Statistical analysis of Red wine and White wine quality

Srujana Guduru

Abstract:

Red wine and white wine data sets contain the quality of wines based on various chemical components. We have analyzed the distribution of red wine and white wine using various summary statistics. Then, we estimated the distribution using empirical cumulative distribution function which is the maximum likelihood estimate of the cumulative distribution. The aim of this project is to compare the distributions of red wine and white wine. We used various methods like bootstrap, permutation testing to estimate the difference of means. We calculated the confidence intervals based on different methods. We also performed hypothesis testing to check whether the hypothesis that both the distributions are same is true. The above methods are the frequentist approaches. We also compared the distributions using Bayesian approach. Finally, we analyzed the quality of red wine and built a model using linear regression bootstrap.

1. Introduction to Data

The wine quality data sets are taken from the UCI Machine learning Repository. The data consists of wine quality at different physicochemical proportions. The columns in the data set are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality.

For our analysis we have considered only alcohol and quality columns.

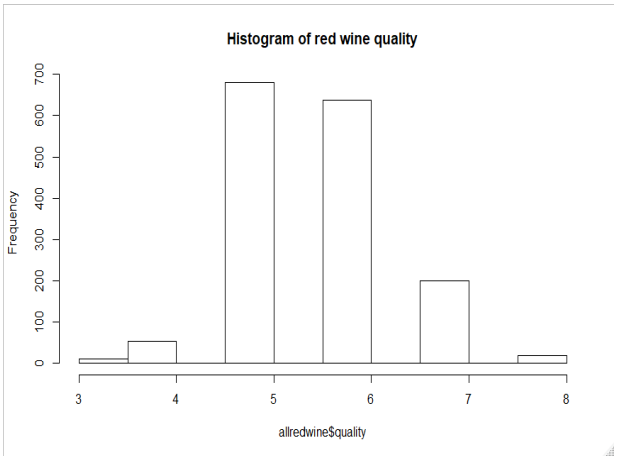

Figure-1: Histogram of red wine quality

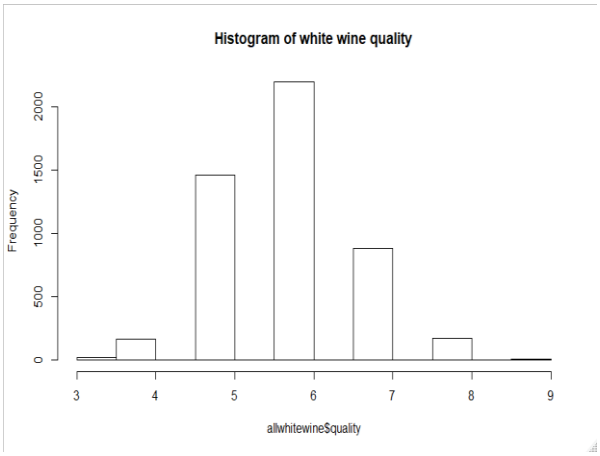

Figure-2: Histogram of white wine quality

2. ECDF:

Empirical cumulative distribution function (ecdf) is the *nonparametric maximum likelihood* estimate of the "underlying population" F. Any functional statistic t(F^) is the nonparametric maximum likelihood estimate of the parameter t(F)

