平方，也就是 4: 1，这里的比值为 4.18，可以看出数据与方差齐次性假设差的不多，可以近似认为满足方差齐次性假设。

注：本题目的代码请见 code 文件夹的 test_five_components.py。

(c)对第七列的数据与第二列的类别做单因素方差分析。写出结论，所用的统计量并且对数据进行可视化分析。

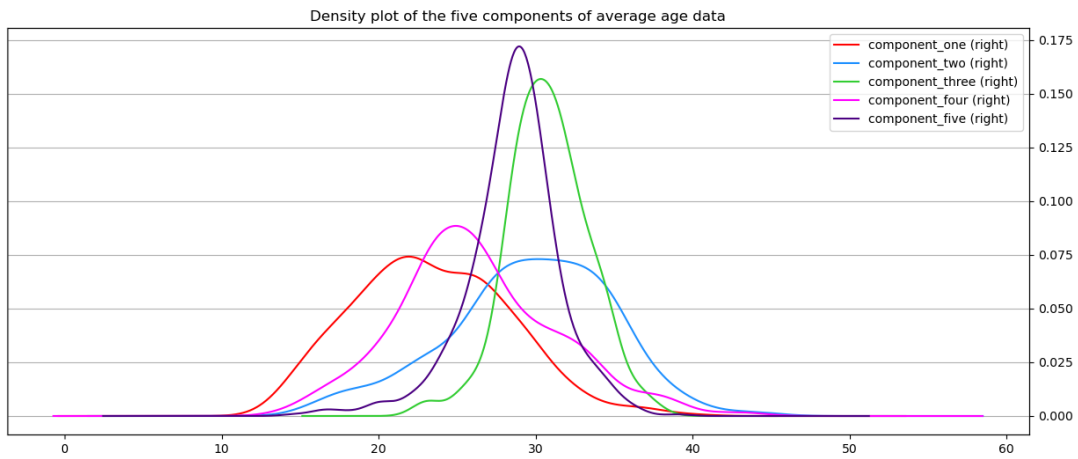
```
运行 ANOVA x
D:\anaconda3\python.exe C:\Users\guin22\Desktop\大数据分析(B)第一次作业—方差分析
\code\ANOVA.py
在这个情境下，得到的F值为171.50703270712037，分子自由度是4，分母自由度是2035
请根据上述信息，查表得到 $\alpha=0.05$ 时F-crit的值2.38
With significance level of 5%,there is highly significant evidence to reject the
null hypothesis
请根据上述信息，查表得到 $\alpha=0.01$ 时F-crit的值3.34
With significance level of 1%,there is highly significant evidence to reject the
null hypothesis

进程已结束，退出代码为 0
```

从结果中我们可以看到，F 的值很大，为 171.51，分子自由度是 4，分母自由度是 2035，查表格的时候只能查到自由度为(4, 1000)的 F-critic 值，但是自由度为(4, 2035)的 F-critic 值应当小于(4, 1000)的 F-critic 值，故得出的结论依然可信。结论为：拒绝原假设，接受备择假设，即第七列的平均年龄对于第二列的类别

而言是有显著差异，即所有类别的平均年龄是不一致的。

在下面做出五个类别的密度图：

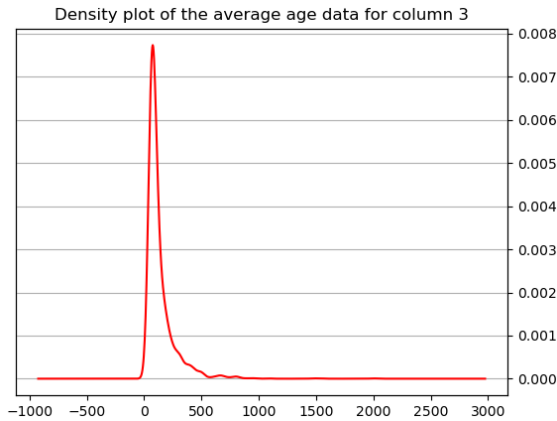


从图中也可以比较明显地看出来第七列的平均年龄对于第二列的类别而言是有显著差异，即所有类别的平均年龄是不一致的。

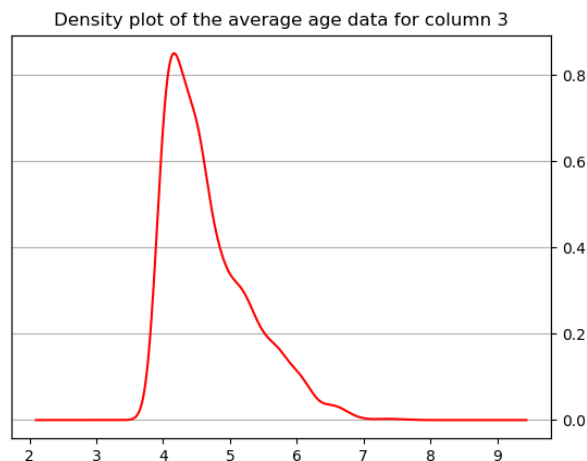
注：本题目的代码请见 `code` 文件夹的 `ANOVA.py`。在经过学习与资料查阅后,发现调用 `scipy.stats.f_oneway` 库可以实现单因素方差分析。采用调用该库的方法的代码请见 `ANOVA_invoking.py` 文件。

