MA 361- Project 2

Hypothesis Testing

Tara Rondinelli, Nicole Dickson, Brighid King

Table of Contents

Introduction:

The level of education among females is not the same around the world. It is important to be aware of these differences because it reflects back on the country. This report focuses on the difference in expected years of education for females across the countries of Europe as well as across two years. Specifically, this report will answer the questions: Is there a difference in mean years of education for females among the countries? Is there a difference in mean years of education for females between the years 2018 and 2020?

The data for this report were obtained from the World's Bank website. Then, the European countries were extracted for the dataset used in this report. Using R, confidence intervals and hypothesis tests will be conducted at the 0.05 significance level to answer the initial questions. The hypothesis test will have two different null hypotheses for each question. The first will be the 35 countries' mean years of education for females being equal to each other, and the alternative hypothesis is at least one of the means is different from the others. If the null hypothesis is rejected, one can conclude with statistical evidence that at least one of the European countries has a different mean years of education for females. If the null hypothesis is not rejected, one can conclude that there is not significant evidence to suggest that European countries have differing mean years of expected education for females. The second hypothesis test will have a null hypothesis where the block (years (2018 and 2020)) means of years of expected education for females are equal to each other, and the alternative hypothesis is that the means are different from each other. If this null hypothesis is rejected, one can conclude with statistical evidence that the two years have different mean expected years of education for females. If the null hypothesis is not rejected, one can conclude that there is not enough statistical evidence to suggest that 2018 and 2020 have different mean expected years of education for females. The first 100(1-α)% confidence intervals will be for the difference between the countries. Albania will be compared to the 35 countries and shown as a representative sample of some of the concluding confidence intervals for countries. The results will show if there is a difference between each country and Albania's means of years of education for females. Depending on the results, it can be concluded whether or not the means from each country significantly differ. The second 100(1-α)% confidence interval will be for the difference between the two means of the years 2018 and 2020. The results will show if there is a significant difference between the two means of years of education for females between the years 2018 and 2020.

Chapter 1: Description of Data

Obtaining the Data

The data were obtained from the World Bank's website. From the initial data extraction, the European countries were removed and placed into the dataset used for this project.

Numerical Descriptions

There are six obvious outliers from the dataset. The approach used to find outliers utilized multiplying the Interquartile Range (IQR) by 1.5. Thus, data points meets the criteria for removal from the dataset if they are outside of (Q_1- (1.5*IQR), Q_3+ (1.5*IQR)) where Q_1 equals the first quartile and Q_3 equals the third quartile.

In the original extracted dataset of 39 European countries, the IQR equals 0.69. Therefore, the data points outside of (11.995, 14.065) were removed. Furthermore, Romania's 2020 data point met criteria for rejection but the 2018 data point did not. For consistency and continuity in analyzing this data, both data points were removed.

Original Data Summary

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
10.98	13.03	13.48	13.25	13.72	13.93

After Removing Outliers Data Summary

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
12.24	13.27	13.53	13.43	13.74	13.93

One should note that 12.24 years is not an outlier according to the original extracted dataset, but it is in the revised version of the data. However, this data set sees heavy concentration around 13.5 years. Thus, this concentration explains the narrow criteria for the slightly differing data points. The outlier removal is solely based on the first examination of the data.

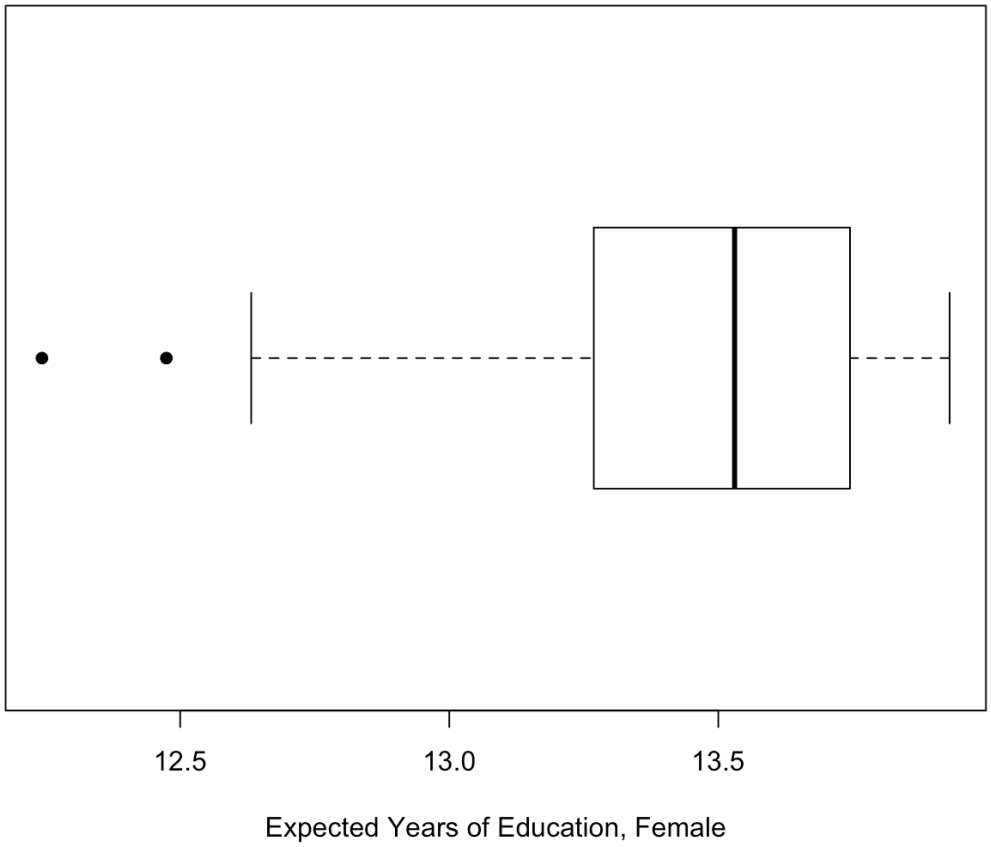
In looking at the data given, this trend makes sense. The countries in Europe will not all see the same years of expected years of education for females, but the majority of them will have similar expectations. Not all countries progress at the same rate, but many try to be competitive with one another. The competition explains the concentration of the data around 13.5 and the lack of

