MATH20811 - Practical Statistics

Coursework 1 Submission

This report aims to analyse the whitewine.csv dataset *(Cortez et al, 2009)* given in Blackboard. This dataset contains measurements on fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality. In this report, we will be looking at white wine's **total sulfur dioxide**, **densities** and **qualities**.

Firstly, we compute the summary statistic and graphical display of the distribution of sulfur dioxide in white wine. As shown below:

	Total Sulfur Dioxide
Min.	9.00
1st Qu.	108.00
Median	134.00
Mean	138.36
3rd Qu.	167.00
Max.	440.00

Table 1: Summary Statistics of Total Sulfur Dioxide

The R code to compute the results of Table 1 is as follows:

```
s <- white_wine$total.sulfur.dioxide
summary(s)

sum <- as.array(summary(s))
row.names(sum) <- "Total Sulfur Dioxide"
library(xtable)
xtable(sum)
```

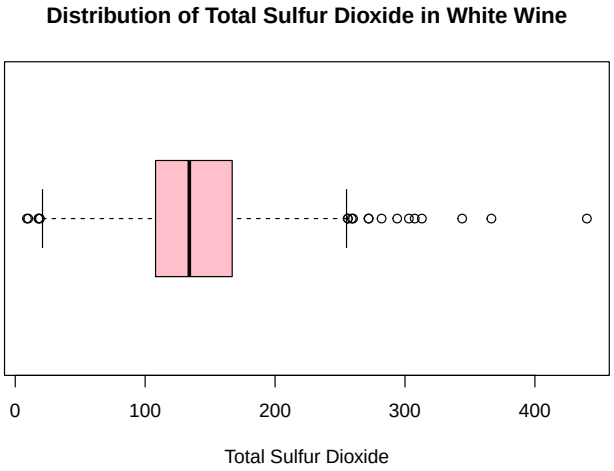

Figure 1: Boxplots of total sulfur dioxide

The R code to compute the results of Figure 1 is as follows:

```
boxplot(s, horizontal = TRUE, main = "Distribution of Total Sulfur
Dioxide in White Wine", col = "pink", xlab = "Total Sulfur Dioxide")
```

Looking at Figure 1, we can see that the box is around the center of the whiskers, hence suggesting that the distribution is symmetrical, this can be proven by finding the skewness of the data by computing the R code:

```
> library(moments)
> skewness(s)
[1] 0.3905902
```

We can see that the skewness is in between -0.5 and 0.5. So it is considered symmetrical. With the exclusion of outliers, the highest level of sulfur dioxide is around 260, which is very far away from the maximum level of sulfur dioxide given in Table 1.

Next, we look at the distributions of the total sulfur dioxide data at the different values of quality using box-plots as follow:

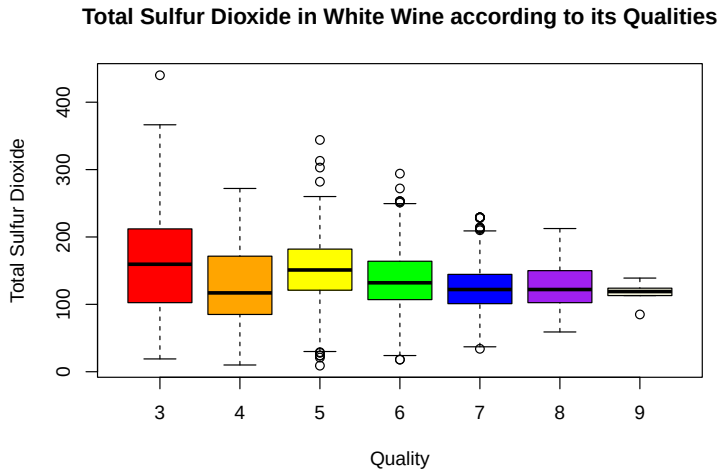

Figure 2: Boxplots of total sulfur dioxide at different white wine qualities

The R code to compute the results of Figure 2 is as follows:

```
q <- white_wine$quality

boxplot(s~q, col = c("red", "orange", "yellow", "green", "blue",
"purple", "beige"), xlab = "Quality", ylab = "Total Sulfur Dioxide",
main = "Total Sulfur Dioxide in White Wine according to its Qualities")
```

All box-plots are approximately symmetrical, but they get smaller when the wine is of higher quality, meaning that the level of sulfur dioxide is more consistent when the quality of the wine is better. When the quality is low, whiskers are longer and few outliers exist.

