MATH20811 - Practical Statistics

Coursework 1 Submission

This report aims to analyse the `white wine` dataset [Cortez *et al*, 2009]. The dataset contains in total of 11-variable-measurements on white wine variants of the Portuguese Vinho Verde wine. In this report, we will only be looking into 3 variables in the dataset:

- `total.sulfur.dioxide`
- `density`
- `quality`

1 Question 1(i)

To start off, we will be exploring the empirical distribution of `total.sulfur.dioxide`. This includes some descriptive statistics as well as plots of the data distribution using boxplot and the Kernel density estimation. The descriptive statistics are summarised below.

Descriptive Statistics							
Measures of location		Measures of dispersion		Measures of shape		Quantiles	
Statistic	Value	Statistic	Value	Statistic	Value	Statistic	Value
Mean	138.36	Standard deviation	42.50	Skewness	0.39	Minimum	9.00
Mode	117.60	Inter-quartile range	59.00	Kurtosis	3.57	Lower Quartile	108.00
Median	134.00	-	-	-	-	Upper Quartile	167.00
-	-	-	-	-	-	Maximum	440.00

Table 1: Descriptive Statistics of total.sulfur.dioxide

The computation of the results in Table 1 is done by using the following R code:

```
white_wine=read.table("white_wine.csv", sep = ";", header = TRUE)
> tsd <- white_wine$total.sulfur.dioxide
> summary(tsd)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    9.0   108.0   134.0   138.4   167.0   440.0 
> sd(tsd)
[1] 42.49806
> IQR(tsd)
[1] 59

> library(e1071)
> skewness(tsd)
[1] 0.3904706
> kurtosis(tsd) #excess kurtosis#
[1] 0.5685873
```

```
> kurtosis(tsd) +3 #kurtosis#
[1] 3.568587

> mymode <- function(x){
+     d <- density(x)
+     return(d$x[which(d$y == max(d$y))])
+ }
> mymode(tsd)
[1] 117.5997
```

In these results, the mean of `total.sulfur.dioxide` in the white wine is 138.4, which is larger than the median, 134.0, and implies that the data may be approximately symmetric since the difference is hardly significant. This statement is confirmed by the value of its skewness that is a positive value but is barely above zero. Another indication of the shape that the data points form is its kurtosis value, 3.57, which is often compared to the kurtosis value of a Normal distribution which is equal to 3. In this case, we can see that the kurtosis is slightly above that value so it is safe to say that the shape of our random sample is approximate to that of a Normal distribution. To observe the dispersion of the data, we take a look at the standard deviation which is 42.5. the large value of the statistic is due to the very large range of values that the data is spread in which is between 9 and 440.

Some plots that might be useful for further dissecting the data would be the boxplot and the kernel density estimate of the random sample shown in Figure 1.