4. 选择另外的三列，画出每一列的经验概率密度函数，并且检验哪一列符合第一题的假设。如果不符合，那么经过对数变换后的结果是怎么样的呢？

答:首先选取第三列的数据进行检验,经验概率密度函数如下图所示，



进行对数变换后的检验结果如下图所示

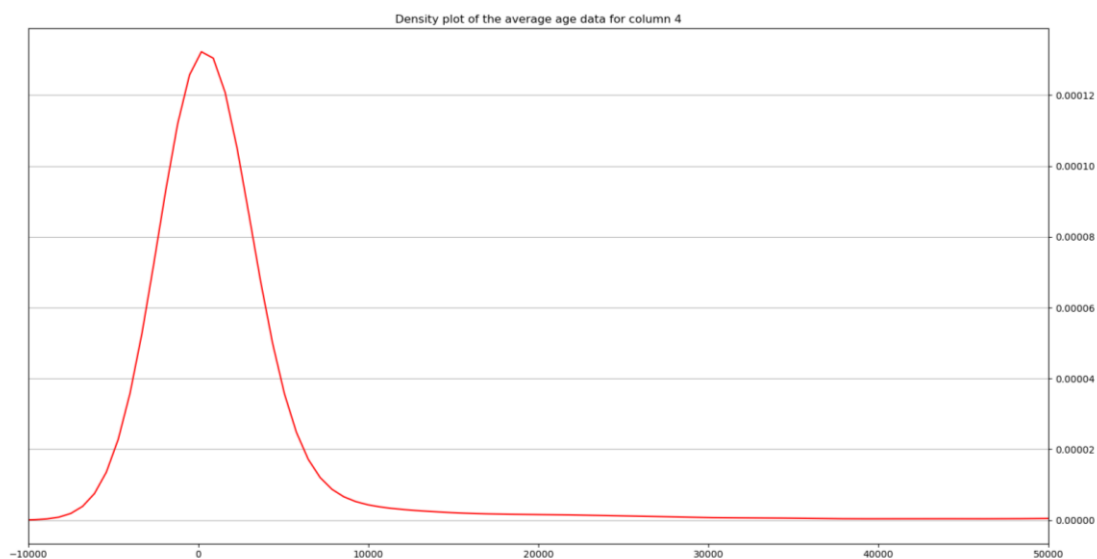


检验结果如下图所示，在没有经过对数变换之前 P-value 太低，并且方差的最大值与最小值的比为 8.31。这说明没有经过对数变换前的数据不服从正态分布的假设，同时也不符合方差齐次性假设。在经过对数变换后，P-value 有了一定的提高，方差的最大值和最小值的比为 3.35，符合方差齐次性假设，可以进行单因素方差分析。

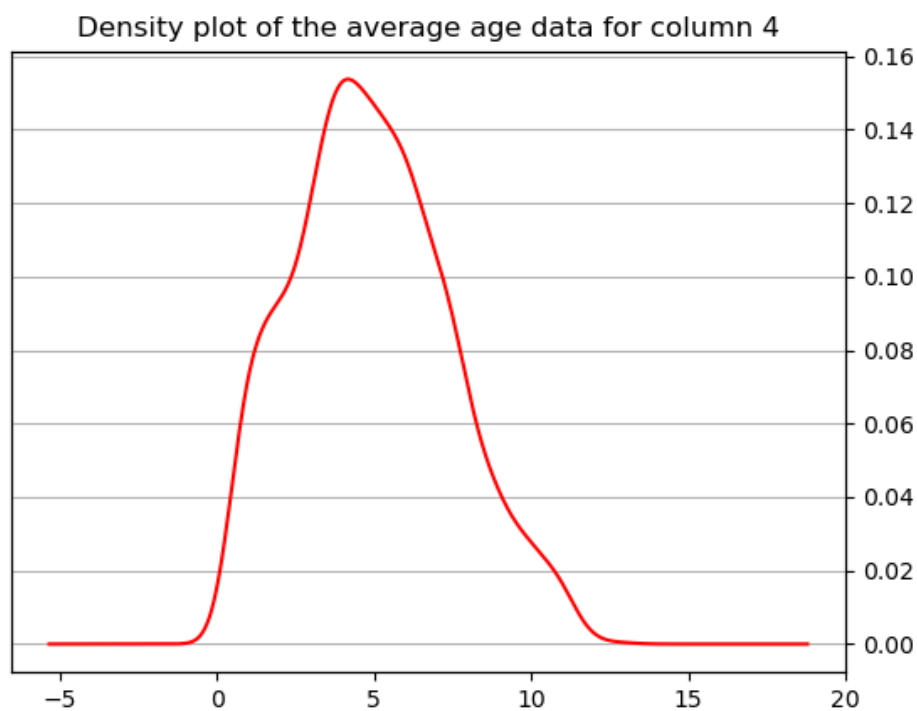
```
D:\anaconda3\python.exe "C:/Program Files/JetBrains/PyCharm Community Edition 2023.2.1/plugins/python-ce/helpers/pycharm/_jb_unittest_runner.py"
--path C:\Users\guin22\Desktop\大数据分析(B)第一次作业-方差分析\code\test_another_three_columns.py
Testing started at 15:50 ...
Launching unittests with arguments python -m unittest C:\Users\guin22\Desktop\大数据分析(B)第一次作业-方差分析\code\test_another_three_columns.py
in C:\Users\guin22\Desktop\大数据分析(B)第一次作业-方差分析\code

这是对第三列数据的检验
scipy.stats.kstest统计检验结果: -----
KstestResult(statistic=0.25911986044718277, pvalue=2.6528954203370566e-121, statistic_location=51, statistic_sign=-1)
scipy.stats.normaltest统计检验结果: -----
NormaltestResult(statistic=1959.9111409630605, pvalue=0.0)
scipy.stats.shapiro统计检验结果: -----
ShapiroResult(statistic=0.5954357385635376, pvalue=0.0)
第一部分的方差为20290.529935150484
第二部分的方差为22965.56828316611
第三部分的方差为8139.369204604921
第四部分的方差为2817.999445061044
第五部分的方差为23425.669499987547
方差的最大值为23425.669499987547，方差的最小值为2817.999445061044
比值为8.31287228996601
这是对第三列数据对数变换后的检验
scipy.stats.kstest统计检验结果: -----
KstestResult(statistic=0.12356146530376777, pvalue=1.3268462281712917e-27, statistic_location=4.624972813285251, statistic_sign=1)
scipy.stats.normaltest统计检验结果: -----
NormaltestResult(statistic=312.4557600563486, pvalue=1.4160995485104785e-68)
scipy.stats.shapiro统计检验结果: -----
ShapiroResult(statistic=0.8993570804595947, pvalue=1.2251247784222103e-34)
第一部分的方差为0.3746085269547691
第二部分的方差为0.5253245716858679
第三部分的方差为0.2994523868799953
第四部分的方差为0.15682625902164868
第五部分的方差为0.40550815147186053
方差的最大值为0.5253245716858679，方差的最小值为0.15682625902164868
比值为3.3497232859029737
```