Then, we look at the contours from a bivariate normal density with approximated parameters of the total sulfur dioxide and density data.

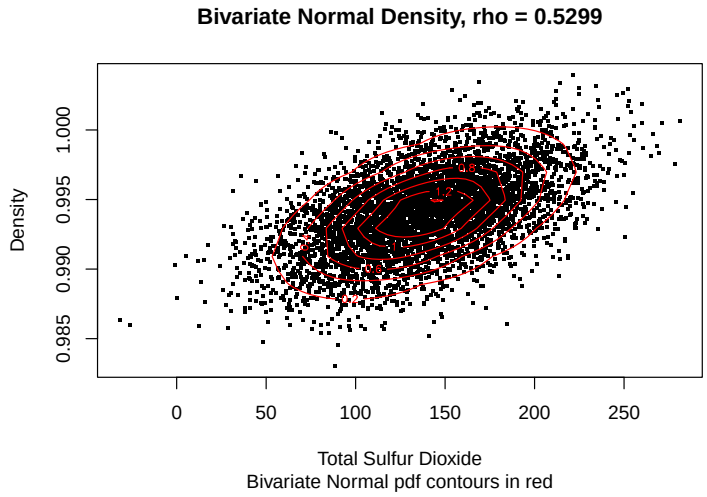

Figure 3: Scatter plot and contour of Total Sulfur Dioxide and white wine's densities

The R code to compute the results of Figure 3 is as follows:

```
rho <- cor(s, d, method = "pearson")

library(MASS)
library(mvtnorm)

twodim.Npdf=function(xlimits, ylimits, mu, covar){
  x.points <- seq(xlimits[1], xlimits[2], length.out = 100)
  y.points <- seq(ylimits[1], ylimits[2], length.out = 100)
  zz <- matrix(0, nrow=100, ncol=100)
  for (i in 1:100) {
    for (j in 1:100) {
      zz[i,j] <- dmvnorm(c(x.points[i],y.points[j]), mean=mu,
      sigma=covar)
    }
  }
  results=list(x.points, y.points, zz)
  return(results)
```

```
}

CN1=function(n, mux,sdx,muy,sdy,rho,text)
{
  means=c(mux, muy)
  C=matrix(c(sdx^2, rho*sdx*sdy, rho*sdx*sdy, sdy^2),nrow=2,ncol=2)
  Nvals<- twodim.Npdf(xlimits=c(9, 440), ylimits=c(0.9, 1.1),
  mu=means, covar=C)
  z=mvrnorm(n, mu=means, Sigma=C)
  x=z[,1]
  y=z[,2]
  plot(x,y,"p",pch=".", cex=3.0)
  contour(Nvals[[1]], Nvals[[2]], Nvals[[3]], col="red",
  xlab="Total Sulfur Dioxide",ylab="Density", main=paste(text),
  sub="Bivariate Normal pdf contours in red", col.sub="red",
  cex.sub=0.8,  cex.lab=1.2, cex.axis=1.0, add=TRUE)
  scor=cor(z, method="pearson")
  print("sample correlation matrix")
  scor
}

CN1(length(s), mean(s), sd(s), mean(d), sd(d), rho, "Bivariate
Normal Density, rho = 0.5299")
```

We also calculated the correlation between the level of sulfur dioxide and density by using Pearson's and Spearman's methods to support the graph above. The results are:

	Pearson	Spearman
Correlation	0.53	0.51

Table 2: Correlation between total sulfur dioxide and density using Pearson's and Spearman's methods

Below is the R code conducted to achieve this result:

```
c1 <- matrix(c(cor(s, d, method = "pearson")))
c3 <- matrix(c(cor(s, d, method = "spearman")))
c5 <- cbind(c1, c3)
colnames(c5) <- c("Pearson", "Spearman")
rownames(c5) <- c("Correlation")
xtable(c5)
```

The correlation between log of total sulfur dioxide and log of density by using both methods is calculated as well:

	Pearson	Spearman
Correlation	0.51	0.56

with the R code as follows:

```
c2 <- matrix(c(cor(log(s), log(d), method = "pearson")))
c4 <- matrix(c(cor(log(s), log(d), method = "spearman")))
c6 <- cbind(c2, c4)
colnames(c6) <- c("Pearson", "Spearman")
rownames(c6) <- c("Correlation")
xtable(c6)
```

Both tests for correlation have similar results. However, Spearman's method for log total sulfur dioxide and log density differs most from the other three results.