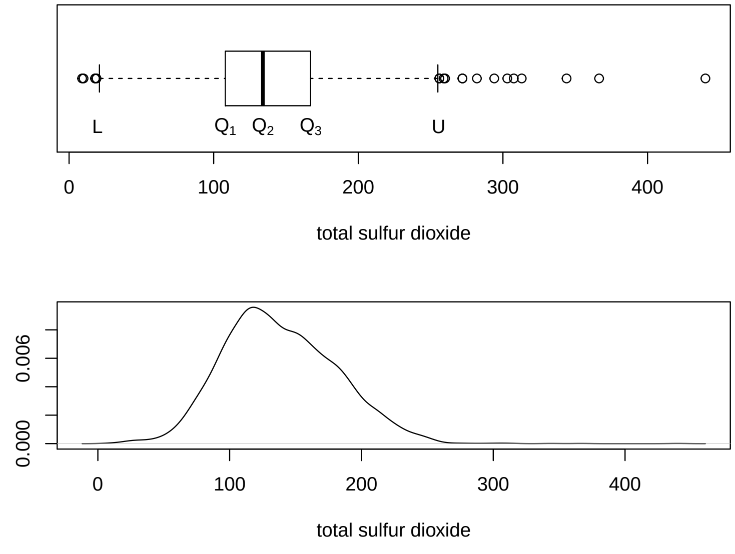

Figure 1: Boxplot and Density Estimate of total.sulfur.dioxide

The whiskers in a boxplot indicates the ranges for the bottom 25% and the top 25%, depending on the interquartile range. Any values outside the range can be regarded as possible outliers. We can see in the boxplot in Figure 1, there are many possible outliers that exceed value U or fall below the L value. These values and plots are obtained using the following R code;

```
> lq <- quantile(tsd)[2]; lq
```

```
25%
108
> med <- quantile(tsd)[3]; med
50%
134
> uq <- quantile(tsd)[4]; uq
75%
167
> L = quantile(tsd)[2] - 1.5* IQR(tsd); L
 25%
19.5
> U = quantile(tsd)[4] + 1.5* IQR(tsd); U
  75%
255.5
> q <- c(L, lq, med, uq, U)
>
> par(mfrow=c(2,1), mai = c(0.8, 0.6, 0.5, 0.3))
> boxplot(tsd, horizontal=TRUE, xlab="total sulfur dioxide")
> text(x = q+0.1, y= 0.65, labels = c(expression(L), expression(Q[1]), expression(Q[2]), expres
> plot(density(tsd), main="", xlab="total sulfur dioxide")
```

The middle 50% of the data, also known as the interquartile range is between 108 and 167. When ignoring the outliers, we can see that the dispersion of the data is fairly symmetrical, which can be compared to the density estimate plotted below it.

2 Question 1(ii)

It is even more useful to evaluate the data by splitting them into groups, in this case, the `quality` grade. To observe the differences of the distribution across the `quality` grade of the white wine, we produce boxplots for each `quality` grade and compare the differences, as we have done in Figure 2.

The main difference we can see between all the boxplots is how the interquartile range for each of them becomes smaller as the `quality` increases, as well as the maximum and minimum that the boxplot takes. It seems as if the median is converging to a certain value as the `quality` grade increases, as though there is a specific range of sulfur dioxide content that determines the quality of the white wine. Another way we can further analyse the boxplots is by observing the skewness and variances of each distribution across the `quality` grades. These values are summarised in Table 2 along with the R codes used to generate it below:

```
> boxplot(total.sulfur.dioxide˜quality, data = white_wine, xlab="quality",
ylab = "total sulfur dioxide", col = "red")

> wwq <- white_wine$quality
> skew <- c(skewness(tsd[wwq==3]), skewness(tsd[wwq==4]), skewness(tsd[wwq==5]),
    skewness(tsd[wwq==6]), skewness(tsd[wwq==7]), skewness(tsd[wwq==8]),
    skewness(tsd[wwq==9]))
```