选取第四列的数据进行检验，经验概率密度函数如下图所示，



进行对数变换后的检验结果如下图所示



检验结果如下图所示，在没有经过对数变换之前 P-value 太低，并且方差的最大值与最小值的比为 32.45。这说明没有经过对数变换前的数据不服从正态分布的假设，同时也不符合方差齐次性假设。在经过对数变换后，P-value 有了一定的提高，方差的最大值和最小值的比为 2.18，符合方差齐次性假设，可以进行单因素方差分析。

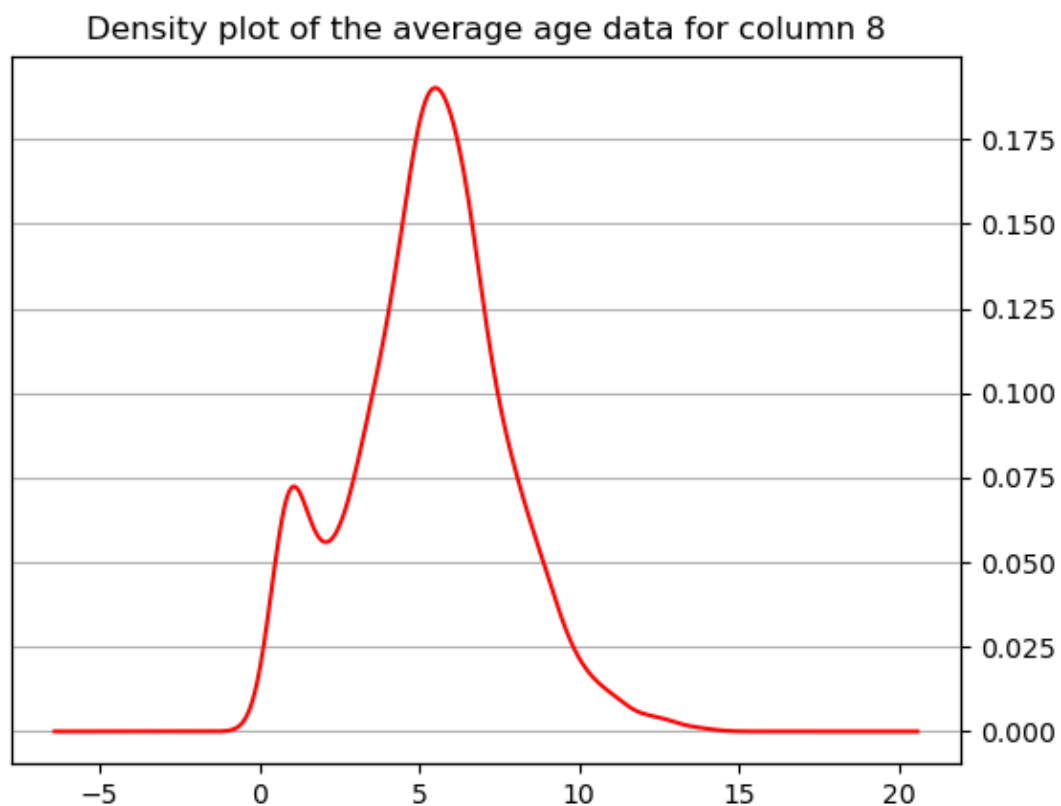
```
这是对第四列数据的检验
scipy.stats.kstest统计检验结果: -----
KstestResult(statistic=0.4145145801650457, pvalue=8.24398e-318, statistic_location=1, statistic_sign=-1)
scipy.stats.normaltest统计检验结果: -----
NormaltestResult(statistic=3947.9225448797833, pvalue=0.0)
scipy.stats.shapiro统计检验结果: -----
ShapiroResult(statistic=0.19944339990615845, pvalue=0.0)
第一部分的方差为485381018.83770496
第二部分的方差为97232493.21849494
第三部分的方差为54935533.594453186
第四部分的方差为19893064.17452834
第五部分的方差为14957304.419518624
方差的最大值为485381018.83770496，方差的最小值为14957304.419518624
比值为32.45110249974615
这是对第四列数据对数变换后的检验
scipy.stats.kstest统计检验结果: -----
KstestResult(statistic=0.04261426380592745, pvalue=0.0011748179598725912, statistic_location=0.6931471805599453, statistic_sign=-1)
scipy.stats.normaltest统计检验结果: -----
NormaltestResult(statistic=57.550476824509, pvalue=3.1847309032145124e-13)
scipy.stats.shapiro统计检验结果: -----
ShapiroResult(statistic=0.982257604598999, pvalue=3.067489584416454e-15)

Ran 0 tests in 0.000s

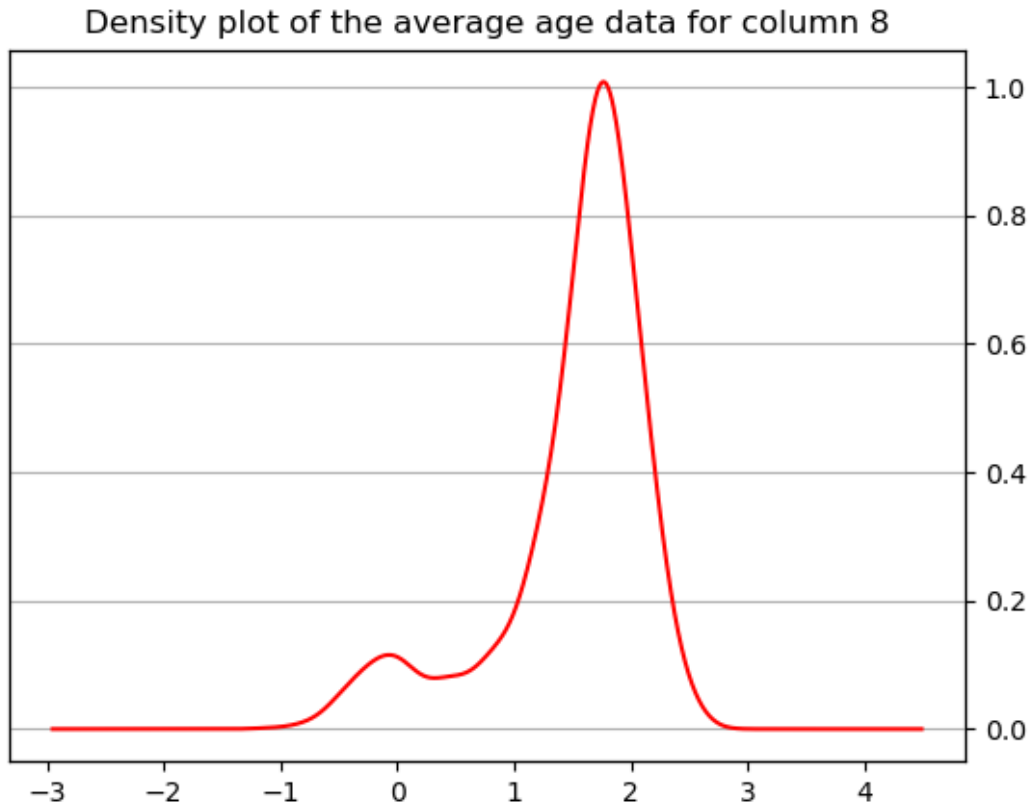
OK
第一部分的方差为8.465636118581726
第二部分的方差为6.984521121383291
第三部分的方差为5.601202556345517
第四部分的方差为5.036996331961192
第五部分的方差为3.875407644081407
方差的最大值为8.465636118581726，方差的最小值为3.875407644081407
比值为2.1844504878114175

进程已结束，退出代码为 0
```