outliers on the right side of the dataset. Some countries will fail to have the proper resources to increase the amount of education for females, which explains the lower outliers.

The overall standard deviation for the data as a whole is approximately 0.425. With respect to the years, the standard deviations are 0.4282 for 2018 and 0.4280 for 2020. As for the countries, the majority of them see no variation, so the standard deviation is zero. However, Albania, Belarus, Montenegro, Serbia, and Ukraine all have very small standard deviations. They are 0.0473, 0.0055, 0.0072, 0.0822, and 0.0005, respectively.

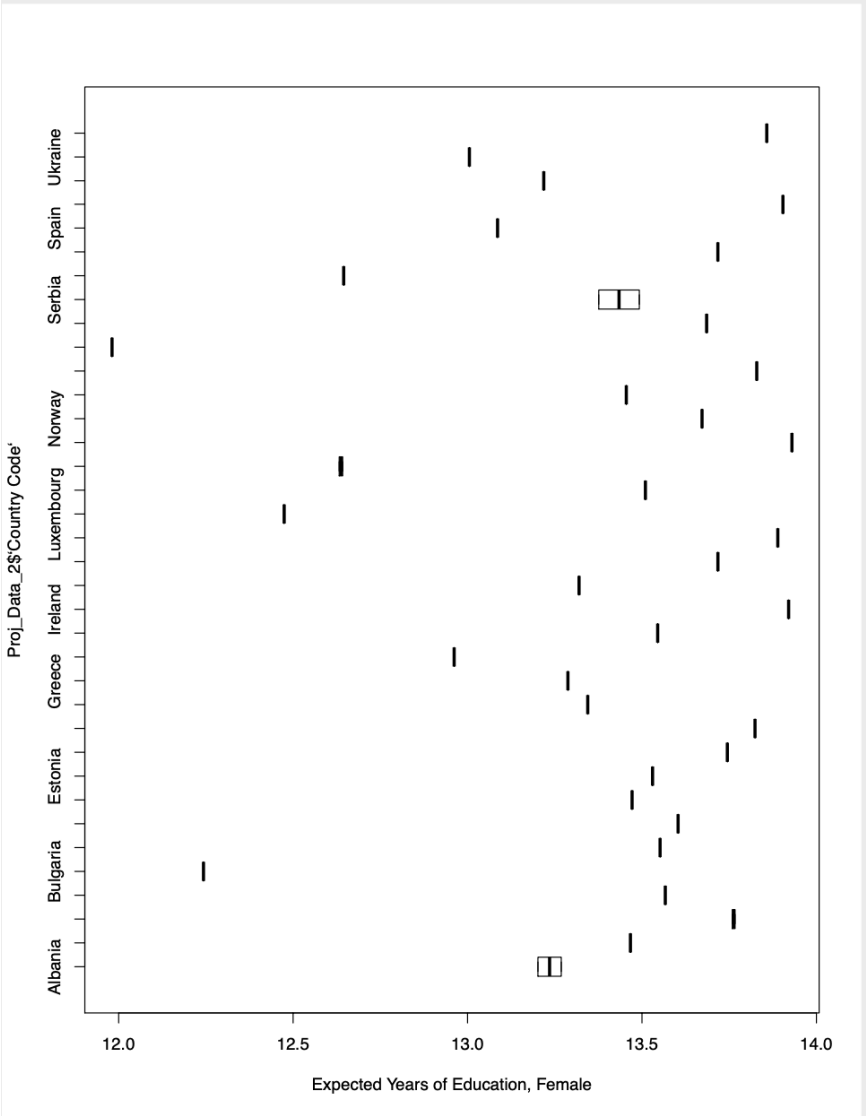
Graphical Descriptions

Data Overall:



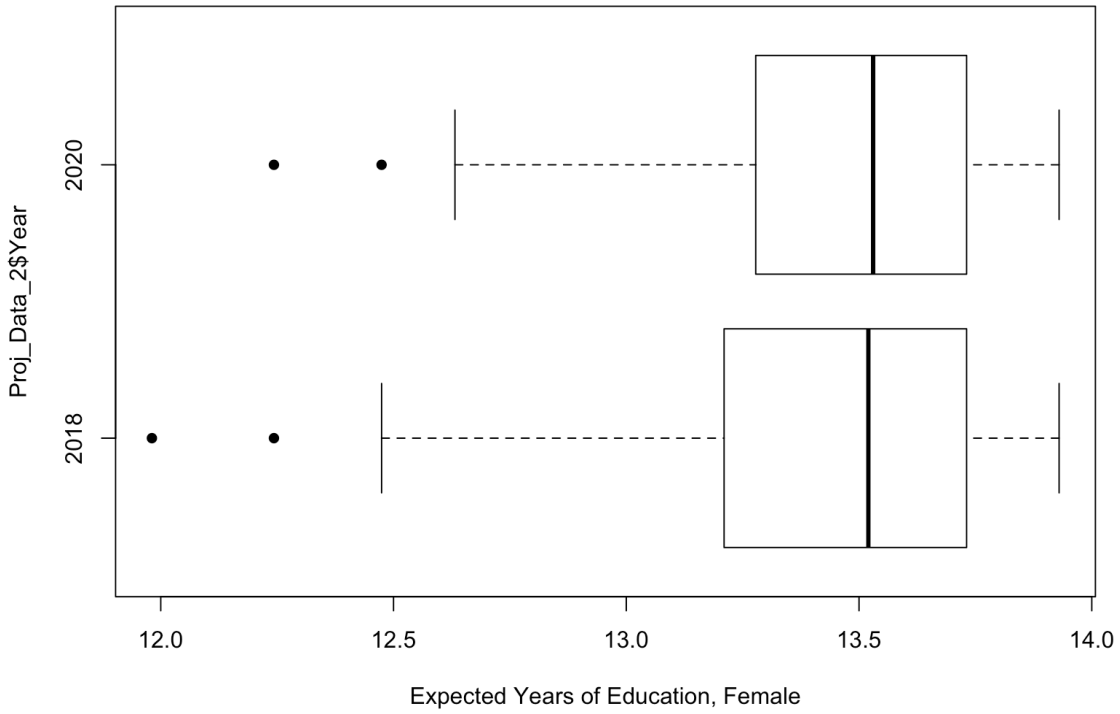
The boxplot above shows that there are two outliers for the overall data set. These outliers were removed. It can also be seen that the mean years of education for females lies around 13.43. The data is skewed left with a minimum of approximately 12.6 and a maximum of approximately 13.93. The IQR is approximately 0.47.

By Country:



The majority of the boxplots for each country above has a mean of around 13.5. This makes seasons because the mean of the overall dataset after combining the data from these countries is 13.43. There are nine countries that have a smaller means closer to 12. With these different means, it can be predicted that the expected years of education for females will be different for at least one of the countries.

By Year:



Shown in the two boxplots above, it can be seen that there are two outliers for both 2018 and 2020. These outliers were removed. It can be seen that the mean years of education for females for both 2018 and 2020 are both very close to 13.5. Both the data from 2018 and 2020 are slightly skewed left with minimums of approximately 12.6 and 12.5 and maximums of approximately 13.8. Since the years have similar means and distributions, it can be predicted that the expected years of education for females will be the same for 2018 and 2020.

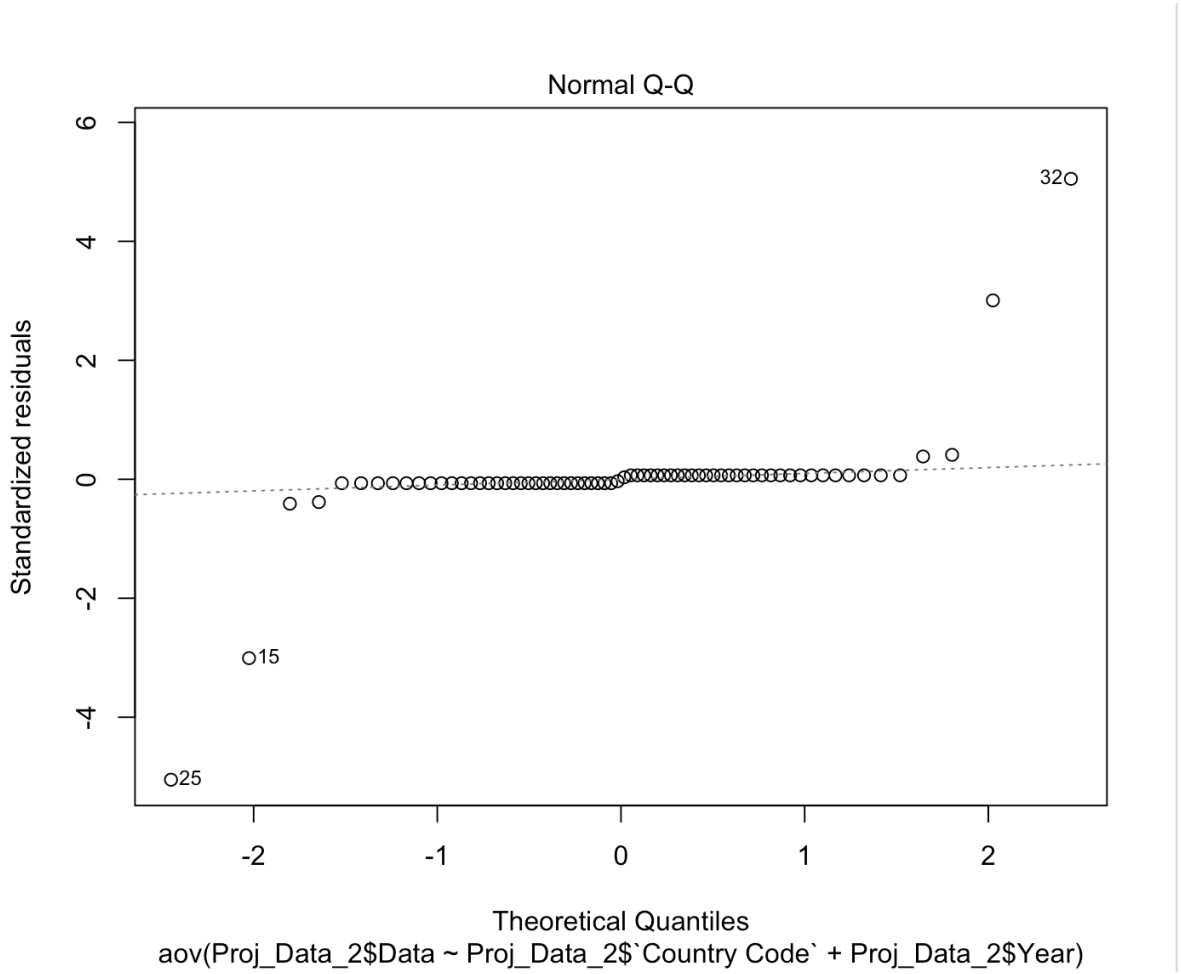
2018:

Minimum	1st Quartile	Median	3rd Quartile	Maximum
12.24310	13.25307	13.53030	13.73116	13.92983

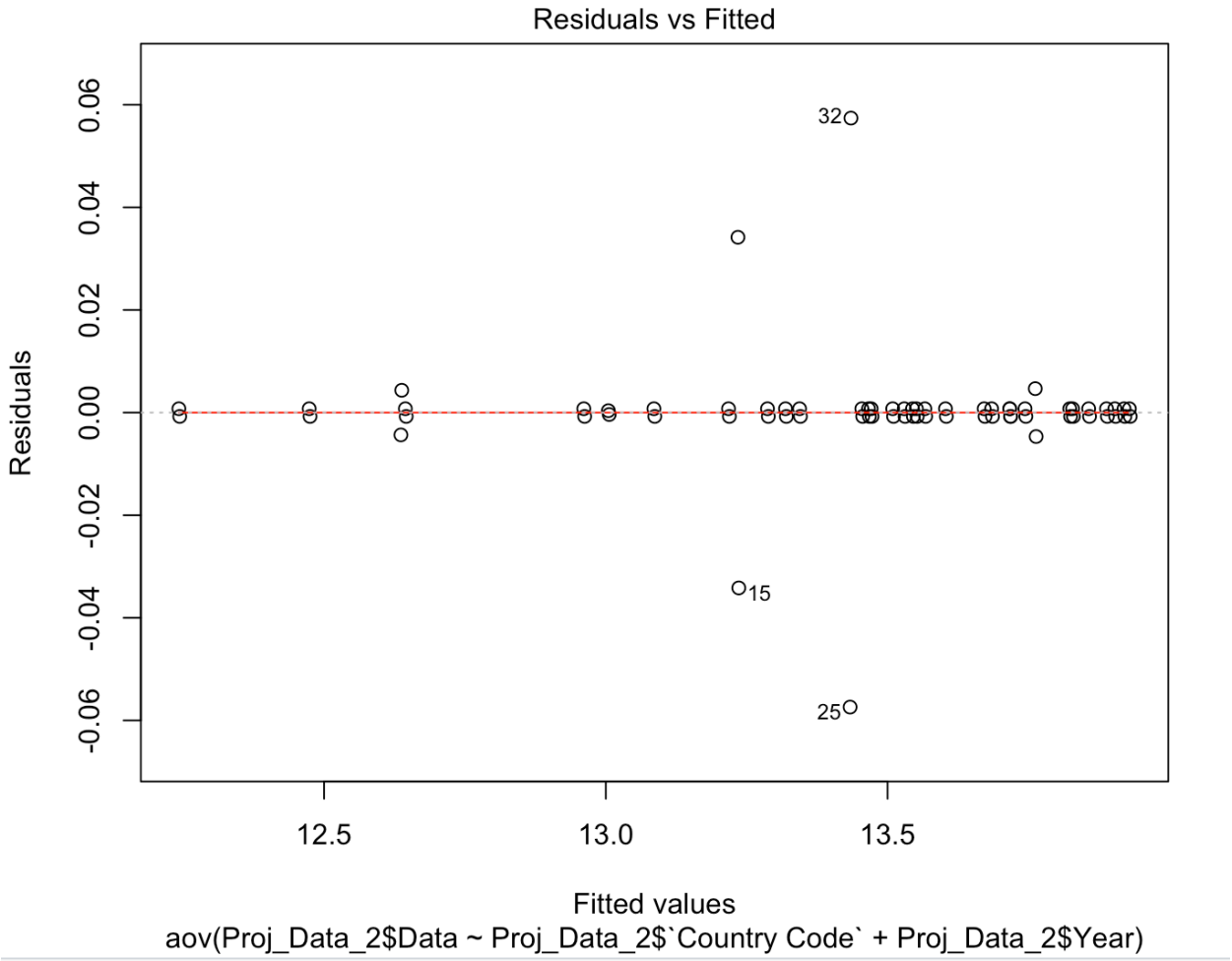
2020:

