# Sai Lakshman Rao Duggirala – M15848995

#### Wine Project Part C

**library**("readxl")

**library**(stats4)

wine<-read.csv(file="D:/Applied Stat methods/winequality-red.csv", header=T, sep = ";")

str(wine)

## 'data.frame': 1599 obs. of 12 variables:

## $ fixed.acidity : num 7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...

## $ volatile.acidity : num 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...

## $ citric.acid : num 0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...

## $ residual.sugar : num 1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...

## $ chlorides : num 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...

## $ free.sulfur.dioxide : num 11 25 15 17 11 13 15 15 9 17 ...

## $ total.sulfur.dioxide: num 34 67 54 60 34 40 59 21 18 102 ...

## $ density : num 0.998 0.997 0.997 0.998 0.998 ...

## $ pH : num 3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...

## $ sulphates : num 0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...

## $ alcohol : num 9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...

## $ quality : int 5 5 5 6 5 5 5 7 7 5 ...

1. Produce summary statistics of “residual.sugar” and use its median to divide the data into two groups A and B. We want to test if “density” in Group A and Group B has the same population mean. Please answer the following questions

**attach**(wine)

summary(residual.sugar)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 0.900 1.900 2.200 2.539 2.600 15.500

Initially, we are calculating the summary statistics of residual sugar.

median\_residual\_sugar<-median(residual.sugar)

median\_residual\_sugar

## [1] 2.2

group\_A <- wine[wine$residual.sugar <= median\_residual\_sugar, ]

group\_B <- wine[wine$residual.sugar > median\_residual\_sugar, ]

1. State the null hypothesis

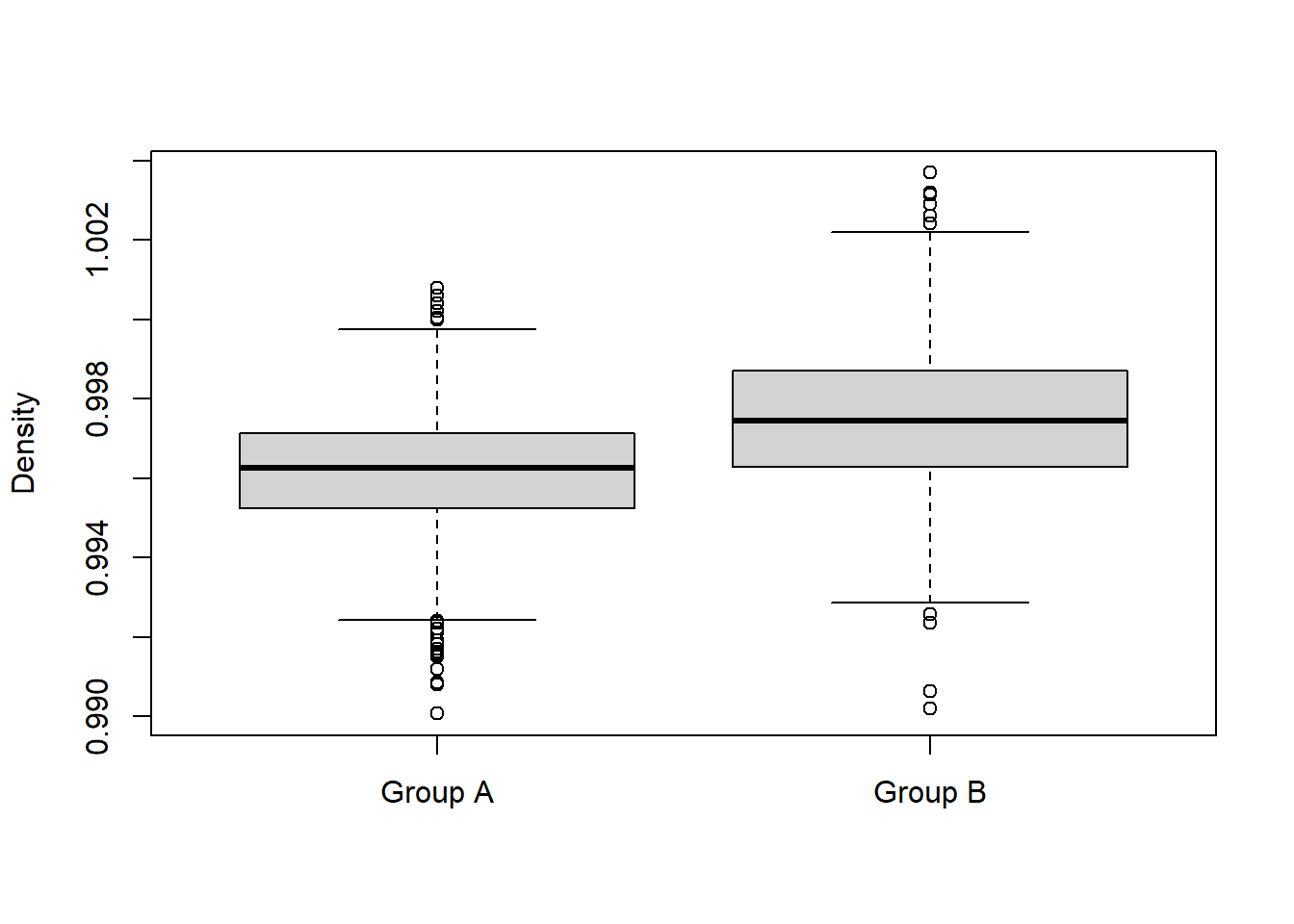
null\_hypothesis <- "The population mean of 'density' in Group A is equal to the population mean of 'density' in Group B."