选取第八列的数据进行检验，经验概率密度函数如下图所示，



进行对数变换后的检验结果如下图所示



检验结果如下图所示，在没有经过对数变换之前 P-value 较低，并且方差的最大值与最小值的比为 2.52。在经过对数变换后，P-value 大大降低，方差的最大值和最小值的比为 8.98。因此，未经过对数变换前的数据符合方差齐次性假设，可以进行单因素方差分析。

```

这是对第八列数据的检验
scipy.stats.kstest统计检验结果: -----
KstestResult(statistic=0.04081852530117569, pvalue=0.002168420496151733, statistic_location=4.57646316072, statistic_sign=-1)
scipy.stats.normaltest统计检验结果: -----
NormaltestResult(statistic=2.206176786494377, pvalue=0.331844632748964)
scipy.stats.shapiro统计检验结果: -----
ShapiroResult(statistic=0.9853329658508301, pvalue=1.2719259510646924e-13)
第一部分的方差为3.0812020687897386
第二部分的方差为5.931338012754853
第三部分的方差为3.539070356811993
第四部分的方差为6.709172410499027
第五部分的方差为2.6575214009839976
方差的最大值为6.709172410499027，方差的最小值为2.6575214009839976
比值为2.524597697694863
这是对第八列数据对数变换后的检验
scipy.stats.kstest统计检验结果: -----
KstestResult(statistic=0.16433558947672117, pvalue=1.3004712467146890e-48, statistic_location=1.461378758398443, statistic_sign=-1)
scipy.stats.normaltest统计检验结果: -----
NormaltestResult(statistic=503.4020153047732, pvalue=4.871259739039577e-110)
scipy.stats.shapiro统计检验结果: -----
ShapiroResult(statistic=0.8538890480995178, pvalue=6.6764024553677314e-40)

第一部分的方差为0.14510425749523234
第二部分的方差为0.24087540844813846
Ren 0 tests in 0.000s
第三部分的方差为0.07949498377532982
第四部分的方差为0.713576680108795
OK
第五部分的方差为0.12619395095713118
方差的最大值为0.713576680108795，方差的最小值为0.07949498377532982
比值为8.976373680703157

进程已结束，退出代码为 0

```

注:本题目的代码请见 `code` 文件夹的 `test_another_three_columes.py`。

5. 对于非正态的数据如何进行单因素方差分析?

(a) 找出并列举出可能的解决方案

(b) 在你所选择的三列上进行单因素方差分析。这些列之间的特征有显著的差异吗? 将你的结果可视化出来。

答:

(a):一般来说,单因素方差分析对数据的正态性有一定的耐受能力,不满足正态性分布假设的情况,但是满足方差齐次性假设(包括松弛后的)的情况,可以直接进行单因素方差分析。

但是,如果正态性分布假设与方差齐次性假设均不满足的情况下,需要对数据进行一定的转换,经过查阅资料、分析探索,转换方法大致如下:

- 1) 平方根变换:对变量取根号
- 2) 对数变换:对变量取自然对数(ln)或者以10为底的对数(log10)。
- 3) 倒数变换:对变量取倒数
- 4) Box-Cox 变换:是一种广义的幂变换,其一般形式通常如下所示:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(y) & \lambda = 0 \end{cases}$$

$\lambda=0$ 时相当于对数变换, $\lambda=2$ 时等同于平方变换, $\lambda=1/2$ 时等同于平方根变换, $\lambda=-1$ 时等同于倒数变换。

(b):在本部分中,我将先对对数变换后的第三、四列,原始的第八列进行与第二列的类别进行单因素的方差分析。然后对对数变换后的第三、四列和原始的第八列进行单因素方差分析。

