Wine Project - Part C

Samarth Sathe

2023-10-10

```r
#Importing the required libraries
library(tidyverse)
library(here)
library(stats4)
library(ggplot2)
```

Importing the dataset

```r
file_path1 <- here("Data", "winequality-red.csv")
data_red_wine <- read.csv(file_path1, sep = ";", header = TRUE, stringsAsFactors = FALSE)
attach(data_red_wine)
data_length <- nrow(data_red_wine)
```

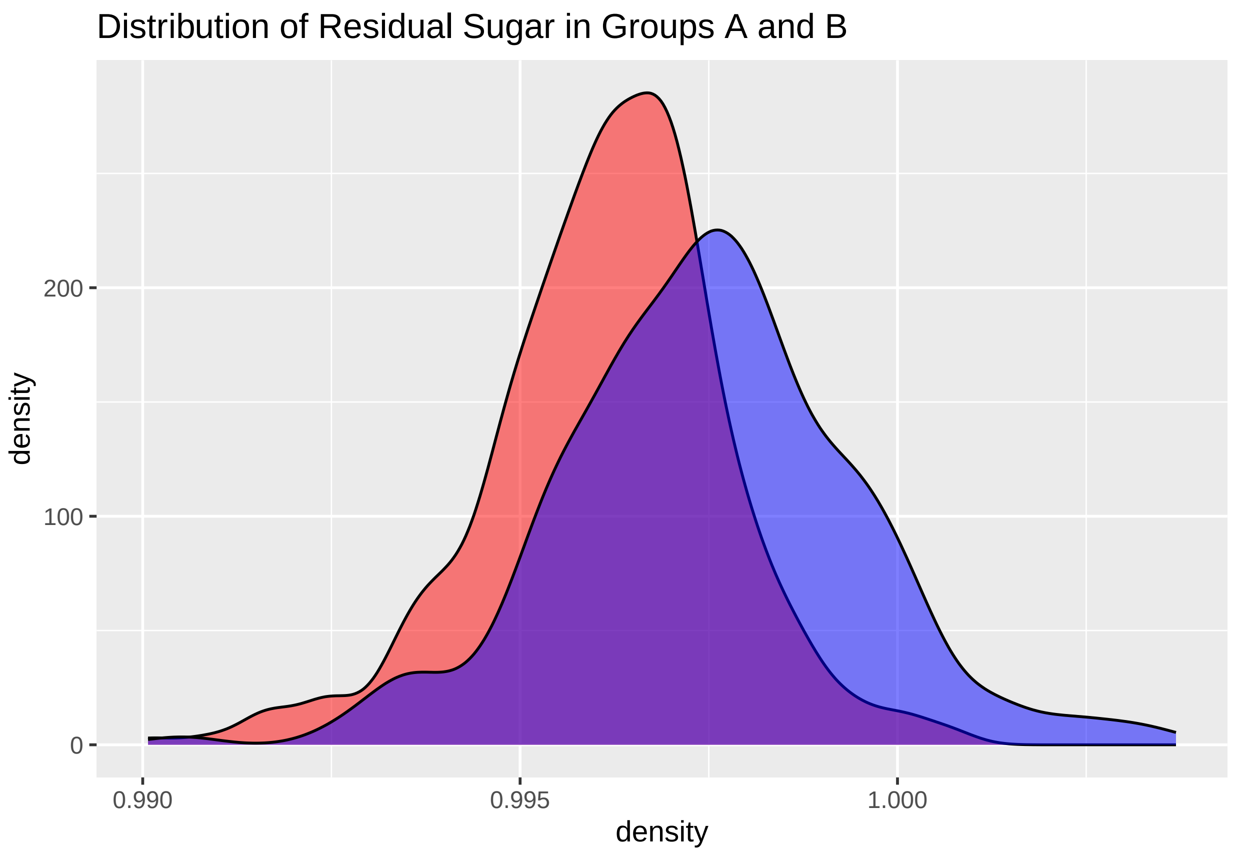
Question 1:Produce summary statistics of "residual.sugar" and use its median to divide the data into two groups A and B. We want to test if "density" in Group A and Group B has the same population mean. Please answer the following questions.

```r
# Divide data into two groups A and B using the median
median_sugar <- median(residual.sugar)
group_A <- data_red_wine %>% filter(residual.sugar <= median_sugar)
group_B <- data_red_wine %>% filter(residual.sugar > median_sugar)
```

Question 1a: State the null hypothesis: The means of group A and group B are expected to be same. Here, null_hypothesis is true difference in means is not equal to 0.

