作业 1：方差分析

（大数据分析课程报告）

姓　　名：肖文韬

学　　号：2020214245

二〇二〇年十月二十五日

第 1 章　HW1 ANOVA

1. (**5 points**) Recall and write down the assumptions which one-way ANOVA are based on.

   ANOVA 是**方差分析（ANalysis Of VAriance**）的缩写，又称 F 检验, 用于三个及以上样本均值差别的显著性检验。它基于以下假设：

   - 所有的数据都是随机采样的。
   - 每个组的方差是一样的（同调性），各组的标准差中最大和最小的比例不超过 2：1。
   - 残差是正态分布。

2. (**5 points**) Focus on two columns columns: Category (Col[2]) and Average Age (Col[7]). Taking feature Average Age as an example, we want to measure whether the average age varied significantly across the categories. Clearly state the null (H0) and the alternative (H1) hypotheses for this task.

   相同 category 的样本属于同一个组。

   - H0: 不是所有组的均值都相等。
   - H1: 所有组的均值都相等（$\mu_1 = \mu_2 = \cdots$)。

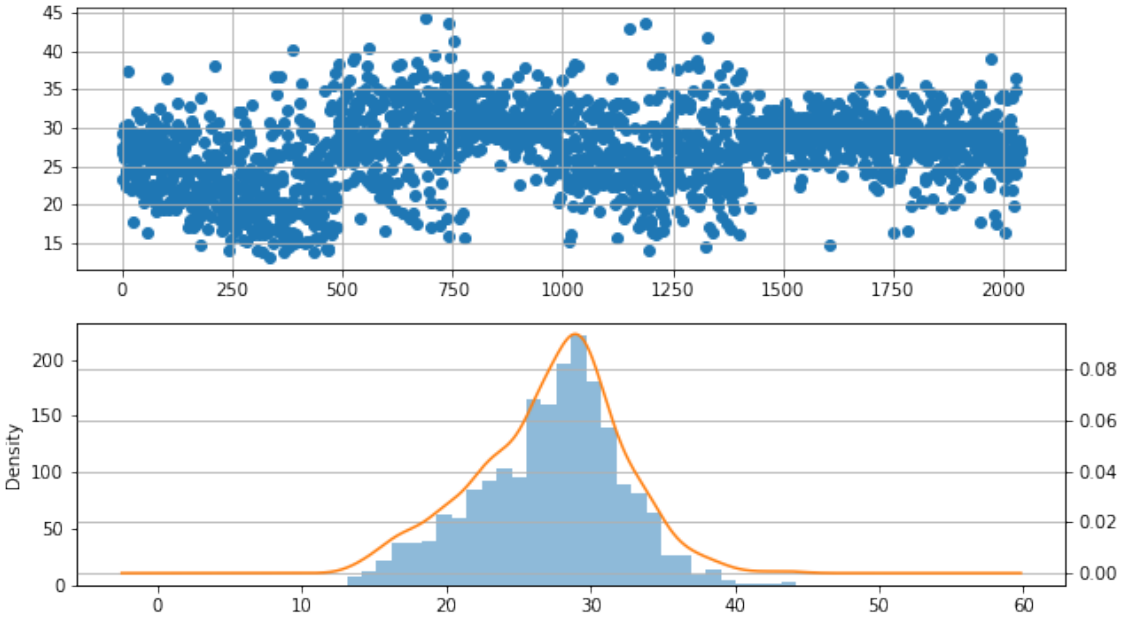

图 1.1　Col[7] 散点图及经验概率密度函数

3. Use your favorite statistics analysis software, like Matlab , R, Excel Excel, SPSS or ⋯

   (a) (**10 points**) Draw the empirical probability density funfction of Col[7], i.e. the empirical pdf of average age. Does the data in this dimension follow Gaussian distribution? Test normality of Col[7].

   Col[7] 的散点图和经验概率密度函数如图 1.1所示。原假设（H0）假设数据符合正态分布。设置显著性水平 $\alpha = 0.01$，通过计算 p-value $= 4.8348 \times 10^{-6} < 0.01$，所以拒绝假设 H_0, col[7] 不符合高斯分布.

   (b) (**10 points**) In Col[7], there are 5 components divided by category labels labels. We denote the data in Col[7] with category i (where $i = 1, \cdots, 5$) as Col[7|categoty=i]. Test the normality of each components and test the homogeneity of variances.

   设置显著性水平 $\alpha = 0.01$，计算得，组 1，2 和 3 符合正态性，组 4 和 5 不符合正态性。同时，计算各组的标准差，最大的与最小的之比为 2.0437 略大于 2：1，所以不符合同调性。具体计算过程见附件代码。

	Source	SS	df	MS	F	P
0	Between	12782.918190	4	3195.73	171.087	3.02173e-126
1	Within	38011.705928	2035	18.679		
2	Total	50794.624118	2039			

图 1.2　Col[7] 方差分析（ANOVA）

(c) (**20 points**) Do the one one-way ANOVA test for Col[7] with categories in Col[2]. Write down your conclusion, supporting statistics, and visualize your data which inspire the process.

结论：在显著性水平 $\alpha = 0.01$ 下，因为计算得到 p-value $\approx 10^{-126} \ll 0.01$，所以拒绝原假设 H_0: 各组的方差相同。其计算过程和具体统计量在图 1.2中展示，具体计算过程见附件代码。

4. (**15 points**) Choose another 3 columns, draw the empirical pdf of each feature columns and test which column follows these assumptions in question 1? How about their corresponding log transformation? 消息数在取 log 之后从原来不符合同掉性变为符合同调性，并且原来五个组均不符合高斯分布，变换之后有两个符合高斯分布了。

5. How to do one one-way ANOVA with the non-normal data data?

(a) (**10 points**) Find and list the possible solutions set.

(b) (**25 points**) Do the one one-way ANOVA on the 3 columns you choose choose. Do these feature columns vary significantly? Visualize the results.

参考文献