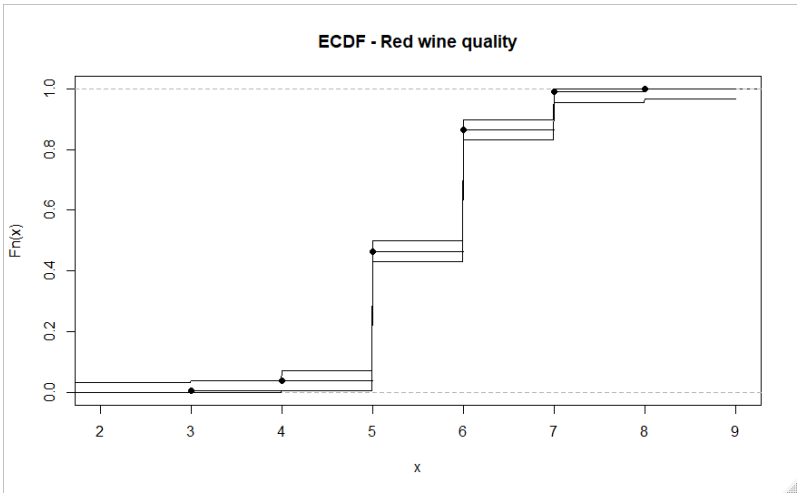

Figure-3: ECDF of Red wine quality

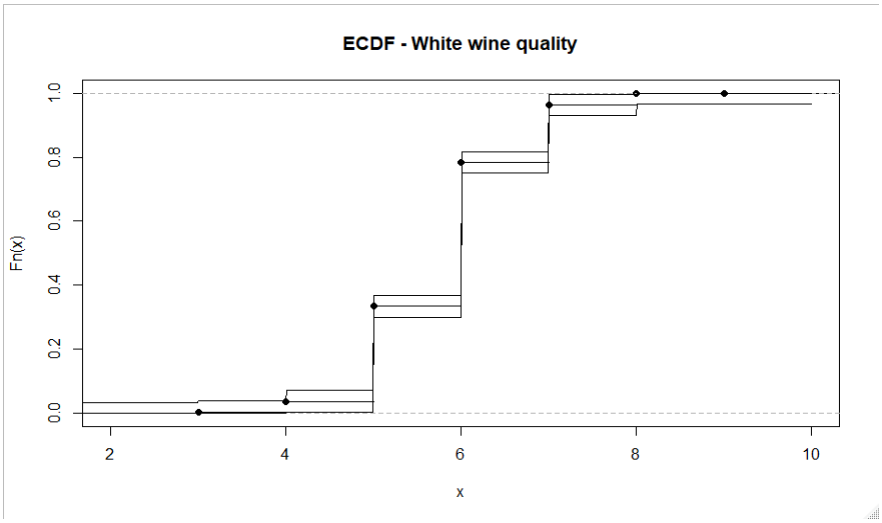

Figure-4: ECDF of White wine quality

The probability of red wine with poor quality i.e., less than 5 is 0.4652908

The probability of White wine with poor quality i.e., less than 5 is 0.3348305

There is high probability for the red wine to have poor quality

Let us compare the distributions of red wine and white wine by their means.

3. Bootstrap estimation of difference of means:

Bootstrap allows us to set confidence intervals on parameters without having to make unreasonable assumptions.

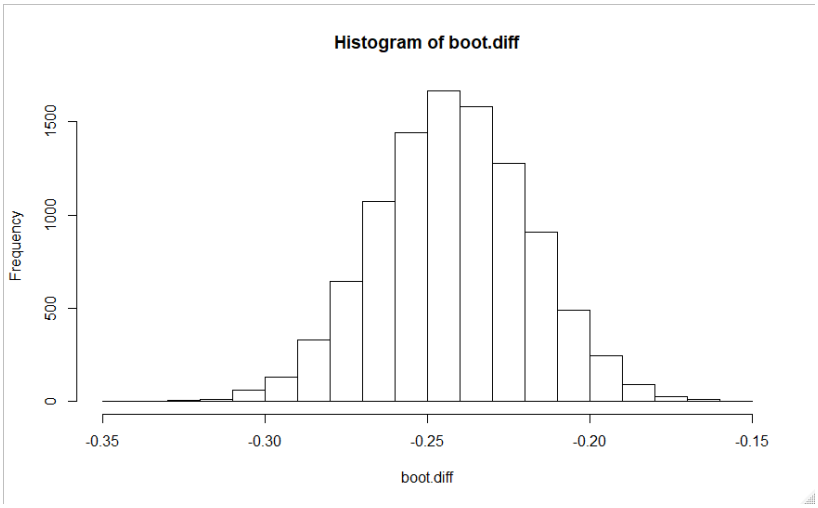

Figure-5: Histogram of difference of means by bootstrap method

Using non-parametric bootstrap to find the difference between the means of the red wine quality and white wine quality.

Calculating the difference of means:

Mean (red wine quality)-Mean (white wine quality):

$$\mu_1 - \mu_2 = -0.2418868$$

Standard error of the difference between the means: 0.02367871

4. Confidence intervals:

95% confidence interval by using Normal method: (-0.2892443, -0.1945294)

Confidence interval using Quantile method: (-0.2882370, -0.1960418)

Quantile method does not take full account of the difference between θ for F and θ,

the true value for $\hat{F}n$. The pivot confidence interval argues that the behavior of $\theta - \theta$ is approximately the same as the behavior of $\theta - \theta^*$.

