**TEAM PROJECT STA 5176**
**DRAFT BY JOVAN BURGOS**

## ➢ Introduction

Wine has been consumed worldwide for a long time. Among the benefits of consuming wine is that the high content of antioxidants has a medicinal and therapeutic value in the cure of cardiovascular diseases. The physicochemical properties of wine vary depending on the environmental conditions of the region in which they are harvested. Some studies indicate that the physicochemical properties of wine may be related to its quality. This work is focused on carrying out a statistical study to try to understand if the physicochemical properties of two types of wine (red and white) cultivated in the City of Napa are different and if these properties are related to the quality of the wine. It was decided to evaluate the characteristics of wines from two different types of grapes grown in the same region in order to understand whether the environmental conditions modify the most outstanding properties of the wine or whether these differences are due to the type of grape. The properties that were evaluated by statistical analysis are: Sulphur Dioxide content, Chlorides, Sugar, Citric Acid, pH and Density. **The statistical tests that were applied to the study of each one of the already mentioned physicochemical properties are:**

**1) One-way ANOVA Test**
**2) Welch-Satterthwaite Two Sample T-Test**
**3) Kruskal-Wallis Test**
**4) Variability Test**
**5) Proportion Test**
**6) Multiple Linear Regression**

- **Statistical test**

## 1. One-Way ANOVA

In order to estimate whether there are significant differences in the contents of Sulphur dioxide the one-way test was performed.

### 1.1. Contain of Sulphur Dioxide

```
wine.Summary <-  wine %>% group_by(type) %>% summarise("total sulfur dioxide mean " =
mean(total.sulfur.dioxide), " total sulfur dioxide stdev" = sd(total.sulfur.dioxide))
pander(wine.Summary, style='rmarkdown')
```

| Type | Total Sulfur Dioxide Mean | Total Sulfur Dioxide Stdev |
|---|---|---|
| Red Wine | 46.47 | 32.9 |
| White Wine | 138.4 | 42.5 |

```
wine_anova_results <- aov(total.sulfur.dioxide ~ type, data = wine)
wine_anova_table <- summary(wine_anova_results)
pander(wine_anova_table, style='rmarkdown')
```

***Analysis of Variance Model***

| | Df | Sum Squ | Mean Squ | F value | Pr(>F) |
|---|---|---|---|---|---|
| **Type** | 1 | 10179301 | 10179301 | 6253 | 0 |
| **Residuals** | 6495 | 10573600 | 1628 | NA | NA |

- **Hypotheses:** $H_0$: $\mu_1 = \mu_2$
  $H_1$: $\mu_1 \neq \mu_2$
- **Test Statistic**:$F_0 = 6252.8$.
- ***P*-value:** $p < 0.0001$. Reject if $p < \alpha$, where $\alpha = 0.05$.
- **Conclusion:** Reject $H_0$. There is sufficient evidence to suggest that the total sulfure dioxide contain in the two wines are different.

## 2. Welch-Satterthwaite's Two Sample T-Test

The mean of chloride and citric acid content in wine samples were compared by Welch-Satterthwaite two sample t-test.

### 2.1. Chloride content

Formally test to determine if the mean of the chlorides contain on red wine is greater than that of white wine. Use the appropriate version of the t-test with the critical at the α = 0.05 level.

```
wine.Summary <-  wine %>% group_by(type) %>% summarise("chlorides mean" = mean(chlorides), "chlorides stdev" = sd(chlorides))
pander(wine.Summary, style='rmarkdown')
```

| type | chlorides mean | chlorides stdev |
|---|---|---|
| Red Wine | 0.08747 | 0.04707 |
| White Wine | 0.04577 | 0.02185 |

```
chlorides_t <- t.test(redwine$chlorides, whitewine$chlorides, paired = FALSE, alternative = "greater")
chlorides_p <- p.value.string(chlorides_t$p.value)
pander(chlorides_t, style='rmarkdown')
```

*Welch Two Sample t-test: redwine$chlorides and whitewine$chlorides (continued below)*

| Test statistic | df | P value | Alternative hypothesis |
|---|---|---|---|
| 34.24 | 1828 | 3.047e-199 * * * | greater |

| mean of x | mean of y |
|---|---|
| 0.08747 | 0.04577 |

- **Hypotheses:**

$$\boldsymbol{H_0:\ \mu_R \leq \mu_W}$$

$$\boldsymbol{H_1:\ \mu_R > \mu_w}$$

- **Test Statistic**: $\boldsymbol{t_0 = 34.24}$
- **P-value**: $p < 0.0001$. Reject if $p < \alpha$, where $\alpha = 0.05$.
- **Interpretation:** Reject $H_0$. There is sufficient evidence to suggest that the contain chlorides in the red wine are higher than the white wine.

