Draft NoteS:

**Statistical Tests:**

**T-Test, two populations, at 95% siginifgance level.**

**Referece:** [**https://support.minitab.com/en-us/minitab/20/help-and-how-to/statistics/basic-statistics/how-to/2-sample-t/interpret-the-results/key-results/**](https://support.minitab.com/en-us/minitab/20/help-and-how-to/statistics/basic-statistics/how-to/2-sample-t/interpret-the-results/key-results/)



###### “ Because the p-value is less than 0.000, which is less than the significance level of 0.05, the decision is to reject the null hypothesis and conclude that the ratings of the countries are different. “

Because the p-value is more than 0.05, which is the significance level, the decision is to accept the null hypothesis, therefore we are unable to conclude that the PPI of Wheat in Ireland and Croatia are different.

To determine whether the difference between the population means is statistically significant, compare the p-value to the significance level. Usually, a significance level (denoted as α or alpha) of 0.05 works well. A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

P-value ≤ α: The difference between the means is statistically significantly (Reject H0)

If the p-value is less than or equal to the significance level, the decision is to reject the null hypothesis. You can conclude that the difference between the population means does not equal the hypothesized difference. If you did not specify a hypothesized difference, Minitab tests whether there is no difference between the means (Hypothesized difference = 0). Use your specialized knowledge to determine whether the difference is practically significant. For more information, go to Statistical and practical significance.

P-value > α: The difference between the means is not statistically significant (Fail to reject H0)

If the p-value is greater than the significance level, the decision is to fail to reject the null hypothesis. You do not have enough evidence to conclude that the difference between the population means is statistically significant. You should make sure that your test has enough power to detect a difference that is practically significant. For more information, go to Power and Sample Size for 2-Sample t.

T-Value DF P-Value

6.31 32 0.000

Test

Null hypothesis H₀: μ₁ - µ₂ = 0

Alternative hypothesis H₁: μ₁ - µ₂ ≠ 0

Key Result: P-Value

In these results, the null hypothesis states that the difference in the mean rating between two countries is 0. Because the p-value is less than 0.000, which is less than the significance level of 0.05, the decision is to reject the null hypothesis and conclude that the ratings of the countries are different.

Step 3: Check your data for problems

Problems with your data, such as skewness and outliers can adversely affect your results. Use the graphs to look for skewness (by examining the spread of each sample) and to identify potential outliers.

**Examine the spread of your data to determine whether your data appear to be skewed.**

**Poisson Distribution** = Elements, in a place, in a period of time.

For Ireland, what is the probability that the Wheat PPP will be greater than the Irish average in a given year?

0.4892

For Ireland, what is the probability that the Wheat PPP will be less than or equal to the Irish average in a given year?

0.5107

We can confirm this is correct by adding these two probabilities together, to get 1. The Wheat PPP in Ireland must either be greater than, less or equal to the average, which covers all possibilities.

**Normal Distribution:**

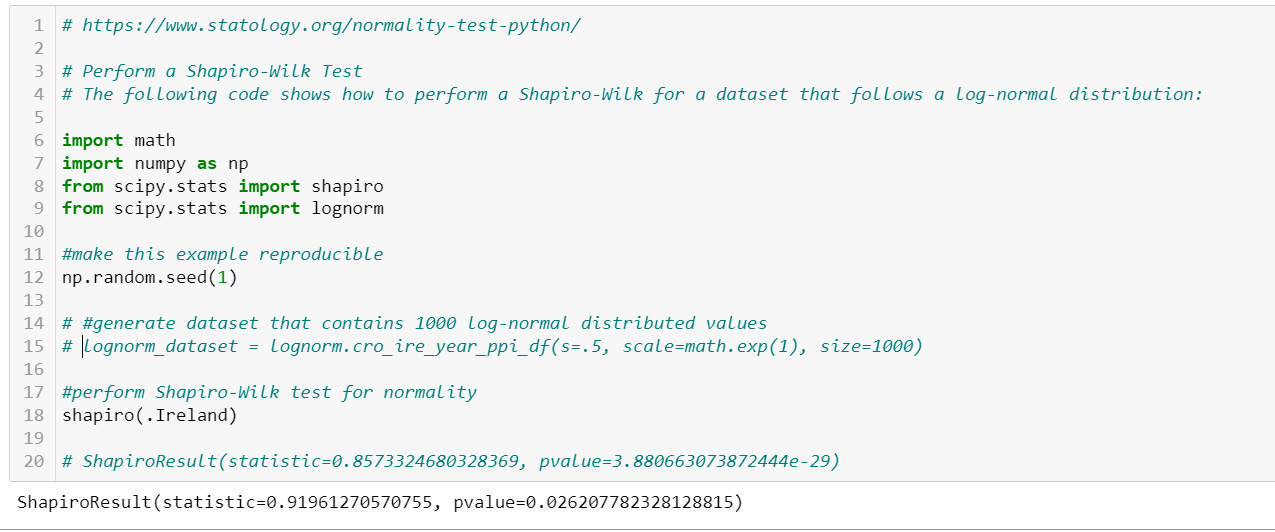
**We need to check the skewness of Ireland**

* The variables with skewness > 1 are highly positively skewed.
* The variables with skewness < -1 are highly negatively skewed.
* The variables with 0.5 < skewness < 1 are moderately positively skewed.
* The variables with -0.5 < skewness < -1 are moderately negatively skewed.
* Variables with -0.5 < skewness < 0.5 are symmetric i.e normally, which is the case for Ireland.

