[redacted] SREENITHI BALASUBRAMANIAN

Question 1: Is there a difference for G3- final grades for these Portuguese students between students with internet from students without internet?

Answer:

- Using the function `byf.shapiro(…)` - **Shapiro-Wilk Normality test**, we get a P-value of 0.00024 and 1.878e-11 < 0.05 (for G3 and internet). This indicates the data is NOT normal. Hence, we choose **non-parametric Mann-Whitney U Wilcoxon test**.
- On performing that, we get a p-value of 0.0324 < 0.05, indicating **there exists a difference in G3 final grades for students with and without internet.**

R Syntax

```
library(RVAideMemoire)

#Check normality
byf.shapiro(as.matrix(df$G3)~internet_num,data=df)



#Mann-Whitney U (Wilcoxon) test - Nonparametric T-Test
wilcox.test(df$G3~df$internet)
```

Output:

```
> library(RVAideMemoire)
> byf.shapiro(as.matrix(df$G3)~internet_num,data=df)

        Shapiro-Wilk normality tests

data:  as.matrix(df$G3) by internet_num

       W   p-value
1 0.9150  0.000247 ***
2 0.9285 1.878e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> wilcox.test(df$G3~df$internet) # where y is numeric and A is A binary factor

	Wilcoxon rank sum test with continuity correction

data:  df$G3 by df$internet
W = 9053, p-value = 0.0324
alternative hypothesis: true location shift is not equal to 0
```

Question 2: Is there a relationship between activities and higher education (higher variable)?

Answer:

- We use chi squared with error correction test to see if there is a relation between two binary variables – activities and higher education.
- We get a p-value of 0.0914 > 0.050, indicating there is **NO RELATION between activities and higher education.**
- Using `shapiro.test(…)` we get p-values for both variables to be less than 0.05, indicating the data is NOT normal. Hence we use **non-parametric tests.**

<u>R Syntax</u>:

```
library(gmodels)

#Check normality
shapiro.test(df$activities_num)
# p-value < 2.2e-16 < 0.05 = NOT NORMAL
shapiro.test(df$higher_num)
# p-value < 2.2e-16 < 0.05 = NOT NORMAL

#Chi-squared test
CrossTable(df$activities_num,df$higher_num,  digits = 2,
expected=TRUE, prop.r=TRUE, prop.c = TRUE, prop.chisq =
FALSE, chisq = TRUE, fisher = TRUE, format="SPSS")
```

Output

```
> library(gmodels)
> shapiro.test(df$activities_num) # p-value < 2.2e-16 < 0.05 = NOT NORMAL

        Shapiro-Wilk normality test

data:  df$activities_num
W = 0.6364, p-value < 2.2e-16

> shapiro.test(df$higher_num) # p-value < 2.2e-16 < 0.05 = NOT NORMAL

        Shapiro-Wilk normality test

data:  df$higher_num
W = 0.2257, p-value < 2.2e-16

> CrossTable(df$activities_num,df$higher_num,  digits = 2, expected=TRUE, prop.r=TRUE,
+            prop.c = TRUE, prop.chisq = FALSE, chisq = TRUE, fisher = TRUE, format="SPSS")

   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  395

                  | df$higher_num
df$activities_num |        1  |        2  | Row Total |
------------------|-----------|-----------|-----------|
                1 |       14  |      180  |      194  |
                  |     9.82  |   184.18  |           |
                  |    7.22%  |   92.78%  |   49.11%  |
                  |   70.00%  |   48.00%  |           |
                  |    3.54%  |   45.57%  |           |
------------------|-----------|-----------|-----------|
                2 |        6  |      195  |      201  |
                  |    10.18  |   190.82  |           |
                  |    2.99%  |   97.01%  |   50.89%  |
                  |   30.00%  |   52.00%  |           |
                  |    1.52%  |   49.37%  |           |
------------------|-----------|-----------|-----------|
     Column Total |       20  |      375  |      395  |
                  |    5.06%  |   94.94%  |           |
------------------|-----------|-----------|-----------|


Statistics for All Table Factors
```

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  3.677104     d.f. =  1     p =  0.05516458

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  2.849511     d.f. =  1     p =  0.09140174


Fisher's Exact Test for Count Data
------------------------------------------------------------
Sample estimate odds ratio:  2.522108

Alternative hypothesis: true odds ratio is not equal to 1
p =  0.0669367
95% confidence interval:  0.8871817 8.185027

Alternative hypothesis: true odds ratio is less than 1
p =  0.9851212
95% confidence interval:  0 6.79846

Alternative hypothesis: true odds ratio is greater than 1
p =  0.04479781
95% confidence interval:  1.024366 Inf



