清华大学电子工程系

大数据分析 B 作业 1

作者: 罗雁天

学号: 2018310742

日期: 2018 年 11 月 7 日

1 Recall and Write down the assumption which one-way ANOVA are based on.

ANSWER

- The data are randomly sampled;
- The variance of each group are assumed equal;
- The residual are normly distributed;

2 Focus on two 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.

ANSWER

The null(H_0) and the alternative (H_1) are:

- H_0: there is no interaction between the average age and the categories;
- H_1: the average age varies significantly across the categories.

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

3.1 Draw the empirical probability density function 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]

ANSWER

The empirical pdf of Col[7] is in Fig.3.1

We also draw the Gaussian distribution in Fig.3.1 for comparison. Intuitively, we can find that the data in Col[7] does not follow Gaussian distribution.

Normality test:

We use Anderson-Darling normality test to test the normality. More detailed, we use function **ad.test()** in **nortest** package in **R** to test the normality. The assumption is:

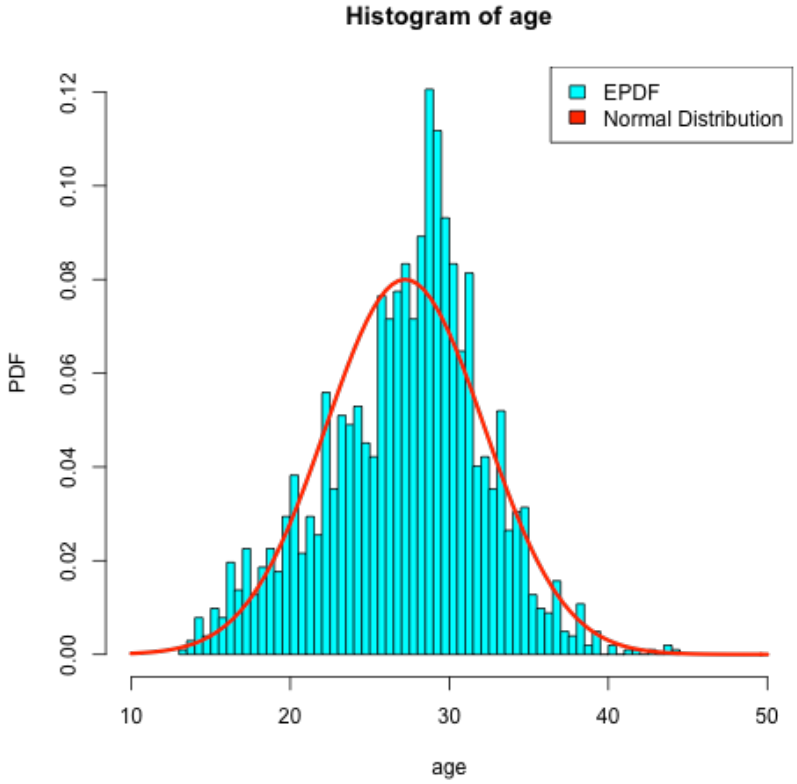

Fig. 3.1. Empirical pdf of Col[7] data

- H_0: the data are normly distributed.
- H_1: the data are not normly distributed.

The result of the test is in Fig.3.2, we can see $A = 9.3826, p - value < 2.2e - 16$, therefore, the data in Col[7] does not follow Gaussian distribution.

```
        Anderson-Darling normality test

data:  age_norm
A = 9.3826, p-value < 2.2e-16
```

Fig. 3.2. The result of normality test

The code for this question is in "code/3a.R"

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

ANSWER

Using the Anderson-Darling normality test like section 3.1, we can get the results(at 5% level) in Table 3.1

Table. 3.1. The results of normality test for each catebory

Category	1	2	3	4	5
p-value	0.02513	0.003444	0.2225	1.088e-05	2.2e-16
Normality	No	No	Yes	No	No

To test the homogeneity of variances, our assumption is:

- H_0: The variances of each group are equal
- H_1: The variances of each group are not equal.

We use Bartlett test of homogeneity of variances, the result is in Fig.3.3, we can see $p\_value < 2.2 \times 10^{-16}$, therefore we reject the null assumption, i.e., the variances of each group is not equal.

```
        Bartlett test of homogeneity of variances

data:  平均年齢 by 群类别
Bartlett's K-squared = 276.32, df = 4, p-value < 2.2e-16
```

Fig. 3.3. The result of the test of homogeneity of variances

The code for this question is in "code/3b.R"

3.3 Do the 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.

ANSWER

Using the function **anova1** in **MATLAB**, we can get the result of the one-way ANOVA test as Fig.3.4

With significance level of 5%, because $p\_value = 1.08209 \times 10^{-126} < 0.05$, wo reject the null assumption. Thus the average age varies significantly across the categories.