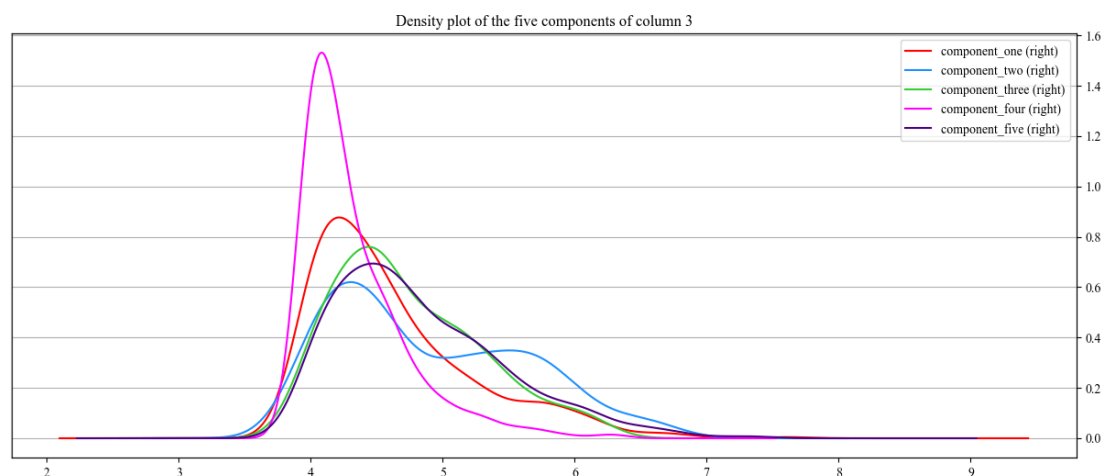
Part1:对数变换后的第三列和第二列的类别进行单因素方差分析

原假设 (H0): 对数变换后的第三列对于第二列的类别而言没有显著差异, 即所有类别的群人数是一致的。

备择假设 (H1): 对数变换后的第三列对于第二列的类别而言是有显著差异, 即所有类别的群人数是不一致的。

```
在这个情境下, 得到的F值为59.7315086425722, 分子自由度是4, 分母自由度是2035  
请根据上述信息, 查表得到 $\alpha=0.05$ 时F-crit的值2.38  
With significance level of 5%, there is highly significant evidence to reject the null hypothesis  
请根据上述信息, 查表得到 $\alpha=0.01$ 时F-crit的值3.34  
With significance level of 1%, there is highly significant evidence to reject the null hypothesis
```

从结果中我们可以看到, F 的值很大, 为 59.73, 分子自由度是 4, 分母自由度是 2035, 查表格的时候只能查到自由度为 (4, 1000) 的 F-critic 值, 但是自由度为 (4, 2035) 的 F-critic 值应当小于 (4, 1000) 的 F-critic 值, 故得出的结论依然可信。结论为: 拒绝原假设, 接受备择假设, 即第三列的群人数对于第二列的类别而言是有显著差异, 即所有类别的群人数是不一致的。



从图中也可以看出来, 第三列的群人数对于第二列的类别而言是有显著差异, 尤其是第四类与其他类别相比差异很大, 即所有类别的群人数是不一致的。

Part2:对数变换后的第四列和第二列的类别进行单因素方差分析

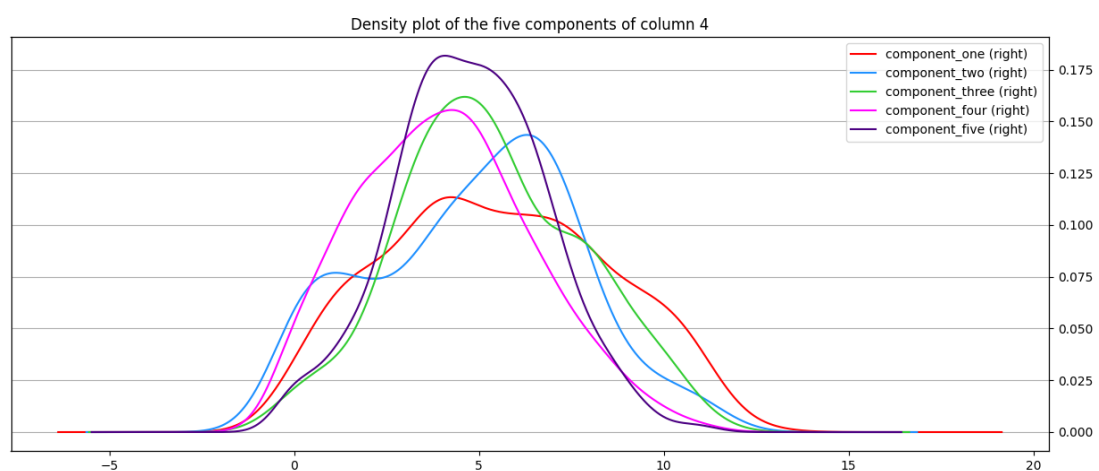
原假设 (H0): 对数变换后的第四列对于第二列的类别而言没有显著差异, 即所有类别的消息数是一致的。

备择假设 (H1): 对数变换后的第四列对于第二列的类别而言是有显著差异, 即所有类别的消息数是不一致的。

```
在这个情境下, 得到的F值为20.011754697728296, 分子自由度是4, 分母自由度是2035  
请根据上述信息, 查表得到 $\alpha=0.05$ 时F-crit的值2.38  
With significance level of 5%, there is highly significant evidence to reject the null hypothesis  
请根据上述信息, 查表得到 $\alpha=0.01$ 时F-crit的值3.34  
With significance level of 1%, there is highly significant evidence to reject the null hypothesis
```

从结果中我们可以看到, F 的值很大, 为 20.01, 分子自由度是 4, 分母自由度是 2035, 查表格的时候只能查到自由度为 (4, 1000) 的 F-critic 值, 但是自由度为 (4, 2035) 的 F-critic 值应当小于 (4, 1000) 的 F-critic 值, 故得出的结论依然可信。结论为: 拒绝原假设, 接受备择假设, 即第四列的消息数对于第二列的类别而言是有显著差异, 即所有类别的消息数是不一致的。

在下面做出五个类别的密度图:



从图中也可以看出来, 第四列的消息数对于第二列的类别而言是有显著差异, 即所有类别的消息数是不一致的。

Part3:第八列和第二列的类别进行单因素方差分析

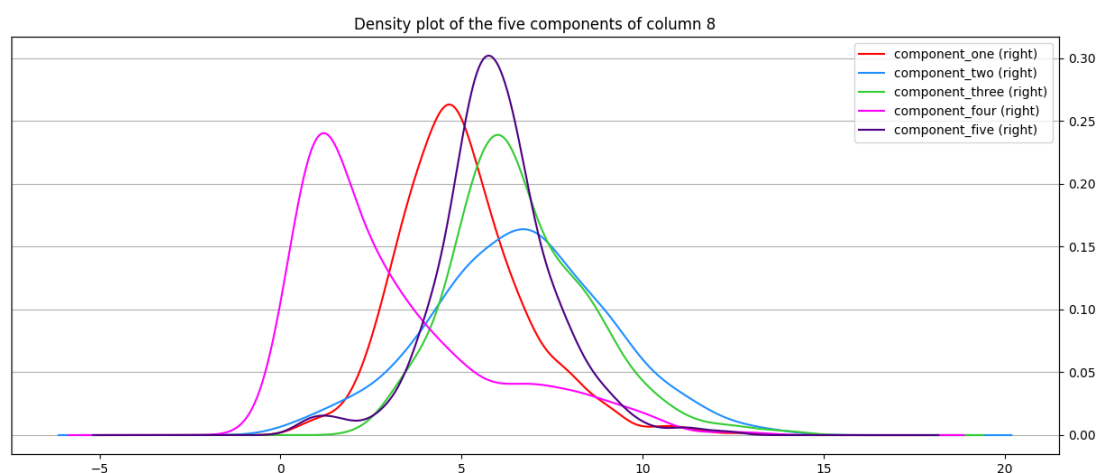
原假设 (H0): 第八列对于第二列的类别而言没有显著差异, 即所有类别的年龄差是一致的。

