Basic Statistics with R

Let's do some basic statistics with the iris dataset

```
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

Descriptive statistics

By using the function `summary`, we'll obtain the basic statistics of the whole iris dataset.

```
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

If, in contrast, we want to know the specific values of the mean, standard deviation, median and range of each class, we can use the functions `mean`, `sd`, `median`, `min` and `max` as follows:

```
m <- aggregate (.~Species, data=iris, FUN=mean)
m
```

```
     Species Sepal.Length Sepal.Width Petal.Length Petal.Width
1     setosa        5.006       3.428        1.462       0.246
2 versicolor        5.936       2.770        4.260       1.326
3  virginica        6.588       2.974        5.552       2.026
```

```
sdev <- aggregate (.~Species, data=iris, FUN=sd)
sdev
```

```
     Species Sepal.Length Sepal.Width Petal.Length Petal.Width
1     setosa    0.3524897   0.3790644    0.1736640   0.1053856
2 versicolor    0.5161711   0.3137983    0.4699110   0.1977527
3  virginica    0.6358796   0.3224966    0.5518947   0.2746501
```

```
median <- aggregate (.~Species, data=iris, FUN=median)
median
```

```
     Species Sepal.Length Sepal.Width Petal.Length Petal.Width
1     setosa          5.0         3.4         1.50         0.2
2 versicolor          5.9         2.8         4.35         1.3
3  virginica          6.5         3.0         5.55         2.0
```

```
minv <- aggregate (.~Species, data=iris, FUN=min)
minv
```

```
     Species Sepal.Length Sepal.Width Petal.Length Petal.Width
1     setosa          4.3         2.3          1.0         0.1
2 versicolor          4.9         2.0          3.0         1.0
3  virginica          4.9         2.2          4.5         1.4
```

```
maxv <- aggregate (.~Species, data=iris, FUN=max)
maxv
```

```
     Species Sepal.Length Sepal.Width Petal.Length Petal.Width
1     setosa          5.8         4.4          1.9         0.6
2 versicolor          7.0         3.4          5.1         1.8
3  virginica          7.9         3.8          6.9         2.5
```

Note that it is not recommended to name objects after function names.

Additional measures such as skewness or kurtosis can also be obtained:

```
#install.packages("moments")
library(moments)
k <- aggregate (.~Species, data=iris, FUN=kurtosis)
k
```

```
     Species Sepal.Length Sepal.Width Petal.Length Petal.Width
1     setosa     2.654235    3.744222     3.804592    4.434317
2 versicolor     2.401173    2.551728     2.925598    2.512167
3  virginica     2.912058    3.519766     2.743528    2.338652
```

```
skew <- aggregate (.~Species, data=iris, FUN=skewness)
skew
```

```
     Species Sepal.Length Sepal.Width Petal.Length Petal.Width
1     setosa    0.1164539  0.03992109    0.1031751   1.2159276
2 versicolor    0.1021896 -0.35186750   -0.5881587  -0.0302363
3  virginica    0.1144447  0.35487781    0.5328219  -0.1255598
```

To measure the correlation among variables, use the `cor()` function:

```
correlations <- cor(iris[,1:4])
correlations
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Hypothesis testing

Hypothesis testing is a process of making statistical decisions consisting on retaining or rejecting a research hypothesis based on measurements of observed samples. The decision is often based on a statistical mechanism called hypothesis testing. Generally, two or more datasets or classes are compared. A research hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets.

The comparison is considered *statistically significant* when the relationship between the data sets would be unlikely under the assumption of the null hypothesis, according to a threshold probability—the **significance level**.

The term significance level (alpha) is used to refer to a pre-chosen probability and the term **P-value** is used to indicate a probability that you calculate after a given study.

If your P value is less than the chosen significance level then you reject the null hypothesis i.e. accept that your sample gives reasonable evidence to support the alternative hypothesis. NOTe that It does NOT imply a “meaningful” or “important” difference.

Remember nothing in statistics is ever absolute and in any type of statistical analysis there is always the randomness factor.

In hypothesis testing there is always some margin of error where it is possible that the test result is not correct. A type I error is the mishap of falsely rejecting a null hypothesis when the null hypothesis is true. The probability of committing a type I error is called the significance level of the hypothesis testing, and is denoted by the Greek letter α .

One of the critical decisions in statistical testing is the choice of the right test. Depending on the experimental conditions, we’ll need to apply one test or another (Table 1). In any case, we can either choose a parametric test or a nonparametric test.

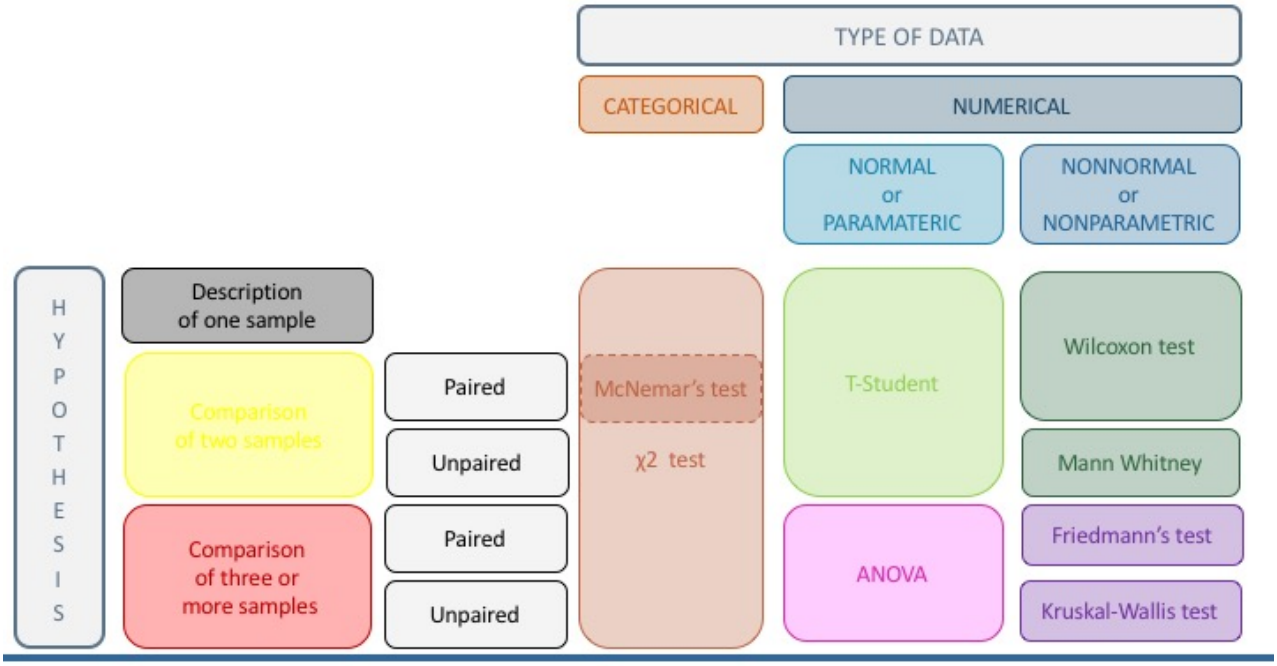

Table 1: Choice of the Statistical Test

Parametric tests are almost always the preferred option because they can be more powerful BUT they make assumptions about the parameters of the population distribution (assumptyion of Gaussian distribution) that are not always satisfied.

For large sample sizes (n>100), This is not a critical issue since parametric tests are robust (p-values will be nearly correct, even if the population is far from a Gaussian population) and nonparametric tests are powerful (as much as parametric tests).

By contrast, for small sample sizes (n<15), parametric tests are not robust (misleading p-values if assumptions not satisfied) and the nonparametric tests are not powerful (weak p-values).

If we wish to compare the Petal Length between two species (e.g. versicolor and virginica), we can use a two-samples t-test.