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	12782.9	4	3195.73	171.51	1.08209e-126
Error	37918.6	2035	18.63		
Total	50701.5	2039			

Fig. 3.4. The result of the ANOVA test

We also draw the box plot of each group as Fig.3.5. Intuitively, we can find that the center lines of the boxes have large differences, which indicate the average age varies significantly across the categories.

The code for this question is in "code/3c.R" and "code/problem3c.m"

4 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?

ANSWER We choose Col[6](性别比), Col[13](夜聊比例), Col[14](图片比例) data in this question. The empirical pdf of each feature columns as in Fig.4.1 and the empirical pdf of their log transformation as in Fig.4.2. In Fig.4.2, we do not show the epdf of their log transformation if the data=0.

Then we test normality (Anderson-Darling normality test) and the homogeneity of variances (Barlett Test) of the data in Col[6], Col[13], Col[14] and their log transformation. With significance level of 5%, the result is in Table4.1

Table. 4.1. The results of normality test and homogeneity of variances test for Col[6,13,14]

Results	Col[6]	Col[13]	Col[14]	log(Col[6])	log(Col[13])	log(Col[14])
Normality p-value	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16
Normality	No	No	No	No	No	No
Homogeneity p-value	2.2e-16	2.2e-16	2.2e-16	2.2e-16	0.0003192	0.504
Homogeneity	No	No	No	No	No	Yes

The code for this question is in "code/4.R"

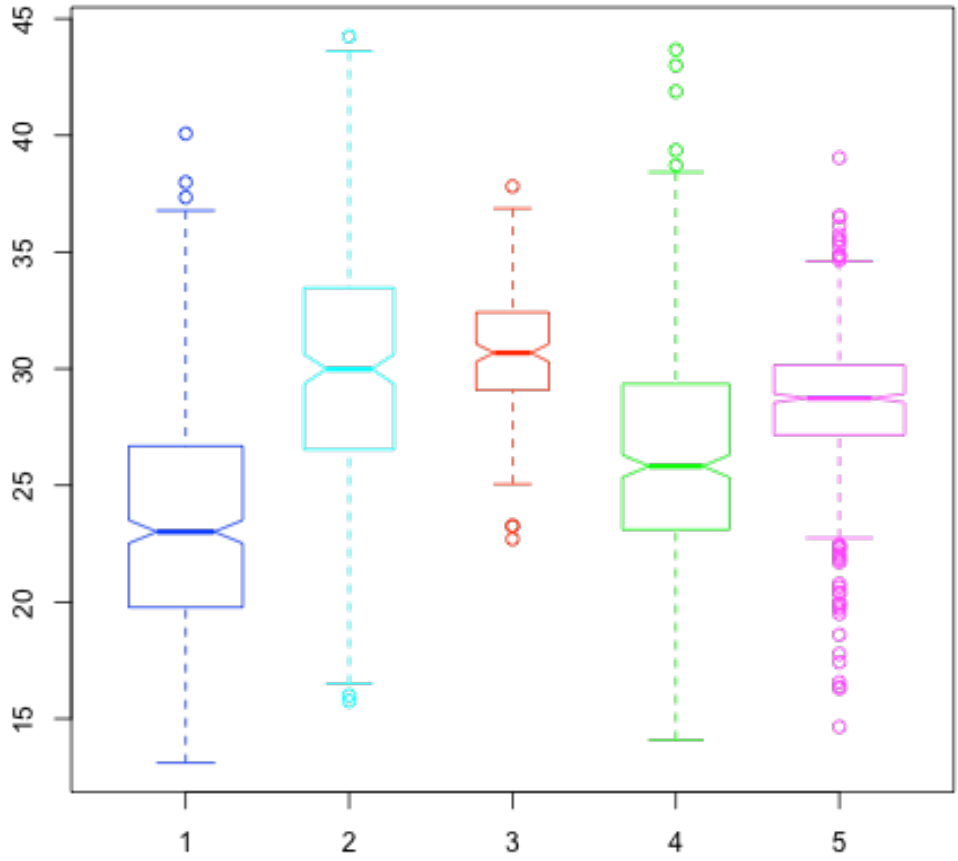

Fig. 3.5. The box plot of each group

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

5.1 Find and list the possible solutions set.

ANSWER

- We can use some algorithms to transform non-normal data into the Gaussian distributed shape.
- We can use the non-parametric Kurskal-Wallis Test, which does not require the normality assumption.
- Last but not least, the one-way ANOVA can tolerate non-normal data (skewed or kurtotic) with a small effect on the Type1 error rate, thus we can use Type1 error rate to do the one-way ANOVA.