备择假设 (H1): 第八列对于第二列的类别而言是有显著差异, 即所有类别的年龄差是不一致的。

```
在这个情境下, 得到的F值为200.24221601479468, 分子自由度是4, 分母自由度是2035  
请根据上述信息, 查表得到 $\alpha=0.05$ 时F-crit的值2.38  
With significance level of 5%, there is highly significant evidence to reject the null hypothesis  
请根据上述信息, 查表得到 $\alpha=0.01$ 时F-crit的值3.34  
With significance level of 1%, there is highly significant evidence to reject the null hypothesis
```

从结果中我们可以看到, F 的值很大, 为 200.24, 分子自由度是 4, 分母自由度是 2035, 查表格的时候只能查到自由度为 (4, 1000) 的 F-critic 值, 但是自由度为 (4, 2035) 的 F-critic 值应当小于 (4, 1000) 的 F-critic 值, 故得出的结论依然可信。结论为: 拒绝原假设, 接受备择假设, 即第八列的年龄差对于第二列的类别而言是有显著差异, 即所有类别的年龄差是不一致的。

在下面做出五个类别的密度图:



从图中也可以看出来, 第八列的年龄差对于第二列的类别而言是有显著差异, 即所有类别的年龄差是不一致的。

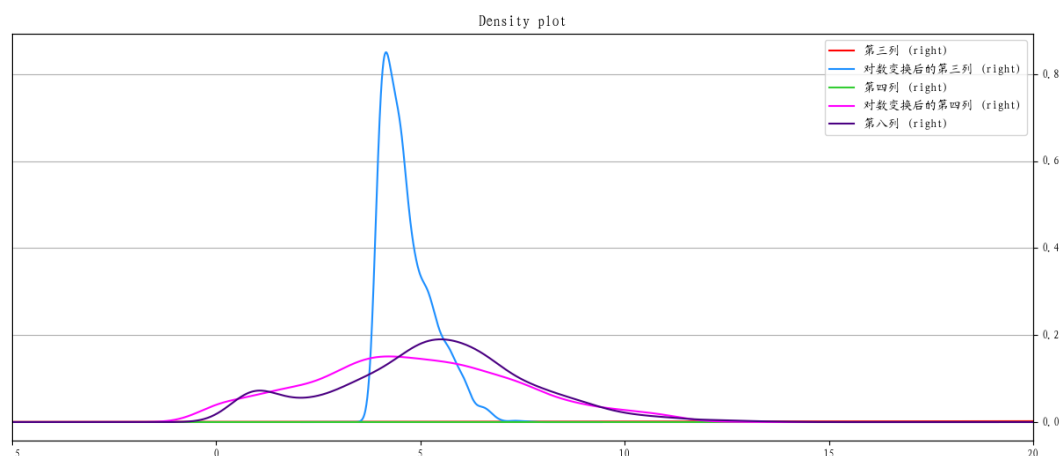
Part4:对数变换后的第三、四列和原始的第八列进行单因素方差分析

原假设 (H0): 对数变换后的第三、四列和原始的第八列之间没有显著性差异。

备择假设 (H1): 对数变换后的第三、四列和原始的第八列之间是有显著性差异。

```
在这个情境下,得到的F值为45.16316491406741,分子自由度是2,分母自由度是6117  
请根据上述信息,查表得到 $\alpha=0.05$ 时F-crit的值3.00  
With significance level of 5%,there is highly significant evidence to reject the null hypothesis  
请根据上述信息,查表得到 $\alpha=0.01$ 时F-crit的值4.63  
With significance level of 1%,there is highly significant evidence to reject the null hypothesis
```

从结果中我们可以看到, F 的值很大, 为 45.16, 分子自由度是 2, 分母自由度是 6117, 查表格的时候只能查到自由度为 (2, 1000) 的 F-critic 值, 但是自由度为 (2, 6117) 的 F-critic 值应当小于 (4, 1000) 的 F-critic 值, 故得出的结论依然可信。结论为: 拒绝原假设, 接受备择假设, 即对数变换后的第三、四列和原始的第八列之间是有显著性差异。