       Minimum expected frequency: 9.822785
```

Question 3: Is there a difference between the G2-second grade and the G3-final grade?

Answer:

- On performing the normality tests for G2 (p-value: 2.084e-07) and G3 (p-value: 8.836e-13) they're both lesser than 0.05, indicating they are **not norma**l. **We use non-parametric tests**. The data is paired.

- Hence, we select the Wilcox test to determine if there is a difference between the two. The p-value using Wilcox test comes up as 0.874 > 0.05, indicating **there is no difference between G2 and G3 final grade.**

R Syntax:

```
# check for normality
shapiro.test(df$G2) # P < 0.05 = NOT NORMAL
shapiro.test(df$G3)  #P < 0.05 = NOT NORMAL


#Paired Test - Nonparametric (Sign Rank test)
wilcox.test(df$G2, df$G3, paired=TRUE)
```

Output

```
> shapiro.test(df$G2) # P < 0.05 = NOT NORMAL

        Shapiro-Wilk normality test

data:  df$G2
W = 0.96914, p-value = 2.084e-07

> shapiro.test(df$G3)  #P < 0.05 = NOT NORMAL

        Shapiro-Wilk normality test

data:  df$G3
W = 0.92873, p-value = 8.836e-13
```

```
> wilcox.test(df$G2, df$G3, paired=TRUE)

	Wilcoxon signed rank test with continuity correction

data:  df$G2 and df$G3
V = 10274, p-value = 0.8748
alternative hypothesis: true location shift is not equal to 0
```

Question 4: Create two categorical variables for G2-second grade and the G3-final grade using a cutoff of 15. Test whether there is a difference between the two categorical variables. Hint: Remember what these categorical variables represent.

Answer:

- The values are put into categories of [0-15) and (15-Inf)
- Using normality tests, we see that p-value for both the categorical variables are less than 0.05 (p-value < 2.2e-16) indicating they are not normal. Hence, we use **non-parametric tests**.
- This data is paired.
- We use McNemar Chi-squared test as the data is paired. We get p-value of 0.096 indicating **there is no difference between G2 and G3 categorical variables.**

R Syntax:

```
# Create two categorical variables for G2-second grade and
the G3-final grade using a cutoff of 15.
# Test whether there is a difference between the two
categorical variables.

df$G2_binary <-cut(df$G2,c(0, 15, Inf), include.lowest =
TRUE, labels=c("0","1"))
df$G2_binary <- as.numeric(df$G2_binary)

df$G3_binary <-cut(df$G3, c(0, 15, Inf), include.lowest =
TRUE, labels=c("0","1"))
```

```
df$G3_binary <- as.numeric(df$G3_binary)


table(df$G2_binary)
table(df$G3_binary)

shapiro.test(df$G2_binary)
shapiro.test(df$G3_binary)

mcnemar.test(df$G2_binary,df$G3_binary)

# ACCORDING TO MCNEMAR TEST, P = 0.096  > 0.05. There is a
no difference
```

Output

```
> df$G2_binary <-cut(df$G2,c(0, 15, Inf), include.lowest = TRUE, labels=c("0","1"))
> df$G2_binary <- as.numeric(df$G2_binary)
> df$G3_binary <-cut(df$G3, c(0, 15, Inf), include.lowest = TRUE, labels=c("0","1"))
> df$G3_binary <- as.numeric(df$G3_binary)
> table(df$G2_binary)

  1   2 
362  33 
> table(df$G3_binary)

  1   2 
355  40 
> shapiro.test(df$G2_binary)

	Shapiro-Wilk normality test

data:  df$G2_binary
W = 0.30799, p-value < 2.2e-16

> shapiro.test(df$G3_binary)

	Shapiro-Wilk normality test

data:  df$G3_binary
W = 0.34449, p-value < 2.2e-16



> mcnemar.test(df$G2_binary,df$G3_binary)

	McNemar's Chi-squared test with continuity correction

data:  df$G2_binary and df$G3_binary
McNemar's chi-squared = 2.7692, df = 1, p-value = 0.09609
```

Question 5: What is the proportion of students receiving extra educational support? • Test whether this proportion is different than 50%?

Answer:

- Proportion of students receiving extra support: ~13% (p = 0.1291)
- 51 out of 395 students receive extra support.
- 95% CI: (0.0017 0.1981) => between 1% and 19.8% of student receive extra support
- This proportion is different from 50% with a p-value < 2.2e-16, which is lesser than 0.05.
- X-squared = 217.34 which indicates the Z-score is sqrt(217.34) = 14.7

R Syntax

```
table(df$schoolsup_num)

51+344

extra_sup <- prop.test(x=51, n=395, p=0.50, correct=FALSE)

extra_sup
```

<u>Output</u>

```
> table(df$schoolsup_num)

  1   2 
344  51 
> 51+344
[1] 395
> extra_sup <- prop.test(x=51, n=395, p=0.50, correct=FALSE)
> #2-YES, 1-NO
> extra_sup