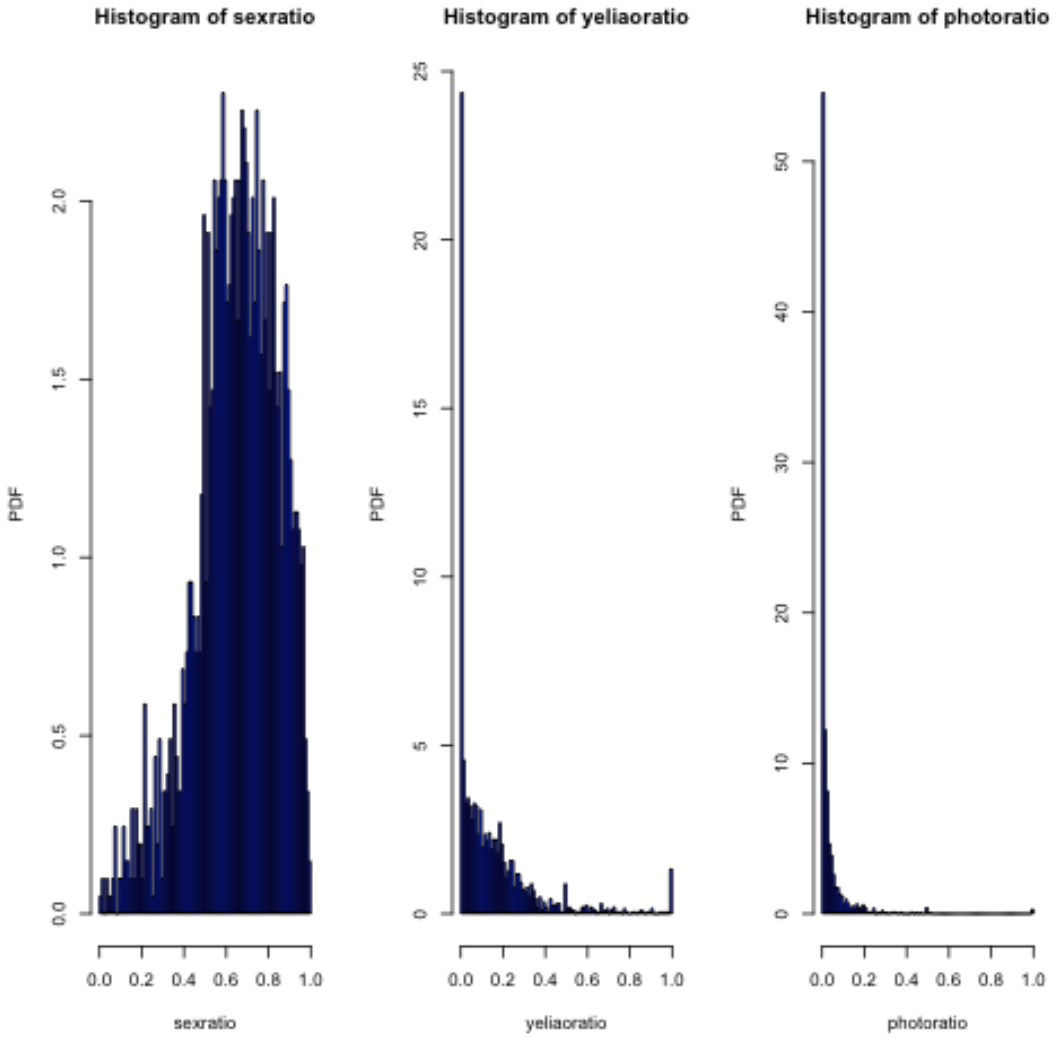

Fig. 4.1. The Empirical pdf of Col[6], Col[13] and Col[14]

5.2 Do the one-way ANOVA on the 3 columns you choose. Do these feature columns vary significantly? Visualize the results.

ANSWER

Like the method in section3.3, we can get the results as followed:

For the data in col[6], the result is in Fig.5.1 and Fig.5.2.With significance level of 5%, because $p\_value = 2.53208 \times 10^{-43} < 0.05$, wo reject the null assumption. Thus the average age varies significantly across the categories. And from the box plot, we can find that the center lines of the boxes have large differences, which indicate the average age varies significantly across the categories.

For the data in col[13], the result is in Fig.5.3 and Fig.5.4.With significance level of 5%, because $p\_value = 9.893 \times 10^{-23} < 0.05$, wo reject the null assumption. Thus the average age varies significantly across the categories. And from the box plot, we can find that the center lines