null\_hypothesis

## [1] "The population mean of 'density' in Group A is equal to the population mean of 'density' in Group B."

b. Use visualization tools to inspect the hypothesis. Do you think the hypothesis is right or not?

boxplot(group\_A$density, group\_B$density, names = c("Group A", "Group B"), ylab = "Density")



The boxplot helped to provide visual support for the hypothesis. We can identify the slight differences in it. In the contest, we conclude to reject the hypothesis.

With further investigation the use of a t-test and finding the p-value, we can confidently state whether the hypothesis is right or wrong.

1. What test are you going to use?

t\_test\_result <- t.test(group\_A$density, group\_B$density)

t\_test\_result

##

## Welch Two Sample t-test

##

## data: group\_A$density and group\_B$density

## t = -14.697, df = 1365.2, p-value < 2.2e-16

## alternative hypothesis: true difference in means is not equal to 0

## 95 percent confidence interval:

## -0.001513022 -0.001156687

## sample estimates:

## mean of x mean of y

## 0.9961490 0.9974838

We performed a t-test statistical method it appears that there is strong evidence to conclude that the population mean of "density" is different between Group A (lower "residual.sugar") and Group B (higher "residual.sugar"). Given the very small p-value and the confidence interval that excludes zero, we can confidently conclude that there is a statistically significant difference in the population mean of "density" between the two groups based on their "residual.sugar" levels. In other words, the data suggests that the level of "residual.sugar" is associated with a difference in the mean "density."

1. What is the p-value?

p\_value <- t\_test\_result$p.value

p\_value

## [1] 1.654816e-45

The generated value is a small p-value indicates that you have strong statistical evidence to conclude that there is a significant difference in the population means of "density" between the two groups based on their "residual.sugar" levels, suggesting an association between these two variables.

1. What is your conclusion?

My conclusion based on the p-value generated is that we will reject the hypothesis, there is statistical evidence to suggest that the population mean of "density" is different between Group A (lower "residual.sugar") and Group B (higher "residual.sugar").

1. Does your conclusion imply that there is an association between “density” and “residual.sugar”?

The conclusion does imply that there is an association between "density" and "residual.sugar." The small p-value indicates that differences in "residual.sugar" levels are associated with differences in "density" means between the two groups. In other words, variations in "residual.sugar" levels appear to have a significant impact on the "density" of the samples, as evidenced by the rejection of the null hypothesis.

1. Produce summary statistics of “residual.sugar” and use its 1st, 2nd, and 3rd quantiles to divide the data into four groups A, B, C, and D. We want to test if “density” in the four groups has the same population mean. Please answer the following questions.

summary(residual.sugar)

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 0.900 1.900 2.200 2.539 2.600 15.500

quantiles <- quantile(wine$residual.sugar, probs = c(0.25, 0.5, 0.75))

quantiles

## 25% 50% 75%

## 1.9 2.2 2.6

We calculated the 1st, 2nd (median), and 3rd quantiles of "residual.sugar" by defining quantiles.