### 2.2. Citric acid content

Formally test to determine if the mean of the citric acid contain on red wine is less than that of white wine. Use the appropriate version of the t-test with the critical at the α = 0.05 level.

```
wine.Summary3 <-  wine %>% group_by(type) %>% summarise("citric acid mean" = mean(citric.acid), "citric acid stdev" = sd(citric.acid))
pander(wine.Summary3, style='rmarkdown')
```

| Type | Citric Acid Mean | Citric Acid stdev |
|---|---|---|
| Red Wine | 0.271 | 0.1948 |
| White Wine | 0.3342 | 0.121 |

```
citricacid_t <- t.test(redwine$citric.acid, whitewine$citric.acid, paired = FALSE, alternative = "less")
citricacid_p <- p.value.string(chlorides_t$p.value)
pander(citricacid_t, style='rmarkdown')
```

*Welch Two Sample t-test: redwine$citric.acid and whitewine$citric.acid (continued below)*

| Test statistic | df | P value | Alternative hypothesis |
|---|---|---|---|
| -12.23 | 2016 | 1.586e-33 * * * | less |

| Mean of x | Mean of y |
|---|---|
| 0.271 | 0.3342 |

- **Hypotheses:**

$$\boldsymbol{H_0}: \boldsymbol{\mu_R \geq \mu_W}$$

$$\boldsymbol{H_1}: \boldsymbol{\mu_R < \mu_w}$$

- **Test Statistic**: $\boldsymbol{t_0 = -12.229.}$
- **P-value**: $p < 0.0001$. Reject if $p < \alpha$, where $\alpha = 0.05$.
- **Interpretation:** Reject$H_0$. There is sufficient evidence to suggest that contain of citric acid in the red wine is less than the white wine.

## 3. Kruskal-Wallis Test

### 3.1. Contain of Alcohol

```
wine.Summary2 <-  wine %>% group_by(type) %>% summarise("alcohol median" = median(alcohol), "alcohol IQR" = IQR(alcohol))
pander(wine.Summary2, style='rmarkdown')
```

| Type | Alcohol Median | Alcohol IQR |
|---|---|---|
| Red Wine | 10.2 | 1.6 |
| White Wine | 10.4 | 1.9 |

```
wine_KW_results <- kruskal.test(alcohol ~ type, data = wine)
pander(wine_KW_results, style='rmarkdown')
```

*Kruskal-Wallis Rank Sum Test: Alcohol by Type*

| Test statistic | df | P value |
|---|---|---|
| 1.783 | 1 | 0.1818 |

- **Hypotheses:**

$$\boldsymbol{H_0}: \mathbf{M_1} = \mathbf{M_2}$$

$$\boldsymbol{H_1}: \mathbf{M_1} \neq \mathbf{M_2}$$

- **Test Statistic**: $\boldsymbol{H = 1.78}$.
- **P-value**: $p = 0.1818$. Reject if $p < \alpha$, where $\alpha = 0.05$.
- **Interpretation**: Fail to reject $H_0$. There is not sufficient evidence to suggest that there is a diferent in the alcohol contain of the two types of wines.

## 4. Variability Test

### 4.1. pH

```
pHvar <- var.test(redwine$pH, whitewine$pH, alternative = "t", conf.level = 0.95, ratio = 1)
pander(pHvar, style='rmarkdown')
```

*F Test to Compare Two Variances: redwine$pH and whitewine$pH (continued below)*

| Test Statistic | Num df | Den df | P value | Alternative hypothesis |
|---|---|---|---|---|
| 1.045 | 1598 | 4897 | 0.2716 | Two.Sided |

| Ratio of Variances |
|---|
| 1.045 |

- **Hypotheses:**

$$\boldsymbol{H_0: \sigma_r = \sigma_w}$$

$$\boldsymbol{H_1: \sigma_r \neq \sigma_w}$$

- **Test Statistic**: $\boldsymbol{F_0 = 1.05.}$
- **P-value**: $p = 0.2716$. Reject if $p < \alpha$, where $\alpha = 0.05$.
- **Interpretation:** Fail to reject$H_0$. There is insufficient evidence to suggest that there is a difference in the variability of pH between red and white wines.

