Non-parametric Tests 1

1. Objectives
- Explain the Mann-Whitney-Wilcoxon test and Kruskal Wallis test procedure
- Provide instructions on how to use R to run the tests.
- Understand and interpret R outputs and make decisions based on p-value.

2. Procedure

We will use the function **wilcox.test** to perform the Mann-Whitney-Wilcoxon test to compare 2 independent samples. The format of the function is:

> ```
> wilcox.test(x,y, data = name of data frame, alternative =
> "two.sided", paired = FALSE, exact = NULL)
> ```

where x and y are numeric vectors of data values.

Here are meanings of other arguments:

`data`	name of your data frame that contains the data for the test.
`alternative`	a character string specifying the alternative hypothesis
`paired`	a logical indicating whether you want a paired-samples test. We are considering independent samples, so choose FALSE.
`exact`	a logical indicating whether an exact p-value should be computed. If this argument is not specified, an exact p-value is computed by default if the samples contain less than 50 finite values and there are no ties. Otherwise, a normal approximation is used.

To conduct Kruskal-Wallis test, we will use the function **kruskal.test( ).**

> ```
> kruskal.test(outcome ~ grouping variable, data = name of data
> frame)
> ```

Here are the meanings of these arguments:

`outcome`	The outcome variable
`Grouping variable`	Categorical variable that defines the groups
`data`	Name of your dataframe that contains the variables for the test.

3. Exercises

Exercise 1. Two fuel additives are being tested to determine their effects on gasoline mileage. Ten (10) cars were tested with additive 1 and 10 cars were tested with additive 2 (the cars are randomly assigned to each group). The following data show the miles per gallon obtained with the two additives. Using α = 0.05, conduct a hypothesis test to see whether there is a significant difference between gasoline mileage for the two additives. Data for this exercise are hypothetical.

Additive 1	Additive 2
17.3	18.7
18.4	17.8
19.1	21.3
16.7	21
18.2	22.1
18.6	18.7
17.5	19.8
15.4	20.7
16.2	20.2
15.1	26

Set up H_o, H_A: We want to see if significant differences exist between 2 additives:

H_O: The two population distributions of gasoline mileage are the same

H_A: The two population distributions are different

Check the assumptions to see whether the MWW test is appropriate.

Because the samples sizes are quite small (10), we produce a stem and leaf display for each sample. Note that when sample sizes are small, it is difficult to determine whether the populations are normal or not.

Remember to create 2 data vectors of Additive 1 and Additive 2, called **add1** and **add2**:

> ➤ `add1<-c(17.3,18.4,19.1,16.7,18.2,18.6,17.5, 15.4, 16.2, 15.1)`
> ➤ `add2<-c(18.7,17.8,21.3,21,22.1,18.7,19.8,20.7,20.2, 26)`

Stem and leaf display for Additive 1:

```
The decimal point is at the |

  15 | 14
  16 | 27
  17 | 35
  18 | 246
  19 | 1
```

Stem and leaf display for Additive 2:

```
The decimal point is at the |

  16 | 8
  18 | 778
  20 | 2703
  22 | 1
  24 |
  26 | 0
```

These figures show that normality should not be assumed here. Also, the samples are independent and random samples (why?).

Then if you write the **wilcox.test()** function:

> ➢ `ex1<-wilcox.test(add1, add2, alternative="two.sided",`
> `paired=FALSE)`

you will have the following ouput:

```
        Wilcoxon rank sum test with continuity correction

data:  add1 and add2
W = 6, p-value = 0.001004
alternative hypothesis: true location shift is not equal to 0
```

Question 1. What is the difference between the choice of **exact = TRUE** and **exact = FALSE**?
Question 2. Let's calculate the test statistic by hand and compare with the value given in R output. Why are they different?

Type in **?wilcox.test** to see the note about test statistic given by R. It states that R produces the same value as Mann-Whitney test statistic, which equals the sum of ranks of the first sample with the minimum value subtracted of $m(m + 1)/2$ (m is sample size of the first one).