Minimum	1st Quartile	Median	3rd Quartile	Maximum

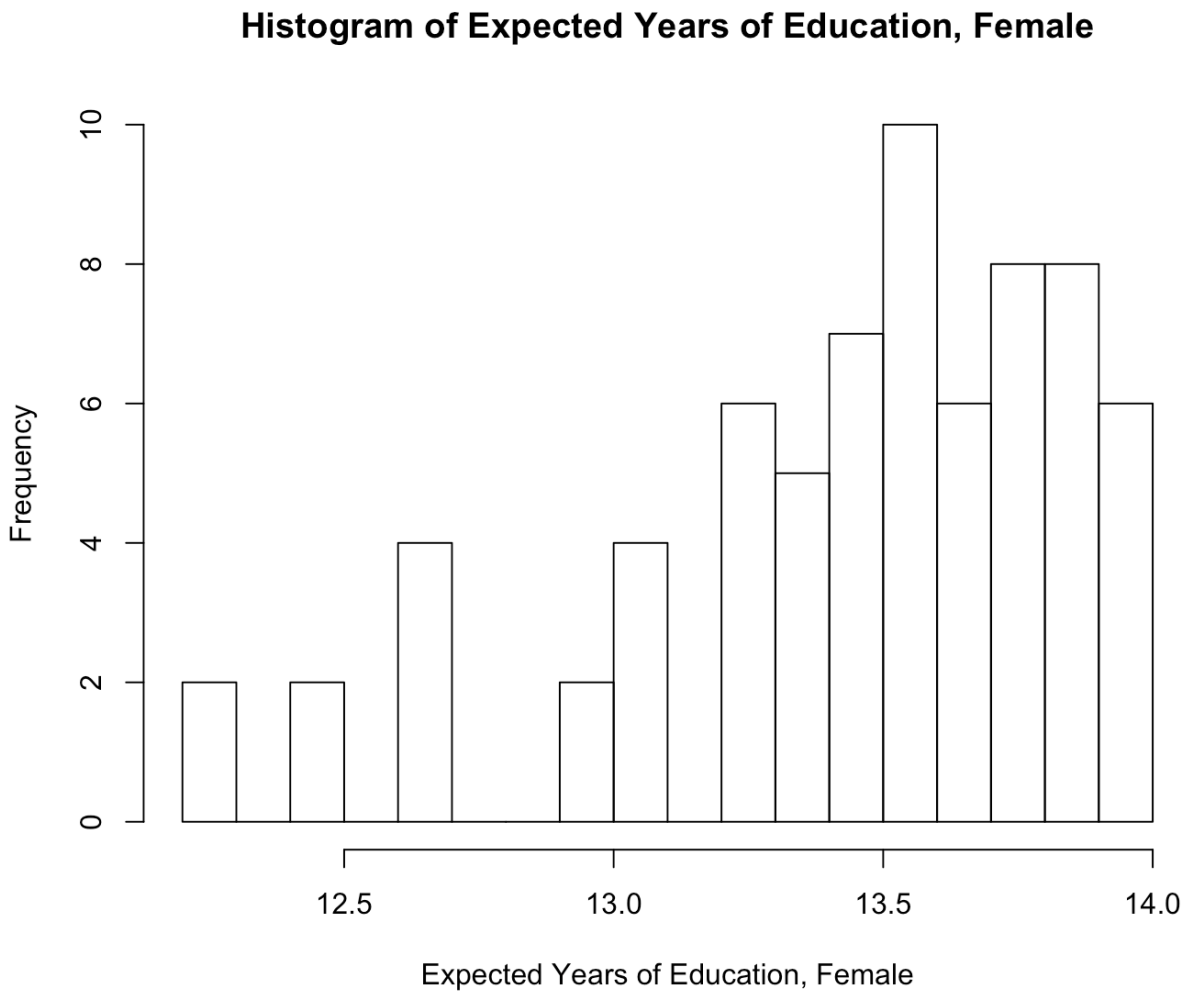
12.24310	13.27814	13.53030	13.73116	13.92983



The European countries obviously have very similar expectations for years of education for females. The majority of residuals fail to deviate far from zero, which creates a QQ plot with a very slight positive incline. Although slight, this positive incline shows us that the data is approximately normal.



Again, the residuals do not deviate far from zero. Since this is the case, it is evident that the data is approximately normal.



This histogram shows the slight left skew of this data and shows that the majority of data points fall between 13.0 years and 14.0 years. However, the dataset is larger than 30, and the slight variation is explained by the background of the set and the data trends seen in the outlier removal. Thus, it is safe to assume normality.

Chapter 2: Confidence Interval: Comparing Means

Assumptions

For the countries, a t-interval was constructed with one degree of freedom since $n \leq 30$ for each country. When looking at 2018 and 2020, because the samples were larger than 30, normality can be assumed and a z-test was used. Independence and collection of the data are also assumed for both tests.

95% Confidence Intervals

Treatments (Countries):
Albania: (12.811, 13.660)
Belarus: (13.713, 13.813)
Montenegro: (12.572,12.702)
Serbia: (12.695, 14.173)
Ukraine: (13.000, 13.010)

The other data points show no variation between the years; therefore, confidence intervals cannot be constructed for those countries. From the data points above, one can see that some of the countries do not overlap when compared to one another; therefore, it is likely that the null hypothesis for treatments (countries) will be rejected.

Blocks (Years):
(2020-2018):
(-0.1991, 0.20208)

2018:
(13.289, 13.573)

2020:
(13.297, 13.581)

Findings and Interpretation from Confidence Interval

The 95% confidence interval of the two means of expected years of education for females years between the years contains zero, suggesting that the two population means are not significantly different.