We then define our four groups as group\_A, group\_B, group\_C, group\_D.

group\_A <- wine[wine$residual.sugar <= quantiles[1], ]

group\_B <- wine[wine$residual.sugar > quantiles[1] & wine$residual.sugar <= quantiles[2], ]

group\_C <- wine[wine$residual.sugar > quantiles[2] & wine$residual.sugar <= quantiles[3], ]

group\_D <- wine[wine$residual.sugar > quantiles[3], ]

1. State the null hypothesis

null\_hypothesis2<-" The population mean of 'density' is the same in all four groups (A, B, C, D)."

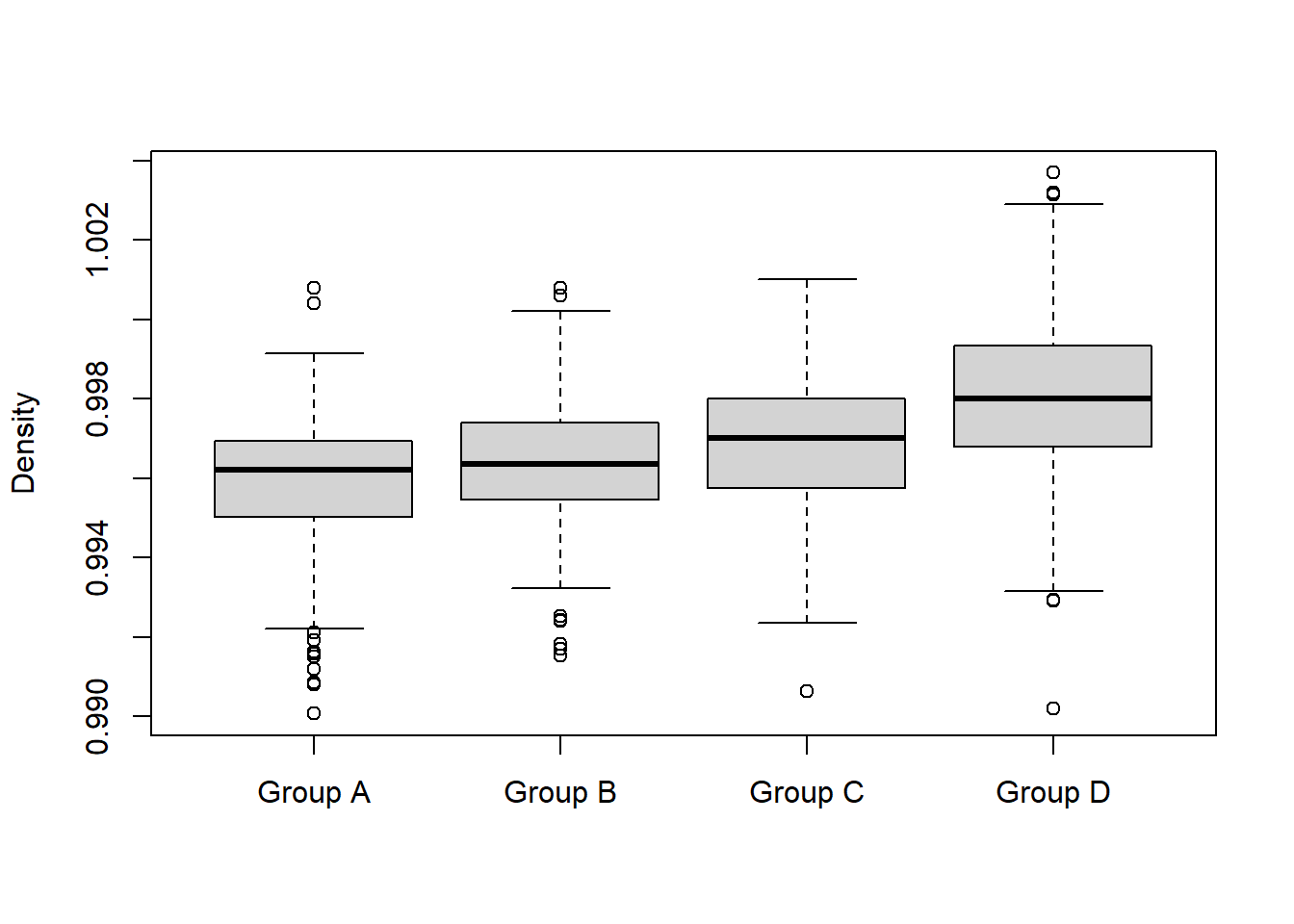
null\_hypothesis2

## [1] " The population mean of 'density' is the same in all four groups (A, B, C, D)."

1. Use visualization tools to inspect the hypothesis. Do you think the hypothesis is right or not?

boxplot(group\_A$density, group\_B$density, group\_C$density, group\_D$density,

names = c("Group A", "Group B", "Group C", "Group D"), ylab = "Density")



We can clearly see the difference in all four groups. The boxplot helped to provide visual support for the hypothesis. In context, we reject the hypothesis.

We further investigate the use of the Anova test as we have 4 groups and by finding the p-value, we can confidently state whether the hypothesis is right or wrong.

1. What test are you going to use?

We are going to use ANOVA test to determine the p-value.

anova\_result <- aov(density ~ factor(group), data = data.frame(density = c(group\_A$density, group\_B$density, group\_C$density, group\_D$density), group = factor(rep(letters[1:4], c(nrow(group\_A), nrow(group\_B), nrow(group\_C), nrow(group\_D))))))

summary(anova\_result)

## Df Sum Sq Mean Sq F value Pr(>F)

## factor(group) 3 0.000996 0.0003321 112.8 <2e-16 \*\*\*

## Residuals 1595 0.004696 0.0000029

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

1. What is the p-value?

p\_value\_anova <- summary(anova\_result)[[1]][["Pr(>F)"]]

p\_value\_anova

## [1] 3.065683e-66

The p-value associated with the ANOVA test you provided is indicated as "3.065693e-66" This notation represents an extremely small p-value. The ANOVA test results strongly suggest that there are significant differences in the means of "density" among the four groups. This implies that the level of "residual.sugar" has a significant impact