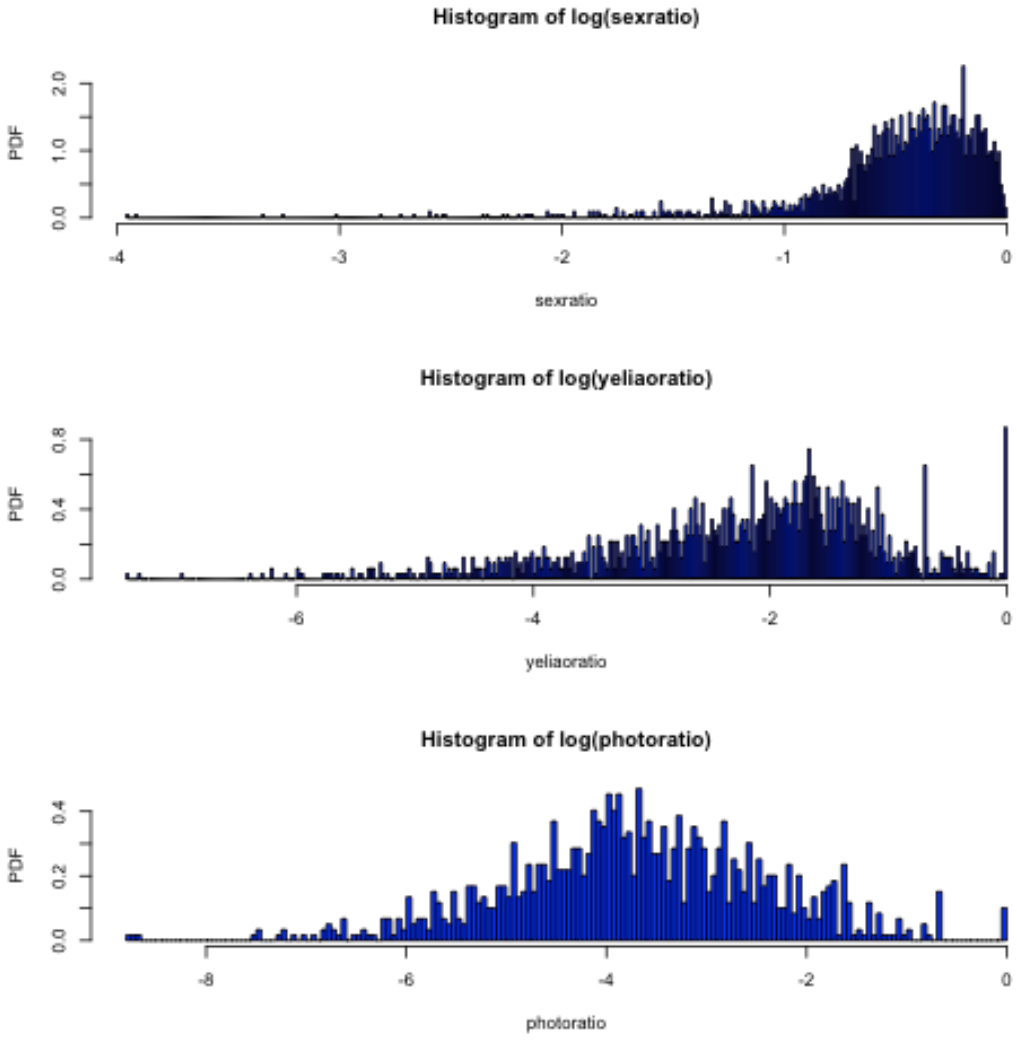

Fig. 4.2. The Empirical pdf of log transformation of Col[6], Col[13] and Col[14]

of the boxes have large differences, which indicate the average age varies significantly across the categories.

For the data in col[14], the result is in Fig.5.5 and Fig.5.6.With significance level of 5%, because $p\_value = 0.006 < 0.05$, wo reject the null assumption. Thus the average age varies significantly across the categories. And from the box plot, we can find that the center lines of the boxes have large differences, which indicate the average age varies significantly across the categories.

The code for this question is in "code/5b.R" and "code/problem5b.m"

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	7.1647	4	1.79117	54.02	2.53208e-43
Error	67.48	2035	0.03316		
Total	74.6447	2039			

Fig. 5.1. The result of ANOVA test on Col[6]