Chapter 3: Hypothesis Test

Assumptions

This two-way ANOVA Test assumes independence between the countries and between the years 2018 and 2020, respectively. Independence is also assumed between the countries and years. Normality is assumed because of the above examination of the data, including n being larger than 30. The variances are assumed to be approximately equal due to their small size. Most of them have zero variation, but a few countries saw a slight change.

Hypothesis Test

Both the rows (Country) and blocks (Year) are tested. Thus, there should be two hypotheses in order to draw proper conclusions.

Parameters: μ_i=mean number of years of education for females for country i=1,...,a; a=35
μ_j=mean number of years of education for females for a particular year j=1,...,b; b=2

Hypotheses:
H_o: $\mu_1 = \ldots = \mu_a$
H_a: at least one pair of means differ
(a = 35)

H_o: $\mu_1 = \ldots = \mu_b$
H_a: at least one pair of means differ
(b = 2)

ANOVA Table:

	SS	df	MS	F	P-value
Treatment (Countries)	12.453	34	0.3363	1377.558	$< 2x10^{-16}$
Blocks (Year)	0.000	1	0.000	0.149	0.702
Error	0.009	34	0.0003		
Total	14.462	69			

Findings and Interpretation from Hypothesis Test

For the first hypothesis test, since $\alpha = 0.05 <$ p-value $= 2 \times 10^{-16}$, the null hypothesis can be rejected at the 0.05 significance level. Therefore, there is statistically significant evidence to conclude that at least one of the European countries has a different mean years of education for females. For the second hypothesis test, since $\alpha = 0.05 >$ p-value $= 0.702$, the null hypothesis cannot be rejected at the 0.05 significance level. Therefore, there is not statistically significant evidence to conclude that the year 2018 and 2020 differ in respect to mean years of education for females.

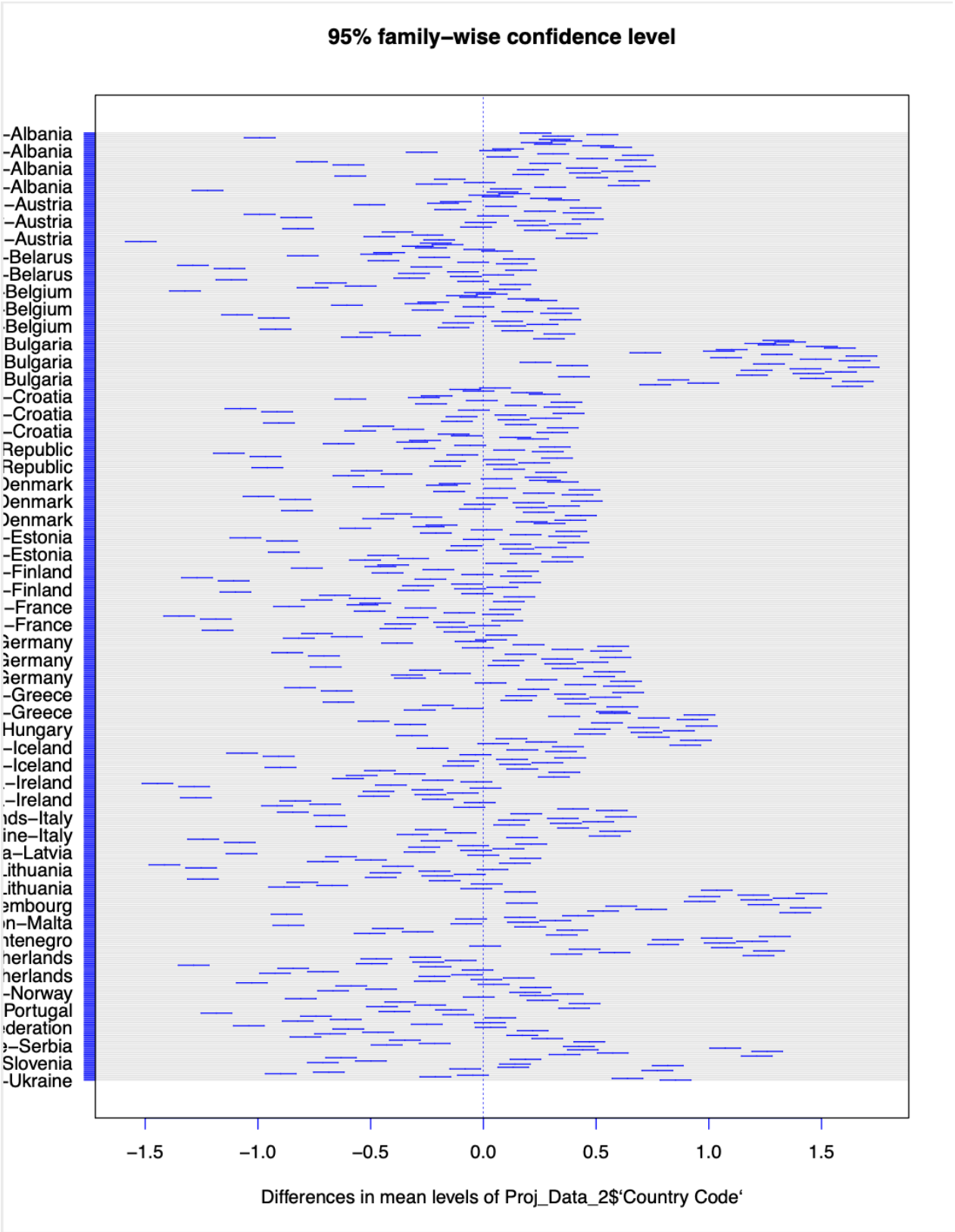
Chapter 4: Confidence Interval

Assumptions:

The confidence intervals were constructed using Tukey's Highest Significant Difference (HSD) approach. This approach assumes that the observations being tested are independent and that there are equal variances within-group across the groups associated with each mean in the test.

