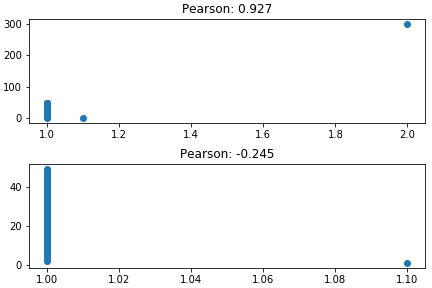
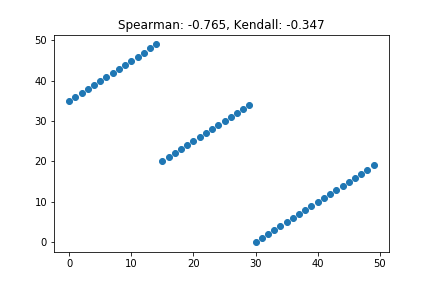
# Question 1

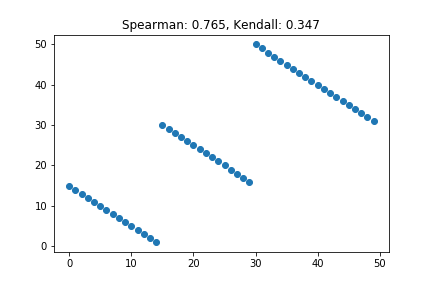
Problem 1: Data in which the but where n-1 points can be selected so that for the vectors restricted to those we have Pearson correlation



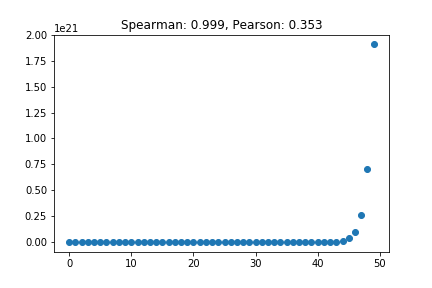
Problem 2:



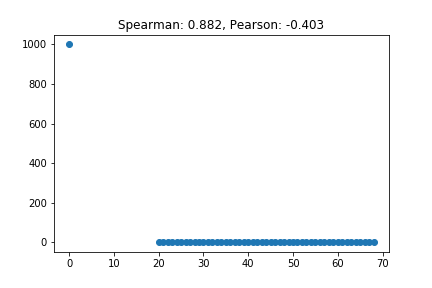
Problem 3:



Problem 4:



Problem 5:



Problem 6:

This situation is impossible.

Explanation:

Assume the situation is possible. Therefore, we can determine that a dataset of 50 points exists so that:

and

In addition, we know that:

From formula, we conclude that . Therefore, the number of concordant pairs is smaller than the number of discordant pairs, and we can deduce that and are independent. In addition, the ranks of and are also independent, resulting in so that , in contradiction to our assumption.