Confidence interval using pivotal method: (-0.2877319, -0.1955367)

5. Hypothesis testing:

Null hypothesis: $\mu_1 = \mu_2$, Mean of red wine quality is equal to the mean of white wine quality.

Alternate hypothesis: $\mu_1 \neq \mu_2$, Mean of red wine quality is different from mean of white wine quality.

Test statistic:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{(s_1^2/n1 + s_2^2/n2)}}$$

$$Z = -10.14936$$

P value for the above statistic is less than 0.00001.

We can reject the null hypothesis with 95% confidence and can conclude that means of red wine quality and white wine quality are different.

6. Permutation testing to compare the distributions:

Permutation tests for comparing two populations is widely used in practice because of flexibility of the test statistic and minimal assumptions.

To perform permutation testing, we have merged the data from both the data sets (red wine and white wine) with labels red and white. Then, we have randomly reassigned the group labels and taken the mean difference of the quality of these new groups. This process is repeated 1000 times to get a distribution of the mean difference of the permuted groups.

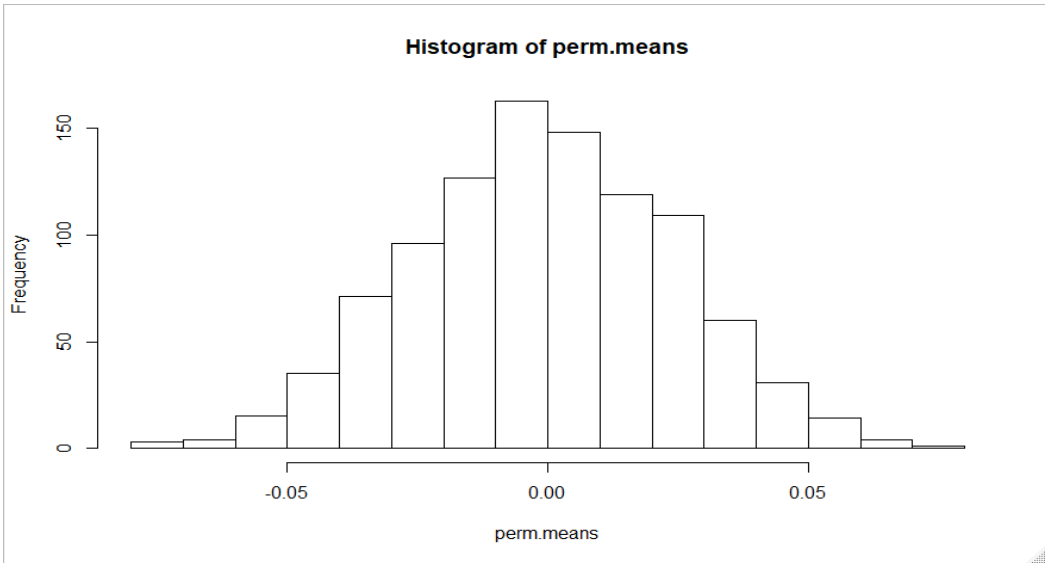

Figure-6: Distribution of permuted mean differences

The number of permuted mean differences that exceeded the true mean difference was 0. As there were 1,000 permutations, the significance level is simply 1/1001, or p = 0.001.

As this is less than 0.05, this means that quality of red wine is different from that of white wine (distribution of both the wines are different).

The main difference between bootstrap and permutation testing is that sampling in bootstrap is done with replacement whereas it is done without replacement in permutation testing.

Permutation method also works for small sample data.

7. Bayesian Analysis:

Using Bayesian approach to compare the distributions of red wine quality and white wine quality.

Prior: We believe that the distributions of red wine quality and white wine quality are standard normal distributions i.e., mean=0 and standard deviation =1.

$$\mu_1 \sim N(0,1)$$

$$\mu_2 \sim N(0,1)$$

We know that when the prior is normal, the posterior is also normal with mean and variances below:

Mean of red wine quality = 5.63

Variance of red wine quality = 0.0004

Mean of white wine quality = 5.87

Variance of white wine quality =0.0006

Based on the Standard normal assumptions, we have the below posterior distribution of difference of means.

We see that the difference in the means is equal to -0.2408082 which is approximately equal to the value we got in frequentist approach.