Confidence Interval:

For the countries, there are several confidence interval comparisons. The chart details some of them, but it certainly explains the rejection of countries as seen by the ANOVA test. Several confidence intervals fail to contain zero, thus explaining the ANOVA test's conclusion that at least one set of means differ.

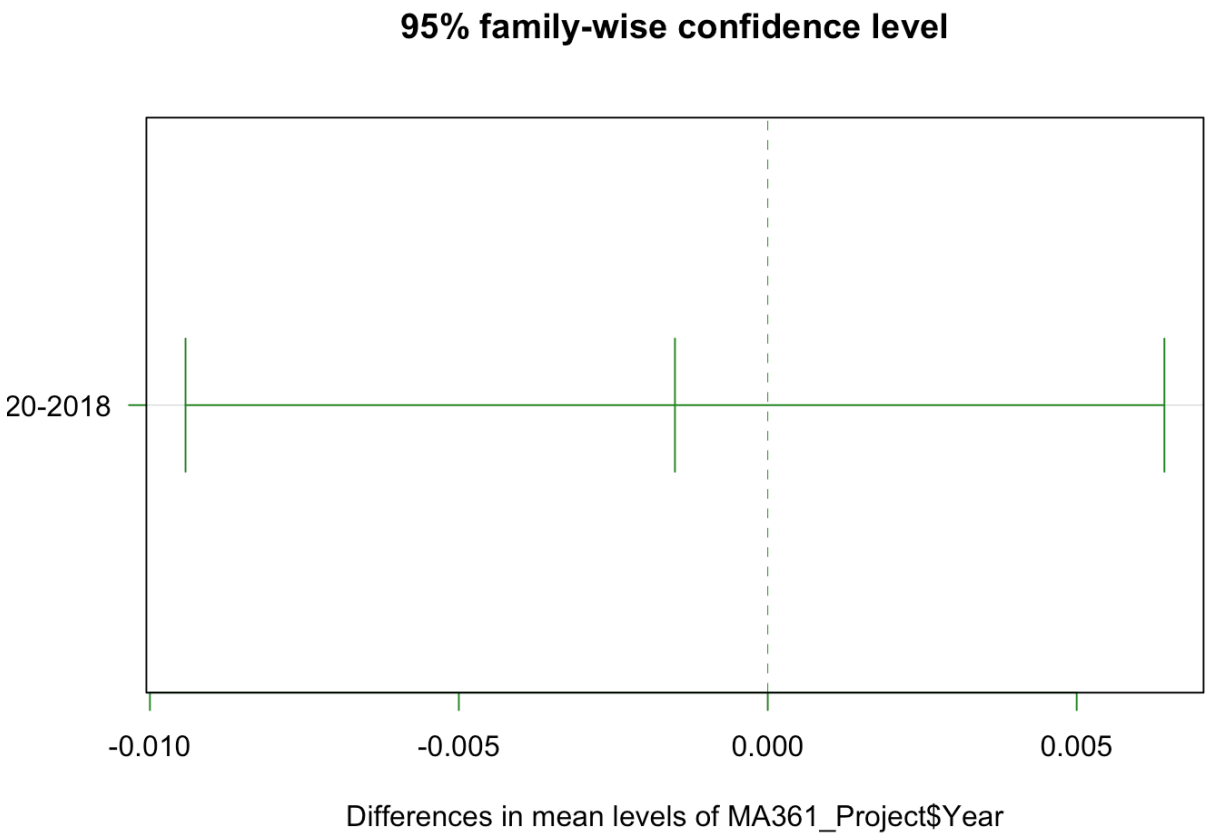

95% family−wise confidence level
−Albania
−Albania
−Albania
−Albania
−Austria
−Austria
−Austria
-Belarus
-Belarus
Belgium
Belgium
Belgium
Bulgaria
Bulgaria
Bulgaria
−Croatia
−Croatia
−Croatia
Republic
Republic
Denmark
Denmark
Denmark
-Estonia
-Estonia
−Finland
−Finland
−France
−France
Germany
Germany
Germany
−Greece
−Greece
Hungary
−Iceland
−Iceland
−Ireland
−Ireland
nds−Italy
ine−Italy
a−Latvia
ithuania
ithuania
embourg
n−Malta
ntenegro
herlands
herlands
−Norway
Portugal
deration
e−Serbia
Slovenia
-Ukraine
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Differences in mean levels of Proj_Data_2$'Country Code'