on the "density" of the samples. The conclusion provided we can say that we reject the null hypothesis. The extremely small p-value (<2e-16) and the significance code indicate strong statistical evidence against the null hypothesis. Therefore, you have sufficient evidence to conclude that the population means of "density" are not the same among the four groups (A, B, C, and D), and there are significant differences in the means.

1. What is your conclusion?

We reject the null hypothesis. There is strong statistical evidence to suggest that the population mean of "density" differs among the four groups. In other words, the level of "residual.sugar" (which defines the groups) has a significant impact on the "density" of the samples.

1. Does your conclusion imply that there is an association between “density” and “residual.sugar”? Compare your result here with that in Question 1. Do you think increasing the number of groups help identify the association? Would you consider dividing the data into 10 groups so as to help the discovery of the association? Why?

In summary, the ANOVA results confirm a strong association between "density" and "residual.sugar," similar to the findings in Question 1. Dividing the data into more groups can help identify more nuanced associations and patterns. The case of dividing the data into four groups (Question 2), provided a more detailed view of how "density" varies across different ranges of "residual.sugar. Dividing the data into more groups can provide additional insights into the relationship, but the choice of the number of groups should be made carefully to balance detail and statistical analysis. This finer division allowed us to identify more specific patterns and associations within the data. So, the granularity observed in Question 2 is a result of the data being divided into more categories, which helps reveal more detailed insights into the relationship between "density" and "residual.sugar." The number can further be increased to get a broader picture of the difference. But we should choose the number of groups based on the application. We cannot create more number of groups as the divergence produced in the mean value of all the groups will increase with number of groups.