从图中也可以明显看出来: 对数变换后的第三、四列和原始的第八列之间是有显著性差异, 也就是说群人数、消息数和年龄差这三个特征之间是有显著性差异。

下面对这三个特征两两进行单因素方差分析。

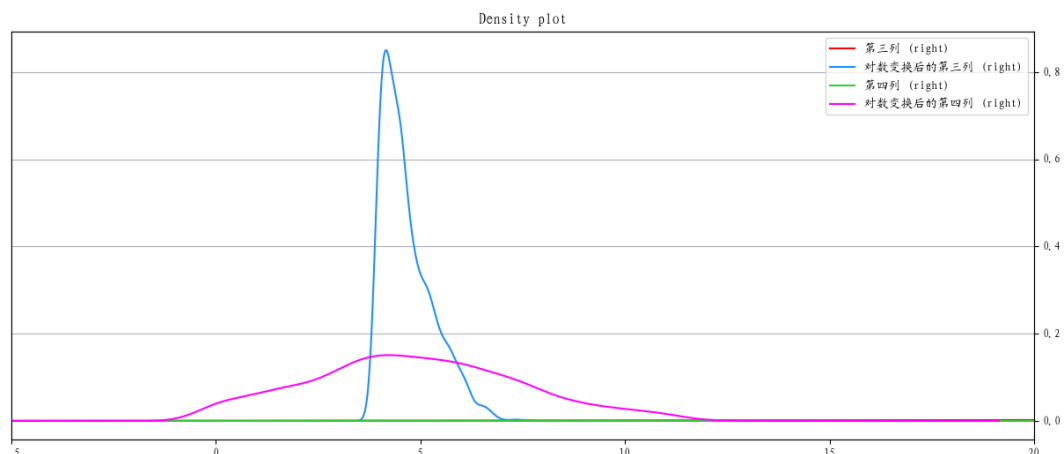
Part5:对数变换后的第三列和对数变换后的第四列进行单因素方差分析

原假设 (H0): 对数变换后的第三和对数变换后的第四列之间没有显著性差异, 即群人数和消息数这两个特征之间没有显著性差异。

备择假设 (H1): 对数变换后的第三和对数变换后的第四列之间有显著性差异, 即群人数和消息数这两个特征之间有显著性差异。

```
在这个情境下, 得到的F值为10.008889122687886, 分子自由度是1, 分母自由度是4078  
请根据上述信息, 查表得到 $\alpha=0.05$ 时F-crit的值3.85  
With significance level of 5%, there is highly significant evidence to reject the null hypothesis  
请根据上述信息, 查表得到 $\alpha=0.01$ 时F-crit的值6.66  
With significance level of 1%, there is highly significant evidence to reject the null hypothesis
```

从结果中我们可以看到, F 的值为 10.01, 分子自由度是 1, 分母自由度是 4078, 查表格的时候只能查到自由度为 (1, 1000) 的 F-critic 值, 但是自由度为 (1, 4078) 的 F-critic 值应当小于 (1, 1000) 的 F-critic 值, 故得出的结论依然可信, 为: 拒绝原假设, 接受备择假设, 对数变换后的第三列和对数变换后的第四列之间有显著性差异, 即群人数和消息数这两个特征之间有显著性差异。



从图中也可以明显看出来: 对数变换后的第三、四列之间是有显著性差异, 也就是说群人数、消息数之间是有显著性差异。

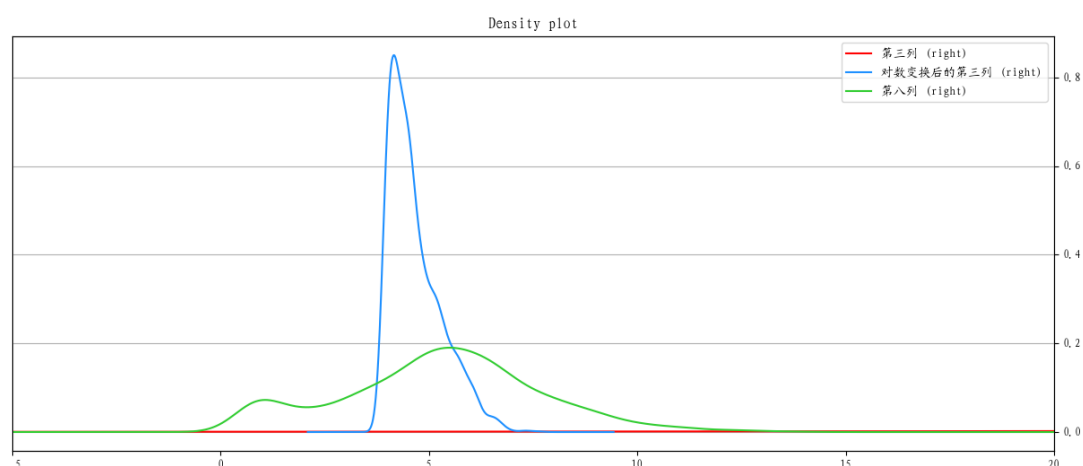
Part6:对数变换后的第三列和第八列进行单因素方差分析

原假设 (H0): 对数变换后的第三列和第八列之间没有显著性差异, 即群人数和年龄差这两个特征之间没有显著性差异。

备择假设 (H1): 对数变换后的第三列和第八列之间有显著性差异, 即群人数和年龄差这两个特征之间有显著性差异。

```
在这个情境下, 得到的F值为118.02576911930151, 分子自由度是1, 分母自由度是4078  
请根据上述信息, 查表得到 $\alpha=0.05$ 时F-crit的值3.85  
With significance level of 5%, there is highly significant evidence to reject the null hypothesis  
请根据上述信息, 查表得到 $\alpha=0.01$ 时F-crit的值6.66  
With significance level of 1%, there is highly significant evidence to reject the null hypothesis
```

从结果中我们可以看到, F 的值为 118.03, 分子自由度是 1, 分母自由度是 4078, 查表格的时候只能查到自由度为 (1, 1000) 的 F-critic 值, 但是自由度为 (1, 4078) 的 F-critic 值应当小于 (1, 1000) 的 F-critic 值, 故得出的结论依然可信。结论为: 拒绝原假设, 接受备择假设, 对数变换后的第三列和对数变换后的第四列之间有显著性差异, 即群人数和年龄差这两个特征之间有显著性差异。



