**Hypothesis Testing with SAS**

**Test of Normality and Two Samples Independent t Test**

Suppose a computer program was run on 20 different files. Each file was available in two different formats: Zip and XML. The execution time (time from the beginning till the end of the program in milliseconds) is recorded on both the file formats for each file. Few examples of execution times of that program are shown in the following table, where each row represents one file format and each column represents a particular file.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | File 1 | File 2 | File 3 | File 4 | File 5 | File 6 | File 7 |
| Zip | 10 | 44 | 65 | 77 | 43 | 44 | 22 |
| XML | 20 | 55 | 75 | 60 | 55 | 88 | 35 |

Find out if there is any significant difference in the execution times of two different file formats when the computer program is run on them.

My Solution:

Null Hypothesis (Ho): there is no significant difference in the execution times of two different file formats.

1. First, we need to check if these samples (of data) paired or independent?

These samples are paired as there is a common association between samples. There is a relationship between “Zip” and “XML” files for each observation that is they are just different formats for the same “file”.

proc import out=FileType

datafile = ""/myfolders/exec\_time\_format\_i.csv"

dbms=csv replace; getnames=yes;

run;

proc print data=FileType;

run;

| **Obs** | **input\_type** | **exec\_time** |
| --- | --- | --- |
| **1** | xml | 20 |
| **2** | xml | 55 |
| **3** | xml | 75 |
| **4** | xml | 60 |
| **5** | xml | 55 |
| **6** | xml | 88 |
| **7** | xml | 35 |
| **8** | xml | 33 |
| **9** | xml | 35 |
| **10** | xml | 80 |
| **11** | xml | 65 |
| **12** | xml | 82 |
| **13** | xml | 47 |
| **14** | xml | 35 |
| **15** | xml | 97 |
| **16** | xml | 110 |
| **17** | xml | 250 |
| **18** | xml | 190 |
| **19** | xml | 110 |
| **20** | xml | 600 |
| **21** | zip | 10 |
| **22** | zip | 44 |
| **23** | zip | 65 |
| **24** | zip | 77 |
| **25** | zip | 43 |
| **26** | zip | 44 |
| **27** | zip | 22 |
| **28** | zip | 66 |
| **29** | zip | 50 |
| **30** | zip | 100 |
| **31** | zip | 55 |
| **32** | zip | 99 |
| **33** | zip | 44 |
| **34** | zip | 23 |
| **35** | zip | 100 |
| **36** | zip | 88 |
| **37** | zip | 200 |
| **38** | zip | 220 |
| **39** | zip | 110 |
| **40** | zip | 551 |

2-Then, we perform the statistical normality test on the data by using the file exec\_time\_format\_i.csv. To do this, we need to first import the file and apply the test of normality. The code snippet for the normality test is shown below but you need to import the file first and write its code.

proc univariate data=FileType normal;

by input\_type;

run;

proc univariate data= FileType;

histogram exec\_time;

by input\_type;

run;

|  |  |
| --- | --- |
| ***The UNIVARIATE Procedure*** | |
| ***Variable: exec\_time*** |

**input\_type=xml**

| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 20 | **Sum Weights** | 20 |
| **Mean** | 106.1 | **Sum Observations** | 2122 |
| **Std Deviation** | 128.719565 | **Variance** | 16568.7263 |
| **Skewness** | 3.32427551 | **Kurtosis** | 12.2966026 |
| **Uncorrected SS** | 539950 | **Corrected SS** | 314805.8 |
| **Coeff Variation** | 121.3191 | **Std Error Mean** | 28.7825697 |

| **Basic Statistical Measures** | | | |
| --- | --- | --- | --- |
| **Location** | | **Variability** | |
| **Mean** | 106.1000 | **Std Deviation** | 128.71956 |
| **Median** | 70.0000 | **Variance** | 16569 |
| **Mode** | 35.0000 | **Range** | 580.00000 |
|  |  | **Interquartile Range** | 62.50000 |