# Question 2

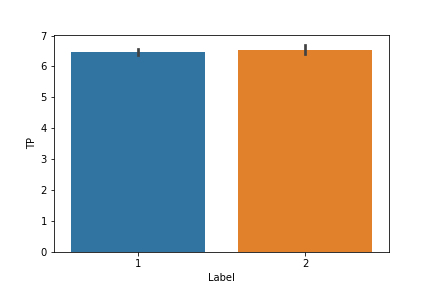
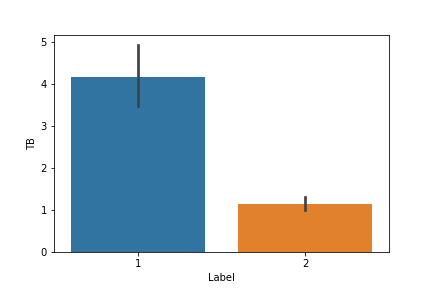
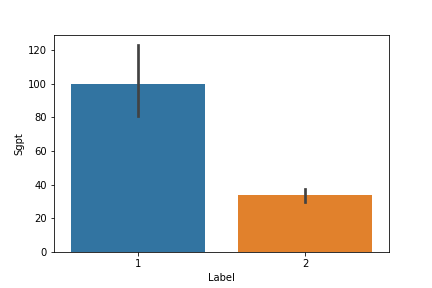
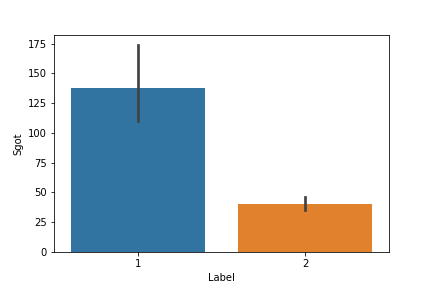
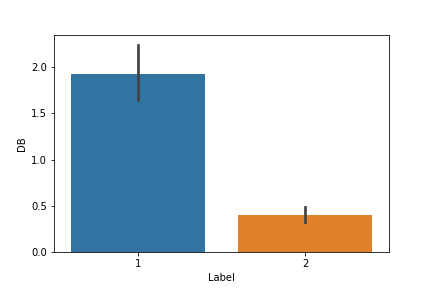
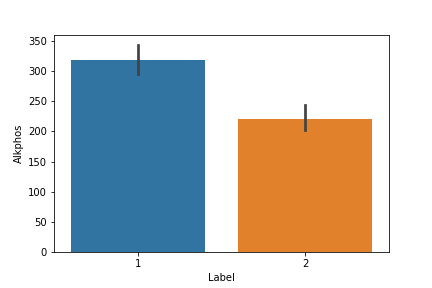
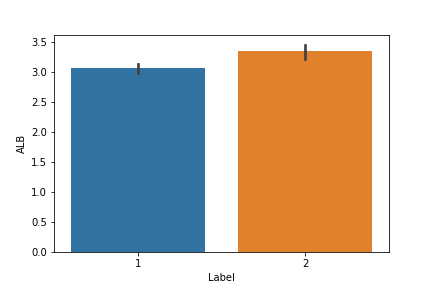
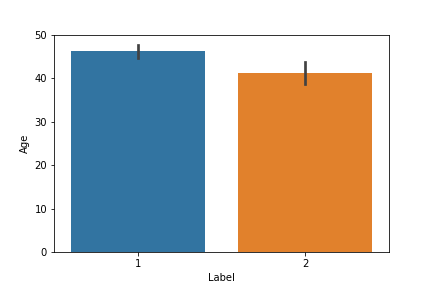
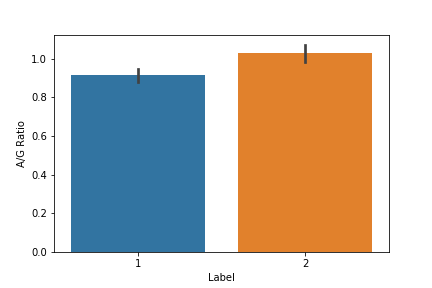
Note: we're using the term "Label" instead of "Selector Field"