Austria-Albania (0.1636872964, 0.299747874)
Belarus-Albania (0.4596505814, 0.595711159)
Belgium-Albania (0.2635904064, 0.399650984)
Bulgaria-Albania (-1.0601019236, -0.924041346)
Croatia-Albania (0.2486377464, 0.384698324)
Czech Republic-Albania (0.3002763464, 0.436336924)
Denmark-Albania (0.1684089364, 0.304469514)
Estonia-Albania (0.2270980564, 0.363158634)
Finland-Albania (0.4414897664, 0.577550344)
France-Albania (0.5207515464, 0.656812124)
Germany-Albania (0.0413280164,) 0.177388594
Greece-Albania (-0.0155281336, 0.120532444)
Hungary-Albania (-0.3415626836, -0.205502106)
Iceland-Albania (0.2418027564, 0.377863334)
Ireland-Albania (0.6171136564, 0.753174234)
Italy-Albania (0.0164104164, 0.152470994)
Latvia-Albania (0.4144130464, 0.550473624)
Lithuania-Albania (0.5860744164, 0.722134994)
Luxembourg-Albania (-0.8288425736, -0.692781996)
Malta-Albania (0.2065959664, 0.342656544)
Montenegro-Albania (-0.6661371486, -0.530076571)
Netherlands-Albania (0.6266265564, 0.762687134)
Norway-Albania (0.3689037064, 0.504964284)
Poland-Albania (0.1518302664, 0.287890844)
Portugal-Albania (0.5257936164, 0.661854194)
Russian Federation-Albania (0.3821407064, 0.518201284)
Serbia-Albania (0.1310415964, 0.267102174)
Slovak Republic-Albania (-0.6582283336, -0.522167756)
Slovenia-Albania (0.4142995564, 0.550360134)
Spain-Albania (-0.2171387036, -0.081078126)
Sweden-Albania (0.6008144064, 0.736874984)
Switzerland-Albania (-0.0847477236, 0.051312854)
Ukraine-Albania (-0.2979153886, -0.161854811)
United Kingdom-Albania (0.5544524864, 0.690513064)

Above are the 95% confidence intervals between Albania and all the other European countries included in the dataset. As one can see, several of these comparisons result in a significant difference. Thus, the null hypothesis for the treatment (countries) is rejected since at least one pair of mean differences does not contain zero within the confidence interval. There is significant statistical evidence to conclude that at least one pair of means of expected years of education for females differ between the countries in Europe.

This visualization only provides some of the confidence intervals, but it shows the inconsistencies between the countries. Thus, it makes sense that the ANOVA test for treatment means (countries) shows at least one pair of means to be different because there are several pairs above that do not contain zero. This utilizes Tukey's Highest Significant Difference (HSD) approach. Tukey's method utilizes the "q" or "studentized range statistic" distribution and displays that several means differ from each other. Thus, the rejection of the null hypothesis for treatments (countries) is supported.

For the years, only one confidence interval is constructed. It contains zero, which supports the conclusion to fail to reject the null hypothesis for blocks (years).



Years using Tukey's HSD:
(-0.00942368, 0.006419279)

The graph above shows the 95% confidence interval for blocks (Years) using Tukey's HSD method. Do not reject the null hypothesis for the blocks since the interval (-0.00942368, 0.006419279) contains zero. Therefore, there is insignificant evidence to suggest that the mean expected years of education for females differs between 2018 and 2020.

Findings and Interpretation from Confidence Intervals

For the first hypothesis test, since $\alpha = 0.05 <$ p-value $= 2 \times 10^{-16}$, the null hypothesis can be rejected at the 0.05 significance level. Therefore, there is statistically significant evidence to conclude that at least one of the European countries has a different mean years of education for females. For the second hypothesis test, since $\alpha = 0.05 >$ p-value $= 0.702$, the null hypothesis fails to be rejected at the 0.05 significance level. Therefore, there is not statistically significant evidence to conclude that the year 2018 and 2020 differ in respect to mean years of education for females.

Conclusion:

The null hypothesis for the treatments is rejected, indicating that at least one mean must be different for the data used in the set. However, there is not enough statistically significant evidence to reject the null hypothesis of the blocks. At a significance level of five percent, we fail to reject the null hypothesis. Thus, one can conclude that there is no significant difference in the means of expected years of education for females between the years 2018 and 2020 for the 35 European countries in the sample.

The confidence intervals visualize the data and support the conclusions of the hypotheses. The graph of the 95 percent confidence intervals show that at least one pair of expected years of education for females is different. Thus, the null hypothesis of the treatments (countries) is rejected at the $\alpha = 0.05$ level. However, the 95 percent confidence interval graph contains zero. Therefore, the null hypothesis fails to be rejected for the blocks (years), and there is insufficient statistical evidence to conclude that the mean expected years of education for females differs between the years 2018 and 2020 for the 35 European countries in the sample.

The results of the hypothesis tests for the two-way ANOVA and the confidence intervals at a significance level of 5 percent both show that there is a significant difference between the average years of education for females in Europe, but there is not a significant difference between the years of 2018 and 2020. Again, the average expected years of education achieved by females across the two years show no significant statistical difference, while the expected mean years of education with respect to countries provides statistical evidence that at least one pair of countries' means differs. This provokes questions about the countries as independent entities. Possible explanations for the significant difference between the countries could be the culture or the economy of the individual countries. However, one should also note that the countries all remain around 11 to 14 expected years of education for females, with the majority of the data falling around 13.5 years. Further research can be conducted in regard to the European countries' backgrounds to explain some of the differences in expected mean years for females.

References

Assumptions of Tukey's test. Passel. (n.d.). https://passel2.unl.edu/view/lesson/2e09f0055f13/15.

World Bank. (2010, July 21). *Gender Statistics*. Gender Statistics (Gender Stats) | Data Catalog. https://datacatalog.worldbank.org/dataset/gender-statistics.

World Bank. (n.d.). *Gender Statistics*. Data Bank. https://databank.worldbank.org/reports.aspx?source=gender-statistics#.