从图中也可以明显看出来: 对数变换后的第三列和第八列之间是有显著性差异, 也就是说群人数和年龄差之间是有显著性差异。

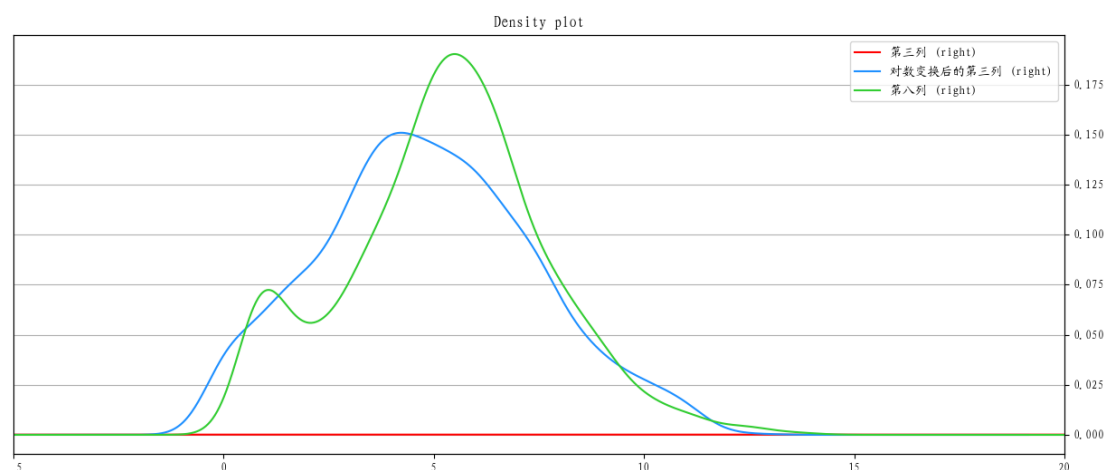
Part7:对数变换后的第四列和第八列进行单因素方差分析

原假设 (H0): 对数变换后的第四列和第八列之间没有显著性差异, 即消息数和年龄差这两个特征之间没有显著性差异。

备择假设 (H1): 对数变换后的第四列和第八列之间有显著性差异, 即消息数和年龄差这两个特征之间有显著性差异。

```
在这个情境下, 得到的F值为28.35251863981804, 分子自由度是1, 分母自由度是4078  
请根据上述信息, 查表得到 $\alpha=0.05$ 时F-crit的值3.85  
With significance level of 5%, there is highly significant evidence to reject the null hypothesis  
请根据上述信息, 查表得到 $\alpha=0.01$ 时F-crit的值6.66  
With significance level of 1%, there is highly significant evidence to reject the null hypothesis
```

从结果中我们可以看到, F 的值为 28.35, 分子自由度是 1, 分母自由度是 4078, 查表格的时候只能查到自由度为 (1, 1000) 的 F-critic 值, 但是自由度为 (1, 4078) 的 F-critic 值应当小于 (1, 1000) 的 F-critic 值, 故得出的结论依然可信。结论为: 拒绝原假设, 接受备择假设, 对数变换后的第四列和第八列之间有显著性差异, 即消息数和年龄差这两个特征之间有显著性差异。



从图中也可以明显看出来: 对数变换后的第四列和第八列之间是有显著性差异, 也就是说消息数和年龄差之间是有显著性差异。

总结

经过上述实验与分析，得到如下结论：

- 第三列的群人数对于第二列的类别而言是有显著差异的
- 第四列的消息数对于第二列的类别而言是有显著差异的
- 第八列的年龄差对于第二列的类别而言是有显著差异的
- 群人数、消息数和年龄差这三个特征之间是有显著性差异的，并且群人数、消息数之间是有显著性差异的；群人数和年龄差之间是有显著性差异的；消息数和年龄差之间是有显著性差异的。

注：本题目的代码请见 `code` 文件夹的 `non_normal_ANOVA_C3-C2.py`、`non_normal_ANOVA_C4-C2.py`、`non_normal_ANOVA_C8-C2.py`、`non_normal_ANOVA_C3-C4-C8.py`、`non_normal_ANOVA_C3-C4.py`、`non_normal_ANOVA_C3-C8.py`、`non_normal_ANOVA_C4-C8.py` 这七个文件。

附录

在本部分中，将展示上述实验通过调用 `scipy.stats.f_oneway` 库实现单因素方差分析的结果。经比较，调用该库的结果与前述实验结果完全一致。

第七列的数据与第二列的类别做单因素方差分析：

```
C:\Users\guin22\anaconda3\envs\nn\python.exe C:\Users\guin22\Desktop\大数据分析(B)第一次作业—方差分析\code\ANOVA_invoking.py
F-value: 171.50703270711966
P-value: 1.0820916064752822e-126
```

注：本结果的代码请见 `ANOVA_invoking.py` 文件

对数变换后的第三列和第二列的类别进行单因素方差分析：

```
C:\Users\guin22\anaconda3\envs\nn\python.exe C:\Users\guin22\Desktop\大数据分析 (B) 第一次作业—方差分析\code\non_normal_ANOVA_invoking.py
result of C3-C2 is as follows
F-value: 59.73150864257041
P-value: 9.493707242830655e-48
```

对数变换后的第四列和第二列的类别进行单因素方差分析：

```
C:\Users\guin22\anaconda3\envs\nn\python.exe C:\Users\guin22\Desktop\大数据分析 (B) 第一次作业—方差分析\code\non_normal_ANOVA_invoking.py
result of C4-C2 is as follows
F-value: 20.011754697728104
P-value: 3.5306714268771483e-16
```

第八列和第二列的类别进行单因素方差分析：

```
C:\Users\guin22\anaconda3\envs\nn\python.exe C:\Users\guin22\Desktop\大数据分析 (B) 第一次作业—方差分析\code\non_normal_ANOVA_invoking.py
result of C8-C2 is as follows
F-value: 200.24221601479533
P-value: 6.3170461722256334e-145
```

对数变换后的第三、四列和原始的第八列进行单因素方差分析：

```
C:\Users\guin22\anaconda3\envs\nn\python.exe C:\Users\guin22\Desktop\大数据分析 (B) 第一次作业—方差分析\code\non_normal_ANOVA_invoking.py
result of C3-C4-C8 is as follows
F-value: 45.1631649140689
P-value: 3.3829223673907525e-20
```

对数变换后的第三列和对数变换后的第四列进行单因素方差分析：

```
C:\Users\guin22\anaconda3\envs\nn\python.exe C:\Users\guin22\Desktop\大数据分析 (B) 第一次作业—方差分析\code\non_normal_ANOVA_invoking.py
result of C3-C4 is as follows
F-value: 10.00888912268803
P-value: 0.0015693164300009381
```

对数变换后的第三列和第八列进行单因素方差分析：

```
C:\Users\guin22\anaconda3\envs\nn\python.exe C:\Users\guin22\Desktop\大数据分析 (B) 第一次作业—方差分析\code\non_normal_ANOVA_invoking.py
result of C3-C8 is as follows
F-value: 118.02576911930521
P-value: 4.012553557349869e-27
```

对数变换后的第四列和第八列进行单因素方差分析：

```
C:\Users\guin22\anaconda3\envs\nn\python.exe C:\Users\guin22\Desktop\大数据分析  
（B）第一次作业—方差分析\code\non_normal_ANOVA_invoking.py  
result of C4-C8 is as follows  
F-value: 28.352518639819174  
P-value: 1.0656232707338136e-07
```

注：上述7个结果的代码请见non_normal_ANOVA_invoking.py文件。