### 4.2. Residual Sugar

```
rsvar <- var.test(redwine$residual.sugar, whitewine$residual.sugar, alternative = "t", conf.level = 0.95, ratio = 1)
pander(rsvar, style='rmarkdown')
```

*F test to compare two variances: redwine$residual.sugar and whitewine$residual.sugar (continued below)*

| Test Statistic | Num df | Den df | P value | Alternative Hypothesis |
|---|---|---|---|---|
| 0.07727 | 1598 | 4897 | 0 * * * | Two Sided |

| Ratio of Variances |
|---|
| 0.07727 |

- **Hypotheses:**

$$\boldsymbol{H_0: \sigma_r = \sigma_w}$$

$$\boldsymbol{H_1: \sigma_r \neq \sigma_w}$$

- **Test Statistic**: $\boldsymbol{F_0 = 0.08.}$
- **P-value**: $p < 0.0001$. Reject if $p < \alpha$, where $\alpha = 0.05$.
- **Interpretation:** Reject $H_0$. There is sufficient evidence to suggest that there is a difference in the variability of residual sugar contain between red and white wines

## 5. Proportion Test

### 5.1. Density

```
wine_density1 <- z.test(wine$density, alternative = "greater", mu = 1, sigma.x = sd(wine$density))
pander(wine_density1, style='rmarkdown')
```

*One-Sample Z-Test: wine$density*

| Test Statistic | P value | Alternative Hypothesis | Mean of x |
|---|---|---|---|
| -142.6 | 1 | greater | 0.9947 |

- **Hypotheses:**

$$\boldsymbol{H_0: \mu \leq 1}$$

$$\boldsymbol{H_1: \mu > 1}$$

- **Test Statistic:** $\boldsymbol{z = -142.5538456.}$
- **P-value:** $p = 1$. Reject if $p < \alpha$, where $\alpha = 0.05$.
- **Interpretation:** Fail to reject $H_0$. There is not sufficient evidence to suggest that the mean density wine is greater than 1.

```
wine_density2 <- z.test(wine$density, alternative = "less", mu = 0.98, sigma.x = sd(wine$density))
pander(wine_density2, style='rmarkdown')
```

*One-sample Z-Test: wine$density*

| Test statistic | P value | Alternative hypothesis | mean of x |
|---|---|---|---|
| 395 | 1 | less | 0.9947 |

- **Hypotheses:**

$$\boldsymbol{H_0: \mu \geq 0.98}$$

$$\boldsymbol{H_1: \mu > 0.98}$$

- **Test Statistic:** $\boldsymbol{z = 395.0437519.}$
- **P-value:** $p = 1$. Reject if $p < \alpha$, where $\alpha = 0.05$.
- **Interpretation:** Fail to reject $H_0$. There is not sufficient evidence to suggest that the mean density wine is less than 0.98.

```
wine_density3 <- z.test(wine$density, alternative = "less", mu = 0.99, sigma.x = sd(wine$density))
pander(wine_density3, style='rmarkdown')
```

*One-sample z-Test: wine$density*

| Test statistic | P value | Alternative hypothesis | mean of x |
|---|---|---|---|
| 126.2 | 1 | less | 0.9947 |

- **Hypotheses:**

$$\boldsymbol{H_0: \mu \geq 0.99}$$

$$\boldsymbol{H_1: \mu > 0.99}$$

- **Test Statistic**: $\boldsymbol{z = 126.2449532.}$
- **P-value**: $p = 1$. Reject if $p < \alpha$, where $\alpha = 0.05$.
- **Interpretation:** Fail to reject $H_0$. There is not sufficient evidence to suggest that the mean density wine is less than 0.99.

## 6. Regression Model

A multiple linear regression was performed to determine if there is a relationship between the physicochemical properties of the wine and its quality.

```
one_model <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlorides
+ total.sulfur.dioxide + density + pH + sulphates + alcohol, data=wine)
one_coef <- coefficients(one_model)
one_anova <- anova(one_model)
one_summary <- summary(one_model)
one_t <- as_tibble(one_summary[[4]])
one_ci <- as_tibble(confint(one_model, level=0.95))
```

The resulting regression model is:

$$\hat{y} = 56.3874229 + 0.0656296x_1 + -1.3840252x_2 + -0.1269532x_3 + 0.0456262x_4 + -0.3399487x_5 + -0.0012266x_6 + -55.6468897x_7 + 0.4649208x_8 + 0.782991x_9$$

### 6.1. ANOVA Table and Significance Test of the Regression Line

The corresponding ANOVA table is as follows:

```
pander(one_anova, style='rmarkdown')
```

