Group member - Arifun Nabi, Abdisha Musa

Samantha Kay Kemper

Project 3

December 20th, 2022

**Part 1 -**

1. 1. **Null hypothesis**- Means for G1 grade and G2 grade are the same for the Portuguese language dataset.

**Research hypothesis**- Means for G1 grade and G2 grade are not the same for the Portuguese language dataset.

**a-level** – 0.05

**Rejection criteria** – If we find statistical support for the research hypothesis (p < a), we reject the null hypothesis.

**Rejected/Accepted (null)** – We reject the null hypothesis.

**Why?** – Because the p value is less than 0.05 which is significant.

**Conclusion** – There is a significance difference in G1 and G2. The G1 is lower than the G2. The mean difference is -0.1710324. **Report**: t (648) = -2.9454, p = 0.003341

**Test all assumptions** – Samples are not independent since it belongs to same dataset. Both the data looks slightly skewed with median data slightly lower. There are also outliers in both the data. However, we are assuming that the data are normally distributed since it looks slightly normally distributed.

**Issues run in** = I tried to transform the G1 column in Portuguese data. I created a new column where I added 2 to get all positive number. Square root the new column and finally log the square root column. When I checked it by looking at the histogram the data looked more skewed to the right.

2. **Null hypothesis**- Means for G1 grade and G2 grade are the same for the mathematics dataset.

**Research hypothesis**- Means for G1 grade and G2 grade are not the same for the mathematics dataset.

**a-level** – 0.05

**Rejection criteria** – If we find statistical support for the research hypothesis (p < a), we reject the null hypothesis.

**Rejected/Accepted (null)** – We fail to reject the null hypothesis.

**Why?** – Because the p value is more than 0.05 which is insignificant.

**Conclusion** –There is no significance difference in G1 and G2. The G1 is slightly higher than the G2. The mean difference is 0.1949367. **Report**: t (394) = 1.9648, p = 0.05014

**Test all assumptions** – Samples are not independent since it belongs to same dataset. Looking at the boxplot, both the data looks normally distributed. However, looking the at the Histogram, both the data looks slightly skewed to the right. However, we will assume that the data are normally distributed since it looks slightly normally distributed.

3. **Null hypothesis**- G3 grades mean in mathematics dataset and G3 mean in Portuguese dataset are same.

**Research hypothesis**- - G3 grades mean in mathematics dataset and G3 mean in Portuguese dataset are not same.

**a-level** – 0.05

**Rejection criteria** – – If we find statistical support for the research hypothesis (p < a), we reject the null hypothesis.

**Rejected/Accepted (null)** – We reject the null hypothesis.

**Why**? – There is significant difference in mean as p value is significantly less than 0.05.

**Conclusion** – There is significant difference as G3 mean in mathematics is lower than G3 mean in Portuguese.

**Report** – t (633.3) = -5.6664 and p = 2.215 e-08

**Test all assumptions** – The samples are independent because two columns belong to two different datasets and there are students who are not in both classes. Both the data looks slightly skewed to the right in the histogram. I tried with adding 1 and sqrt it in a new column. Finally, create a new column by log the square root. However, I still get same p value. We will assume that both data is normally distributed since its approximately normally distributed.

4. **Null hypothesis**- The means of G3 grades for both Portuguese and Mathematics for students appearing in both datasets are equal.

**Research hypothesis**- The means of G3 grades for both Portuguese and Mathematics for students appearing in both datasets are not equal.

**a-level** – 0.05

**Rejection criteria** – If we find statistical support for the research hypothesis (p < a), we reject the null hypothesis.

**Rejected/Accepted (null)** – We reject the null hypothesis.

**Why?** – There is significant difference as p value is significantly less than 0.05.

**Conclusion** – There is significant difference between G3 grade in math and G3 grade in Portuguese. The G3 mathematics grade is lower than G3 Portuguese grade.

**Report -** t (381) = -9.9771 and p < 2.2e-16

**Test all assumptions** – The two samples are not independent since the means for two columns belong to a combined dataset and students are from both classes. The histogram looks close to normally distributed.

**Part 2-**

1. 1. **Null Hypothesis**-There is no effect of weekend and workday alcohol consumption on G3 grades for the mathematics dataset.

**Research hypothesis**- There is effect of weekend and workday alcohol consumption on G3 grades for the mathematics dataset.

a-level – 0.05

**Rejection criteria** – If we find statistical support for the research hypothesis (Pr(>F) < a), we reject the null hypothesis.