**A/G Ratio** has 4 missing values, which are populated by the median of the attribute – 0.93

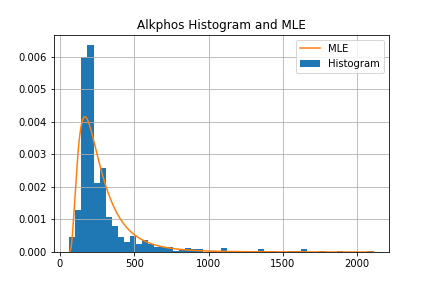
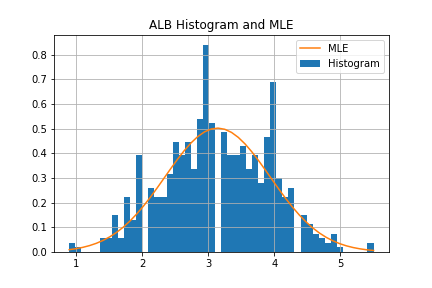
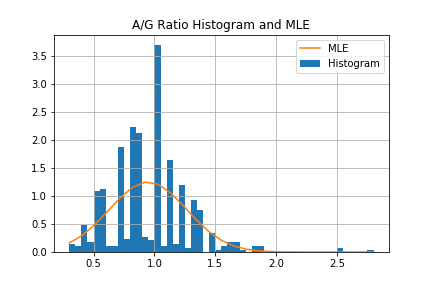
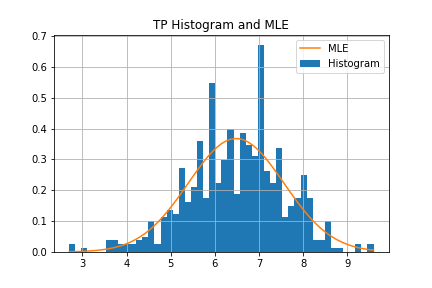
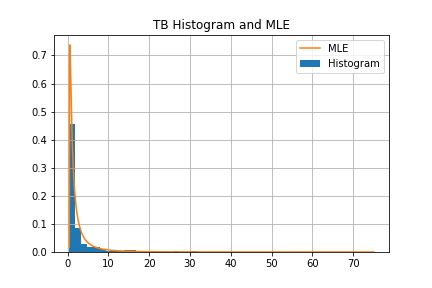
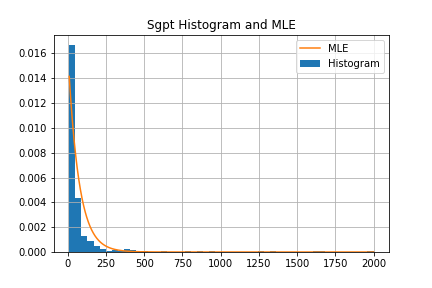
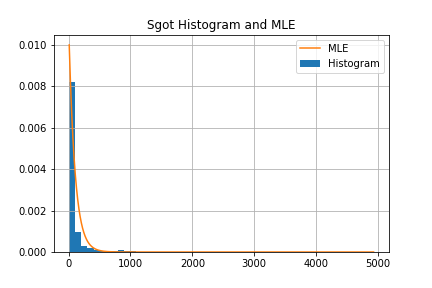
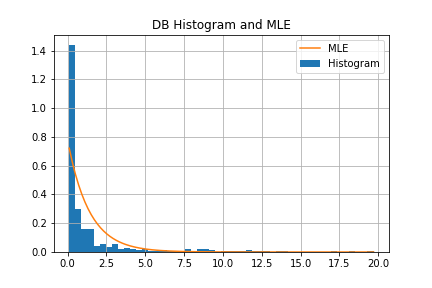
We observe the following:

1. In Label column: 1 – liver patient, 2 – non liver patient

Confidence Interval



Histograms and MLE



Correlations

Number of tests: 36

Before consider multiple-testing, number of correlations found: 24

After consider multiple-testing, number of correlations found: 21

Correlation found between Age and TP. p-value=5.2e-06, ro=-0.19

Correlation found between Age and ALB. p-value=6.8e-11, ro=-0.27

Correlation found between Age and A/G Ratio. p-value=1.4e-07, ro=-0.22

Correlation found between TB and DB. p-value=7.9e-185, ro=0.87

Correlation found between TB and Alkphos. p-value=4.8e-07, ro=0.21

Correlation found between TB and Sgpt. p-value=1.8e-07, ro=0.21

Correlation found between TB and Sgot. p-value=6.1e-09, ro=0.24

Correlation found between TB and ALB. p-value=5.9e-08, ro=-0.22

Correlation found between TB and A/G Ratio. p-value=5.2e-07, ro=-0.21

Correlation found between DB and Alkphos. p-value=9.4e-09, ro=0.23

Correlation found between DB and Sgpt. p-value=1.1e-08, ro=0.23

Correlation found between DB and Sgot. p-value=2.8e-10, ro=0.26

Correlation found between DB and ALB. p-value=2.4e-08, ro=-0.23

Correlation found between DB and A/G Ratio. p-value=1.1e-06, ro=-0.2

Correlation found between Alkphos and Sgot. p-value=5e-05, ro=0.17

Correlation found between Alkphos and ALB. p-value=6e-05, ro=-0.17

Correlation found between Alkphos and A/G Ratio. p-value=1.1e-08, ro=-0.23

Correlation found between Sgpt and Sgot. p-value=1.4e-126, ro=0.79

Correlation found between TP and ALB. p-value=1.9e-122, ro=0.78

Correlation found between TP and A/G Ratio. p-value=1.1e-08, ro=0.23

Correlation found between ALB and A/G Ratio. p-value=2.6e-82, ro=0.69

Number of tests: 36

Before consider multiple-testing, number of correlations found: 20

After consider multiple-testing, number of correlations found: 13

Healthy: Correlation found between Alkphos and Sgpt only in healthy group

Number of tests: 36

Before consider multiple-testing, number of correlations found: 22

After consider multiple-testing, number of correlations found: 19

Sick: Correlation found between Age and A/G Ratio only in sick group

Number of tests: 36

Before consider multiple-testing, number of correlations found: 21

After consider multiple-testing, number of correlations found: 19

Male: Correlation found between Age and TP only in Males group

Male: Correlation found between Alkphos and ALB only in Males group

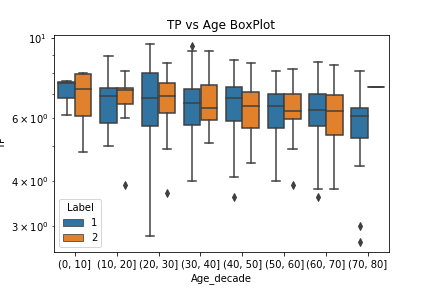
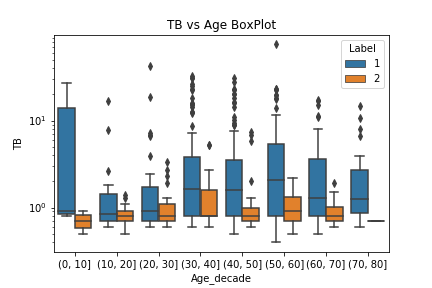
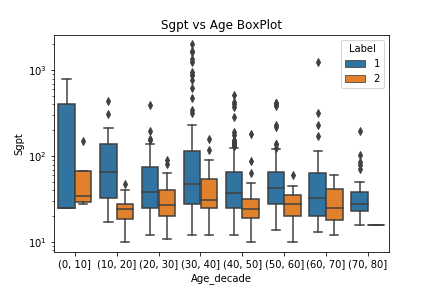
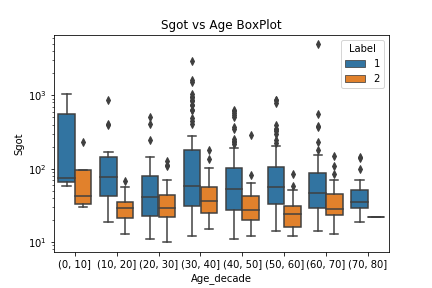
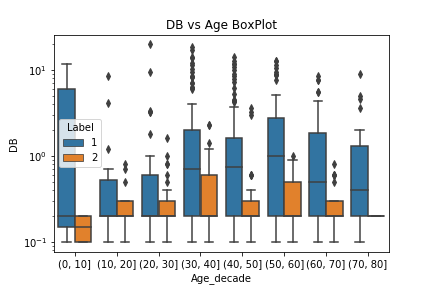
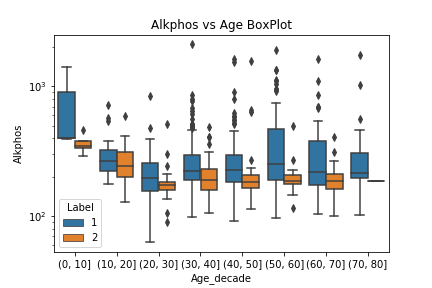
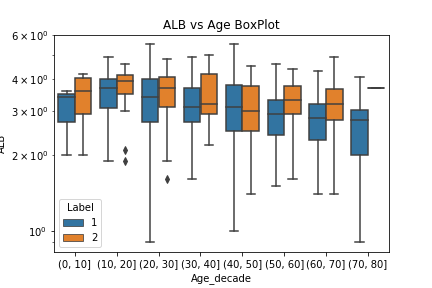
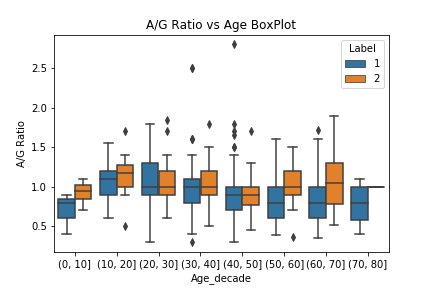
Number of tests: 36