| **Tests for Location: Mu0=0** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Student's t** | **t** | 3.686259 | **Pr > |t|** | 0.0016 |
| **Sign** | **M** | 10 | **Pr >= |M|** | <.0001 |
| **Signed Rank** | **S** | 105 | **Pr >= |S|** | <.0001 |

| **Quantiles (Definition 5)** | |
| --- | --- |
| **Level** | **Quantile** |
| **100% Max** | 600.0 |
| **99%** | 600.0 |
| **95%** | 425.0 |
| **90%** | 220.0 |
| **75% Q3** | 103.5 |
| **50% Median** | 70.0 |
| **25% Q1** | 41.0 |
| **10%** | 34.0 |
| **5%** | 26.5 |
| **1%** | 20.0 |
| **0% Min** | 20.0 |

| **Extreme Observations** | | | |
| --- | --- | --- | --- |
| **Lowest** | | **Highest** | |
| **Value** | **Obs** | **Value** | **Obs** |
| 20 | 1 | 110 | 16 |
| 33 | 8 | 110 | 19 |
| 35 | 14 | 190 | 18 |
| 35 | 9 | 250 | 17 |
| 35 | 7 | 600 | 20 |



|  |  |
| --- | --- |
| ***The UNIVARIATE Procedure*** | |
| ***Variable: exec\_time*** |

**input\_type=zip**

| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 20 | **Sum Weights** | 20 |
| **Mean** | 100.55 | **Sum Observations** | 2011 |
| **Std Deviation** | 118.819844 | **Variance** | 14118.1553 |
| **Skewness** | 3.18404029 | **Kurtosis** | 11.5627646 |
| **Uncorrected SS** | 470451 | **Corrected SS** | 268244.95 |
| **Coeff Variation** | 118.169909 | **Std Error Mean** | 26.5689248 |

| **Basic Statistical Measures** | | | |
| --- | --- | --- | --- |
| **Location** | | **Variability** | |
| **Mean** | 100.5500 | **Std Deviation** | 118.81984 |
| **Median** | 65.5000 | **Variance** | 14118 |
| **Mode** | 44.0000 | **Range** | 541.00000 |
|  |  | **Interquartile Range** | 56.00000 |

| **Tests for Location: Mu0=0** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Student's t** | **t** | 3.784496 | **Pr > |t|** | 0.0013 |
| **Sign** | **M** | 10 | **Pr >= |M|** | <.0001 |
| **Signed Rank** | **S** | 105 | **Pr >= |S|** | <.0001 |

| **Quantiles (Definition 5)** | |
| --- | --- |
| **Level** | **Quantile** |
| **100% Max** | 551.0 |
| **99%** | 551.0 |
| **95%** | 385.5 |
| **90%** | 210.0 |
| **75% Q3** | 100.0 |
| **50% Median** | 65.5 |
| **25% Q1** | 44.0 |
| **10%** | 22.5 |
| **5%** | 16.0 |
| **1%** | 10.0 |
| **0% Min** | 10.0 |

| **Extreme Observations** | | | |
| --- | --- | --- | --- |
| **Lowest** | | **Highest** | |
| **Value** | **Obs** | **Value** | **Obs** |
| 10 | 21 | 100 | 35 |
| 22 | 27 | 110 | 39 |
| 23 | 34 | 200 | 37 |
| 43 | 25 | 220 | 38 |
| 44 | 33 | 551 | 40 |



After running the normality test and as the sample size is less than 2000, we need to look at the p values for Shapiro test:

For input\_type=zip : p Value <0.0001

For input\_type=xml : p Value <0.0001

Then, these are not normally distributed data.

3- These are Paired samples with not normal distribution, then the statistical test we are going to perform on this data is Wilcoxon Signed Rank test.