**Rejected/Accepted (null)** – We fail to reject the null hypothesis

**Why**? – There is no statistical significance since Pr(>F) > 0.05

**Conclusion** – Best fit model is the additive model. Summary of the model does not show statistical significance for weekday and weekend alcohol consumption which is why we do not need to run a post hoc TukeyHSD. Pr>F 0.0.568, 0.262 respectively.

**Report** - There is no significant difference in effect on G3 grade by weekend (f(4) = 0.736, p>0.05) and workday (f(4) = 1.319, p>0.05) alcohol consumption and the interaction between these terms was also not significant.

**Test all assumptions** – The response variable looks slightly normally distributed in the histogram. The qq plot shows that the residuals are normally distributed. There is no outlier in the box plot. However, the sample data are dependent. The var test and bartlett test shows that it does not have equal variance.

2. **Null hypothesis**- There is no effect of weekend and workday alcohol consumption on G3 grades for the Portuguese dataset.

**Research hypothesis**- There is effect of weekend and workday alcohol consumption on G3 grades for the Portuguese dataset.

**a-level** – 0.05

**Rejection criteria** – – If we find statistical support for the research hypothesis (Pr(>F) < a), we reject the null hypothesis.

**Rejected/Accepted (null)** – We reject the null hypothesis

**Why?** – There is significance difference found

**Conclusion** – Best fit model is the additive model. Summary of the mode show there is significant difference for weekend and workday alcohol consumption Pr>F 0.000188, 0.006811 respectively.

**Test all assumptions** – The response variable looks slightly normally distributed in the histogram. The qq plot shows that the residuals are normally distributed. However, there is outlier in the boxplot. The sample data are dependent. The var test and bartlett test shows that it does not have equal variance.

**Report** - There is significant difference in effect on G3 grade by weekend (f(4) = 5.624, p<0.001) and workday (f(4) = 3.573, p<0.001) alcohol consumption, though the interaction between these terms was not significant.

**TukeyHSD** - A TukeyHSD post-hoc test revealed significant pairwise differences between weekend 4 and 1 (-1.33 under 4), weekend 5 and 1 (-1.80 under 1), weekend 4 and 2 (-1.23 under 2), weekend 5 and 2 (-1.70 under 5) alcohol consumption, and workday alcohol consumption of 4 and 1 (-2.39 under 4).

3. **Null Hypothesis**-There is no effect of weekend and workday alcohol consumption on G3 grades for mathematics and Portuguese datasets.

**Research hypothesis**- There is effect of weekend and workday alcohol consumption on G3 grades for mathematics and Portuguese datasets.

**a-level** – 0.05

**Rejection criteria** – If we find statistical support for the research hypothesis (Pr(>F) <a), we reject the null hypothesis.

**Rejected/Accepted (null)** – We fail to reject the null hypothesis on the mathematics dataset. However, we reject the null hypothesis on Portuguese dataset.

**Why?** – For mathematics dataset, there is no statistical significance since Pr(>F) > 0.05. However, on Portuguese dataset, there is significance difference found.

**Conclusion** – The mathematics data shows that there is no effect of weekend and workday alcohol consumption on G3 grades as the Pr(>F) > 0.05 which is why we fail to reject the null hypothesis. However, in the Portuguese data, we reject the null hypothesis because there was significant effect of weekend and workday alcohol consumption on G3 grades. The TukeyHSD post hoc result shows that there was significant difference between weekend 4 and 1 (-1.33 under 4), weekend 5 and 1 (-1.80 under 1), weekend 4 and 2 (-1.23 under 2), weekend 5 and 2 (-1.70 under 5) alcohol consumption, and workday alcohol consumption of 4 and 1 (-2.39 under 4). I conclusion we can say that the alcohol consumption on weekday and workday does not have any effect on the G3 grade of Mathematics. However, it has significant effect on G3 grade of Portuguese.

**Test all assumptions** – For mathematics dataset, the response variable looks slightly normally distributed in the histogram and there is no outlier in the box plot. However, on Portuguese dataset, the response variable looks slightly normally distributed in the histogram but there is outlier in the boxplot. Both the data shows normal distribution on residual in qq plot. The sample data are dependent for all the datasets. On both the datasets, the samples do not have equal variance.

**Part 3 -**

1. 1**. Null hypothesis**- There is no correlation between age and G3 grade for Mathematics datasets.

**Research hypothesis**- There is correlation between age and G3 grade for Mathematics datasets.

**a-level** – 0.05

**Rejection criteria** – If we find statistical support for the research hypothesis (p < a), we reject the null hypothesis.