Before consider multiple-testing, number of correlations found: 25

After consider multiple-testing, number of correlations found: 14

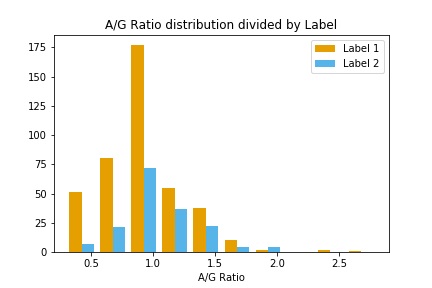
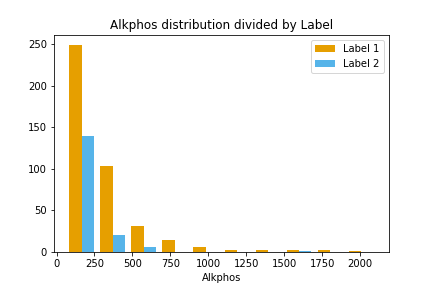
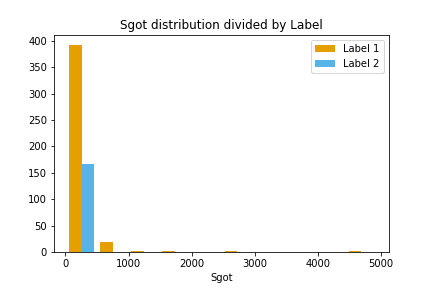
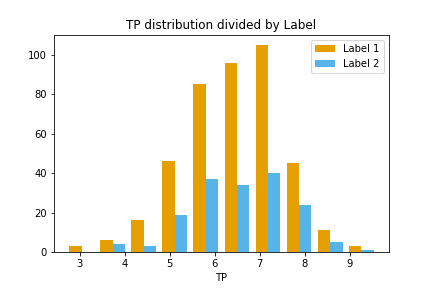
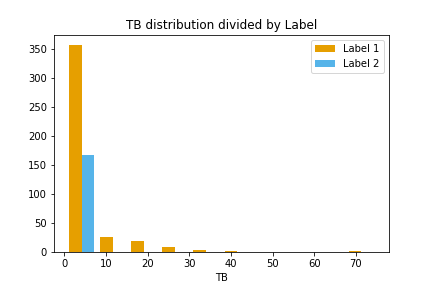
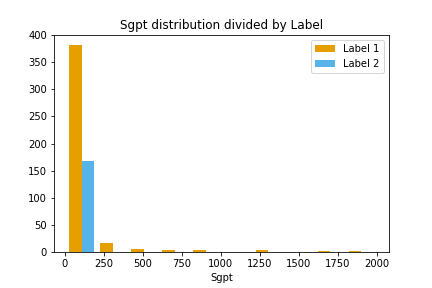
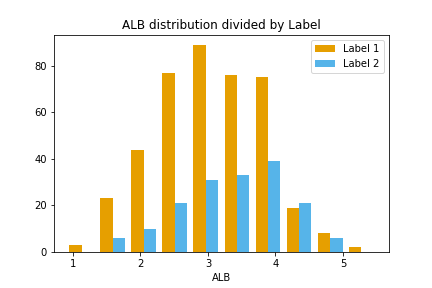
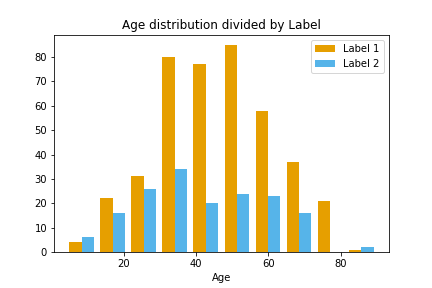
Female: Correlation found between Alkphos and Sgpt only in Females group

