**Final Data Analysis Project**

**Ryan Johnson**

**Introduction/Overview:**

Throughout my life I have struggled with mental health. Whether it was heightened anxiety as a kid, or depression symptoms as a young adult, mental health has been a constant struggle in my life. Mental health has seemed to become a more accepted topic of discussion in recent years, especially during the COVID-19 pandemic, where many have experienced depression symptoms due to the isolation. Due to this, I thought it would be interesting to explore mental health data through my data analysis. While this is a very broad topic, I first had the idea of exploring mental health and social media, seeing if there was a discernible link between social media usage and mental health issues. I was not able to find a good dataset for analysis, so I turned to something much more tragic: suicide rates. Suicide rates are something everyone should be invested in, as checking in on your friends, family and colleagues should be a common practice during these difficult times. I believe making yourself open to just listen to other people and show them someone cares can go a long way in preventing such a horrific outcome. And while I have never thought of committing suicide myself, I do sometimes have to remind myself that talking about my anxiety and depression symptoms does help, and someone always cares about me no matter what. Sadly, not everyone is able to overcome these demons, as mental health is an extremely difficult aspect to have control of. I was able to find a dataset on suicide rates that I was able to analyze.

**Statement of Hypothesis(es):**

With this analysis, we will look at suicide rate change over time. For this, we will have 3 distinct tests we will run. The first one will compare the suicide rates of 2000 to 2010 to see if there is a statistically significant difference. I predict there will be a statistically significant increase in suicide rates from 2000 to 2010. The hypotheses for this test are:

Null Hypothesis: There is no difference between 2000 and 2010 in terms of suicide rates.

Alternative Hypothesis: There is a statistically significant difference in the suicide rates between 2000 and 2010.

Our next test will compare the suicide rates between 2010 and the latest year in the sample, 2016. I predict there will be a statistically significant increase in suicide rates from 2010 to 2016. The hypotheses for this test are:

Null Hypothesis: There is no difference between 2010 and 2016 in terms of suicide rates.

Alternative Hypothesis: There is a statistically significant difference in the suicide rates between 2010 and 2016.

For the final test, we will look at sucide rates of the latest year in the sample, 2016, and see if sex has an impact on suicide rates. I predict sex will not have an impact on suicide rate. The hypotheses for this test are:

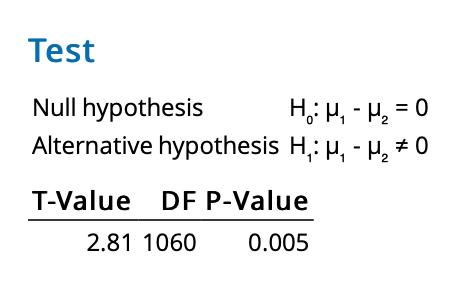
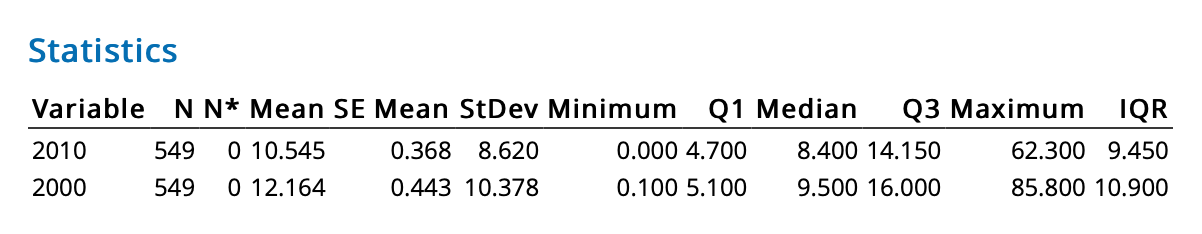
Null Hypothesis: There is no effect of this variable (sex) on the response variable (2016 suicide rate).

Alternative Hypothesis: There is a statistically significant effect of this variable (sex) on the response variable (2016 suicide rate).