Then if you want to express test statistic as sum of ranks, you must add $m(m + 1)/2$ to the value given in R output. If you want this sum of ranks to appear in the output at the end, try the following commands:

> ➢ `n1<-length(add1)`
> ➢ `ex1$statistic<-ex1$statistic+n1*(n1+1)/2`
> ➢ `names(ex1$statistic)<- "T.W"`
> ➢ `ex1`

The new output will appear:

```
        Wilcoxon rank sum test with continuity correction

data:  add1 and add2
T.W = 61, p-value = 0.001004
alternative hypothesis: true location shift is not equal to 0
```

Important note: We should not use R to run Mann-Whitney-Wilcoxon test in case a sample size is smaller than 7 and there are ties.

Exercise 2. In the consumer preference study, one group of consumers was "primed" to images that would appear on product labels, while the other group of consumers was not "primed." All consumers were then asked their attitude toward a product on a seven-point scale (from 1 = dislike very much to 7

= like very much). Carry out the Mann-Whitney-Wilcoxon procedure to test if "primed" consumers are associated with systematically higher preference scores. What do you conclude?

What type of hypothesis is this? What is the type of the data we wish to compare?

At first, we must import file **Brandpreference.csv** into R:

```
➢ brand<-read.table("Brandpreference.csv",header=TRUE,sep = ",")
➢ brand
```

The second command is to check how the data are presented. You will see 2 groups of primed and not primed customers presented by index number 0 and 1. So you must separate these 2 groups first, using **subset()** function:

```
➢ primed<-subset(brand, Primed == 1)
➢ notprimed<-subset(brand, Primed == 0)
```

Note that the samples are independent and we do not need to consider normality assumption (why?).

Then, you're expected to produce the following output:

```
        Wilcoxon rank sum test with continuity correction

data:  primed$Preference and notprimed$Preference
T.W = 600.5, p-value = 0.00042
alternative hypothesis: true location shift is greater than 0
95 percent confidence interval:
 0.99994      Inf
sample estimates:
difference in location
             1.00004
```

The code to produce the above output is:

```
➢ ex2 <- wilcox.test(primed$Preference, notprimed$Preference,
  alternative="greater", paired=FALSE, exact=FALSE)
➢ ex2
```

Question 3: Try setting **exact = TRUE** in the above code and see what happens.

HOMEWORK

Exercise 3. "Conservationists have despaired over destruction of tropical rainforest by logging, clearing, and burning". These words begin a report on a statistical study of the effects of logging in **Borneo.csv**. Does logging significantly reduce the number of species in a plot after 8 years?

The R output for the test is provided below. Try to demonstrate that the populations are not normal and that the assumptions of the Mann-Whitney-Wilcoxon test are satisfied.

```
    Wilcoxon rank sum test with continuity correction

data:  logged$Species and unlogged$Species
T.W = 72, p-value = 0.029
alternative hypothesis: true location shift is less than 0
```

Exercise 4. A sample of 20 engineers employed with a company for three years has been rank ordered with respect to managerial potential. Some of the engineers attended the company's management-development course, others attended an off-site management-development program at a local university, and the remainder did not attend any program. Use the following rankings and $\alpha = 0.025$ to test for a significant difference in the managerial potential of the three groups. Remember to state the hypotheses, and carry out other necessary steps for the test.

Firstly, you must import file **Programs.csv** as program data frame into R. Then just apply the function **kruskal.test**. You're expected to produce the following output:

```
    Kruskal-Wallis rank sum test

data:  RankOfManagerialPotential by Program
Kruskal-Wallis chi-squared = 12.611, df = 2, p-value = 0.001826
```

Exercise 5. In Exercise 3 you compared the number of tree species in plots of land in a tropical rain forest that had never been logged with similar plots nearby that had been logged 8 years earlier. The researchers also counted species in plots that had been logged just 1 year earlier. Check the normal assumptions, are there features that might prevent use of ANOVA? Use the Kruskal-Wallis test to compare the distributions of tree counts. State hypotheses, the test statistic and its p-value, and your conclusions.

Import file **Borneo3.csv** as **borneo3** data frame into R.

The R output for the test is provided below:

```
    Kruskal-Wallis rank sum test

data:  species by groups
Kruskal-Wallis chi-squared = 9.4382, df = 2, p-value = 0.008923
```