BoxPlots



Original question

Our original question is: For each numerical feature, plot one histogram with each label (live patient, non liver patient)



# Question 3

## Part 1

Derive expressions:

Claim 1:

Claim 2:

Claim 3:

In conclusion we derived

In conclusion, we derived

In conclusion, we derived

## Part 2

Problem 1

We denote

1. - all the winners
2. - winning numbers
3. – the number chosen by the player

Probability of winning the big prize:

Probability of winning the second prize:

Problem 2

Problem 3

Problem 4

We can analyze the scenario in the question using a binominal model, but this will be incorrect. The values of the binominal will not result in correct probability values.

Explanation:

The lotto card has 33 options from which we can choose 7 numbers. There are 6 winning numbers. Therefore, each of the numbers we chose is a winning number with probability .

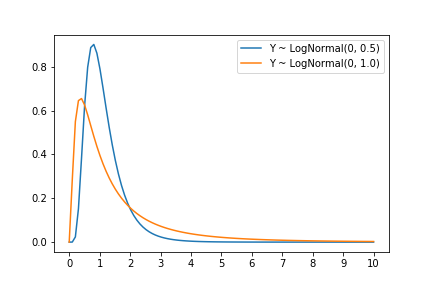
Assume that choosing 7 numbers is the same as having a sequence of 7 Bernoulli trials with .

In a binominal model, we assume that we can repeat the trial as many times as we want since the model is memory-less and the trials are independent of each other. In our scenario, it is not the case since each winning number cannot be selected again.

As evidence we show that calculating the variance under the assumption of a binominal model is incorrect:

# Question 4

Problem 1



Problem 2

Problem 3

We denote

We calculate the variance of Y1

𝑒𝑠 𝑠 is 4.2% 0rived ch are populated by the mean of the attribute he following steps:

We calculate the CDF so that

The probability is 0.565%

Problem 4

We denote

We calculate the variance of Y2

4.6708𝑒𝑠 𝑠 is 4.2% 0rived ch are populated by the mean of the attribute he following steps:

We calculate the CDF so that

The probability is 0.986%

Problem 5

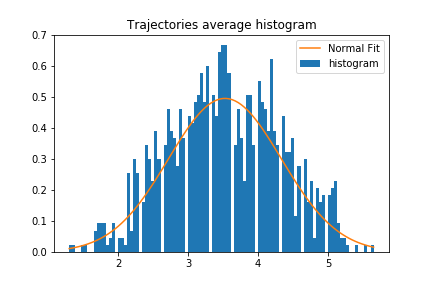
In order to calculate the IQR of a log normal distribution, we will use the following steps:

1. Use so that and
2. Given , find z in Z-table
3. Solve for positive and negative z values, and denote accordingly

# Question 5

Problem A

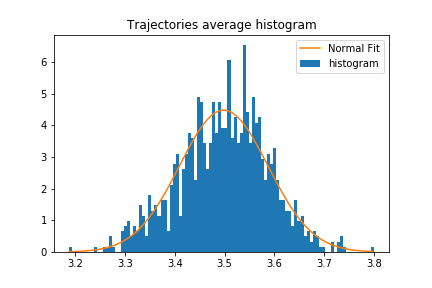
1. We expect the average to be 3.5



1. The distribution looks like a Normal distribution. The empirical mean and std are

Problem B

1. We expect the average to be 3.5



1. The distribution looks like a Normal distribution. The empirical mean and std are

Problem C

See histograms in problems A and B

Problem D (Bonus)

We assume that for a large enough the distribution of Markov chains is similar to a normal distribution. Therefore, the CLT can be applied on Markov chains for a large .