	1-sample proportions test without continuity correction

data:  51 out of 395, null probability 0.5
X-squared = 217.34, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.09958378 0.16578847
sample estimates:
        p 
0.1291139 
```

Question 6: What is the proportion of males and females in extracurricular activities? • Construct a 95% Confidence Interval

Answer:
For our question:

- PROPORTION OF ALL MALE IN EXTRA CURRICULAR = 47.76% (0.4776)
- PROPORTION OF ALL FEMALE IN EXTRA CURRICULAR ACTIVITIES = 52.2% (0.5222)
- Out of 201 students who are in extracurricular activities, 96 are female and 105 are male.

95% confidence intervals:

- Female: (0.4096278 0.5464358) - 40.9 - 54.6% of female students are in extra curricular activities
- Male: (0.4535642 0.5903722) - 45.3% - 59.0% of male students are in extra curricular activities

R Syntax:

```
table(df$sex, df$activities)


table(df[df$activities=='yes',]$sex)
#female
96 + 105
female_yes<- prop.test(x=96, n=201, p=0.5, correct=FALSE)

#MALE
male_yes<-  prop.test(x=105, n=201, p=0.5, correct=FALSE)

female_yes
male_yes
```

Output

```
> # What is the proportion of males and females in extracurricular activities?
> # • Construct a 95% Confidence Interval
> table(df$sex, df$activities)

     no yes
  F 112  96
  M  82 105
> table(df[df$activities=='yes',]$sex)

  F   M
 96 105
> #female
> 96 + 105
[1] 201
```

```
> female_yes<- prop.test(x=96, n=201, p=0.5, correct=FALSE)
> #MALE
> male_yes<-  prop.test(x=105, n=201, p=0.5, correct=FALSE)
> female_yes

        1-sample proportions test without continuity correction

data:  96 out of 201, null probability 0.5
X-squared = 0.40299, df = 1, p-value = 0.5256
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4096278 0.5464358
sample estimates:
        p
0.4776119

> male_yes

        1-sample proportions test without continuity correction

data:  105 out of 201, null probability 0.5
X-squared = 0.40299, df = 1, p-value = 0.5256
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4535642 0.5903722
sample estimates:
        p
0.5223881
```

Question 7: What data cleaning techniques were required to prepare this data for analysis? Are there any changes you would make about how the data was collected or asked?

Answer:

- There were **no missing values** to eliminate or fill.
- Variables with “yes” or “no” values – categorical variables were changed to numerical 1s and 0s.
- Variables changed (according to what was needed for the questions): sex, activities, internet, higher, schoolsup
- For converting G2 and G3 into two categories (0<=15<=Inf), the cut function was used.
- This dataset preserves the privacy of those from whom the data was collected. It does not mention anything about a person's identity.
- This dataset talks about grades for a group of Portuguese students. From their age range, it looks like they are in high school. I would reach out and collect data from high schoolers belonging to other regions and other schools.
- This sample is small and specific to one school and one region, so we cannot generalize well. If we do cluster sampling for a bunch of schools, it’d be better to generalize any statistical conclusion – for example when using regression.
- It did not mention how the data was collected: but I’d send out an online survey to high schoolers belonging to different regions and include another column “region” to indicate which country or city they are from.

Question 8: What are some limitations of these study questions?

Answer:

- The data is not normally distributed. The data can be highly skewed with a lot of outliers - for ex: the number of students with grades below a certain threshold could be less.

- The group is small – only one set of Portuguese high school students are involved. So, generalizing statistical findings will be difficult.
- Also there is no historical data available - we cannot see the change of values over time. If this data was collected repeatedly over several years, we can see how values change for different samples over time.

Question 9: What do we learn about these students from the answers to questions 1-6? Note: Summarize these findings as conclusions, no statistical terms should be mentioned here.

- We can see there is a significant difference in the final grades for students having internet and those without internet. This could indicate the importance of an internet connection in receiving a better grade - students may not have been able to access materials online without an internet connection. But it could also have the opposite effect – students with internet could have received a lesser grade.
- It tells us that physical activity has no association with someone wanting to pursue higher education.
- It tells us there is no significant difference between final grades and grades received during the second period (G2)
- It tells us that there is a higher proportion of male students who are in extra curricular activities compared to female students
- It tells us that only a small proportion of students receive extra financial support.

Question 10: What would be future work or additional research questions to study based upon the present findings of questions 1-6?

- Does extra curricular activities influence a student's final grades?
- Does wanting to pursue higher education influence a student's final grade?
- Does having extra financial support have an influence on a students grade?

- Based on previous grades (G1 and G2), presence of internet connection, extra curricular activities and the want to pursue higher education, can we predict the range of final grade G3?
- Is there a relationship between receiving financial support and higher education?