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

```r
ggplot() +
  geom_density(data = group_A, aes(x = density), fill = "red", alpha = 0.5) +
  geom_density(data = group_B, aes(x = density), fill = "blue", alpha = 0.5) +
  labs(title = "Distribution of Residual Sugar in Groups A and B") +
  scale_fill_manual(values = c("red", "blue"), labels = c("Group A", "Group B"))
```



Question 1c: What test are you going to use?

```r
# Perform a t-test to compare the means of the two groups
t_test <- t.test(group_A$density, group_B$density)
t_test
```

```
##
##  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
```

Question 1d: What is the p-value?

```r
# P-value
p_value <- t_test $p.value
p_value
```

```
## [1] 1.654816e-45
```

```r
# Print the p-value and conclusion
cat("P-value:", p_value, "\n")
```

```
## P-value: 1.654816e-45
```

Question 1e: What is your conclusion?

```
## Conclusion: Reject the null hypothesis. The p_value is smaller than 0.05 and there is a significant
```

Question 1f: Does your conclusion imply that there is an association between "density" and "residual.sugar"?

Yes, if the means are significantly different, there is an association.

Question 2: 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.

Question 2a: State the null hypothesis

```r
summary(residual.sugar)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

```r
Q1 <- quantile(residual.sugar, 0.25)
Q2 <- quantile(residual.sugar, 0.5)
Q3 <- quantile(residual.sugar, 0.75)

rs.group <- NULL

for (i in 1:length(residual.sugar)){
  if(residual.sugar[i]<=Q1) rs.group[i] <- 'A'
  else if(residual.sugar[i]<=Q2) rs.group[i] <- 'B'
  else if (residual.sugar[i]<=Q3) rs.group[i] <- 'C'
  else rs.group[i] <- 'D'

}
table(rs.group)
```

```
## rs.group
##   A   B   C   D
## 464 419 361 355
```

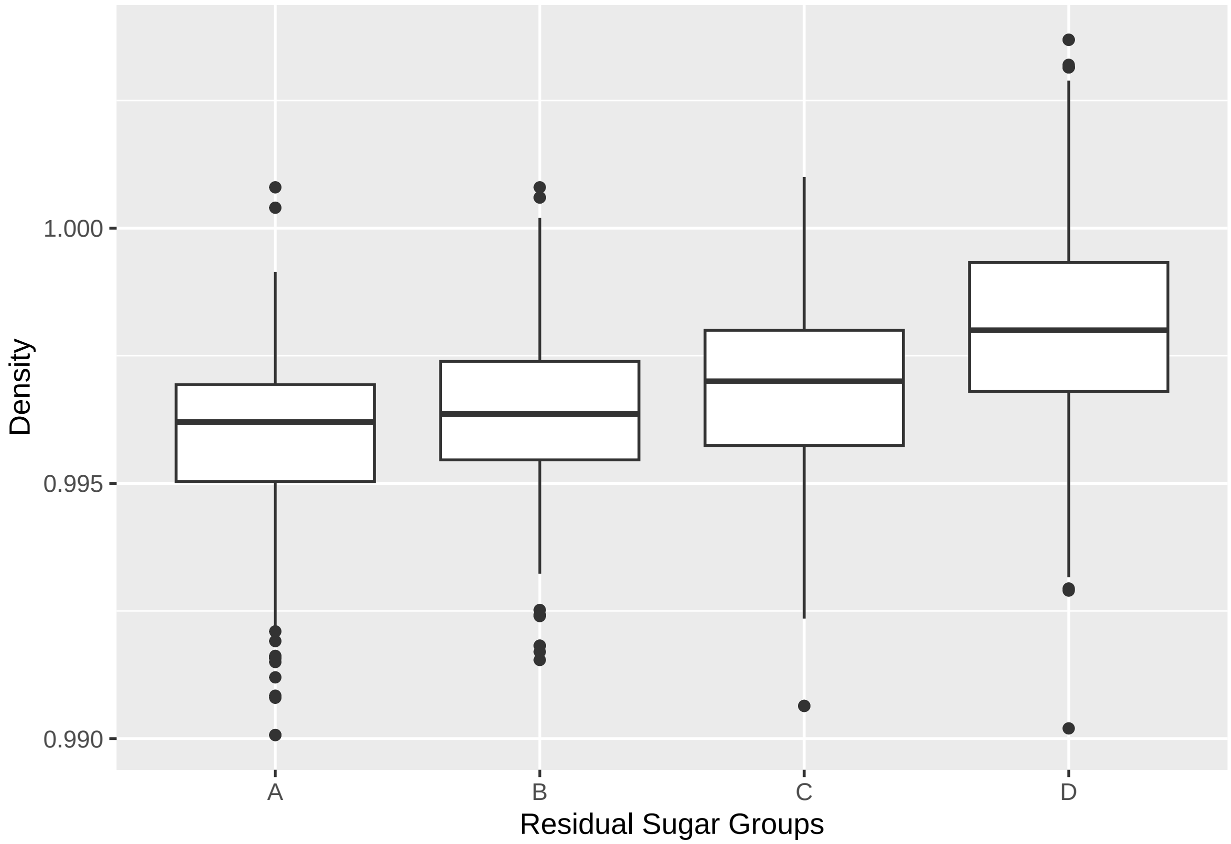
```r
# Calculate quantiles of "residual.sugar"
quantiles_sugar <- quantile(residual.sugar, probs = c(0, 0.25, 0.5, 0.75, 1))
```

Question 2a: State the null hypothesis

No difference between groups is expected. The population mean of residual sugar is the not same across all four groups.

Question 2b: Use visualization tools to inspect the hypothesis. Do you think the hypothesis is right or not?