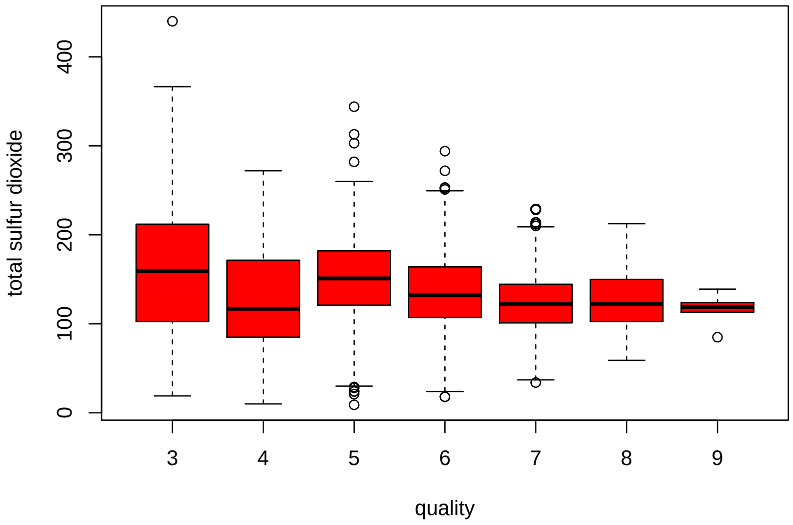

Figure 2: Distribution of total sulfur dioxide by its quality

Quality grade	Skewness	Variance
3	0.81	11611.86
4	0.21	2782.96
5	-0.03	1943.59
6	0.36	1704.55
7	0.50	1072.10
8	0.54	1089.42
9	-0.44	393.00

Table 2: Summarised skewness and variances of total.sulfur.dioxide across its quality grade

```
> vary <- c(var(tsd[wwq==3]), var(tsd[wwq==4]), var(tsd[wwq==5]),
    var(tsd[wwq==6]), var(tsd[wwq==7]), var(tsd[wwq==8]), var(tsd[wwq==9]))
> cbind(skew, vary)
            skew      vary
[1,]  0.81071555 11611.858
[2,]  0.20641772  2782.960
[3,] -0.03170024  1943.592
[4,]  0.35591447  1704.552
[5,]  0.50036525  1072.103
[6,]  0.53693167  1089.418
[7,] -0.43928052   393.000
```

From the listed skewness of each data points at different `quality` grades, we notice that they are all close to zero, which implies that they are fairly symmetric. The `quality` grade that had the highest skewness value is `quality` grade 3, which is clearly seen in its boxplot in Figure 2. On the other hand, the variances across the `quality` grades are all very large, which is due to the large spread of the `total.sulfur.dioxide` data. `Quality` grade 3 also has the largest variance compared to the rest. It may be important to note that the distribution of `total.sulfur.dioxide` at `quality` grades 5, 6 and 7 contains the most possible outliers in the data that may add to the variances being so large.

3 Question 1(iii)

Then, we take a look at the scatterplot of `total.sulfur.dioxide` and `density` alongside the contours from a Bivariate Normal density to observe the suitability of the bivariate Normal distribution being the probability model for the aforementioned data pairs. Using the subsequent R codes, we are able to produce the plot in Figure 3 on page 6,

```
> wwd<-white_wine$density
>
> 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(0,400), ylimits = c(0.99, 1.04), mu = means,
    covar = C)
+ contour(NVals[[1]], NVals[[2]], NVals[[3]], col = "red", xlab = "x (total
    sulfur dioxide)", ylab = "y (density)", main = paste(text),
    sub = "Bivariate Normal Contours in Red", col.sub = "red", cex.sub = 0.8,
    cex.lab = 1.2, cex.axis = 1.0)
+ points(tsd, wwd, "p", pch = ".", cex = 1.0)
+ scor = cor(tsd,wwd, method = "pearson")
+ print("sample correlation matrix")
+ scor
+ }
>
> par(mfrow = c(1,1), mai = c(1.3, 0.9, 0.9, 0.5))
> set.seed(11)
> CN1(4894, mean(tsd), sd(tsd), mean(wwd), sd(wwd),0.50, "")
[1] "sample correlation matrix"
[1] 0.5298813
```