**Rejected/Accepted (null)** – We reject the null hypothesis.

**Why?** – We would reject the null since p<0.05. This shows there is statistically significant negative correlation between age and G3 grade on Mathematics datasets.

**Conclusion** – The two variables are strongly negative correlated, and the relationship was significant, r (393) = -0.1615794, p = 0.001271

**Test all assumptions** – Variables are quantitative. The scatter plot shows linear relationship, and it does not look there is any big outliers. The box plot shows equal variance. The qq plot and histogram shows that the samples are somewhat normally distributed. The two pairs are also related pair as it’s in the same data set and two variables are from same individual student.

2. **Null hypothesis-** There is no correlation between age and G3 grade for Portuguese datasets.

**Research hypothesis**- There is correlation between age and G3 grade for Portuguese datasets.

**a-level** – 0.05

**Rejection criteria** – If we find statistical support for the research hypothesis (p < a), we reject the null hypothesis.

**Rejected/Accepted (null)** – We reject the null hypothesis.

**Why?** – We would reject the null since p<0.05. This shows there is statistically significant negative correlation between age and G3 grade on Portuguese dataset.

**Conclusion** – The two variables are strongly negative correlated, and the relationship was significant, r (647) = -0.1065054, p = 0.006612

**Test all assumptions** – Variables are quantitative. The scatter plot shows linear relationship, and it does not look there is any big outliers. The box plot shows equal variance. The qq plot and histogram shows that the samples are somewhat normally distributed. The two pairs are also related pair as it’s in the same data set and two variables are from same individual student.

3. **Null hypothesis**- There is no correlation between mathematics and Portuguese Grade for students who are in both classes.

**Research hypothesis**- There is correlation between mathematics and Portuguese Grade for students who are in both classes.

**a-level** – 0.05

**Rejection criteria** – If we find statistical support for the research hypothesis (p < a), we reject the null hypothesis.

**Rejected/Accepted (null)** – Rejected

**Why?** – We would reject the null since p<0.05. This shows there is statistically significant positive correlation between age and G3 grade on Mathematics datasets.

**Conclusion** – The two variables are somewhat positively correlated, and the relationship was significant, r (380) = 0.4803494, p < 2.2e-16.

**Test all assumptions** – Variables are quantitative. The scatter plot shows somewhat linear relationship, and it does not look there is any big outliers. The box plot shows equal variance. Both the variables look somewhat normally distributed in histogram. The two pairs are also related pair as it’s in the same data set and two variables are from same individual student.

**Part 4 -**

1. 1. **Null Hypothesis**-There is no relationship between age and G3 grade for Mathematics datasets.

**Research hypothesis**- There is relationship between age and G3 grade for Mathematics datasets.

**a-level** – 0.05

**Predictor variable= age from Mathematics dataset.**

**Response variable = Grade from Mathematics dataset.**

**Rejection criteria** – If we find statistical support for the research hypothesis (p < a), we reject the null hypothesis.

**Rejected/Accepted** (null) – We reject the null hypothesis.

**Why?** – f statistic and p value are less than alpha. P = 0.001271

**Conclusion** – Linear regression model: G3 = 20.1011 + (-0.5801) \*(age). The relationship was significant, and we can conclude that there is relationship between age and G3 grade. We can also conclude that the model does a statistical significance job of predicting the response variable.

**Test all assumptions** – The scatter plot shows linear relationship, and it does not look there is any big outliers. We are assuming that the variables are independence. The plot looks like residuals are random thus it has homoscedasticity. QQ plot shows the residual are normally distributed.

**Part 5-**

1. 1. **Null hypothesis**- Student who have internet access and who doesn’t have the same effect on mathematics students G3 grade.

**Research hypothesis**- At least one type of internet access has different effect on G3 grade effect on mathematics students.

**a-level** – 0.05

**Rejection criteria** – If we find statistical support for the research hypothesis (Pr(>F) < a), we reject the null hypothesis.

**Rejected/Accepted (null)** – We fail to reject the null hypothesis.

**Why?** – There is no significant difference found between student who has internet

access and who doesn't. Pr(>F) > 0.05 which is why we fail to reject the null hypothesis.

**Conclusion** – We conclude that there is not significance difference between student who has internet access and who doesn't. The effect on students Mathematics G3 grade is the same for both student who does and who doesn't.

**Test all assumptions** – The response variable looks slightly normally distributed in the histogram. The qq plot shows that the residual is normally distributed. There is no outlier in the box plot. However, the sample data are dependent. Bartlett test shows that there is equal variance.