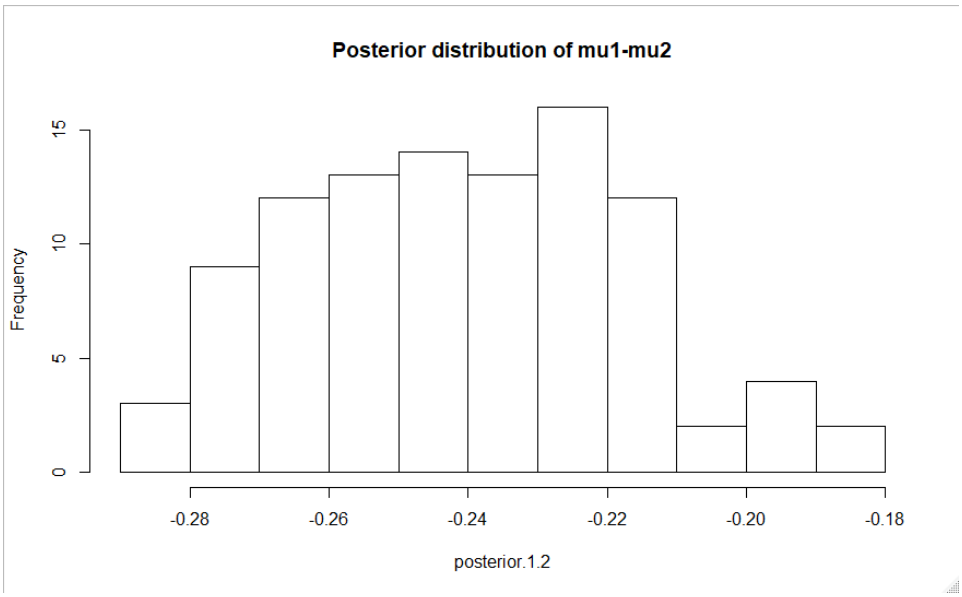

Figure-7:Posterior distribution of difference of means

We have seen that both the red wine and white wine have different distributions and now we will analyze the red wine quality based on the alcohol content using linear regression bootstrap.

8. Regression analysis of red wine using bootstrap:

Red wine quality is linearly dependent on alcohol content. We can build a linear model using linear regression to predict the quality of the new data using alcohol content.

Y= quality

X= alcohol

Linear model: $Y = \beta_0 + \beta_1 * X$

Summary of the linear model:

Coefficients	Estimate	Standard Error	t-value	Pr(>|t|)
Intercept	1.87497	0.17471	10.73	<2e-16 ***
Redwine$alcohol	0.36084	0.01668	21.64	<2e-16 ***

Using bootstrap:

lm.boot:Bootstrapping of linear model fits (using lm). Bootstrapping can be done by either resampling rows of the original data frame or resampling residuals from the original model fit.

Summary of the linear fit bootstrap:

	Coefficients	Bootstrap SD's
Intercept	1.8750	0
Redwine$alcohol	0.3608	0

By using both the methods the methods of resampling, the summary obtained is same.

So, the estimated linear model is $Y = 1.875 + 0.3608 * X$

9. Summary and conclusion:

First, we estimated the distribution of quality by using empirical cumulative distribution function. Then, from the bootstrap method and permutation method, we found that there is a difference in the means of both the distributions. Based on the hypothesis testing we rejected the hypothesis that both the distributions are equal. We also validated the result using Bayesian approach which also gave the same result.

By the various methods, we concluded that the red wine and white wine have different distributions.

Finally, we have built a linear regression model of red wine quality based on alcohol content using bootstrap.

APPENDIX- R CODE

```
library(car)
allredwine <- read.csv("winequality-red.csv",header=TRUE,sep=";")
allwhitewine <- read.csv("winequality-white.csv",header=TRUE,sep=";")
cor(allwhitewine)
redwine<-allredwine[,c('alcohol','quality')]
whitewine<-allwhitewine[,c('alcohol','quality')]
```

##ECDF of red wine along with 95% confidence interval

```
red_quality.ecdf <-ecdf(redwine$quality)
plot(red_quality.ecdf,main='ECDF - Red wine quality')
Alpha=0.05
n1=length(redwine$quality)
Eps=sqrt(log(2/Alpha)/(2*n1))
grid<-seq(0,9, length.out = 1000)
lines(grid, pmin(red_quality.ecdf(grid)+Eps,1))
lines(grid, pmax(red_quality.ecdf(grid)-Eps,0))
```