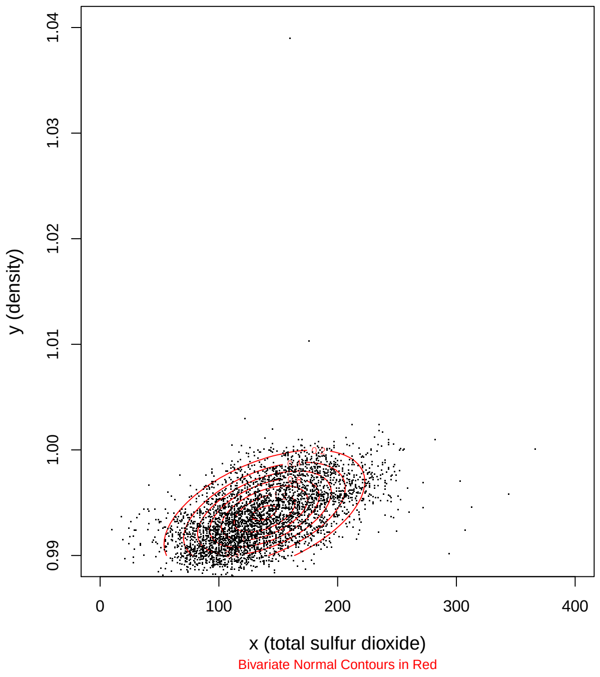

Figure 3: Scatterplot of total.sulfur.dioxide against density data superimposed by the contours of a Bivariate Normal distribution

The parameters that were chosen to create the contours are based off of means and variances of the `total.sulfur.dioxide` and `density` while the rho was estimated based on the correlation coefficient calculated in question 2. Based on the plot, we can see that the principal axis of the contours has a positive slope , which could be due to the value of rho chosen. The contours are slightly narrowed at the sides. Most of the data points are concentrated in the contours where, similar to the boxplot, the data points outside of it could be possible outliers. It is worth noting that they have affected the variances of each variable, and could affect the exact position of the contour plot. That said, a lot of the data points seem to be more concentrated in the bottom left of the contour plot. Another useful observation would be the position of the mode that was calculated in Table 1. The mode, 117.6, seems to be close to the center of the contours, where the region has the highest probability in the Normal distribution. That corresponds with the definition of the mode being the data that is most likely to occur. All in all, the Bivariate Normal distribution is a suitable probability model for the data pairs despite the outliers affecting the data.

4 Question 2

Next, we will be exploring the correlation coefficients between two pairs of data,

- `total.sulfur.dioxide` and `density`,
- `log(total.sulfur.dioxide)` and `log(density)`

using two methods, pearson's and spearman's.

```
> cor(tsd, wwd, method = "pearson")
[1] 0.5298813
> cor(tsd, wwd, method = "spearman")
[1] 0.5638241
> cor(log(tsd), log(wwd), method = "pearson")
[1] 0.5055141
> cor(log(tsd), log(wwd), method = "spearman")
[1] 0.5638241
```

Judging the results of the four sample correlation coefficients, they all fall around the same value. Between the `total.sulfur.dioxide` and `density`, both the spearman's and pearson's method shows a positive relationship between the two variables. The spearman's correlation coefficient however, is slightly larger than the pearson's. Both of the values show some monotonic or linear relationship but none of them show any significantly strong relation between the `density` of the white wine and the `total.sulfur.dioxide` content. Since all the values are taken into account in this calculation, the possible outliers that have been previously mentioned could be the cause of this. However, referring to the scatter plot in Figure 3, the data points alone don't show any significant linear or monotonic relationship at face value.

Likewise, the correlation coefficients for the `log(total.sulfur.dioxide)` and `log(density)` is also fairly similar. However, the difference between spearman's and pearson's correlation coefficients is slightly larger compared to those of the original data. Similar to the first set of data we calculated the correlation for, both values don't show whether they have strong or weak linear or monotonic relation to each other. It is useful to note that the main aim of log transforming the data points in general is to deal with the skewness and make the data become closer to following a Normal distribution in order to allow the statistical analysis to be more valid. This means it deals with the skewness of the data at hand, to make it closer to zero. However, we have to question the practicality of doing this considering that we have already established in Section 3, Question 1(iii) that a Bivariate Normal distribution is a suitable probability model for these set of data.