*Analysis of Variance Table*

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **fixed.acidity** | 1 | 29.17 | 29.17 | 53.44 | 2.981e-13 |
| **volatile.acidity** | 1 | 322.3 | 322.3 | 590.5 | 6.656e-125 |
| **citric.acid** | 1 | 0.3516 | 0.3516 | 0.644 | 0.4223 |
| **residual.sugar** | 1 | 42.01 | 42.01 | 76.96 | 2.199e-18 |
| **chlorides** | 1 | 62.74 | 62.74 | 114.9 | 1.365e-26 |
| **total.sulfur.dioxide** | 1 | 137.9 | 137.9 | 252.5 | 8.085e-56 |
| **density** | 1 | 339.3 | 339.3 | 621.5 | 4.39e-131 |
| **pH** | 1 | 194.8 | 194.8 | 356.9 | 1.551e-77 |
| **sulphates** | 1 | 144.8 | 144.8 | 265.2 | 1.809e-58 |
| **alcohol** | 1 | 139.6 | 139.6 | 255.8 | 1.689e-56 |
| **Residuals** | 6486 | 3541 | 0.5459 | NA | NA |

- **Hypotheses:**

$$H_0\colon \beta_i = 0$$

$$H_1\colon \text{At least one is not } 0.$$

- **Test Statistic:** $F_0 = 53.44.$

- **P-value**: $p = 0.4223$. Reject if $p < \alpha$, where $\alpha = 0.05$.
- **Interpretation:** Comparing every p value of the previous table with the alpha, can be check that with at least one of them are greater than alpha, by this reason we fail to reject $H_0$. There is not sufficient evidence to suggest that the regression line is significant.

## 6.2. Significance Test of the Individual Predictor

```
pander(one_t, style='rmarkdown')
```

| Estimate | Std. Error | t value | P-Value |
|---|---|---|---|
| 56.39 | 11.95 | 4.718 | 2.427e-06 |
| 0.06563 | 0.01565 | 4.195 | 2.767e-05 |
| -1.384 | 0.07742 | -17.88 | 8.209e-70 |
| -0.127 | 0.07997 | -1.588 | 0.1124 |
| 0.04563 | 0.005174 | 8.819 | 1.468e-18 |
| -0.3399 | 0.3338 | -1.018 | 0.3085 |
| -0.001227 | 0.0002283 | -5.373 | 8.006e-08 |
| -55.65 | 12.2 | -4.563 | 5.135e-06 |
| 0.4649 | 0.09075 | 5.123 | 3.088e-07 |
| 0.783 | 0.07646 | 10.24 | 2.013e-24 |
| 0.2688 | 0.01681 | 15.99 | 1.689e-56 |

- **Hypotheses:**

$$\boldsymbol{H_0:\ \beta_i = 0}$$

$$\boldsymbol{H_1:\ At\ least\ one\ is\ \ not\ 0.}$$

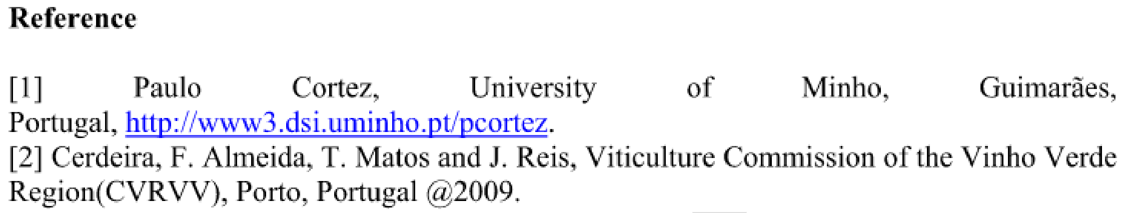
- **Test Statistic**: $\boldsymbol{t_0 = -1.59.}$
- **P-value**: $p = 0.1124$. Reject if $p < \alpha$, where $\alpha = 0.05$.
- **Interpretation:** Comparing every p value of the previous table with the alpha, at least one of them is greater than alpha, by this reason fail to reject $H_0$. There is not sufficient evidence to suggest that the regression line is significant.

➢ **Results**

Statistical analyses revealed significant differences (p < 0.05) in Sulphur Dioxide content. They also indicated that the Citric Acid Content in red wine is lower than that in white wine. Chloride content is higher in red wines. For the alcohol content, no evidence was found that the content is different for wines (p>0.05). In terms of Variability, no significant differences were found in the pH of the wines, but differences were found in the residual sugar content. Finally, multivariable linear regression did not provide evidence of a relationship between physicochemical properties and wine quality.

## Conclusion

The statistical study allowed us to infer that perhaps the physicochemical properties that are dependent on the region are the alcohol content and pH, since no significant differences were found between white and red wines. On the other hand, significant differences were found in the content of Sulphur Dioxide, Chlorides, Citric Acid and Residual Sugar content, which could mean that the type of grape is more related to these properties of the wine than the region of cultivation. However, a more complete study and analysis is considering more regions and different grape types. That would be necessary to draw more reliable conclusions. Finally, it seems that the quality associated with each type of crude does not have a direct relationship with the physicochemical properties evaluated in this study.



### Reference

[1] Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez.
[2] Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009.