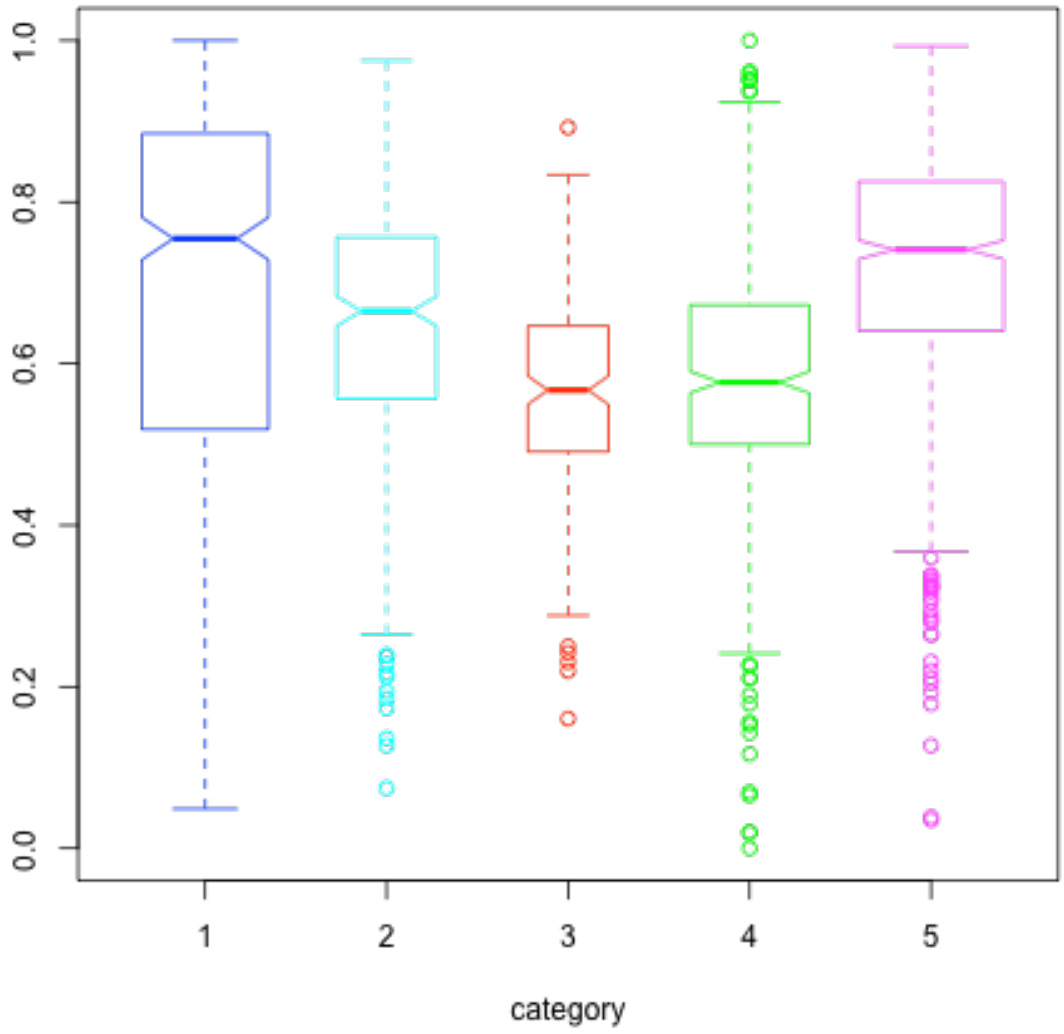

Fig. 5.2. The box plot of each group on Col[6]

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	3.7215	4	0.93038	28.08	9.893e-23
Error	67.4316	2035	0.03314		
Total	71.1531	2039			

Fig. 5.3. The result of ANOVA test on Col[13]