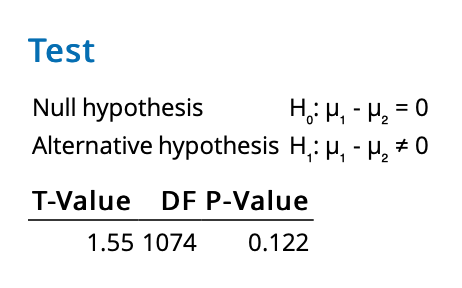
**Data & Analysis Methodology:**

My dataset can be found on Kaggle.com with the title “Mental Health and Suicide Rates.” The author of this dataset is Twinkle Khanna, and she had 4 datasets but I only used the first one, Age-standardized suicide rates, for my analysis. She obtained this data from the WHO (World Health Organization) which conducted a study determining estimates for the suicide rates in each country and age group, by year with their goal being to reduce suicide rates by ⅓ by 2030. This study was titled “Suicide in the World: Global Health Estimates.” Since these are estimates, I did not consider them to be the true population data, rather estimates based on sampling. I also looked up suicide rates in 2016 and found different figures between other sources such as the CDC and macrotrends.net, leading me to believe the true figures of the entire population are unknown. In this dataset, there are 6 variables: country, sex, 2016 suicide rate, 2015 suicide rate, 2010 suicide rate and 2000 suicide rate. For the years, they all present the suicide rates in percentages, and the sex column includes 3 values: male, female and both sexes. For this study, she only included sex based on the binary scale, so that is what I used for my analysis. Each country is repeated 3 times to account for the 3 sex categories. In order to do my analysis, I converted the sex categories to numeric data. I coded this variable as follows: Male=1, Female=2, and Both sexes=3. For each of my analyses, I used Minitab because I was running statistical tests. For my test to determine if there was a statistically significant difference between suicide rates in 2000 and 2010, I used a 2-sample t-test to compare the samples. This test determines if there is a statistically significant difference between 2 variables, and I used a 95% confidence interval. For the next test comparing the suicide rates of 2010 and 2016, I also used a 2-sample t-test. This would allow me to determine if there is a statistically significant difference between the variables, based on a 95% confidence interval. For the final test, I wanted to determine if there was a statistically significant impact of sex on suicide rate. For this test, I ran a single regression which would allow me to determine if there was a statistically significant impact of sex on suicide rate in 2016, again using a 95% confidence interval. I used the 2016 suicide rates since they were the most recent statistics.

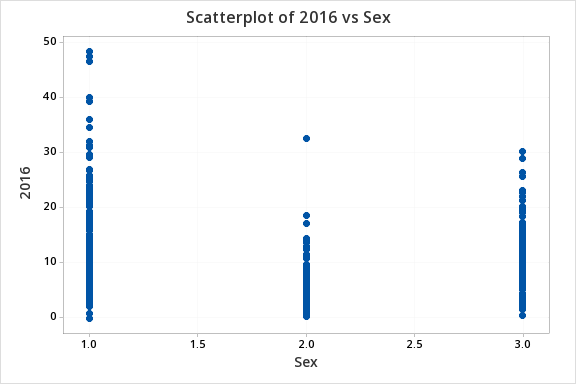
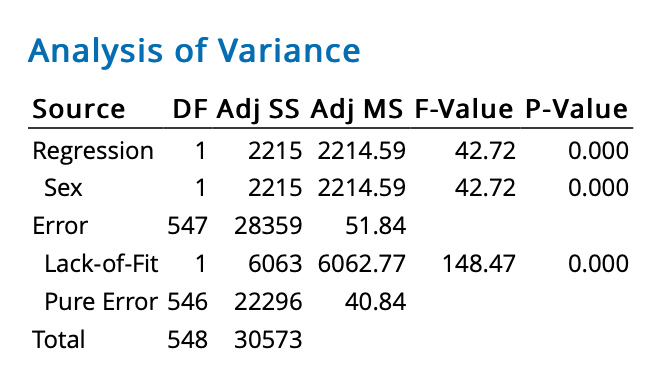
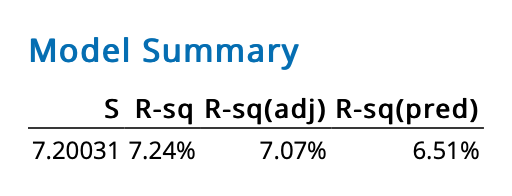
**Results:**