4- we run the following codes for the appropriate statistical test and find the p-value to answer if the difference is significant or not.

proc import out=my\_data

datafile = "/myfolders/exec\_time\_format\_p.csv "

dbms=csv replace; getnames=yes;

run;

proc print data=my\_data;

run;

data diff\_columns\_data;

set my\_data;

diff = zip\_type\_exec\_time - xml\_type\_exec\_time;

run;

proc print data=diff\_columns\_data;

run;

proc univariate data = diff\_columns\_data;

var diff;

run;

| **Obs** | **zip\_type\_exec\_time** | **xml\_type\_exec\_time** |
| --- | --- | --- |
| **1** | 10 | 20 |
| **2** | 44 | 55 |
| **3** | 65 | 75 |
| **4** | 77 | 60 |
| **5** | 43 | 55 |
| **6** | 44 | 88 |
| **7** | 22 | 35 |
| **8** | 66 | 33 |
| **9** | 50 | 35 |
| **10** | 100 | 80 |
| **11** | 55 | 65 |
| **12** | 99 | 82 |
| **13** | 44 | 47 |
| **14** | 23 | 35 |
| **15** | 100 | 97 |
| **16** | 88 | 110 |
| **17** | 200 | 250 |
| **18** | 220 | 190 |
| **19** | 110 | 110 |
| **20** | 551 | 600 |

| **Obs** | **zip\_type\_exec\_time** | **xml\_type\_exec\_time** | **diff** |
| --- | --- | --- | --- |
| **1** | 10 | 20 | -10 |
| **2** | 44 | 55 | -11 |
| **3** | 65 | 75 | -10 |
| **4** | 77 | 60 | 17 |
| **5** | 43 | 55 | -12 |
| **6** | 44 | 88 | -44 |
| **7** | 22 | 35 | -13 |
| **8** | 66 | 33 | 33 |
| **9** | 50 | 35 | 15 |
| **10** | 100 | 80 | 20 |
| **11** | 55 | 65 | -10 |
| **12** | 99 | 82 | 17 |
| **13** | 44 | 47 | -3 |
| **14** | 23 | 35 | -12 |
| **15** | 100 | 97 | 3 |
| **16** | 88 | 110 | -22 |
| **17** | 200 | 250 | -50 |
| **18** | 220 | 190 | 30 |
| **19** | 110 | 110 | 0 |
| **20** | 551 | 600 | -49 |

| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 20 | **Sum Weights** | 20 |
| **Mean** | -5.55 | **Sum Observations** | -111 |
| **Std Deviation** | 23.8337995 | **Variance** | 568.05 |
| **Skewness** | -0.3507587 | **Kurtosis** | -0.2551948 |
| **Uncorrected SS** | 11409 | **Corrected SS** | 10792.95 |
| **Coeff Variation** | -429.43783 | **Std Error Mean** | 5.32939959 |

| **Basic Statistical Measures** | | | |
| --- | --- | --- | --- |
| **Location** | | **Variability** | |
| **Mean** | -5.5500 | **Std Deviation** | 23.83380 |
| **Median** | -10.0000 | **Variance** | 568.05000 |
| **Mode** | -10.0000 | **Range** | 83.00000 |
|  |  | **Interquartile Range** | 28.50000 |

| **Tests for Location: Mu0=0** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Student's t** | **t** | -1.04139 | **Pr > |t|** | 0.3108 |
| **Sign** | **M** | -2.5 | **Pr >= |M|** | 0.3593 |
| **Signed Rank** | **S** | -16.5 | **Pr >= |S|** | 0.5215 |

| **Quantiles (Definition 5)** | |
| --- | --- |
| **Level** | **Quantile** |
| **100% Max** | 33.0 |
| **99%** | 33.0 |
| **95%** | 31.5 |
| **90%** | 25.0 |
| **75% Q3** | 16.0 |
| **50% Median** | -10.0 |
| **25% Q1** | -12.5 |
| **10%** | -46.5 |
| **5%** | -49.5 |
| **1%** | -50.0 |
| **0% Min** | -50.0 |

| **Extreme Observations** | | | |
| --- | --- | --- | --- |
| **Lowest** | | **Highest** | |
| **Value** | **Obs** | **Value** | **Obs** |
| -50 | 17 | 17 | 4 |
| -49 | 20 | 17 | 12 |
| -44 | 6 | 20 | 10 |
| -22 | 16 | 30 | 18 |
| -13 | 7 | 33 | 8 |

The p-value is 0.5215. As the p-value is larger than 0.05, there is NO significant difference between two samples and we can NOT reject Null hypothesis.