**3. (Formal Statistical Test) Perform a Shapiro-Wilk Test.**

* If the p-value of the test is greater than α = .05, then the data is assumed to be normally distributed.

Ireland – Test says Ireland data is not normally distributed (not sure if correct, looks normally distributed)



**Wilcoxon signed-rank test,** also known as Wilcoxon matched pair test is a non-parametric hypothesis test that compares the median of two paired groups and tells if they are identically distributed or not.

In this example, the Wilcoxon Signed-Rank Test uses the following null and alternative hypotheses:

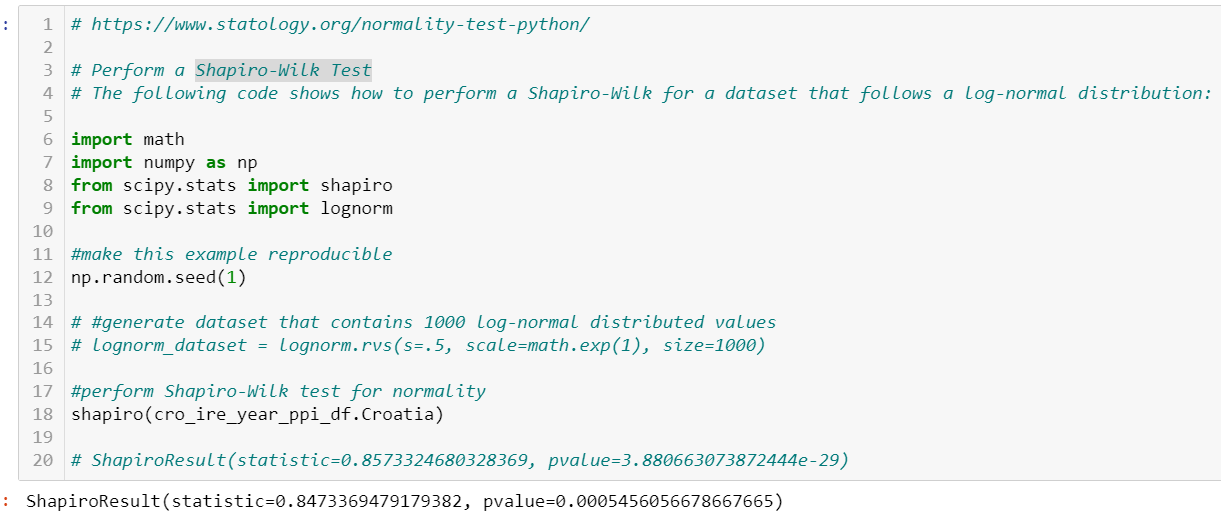
H0: The true mean is equal between the two groups

HA: The true mean is not equal between the two groups

Since the p-value is less than 0.05, we reject the null hypothesis.

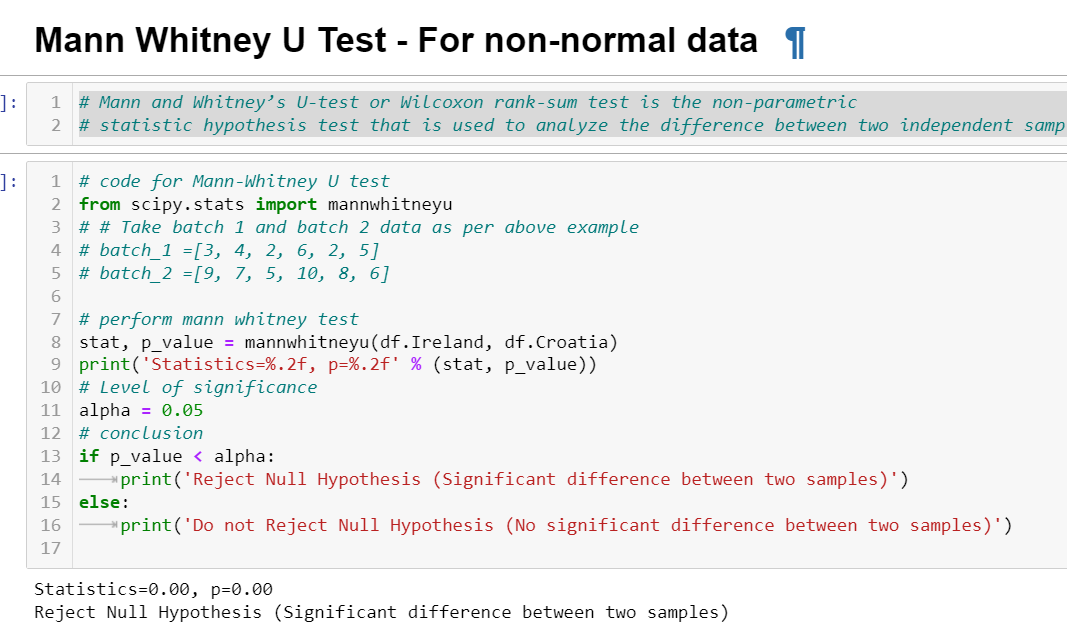
We have sufficient evidence to say that the true mean is not equal between the two countries.