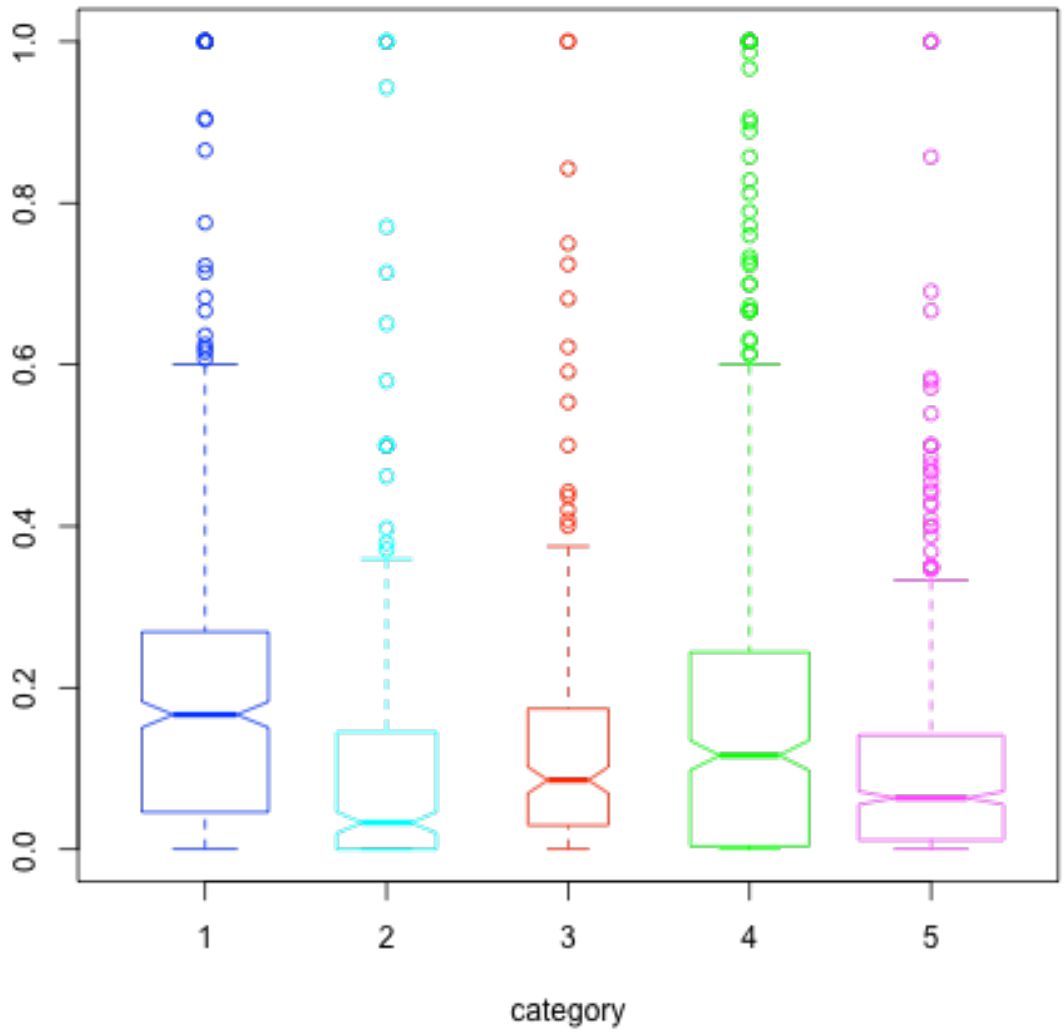

Fig. 5.4. The box plot of each group on Col[13]

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	0.1286	4	0.03216	4.96	0.0006
Error	13.2057	2035	0.00649		
Total	13.3344	2039			

Fig. 5.5. The result of ANOVA test on Col[14]

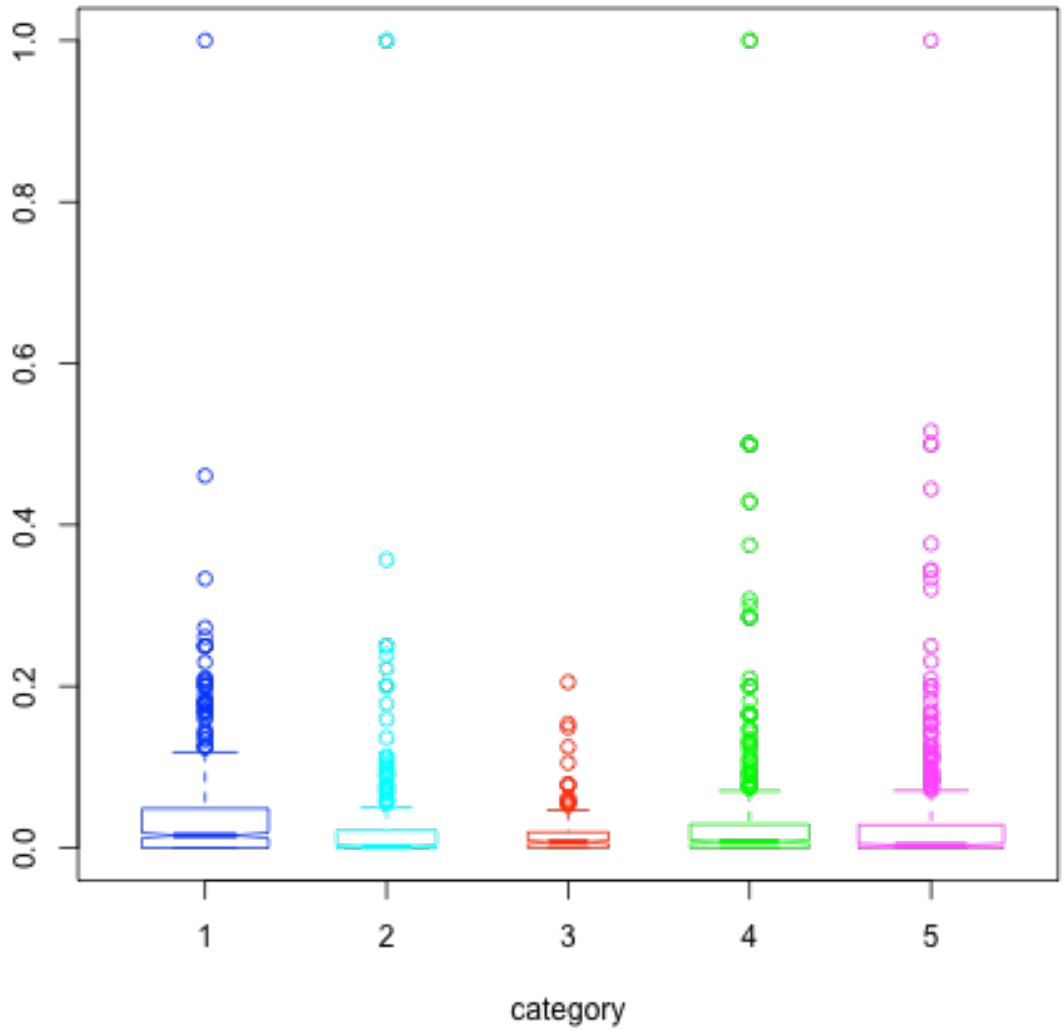

Fig. 5.6. The box plot of each group on Col[14]

6 Redo the ANOVA test in question 3 c) by sampling 10% data (i.e. around 200 groups). Repeat 10 times and compute the mean and standard deviation of the supporting statistics (F value). Compare at least two sampling strategies. Which sampling method is more stable? How are the results compared to the results without sampling? Why?