```
tt <- t.test(iris$Petal.Length[which(iris$Species=="virginica")], iris$Petal.L
ength[which(iris$Species=="versicolor")])
tt
```

```

    Welch Two Sample t-test

data:  iris$Petal.Length[which(iris$Species == "virginica")] and iris$Petal.Le
ngth[which(iris$Species == "versicolor")]
t = 12.604, df = 95.57, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.08851 1.49549
sample estimates:
mean of x mean of y
    5.552     4.260
```

```
tt$p.value
```

```
[1] 4.900288e-22
```

As we can see, with a significant p-value, we can reject the null hypothesis of no differences between the samples, and state that the Petal Length differs significantly between the versicolor and the virginica species.

The equivalent nonparametric test is the Mann-Whitney U-test:

```
ut <- wilcox.test(iris$Petal.Length[which(iris$Species=="virginica")], iris$Pe
tal.Length[which(iris$Species=="versicolor")])
ut
```

```

    Wilcoxon rank sum test with continuity correction

data:  iris$Petal.Length[which(iris$Species == "virginica")] and iris$Petal.Le
ngth[which(iris$Species == "versicolor")]
W = 2455.5, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

```
ut$p.value
```

```
[1] 9.133545e-17
```

We observe that they lead to the same conclusion (which is a good sign)

If, by contrast, we want to compare the Petal Length among the three species, we'll need to apply an ANOVA test.

```
at <- aov(data=iris, Petal.Length~Species)
summary(at)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  437.1  218.55    1180 <2e-16 ***
Residuals   147   27.2    0.19
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(at)[[1]][["Pr(>F)"]][1]
```

```
[1] 2.856777e-91
```

When comparing more than two tests, it is often useful to perform a post-hoc test

```
TukeyHSD(at)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Petal.Length ~ Species, data = iris)

$Species
                      diff     lwr     upr p adj
versicolor-setosa    2.798 2.59422 3.00178     0
virginica-setosa     4.090 3.88622 4.29378     0
virginica-versicolor 1.292 1.08822 1.49578     0
```

The equivalent nonparametric test is the Kruskal-Wallis W-test:

```
wt <- kruskal.test(data=iris, Petal.Length~Species)
summary(wt)
```

```
          Length Class  Mode
statistic 1      -none- numeric
parameter 1      -none- numeric
p.value   1      -none- numeric
method    1      -none- character
data.name 1      -none- character
```

```
wt$p.value
```

```
[1] 4.803974e-29
```

```
#install.packages("PMCMR")
library(PMCMR)
posthoc.kruskal.nemenyi.test(x=iris$Petal.Length, g=iris$Species, method="Tuke
y")
```

```
Warning in posthoc.kruskal.nemenyi.test.default(x = iris$Petal.Length, g =
iris$Species, : Ties are present, p-values are not corrected.
```

```

    Pairwise comparisons using Tukey and Kramer (Nemenyi) test
                  with Tukey-Dist approximation for independent samples

data:  iris$Petal.Length and iris$Species

           setosa  versicolor
versicolor 1.4e-08 -
virginica  < 2e-16 8.6e-08

P value adjustment method: none
```

In both cases, the overall p-value and the pairwise p-values are highly significant, confirming that

there are significant differences among the iris species.