3. Create a 2 by 4 contingency table using the categories A, B, C, D of “residual.sugar” and

the binary variable “excellent” you created in Part B. Note that you have two factors: the

categorical levels of “residual.sugar” (A, B, C and D) and an indicator of excellent wines

(yes or no).

wine$factor\_residual\_sugar <- cut(wine$residual.sugar, breaks = quantile(wine$residual.sugar, probs = c(0, 0.25, 0.5, 0.75, 1)), labels = c("A", "B", "C", "D"))

*# Create a 2x4 contingency table*

contingency\_table <- table(wine$factor\_residual\_sugar, excellent\_wine)

*# Rename the rows for better labelling*

rownames(contingency\_table) <- c("A", "B", "C", "D")

*# Rename the columns for better libelling*

colnames(contingency\_table) <- c(0,1)

*# Display the contingency table*

print(contingency\_table)

## excellent\_wine

## 0 1

## A 409 53

## B 367 52

## C 308 53

## D 296 59

a. Use the Chi-square test to test if these two factors are correlated or not;

chi\_square\_result <- chisq.test(contingency\_table)

chi\_square\_result

##

## Pearson's Chi-squared test

##

## data: contingency\_table

## X-squared = 5.4033, df = 3, p-value = 0.1445

The Chi-squared statistic (X-squared) is 5.4033 and degrees of freedom (df) is 3. The p-value associated with the Chi-squared statistic is 0.1445. To determine whether these two factors, "residual.sugar" categories and the excellence of the wine, are correlated or not, we typically compare the p-value to a chosen significance level (alpha), which is often set at 0.05. In this case, the p-value (0.1445) is greater than the significance level (0.05). Therefore, we do not have strong evidence to reject the null hypothesis. Based on the Chi-square test at a significance level of 0.05, we do not find sufficient evidence to conclude that "residual.sugar" categories and the excellence of wine are correlated. In other words, there is no significant association between these two factors based on the Chi-square test result.

1. Use the permutation test to do the same and compare the result to that in (a);

*chi\_square\_result <- chisq.test(contingency\_table)*

chi\_square\_result

*# Observed chi-squared statistic*

observed\_statistic <- chi\_square\_result$statistic

*# Number of permutations for the test*

n\_permutations <- 1000 *# You can increase this for more accuracy*

*# Initialize an empty vector to store permuted statistics*

permuted\_statistics <- numeric(n\_permutations)

*# Perform the permutation test*

**for** (i **in** 1:n\_permutations) {

*# Permute the 'excellent\_wine' variable*

permuted\_excellent\_wine <- sample(excellent\_wine)

*# Recreate the contingency table with permuted 'excellent\_wine'*

permuted\_contingency\_table <- table(wine$factor\_residual\_sugar, permuted\_excellent\_wine)

*# Perform Chi-square test on permuted data*

permuted\_chi\_square\_result <- chisq.test(permuted\_contingency\_table)

*# Store the permuted chi-squared statistic*

permuted\_statistics[i] <- permuted\_chi\_square\_result$statistic

}

*# Calculate the p-value by comparing observed statistics to permuted statistics*

p\_value\_permutation <- sum(permuted\_statistics >= observed\_statistic) / n\_permutations

*# Print the p-value*

print(p\_value\_permutation)

## [1] 0.135

Both the chi-square test (part a) and the permutation test (part c) yield p-values greater than the conventional significance level of 0.05. The p-values from both tests suggest that there isn't strong evidence to reject the null hypothesis that "residual.sugar" and the excellence of wine are independent. Based on these results, we might not conclude that "residual.sugar" is a significant factor contributing to the excellence of wine because the p-values are not small enough to reject the null hypothesis. Keep in mind that statistical significance does not necessarily imply practical significance or a strong relationship. The conclusion should also consider domain knowledge and the context of your research. In this case, the data does not provide strong evidence of a significant association between "residual.sugar" and wine excellence, but other factors may still be contributing.

c. Can you conclude that “residual.sugar” is a significant factor contributing to the excellence of wine? Why?

Based on the results obtained from both the chi-square test and the permutation test, we cannot conclude that "residual.sugar" is a significant factor contributing to the excellence of wine.

**Chi-Square Test Result (Part a):**

The chi-square test yielded a p-value of approximately 0.1445.

**Permutation Test Result (Part C):**

The permutation test yielded a p-value of 0.135.

In both tests, the p-values are greater than the conventional significance level of 0.05. This suggests that there isn't strong statistical evidence to reject the null hypothesis, which assumes that "residual.sugar" and the excellence of wine are independent of each other.

Therefore, we cannot conclude that "residual.sugar" is a significant factor contributing to the excellence of wine based on the available data and the results of these statistical tests. Other factors may have a more substantial impact on wine excellence or the relationship between "residual.sugar" and wine excellence may be more complex and require further investigation.