5 Question 3

In addition, a hypothesis test at 5% significance level on the true correlation coefficient, ρ_1 is to be carried out, with H_0 and H_A denoting the null hypothesis and alternate hypothesis respectively.

$$H_0 : \rho_1 = 0.5 \text{ vs } H_A : \rho_1 \neq 0.5$$

Based on the alternate hypothesis, we will be conducting a two tailed-test using the Fisher's z-transform, which states that the statistic

$$z = \sqrt{n-3}(\frac{1}{2}log\frac{1+r}{1-r} - \frac{1}{2}log\frac{1+\rho}{1-\rho}) \sim N(0,1)$$

This test is conducted using the following R code, where `zts` denotes the statistic that approximately follows the Standard Normal distribution.

```
> #zts is the z statistic
> n1 <- 4898
> r <- cor(tsd, wwd, method = "pearson")
> rho <- 0.5
> zts <- sqrt(n1-3)*(atanh(r)-atanh(rho))
> zts
[1] 2.845727
```

There are two ways to evaluate this statistic, we can compare it to the critical value on the standard normal distribution, or we can compute the p-value to this statistic and compare it with the significance level that the test is carried out at. We will be using the latter. The next lines of R codes show how we can compute the p-value, denoted by `pval`

```
> #pvalue
> pval <- 2*(1-pnorm(q = zts, mean = 0, sd = 1)); pval
[1] 0.004431021
>
```

Compared to the 0.05 significance level, the p-value is a lot smaller. Therefore, there is sufficient evidence for us to reject the null hypothesis in favour of the alternate one. The corresponding confidence intervals are computed in R by;

```
> #confidence interval
> ztwot <- qnorm (p = 0.025); ztwot
[1] -1.959964
> u <- exp(-2*ztwot/sqrt(n1-3))
> v <- exp(2*ztwot/sqrt(n1-3))
> lb <- ((1 + r - (1-r)*u)/(1 + r + (1-r)*u))
> ub <- ((1 + r - (1-r)*v)/(1 + r + (1-r)*v))
> ci <- c(lb, ub)
> ci
[1] 0.5094349 0.5497297
```

We can see that the confidence interval, `ci`, agrees with the `pval` that is less than the significance level, as they do not contain the null hypothesis value. The upper bound and lower bound of it agrees with the confidence interval calculated using the `cor.test` function.

```
> cor.test(tsd, wwd, method = "pearson")

Pearson's product-moment correlation

data:  tsd and wwd
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
```

6 Question 4

Fisher’s z-transform stated in (1) that was used in the question above is a very useful tool in testing hypotheses when the null hypothesis is equal to a value that is not 0. Theoretically, the transform approximately follows a Normal distribution with the following parameters (2);

$$Z = \frac{1}{2} log \frac{1+R}{1-R} \tag{1}$$

$$Z \sim N(\frac{1}{2} log \frac{1+\rho}{1-\rho}, \frac{1}{\sqrt{n-3}}) \tag{2}$$

In this section of the report we will be looking at whether the sampling distribution of Fisher’s z-transform follows this distribution by comparing it with the probability density function of the Normal distribution with the parameters in (2) under the assumption that the true correlation parameter equals zero. We will create a simulation that produces the Fisher’s z-transform sample data with a given sample size N, from random sample data pairs (x, y) that are independent and both follow some Normal distribution. Then, we superimpose its density to the probability density function of the Normal distribution with parameters in (2) for comparison using the following R code lines

```
> fishersz = function(N, n, mux, muy, sdx, sdy){
+ z = 0
+ r = 0
+ for (i in 1:N) {
+ x = rnorm(n, mux, sdx)
+ y = rnorm(n, muy, sdy)
+ r[i] = cor(x, y)
+ z[i] = log((1 + r[i])/(1 - r[i]))/2
+ }
+ zx<- seq(from = -1, to = 1, length.out = 100)
+ szpdf <- dnorm(x= zx, mean(log((1 + r)/(1 - r))/2), sqrt(1/(n-3)))
+ plot( zx, szpdf, "l", xlab = "z", ylab = "P(Z = z)")
+ lines(density(z), col = "red", cex = 0.1)
+ legend(’topright’, c(’z-transform sample distribution’,’Normal pdf’), lty=1,
    col=c(’red’,’black’),bty=’n’,cex= 0.75)
+ lines(density(z),col="red")
+ xycor = cor(x, y, method = "pearson")
+ print("sample correlation coefficient")
+ xycor
+ }
>
> fishersz(1000, 100, 5, 3, 6, 2)
[1] "sample correlation coefficient"
[1] 0.04511813
```