ANSWER

Here we use Simple Random Sampling, Stratified Random Sampling and Systematic Random Sampling for comparison. The results of 10 times F_values, means and standard deviations are shown in Table6.1

Table. 6.1. F values, means and stds of 10 times for 3 sampling strategies

Samplings	1	2	3	4	5	6	7	8	9	10	mean	std
Simple	31.78	17.11	17.37	34.08	15.97	12.18	16.42	15.49	12.65	29.97	20.30	8.26
Stratified	17.29	22.13	20.39	12.39	15.83	15.49	18.48	17.19	9.69	23.62	17.25	4.23
Systematic	28.44	11.38	11.38	11.38	13.58	16.65	20.65	25.24	9.58	25.24	17.35	6.99

From Table6.1, we can find that Stratified Random Sampling is more stable because it has smallist standard deviation. Compared to the results without sampling in Fig.3.4(F0=171.51), the sampling results are about 10% of the F0. Recall the F formula:

$$F = \frac{MS_{Between}}{MS_{Within}} = \frac{SS_b \cdot db_w}{SS_w \cdot df_b} \tag{6.1}$$

In the calculation of sampling, SS_b and SS_w calculations only consider 10% of total data, df_b is still 4 and $db_w = 204 - 5 = 199$ (nearly 10% of 2040-5). Therefore, the F-stats with sampling is nearly 10% of the F-stats without sampling.

The code for this question is in "code/problem6.m"

7 Choose any two categories, and classify them by logistical regression, or you can try multi-label classification on all categories.

ANSWER

First we choose category 1 and category 2 for logistical regression. We assign $label = 0$ if the group is in category 1 else assign $label = 1$. We use cross_entropy in equation 7.1 as our loss function and we use Gradient Descent to optimize weight w in logistical regression. We set

$learningrate = 0.1$ and $iter_num = 2000$, we get the accurcay curve and the loss curve in Fig.7.1. And the last accuracy is 76.9%.

$$loss = -\sum_{i=1}^{n}(y_i \log p_i + (1 - y_i) \log(1 - p_i)) \tag{7.1}$$

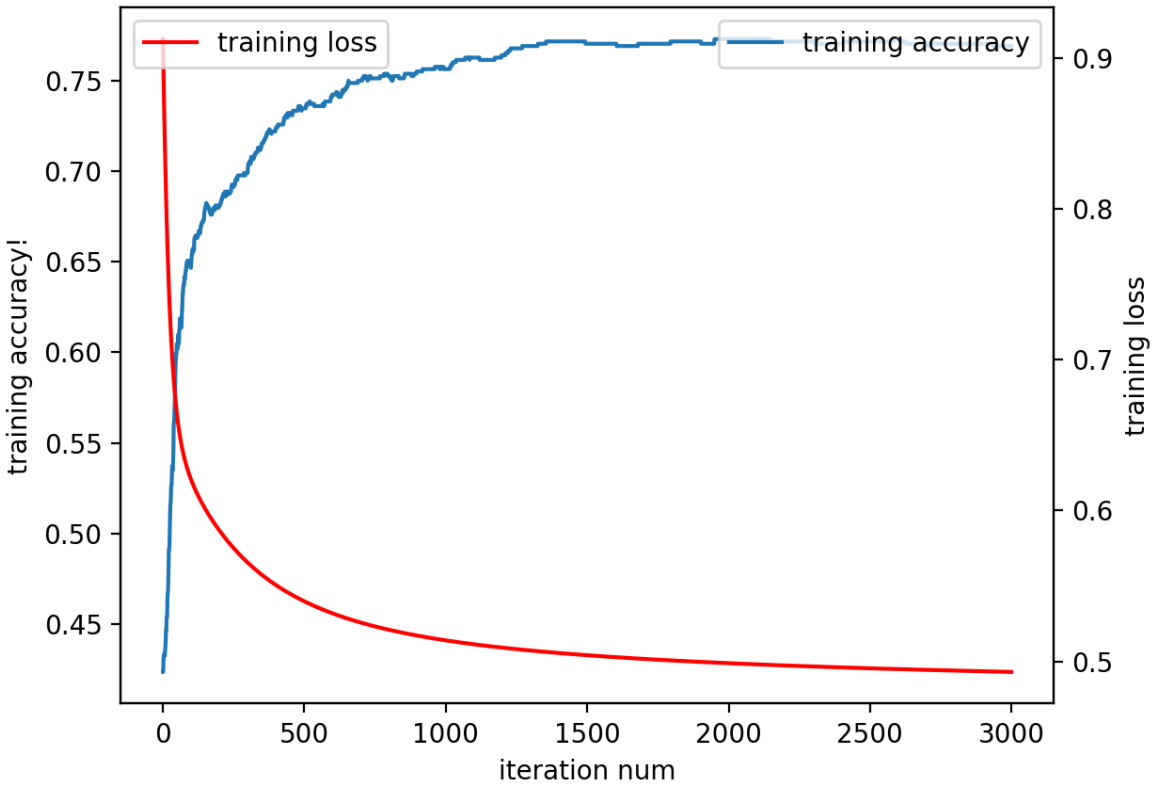

Fig. 7.1. The training accuracy and training loss of logistical regression on category 1 and coregory 2