##ECDF of white wine along with 95% confidence interval

```
white_quality.ecdf <-ecdf(whitewine$quality)
plot(white_quality.ecdf,main='ECDF - White wine quality')
Alpha=0.05
n2=length(whitewine$quality)
Eps=sqrt(log(2/Alpha)/(2*n2))
grid<-seq(0,10, length.out = 1000)
lines(grid, pmin(white_quality.ecdf(grid)+Eps,1))
lines(grid, pmax(white_quality.ecdf(grid)-Eps,0))
## Probability of poor quality
red_quality.ecdf(5)
```

```
white_quality.ecdf(5)
# Bootstrap estimation for difference of means:
x1_bar <- mean(redwine$quality)
x2_bar <- mean(whitewine$quality)
mean_diff <- x1_bar-x2_bar
```
 # Non-parametric bootstrap
```
library(bootstrap)
cat("bootstrapping the difference of means of redwine and white wine:\n")
cat("bootstrapping is done independently for the two groups\n")
red.boot <- bootstrap(redwine$quality, 10000, mean)
white.boot <- bootstrap(whitewine$quality, 10000, mean)
boot.diff <- red.boot$thetastar - white.boot$thetastar
abline(v=0, col="red2")
se.boot<- sd(boot.diff)
```
##Confidence intervals
```
normal.ci<-c(mean_diff-2*se.boot, mean_diff+2*se.boot)i
quantile.ci<-quantile(boot.diff,c(0.025, 0.975))
pivatol.ci<-c(2*mean_diff-quantile(boot.diff,0.975), 2*mean_diff-quantile(boot.diff,0.025))
```
##Hypothesis testing:
```
#Test statistic:
sigma.hat<-sqrt((sd(redwine$quality)^2/n1)+(sd(whitewine$quality)^2/n2))
z<- mean_diff/(sigma.hat)
p.value=2*(1-pnorm(abs(z)))
x<-(1-pnorm(10.14936))
## Bootstrap on Regression model
library(simpleboot)
install.packages('simpleboot')
```

```
linearmodel <- lm(redwine$quality~redwine$alcohol)
summary(linearmodel)
R=1000
modelboot <- lm.boot(linearmodel, R, rows = TRUE, new.xpts = NULL, ngrid = 100,
      weights = NULL)
summary(modelboot)
modelboot11 <- lm.boot(linearmodel, R, rows = FALSE, new.xpts = NULL, ngrid = 100,
                weights = NULL)
summary(modelboot11)
plot(redwine$alcohol,redwine$alcohol)
##Permutation test to compare the means:
label<-'red'
redwine_m <- cbind(redwine,label)
label<-'white'
whitewine_m <- cbind(whitewine,label)
wine_m <- rbind(redwine_m,whitewine_m)
## diff test
diff.test <- function(group, quality) {
      resampled.group <- sample(group)
      mean(quality[resampled.group == "red"]) -
            mean(quality[resampled.group == "white"])}
perm.means <- replicate(1000, diff.test(wine_m$label, wine_m$quality))
hist(perm.means)
pvalue <- mean(abs(perm.means) > abs(diff.means))
```

Bayesian analysis:

```
#posterior of mu1
lb.1=1
```

```
Ix.1 = n1/var(redwine$quality)
post1.mean=((mean(redwine$quality))*Ix.1)/(Ib.1+Ix.1)
post1.var = 1/(Ib.1+Ix.1)
posterior.1 = rnorm(100,post1.mean,sqrt(post1.var))
#posterior of mu2
Ib.2=1
Ix.2 = n2/var(whitewine$quality)
post2.mean=((mean(whitewine$quality))*Ix.2)/(Ib.2+Ix.2)
post2.var = 1/(Ib.2+Ix.2)
posterior.2 = rnorm(100,post2.mean,sqrt(post2.var))
#posterior of mu1-mu2
post3.mean = post1.mean-post2.mean
post3.var = post1.var+post2.var
posterior.1.2 = rnorm(100,post3.mean,sqrt(post3.var))
hist(posterior.1.2,main="Posterior distribution of mu1-mu2")
mean(posterior.1.2)
```