The function requires user to input the values of the sample size N of the Fisher’s z-transform sample and the parameters of each x and y distributions with n numbers of random samples. When we run the simulation, an example of the plot it produces is presented below and it prints the sample correlation coefficient shown in the last lines in the R code above, when 100 random samples are taken from $x \sim N(5,6)$ and $y \sim N(3,2)$ to produce 1000 Fisher's z-transform samples.

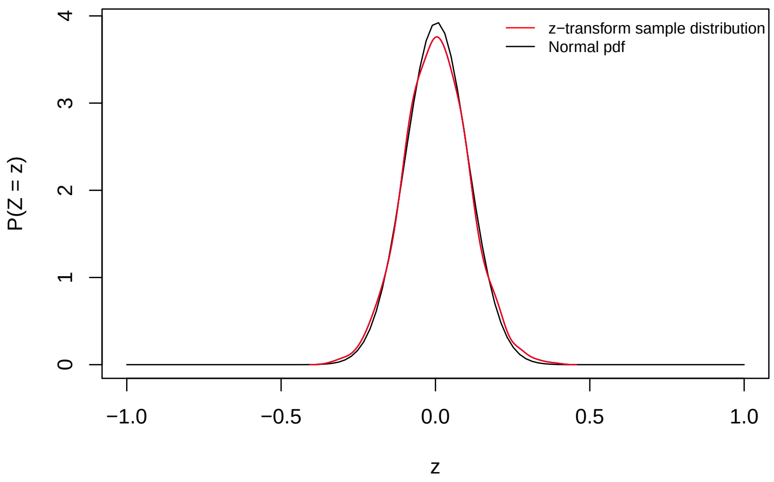

Figure 4: Plot comparison example between Fisher's z-transform's density and Normal distribution density function

From Figure 4, we can see that the density of Fisher's z-tranform and the Normal distribution mentioned in (2) is approximately the same. Therefore we can confirm that the Fisher's z-transform does follow the Normal distribution with parameters aforementioned.

7 Question 5

Finally, we complete our analysis by testing the following null hypothesis to further dissect this data.

H_0: there is no monotonic association between `total.sulfur.dioxide` and `density` vs H_A: there is monotonic association between these two variables.

The R codes used to carry out the test are as follows,

```
> cor.test(tsd, wwd, method = "spearman", exact = FALSE)

Spearman's rank correlation rho

data:  tsd and wwd
S = 8542141243, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
```

```
sample estimates:
      rho
0.5638241
```

The `exact` argument is a logical value that gives us the exact value of the p-value by considering all n! permutations when it is true and calculating the rho for each pairs of ranks. From the code, we have chosen the `exact` logical value to be false. This is due to the existence of ties in the data, so the exact p-value cannot be computed. Instead, the rough upper bound for the p-value is calculated using the following test statistic and obtained from the cumulative distribution function of the student t distribution with $n-2$ degrees of freedom.

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

From the p-value which is approximately 2.2×10^{-16}, there is highly significant evidence to reject the null hypothesis, in favour of the alternate hypothesis. We can conclude that there *is* monotonic association between the two variables.

References

[Cortez *et al*, 2009] P. Cortez, A. Cerdiera, F. Almeida, T. Matos and J. Reis. *Modeling win preferences by data mining from physicochemical properties.* In Decision Support Systems, Elsevier, 47(4):547-553. ISSN:0167-9236 [Cortez *et al*, 2009]