Then we only take the Col[7](平均年齢), Col[8](年齢差) of category 1, 2 and we classify them by logistical regression and we draw the classification border of the two categories as shown in Fig.7.2. And the last accuracy is 75.8%, the accurcay curve and the loss curve in Fig.7.3. From the result, 75.8% is nearly 76.9%, we can find the two categories mostly differ in age.

The code for this question is in "code/problem7.py"

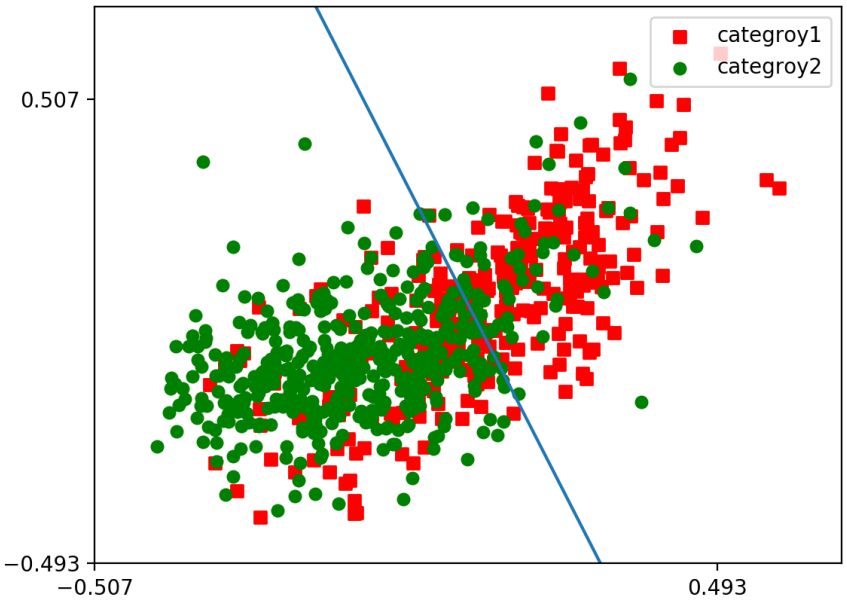

Fig. 7.2. The classification border of category 1,2 on Col[7], Col[8]

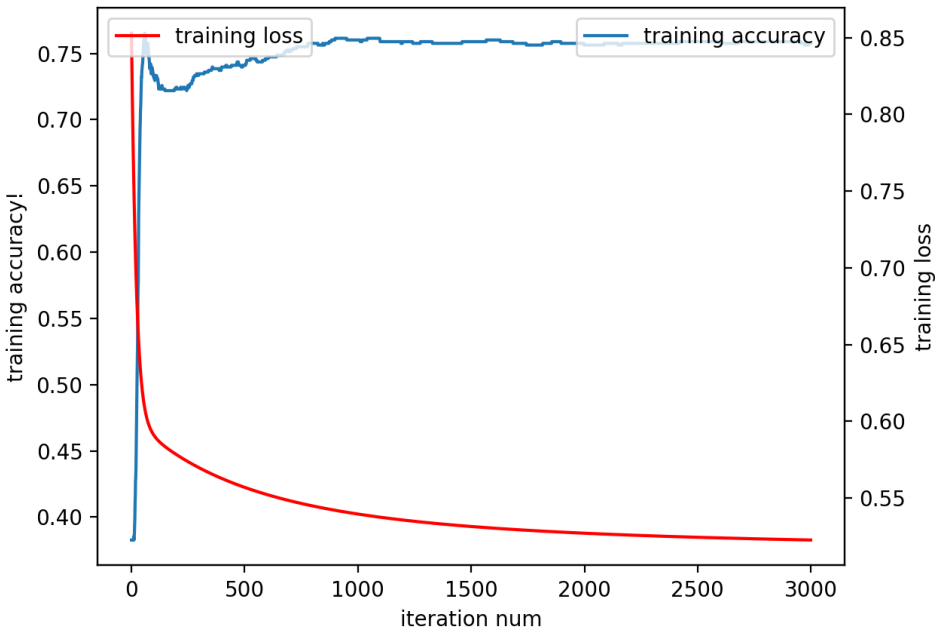

Fig. 7.3. The training accuracy and training loss of logistical regression on category 1 and coregory 2 with only Col[7], Col[8]