```r
# Visualize the distribution of residual sugar in the four groups
data_red_wine %>%
  ggplot(aes(x = rs.group, y = density)) +
  geom_boxplot() +
  labs(x = "Residual Sugar Groups", y = "Density")
```



Question 2c: What test are you going to use?

```r
anova_result <- summary(aov(density ~ rs.group))
anova_result
```

```
##               Df   Sum Sq   Mean Sq F value Pr(>F)
## rs.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
```

Question 2d: What is the p-value?

The p-value is 2 x 10ˆ(-16)

Question 2e: What is your conclusion?

As the p-value is lesser than 0.05, the null hypothesis is rejected.

Question 2f: 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?

Since the null hypothesis is rejected, it indicates an association between the density and residual sugar variables. The consistent conclusion obtained when increasing the number of groups suggests that dividing the data into 10 groups will likely yield the same result.

Question 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).

```
# Create a binary variable for "excellent"
data_red_wine$excellent <- as.integer(data_red_wine$quality >= 7)

# Create a 2 by 4 contingency table
residual_sugar_group <- ifelse(data_red_wine$residual.sugar <= quantiles_sugar[2], "A",
                               ifelse(data_red_wine$residual.sugar <= quantiles_sugar[3], "B",
                                      ifelse(data_red_wine$residual.sugar <= quantiles_sugar[4], "C", "

contingency_table <- table(
  factor(residual_sugar_group),
  factor(data_red_wine$excellent)
)
```

Question 3a: Use the Chi-square test to test if these two factors are correlated or not.

```
# Perform a Chi-square test
chi_square_result <- chisq.test(contingency_table)

# Print the results
cat("Chi-square Test:\n")
```

```
## Chi-square Test:
```

```
cat("Chi-square Statistic:", chi_square_result$statistic, "\n")
```

```
## Chi-square Statistic: 5.499973
```

```
cat("P-value:", chi_square_result$p.value, "\n")
```

```
## P-value: 0.1386402
```

Question 3b: Use the permutation test to do the same and compare the result to that in (a).

```
# Perform a permutation test
permutation_test_result <- chisq.test(contingency_table, simulate.p.value = TRUE, B = 10000)

cat("\nPermutation Test:\n")
```

```
##
## Permutation Test:
```

```r
cat("Chi-square Statistic:", permutation_test_result$statistic, "\n")
```

```
## Chi-square Statistic: 5.499973
```

```r
cat("P-value:", permutation_test_result$p.value, "\n")
```

```
## P-value: 0.1349865
```

The p-value obtained from the chi square test is 0.1386402 while the p-value obtained from the permutation test is 0.1349865 Both p-values are above 0.05.

Question 3c: Can you conclude that "residual.sugar" is a significant factor contributing to the excellence of wine? Why?

Given that both p-values exceed the 0.05 threshold, we can conclude that we fail to reject the null hypothesis. This indicates that the null hypothesis, which suggests that residual sugar does not significantly contribute to the excellence of the wine, is substantiated by the data.