Croatia: Test says Croatia data is not normally distributed



# Mann Whitney U Test – For non-normal data

Mann and Whitney’s U-test or Wilcoxon rank-sum testis the non-parametric statistic hypothesis test that is used to analyze the difference between two independent samples of ordinal data.

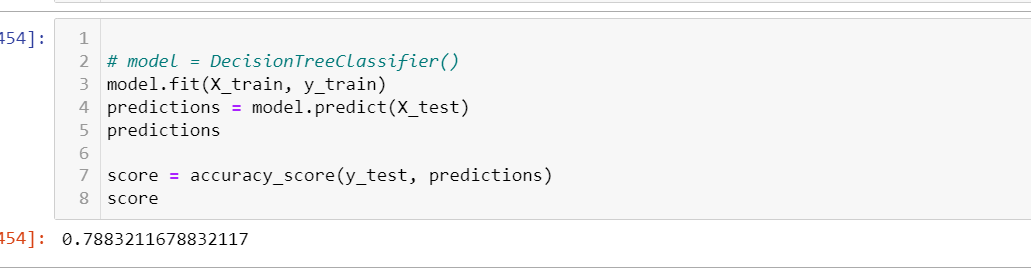


ANOVA Test:

The null hypothesis (H0) is that there is no difference between the groups and equality between means (walruses weigh the same in different months). The alternative hypothesis (H1) is that there is a difference between the means and groups (walruses have different weights in different months.



Decision Tree



Under Machine learning model 2 : should standardise data before modelling…..

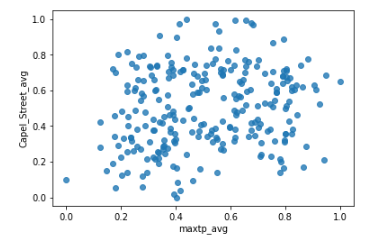
**Transforming the Data:**

Machine learning algorithms often perform better when numerical input variables are scaled to a standard range. This includes algorithms that use a weighted sum of the input, like linear regression, and algorithms that use distance measures, like k-nearest neighbours.

The two most popular techniques for scaling numerical data prior to modelling are normalization and standardization.

* Normalization scales each input variable separately to the range 0-1, which is the range for floating-point values where we have the most precision. This is the method used in this analysis.
* Standardization scales each input variable separately by subtracting the mean (called centring) and dividing by the standard deviation to shift the distribution to have a mean of zero and a standard deviation of one. (Müller and Guido, 2017)

The scatter plot below shows the data after it has been scaled. Notice both the x-axis and the y-axis range from 0-1. From this graph, there is no obvious relationship between the temperature and Capel Street footfall data.



**Approach 2: Machine Learning: Simple Linear Regression:**

Linear Regression is a machine learning algorithm based on supervised learning. Supervised learning is a machine learning technique that is defined by its use of labelled datasets to train algorithms that classify data or predict outcomes accurately. As input data is fed into the model, its weights are adjusted until the model has been fitted appropriately. (Severance, 2016)

Linear regression is used to identify the relationship between a dependent variable and one or more independent variables and is useful for finding out the relationship between variables and forecasting. Linear regression performs the task to predict a dependent variable value (y, in this case pedestrian footfall on Capel Street) based on a given independent variable (x, in this case the temperature).

There are two types of Linear Regressions

* Simple Linear Regression: Where there is only one independent variable. In the formula below, bo is the intercept, b1 is the coefficient or slope, x is the independent variable and y is the dependent variable.



* Polynomial Regression or Multiple Linear Regression: This is a linear regression with multiple independent variables. In the formula below, bo is the intercept, b1,b2,b3,b4…,bn are coefficients or slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.



This model is a simple linear regression using the temperature variable as the independent variable and pedestrian footfall as the dependent variable.

**Building the Model:**

Our data is broken down into train and test data. The Train data usually accounts for a minimum of 70% of the data. 70% of the dataset is randomly fed in to train the model, and the remaining 30% is used to test the model. The 70% of the data fed in to train the model is random, so the training and test datasets will change each time the experiment is run, which will lead to slightly different outputs.

* Train Dataset: Used to fit the machine learning model
* Test Dataset: Used to evaluate the fit machine learning model

The objective is to estimate the performance of the machine learning model on new data: data not used to train the model. The model is expected to fit available data with known inputs and outputs, then make predictions on new data in the future.