For our first hypothesis, we obtain a p-value of 0.005 when running the 2-sample t-test between 2000 and 2010 suicide rates. This p-value tells us that there is a 0.5% chance of obtaining a random sample with these characteristics if our null, that there is no difference between 2010 and 2016 suicide rates, is true. So, we can reject our null hypothesis and determine with 95% confidence that there is a statistically significant difference between suicide rates in 2000 as compared to 2010. I also ran the descriptive statistics on both 2000 and 2010 suicide rates, and found that the mean for 2000 was 12.164% compared to 10.545% in 2010. The median for 2000 was 9.5% while the median for 2010 8.4%. This tells us that there seems to be a statistically significant decrease in suicide rates from 2000 to 2010.



For our next hypothesis, I ran the 2-sample t-test between 2010 and 2016 suicide rates, and the p-value I obtained was 0.122. This tells us that there is a 12.2% chance of obtaining a random sample with these characteristics if the null, that there is no difference between suicide rates in 2010 and 2016, is true, therefore we can not comfortably reject the null with 95% confidence because the p-value is above 0.05. There is no statistically significant difference between suicide rates in 2010 and 2016, according to our sample.



For testing the final hypothesis, I used a regression analysis test between sex as the predictor variable and 2016 suicide rate as the response variable. I obtained a p-value of 0 which tells us that there is a 0% chance of obtaining a sample with these characteristics if the null, that sex has no impact on 2016 suicide rates, was true. Therefore, we can conclude with 95% confidence that sex has a statstically significant impact on 2016 suicide rate. This test also gave us the R-squared figure, which explains how much of the change in the response variable (2016 suicide rate), can be explained by the predictor variable (sex). The R-squared value was 7.24% meaning that only 7.24% of the change in 2016 suicide rate is explained by the sex of a person on the binary scale. So while there is a statistically significant impact, it is not a strong predictor. I also ran the scatterplot of sex as the predictor and 2016 rate as the response. This scatterplot shows the rates and is clustered into 3 sex values with 1 being male, 2 being female and 3 being both sexes. From this graph, it seems to show females are clustered at a lower rate of suicide than males, which goes against my prediction that there was no difference in suicide rate based upon gender. This also is only for 2016, so the other years may produce different results.

**Discussion and Conclusion:**

In this analysis, we conducted statistical tests on suicide rate data. We used 2-sample t-tests to determine if there was a statistically significant difference between 2000 and 2010 rates and between 2010 and 2016 rates, respectively. We found that there was a statistically significant difference in rates between 2000 and 2010, as there seemed to be a decline in suicide rates from 2010 to 2016. We did not find a statistically significant difference in rates between 2010 and 2016 however. We also used regression analysis to determine if sex had a statistically significant influence on suicide rates in 2016. We found that it did have a statistically significant impact, but it was not a strong predictor. However, the scatterplot between the predictor sex and response 2016 suicide rate showed us that males were clustered at a higher suicide rate than females, but this was only for the 2016 data, so other years may have different findings. Some limitations of these analyses are that we did not include data from the past 5 years as well as other factors such as age, income, mental health disorders and many other variables that could be reasonably concluded to have an impact on the results. These rates should continue to be examined to determine what is causing suicides and if there are better predictors we can use in order to prevent these tragedies. Another huge variable would be COVID-19’s impact on these rates, as the isolation may have had a negative impact on these rates. Further testing should be done on this complex topic, with more of the variables discussed above included.