We will now perform a DIY hypothesis testing using Pearson's correlation coefficient for H0 : rho1 = 0.6 vs HA : rho1 $\neq$ 0.6 at 5% significance level using Fisher's Z-transform

```
> rho1 <- c1
> n = length(s)
> Z = function(i)
+   log((1+i)/(1-i))/2

> z = sqrt(n-3)*(Z(rho1) - Z(0.6))

> pnorm(n, Z(0.6), sqrt(1/(n-3)))
[1] 1
> pnorm(n, 0, 1)
[1] 1
```

Both Fisher's Z-transform and Normal distribution have a p-value of 1, hence H0 is true as there is insufficient evidence to reject it.

A calculation for the 95% confidence interval based on the Fisher's Z-transform can also be calculated by:

```
> u <- exp(-2*qnorm(0.975)/sqrt(n-3))
> v <- exp(2*qnorm(0.975)/sqrt(n-3))
>
> ci.lower <- (1+rho1-(1-rho1)*v)/(1+rho1+(1-rho1)*v)
> ci.upper <- (1+rho1-(1-rho1)*u)/(1+rho1+(1-rho1)*u)

> ci.lower;ci.upper
          [,1]
[1,] 0.5094349
          [,1]
[1,] 0.5497297
>
> cor.test(s, d, method = "pearson")

 Pearson's product-moment correlation

data:  s and d
t = 43.719, df = 4896, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5094349 0.5497297
sample estimates:
      cor
0.5298813
```

The 95% confidence interval turns out to be (0.5094349, 0.5497297), which is proven to be correct by computing the correlation test function in R.

We will now verify via simulation that the distribution of Fisher's Z-transform statistic, z, for a given sample size n, is approximately Normal.

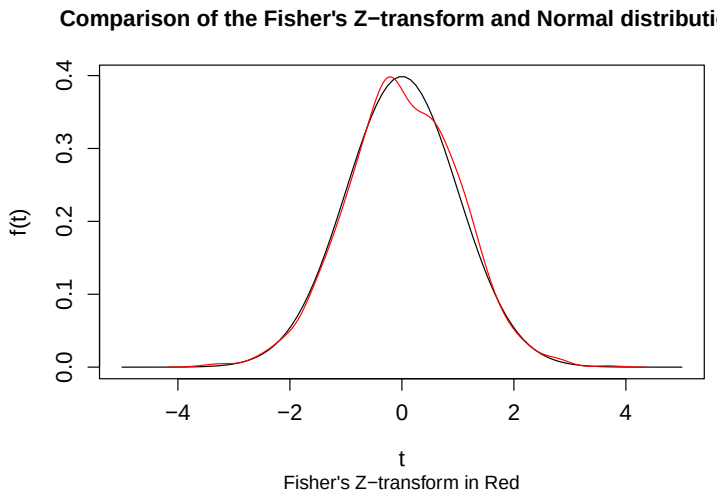

Figure 4: Comparison of Fisher's Z-transform and Normal distribution

```
N = 1000
t = 0
set.seed(1)
for (i in 1:N)
{
  x = rnorm(n)
  y = rnorm(n)
  r = cor(x, y)
  t[i] = sqrt(n-3)*(atanh(Z(r) - atanh(Z(0))))
}
tx = seq(-5, 5, length.out = 100)
plot(tx, dt(tx, n-2), type = "l", cex.lab = 1.2, cex.axis = 1.2,
xlab = "t", ylab = "f(t)", main = "Comparison of the Fisher's
Z-transform and Normal distribution", sub = "Fisher's Z-transform
in Red")
lines(density(t), col = "red")
```

The red graph, as shown in Figure 4, is the Fisher's Z-transform, it can be seen that it is very similar to the normal distribution with the true correlation parameter assumed to be zero.

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.