Proyecto Rstudio

Celina Villalobos

2023-10-17

Información general

```
#DATA:

Estanque_plantas <- read.csv("~/RStudio/CursoInnovak/Materiales/Estanque_plantas.csv")
Set_datos <- read.csv("~/RStudio/CursoInnovak/Proyecto_1_RMarkdown/Set de datos proyecto.csv")

#Librerias

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.3     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:purrr':
##
##     some
##
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(agricolae)
library(devtools)
```

```
## Loading required package: usethis
```

```r
devtools::install_github('yihui/tinytex')
```

```
## Skipping install of 'tinytex' from a github remote, the SHA1 (3495cb88) has not changed since last in
##   Use 'force = TRUE' to force installation
```

Ejercicio de correlacion

Usando los datos "modernos" de la tabla Estanques_Biomasa, determinar si existe una correlacion entre la biomasa de dos especies acuaticas de plantas en estanques de Alaska: *Carex*, *Arctophila*

- Recuerden revisar si los datos cumplen todas las suposiciones de una correlacion.
- Reporten el coeficiente de correlacion y su p-value
- Expliquen que significan esos valores y denle una interpretacion a los resultados.

```r
#Seleccion de columnas a trabajar
Datos_filtrados <- Estanque_plantas[,c("Era","Artophila","Carex")] #Filtramos los datos que...
#...solo seran necesarios para realizar la correlación.

#Eliminacion de datos no necesarios
Artophila <- Estanque_plantas[-c(20:26,56:62,80:87,27:36),] #Eliminamos los datos que nos...
#...estorban para proseguir con la correlación.

#Correlacion de variables numericas
cor.test(Artophila$Artophila, Artophila$Carex) #Realizamos la correlación de la biomasa...
```

```
##
##  Pearson's product-moment correlation
##
## data:  Artophila$Artophila and Artophila$Carex
## t = 3.6467, df = 17, p-value = 0.001996
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.2979541 0.8584059
## sample estimates:
##       cor
## 0.6625044
```

```
#...entre las dos especies de plantas acuaticas que se encuentran en...
#...estanques de Alaska.
```

CONCLUSION:

Observo que el valor de p-value es < 0.05, lo cual nos dice que, existe significancia entre los valores. Obtenemos como resultado un coeficiente de correlacion de 0.66, esto nos dice que, la correlacion que existe entre estas dos variables numericas es una **correlacion positiva** ya que el valor se acerca al 1, por lo cual, existe relacion entre la biomasa de las dos especies de plantas que se encuentran en los estanques de Alaska relacionado a los datos de la era **moderna**.

Proyecto

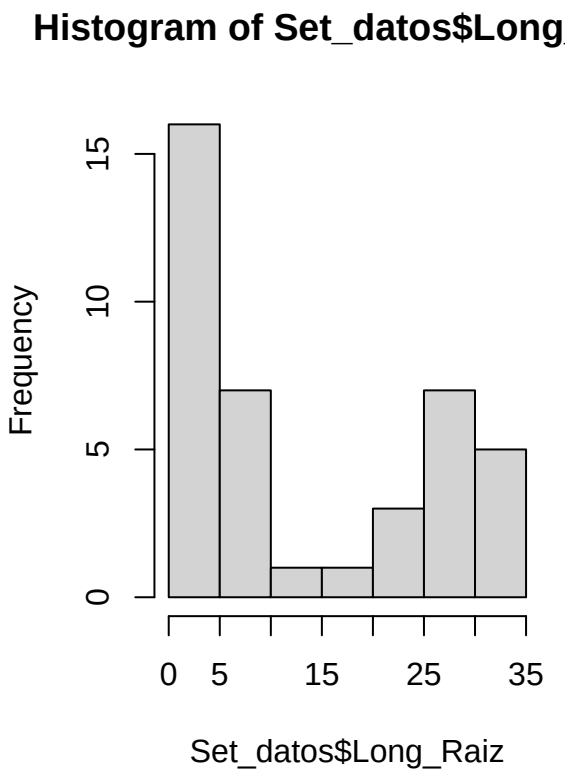
Normalidad:

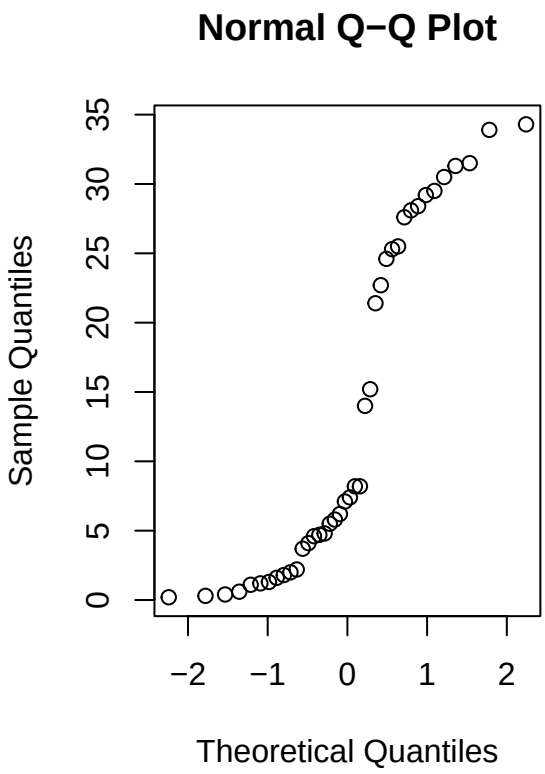
Para iniciar, hay que revisar si nuestro set de datos cuenta con *datos normales*, esto quiere decir que; con ayuda de shapiro, revisaremos si la distribución de los datos se ecuentra normal, por lo tanto, nuestro valor p-value debera ser mayor que el nivel de significancia que es 0.05.

```
# Primera variable
shapiro.test(Set_datos$Long_Raiz)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Set_datos$Long_Raiz
## W = 0.83328, p-value = 3.624e-05
```

```
par(mfrow=c(1,2))
hist(Set_datos$Long_Raiz)
qqnorm(Set_datos$Long_Raiz)
```



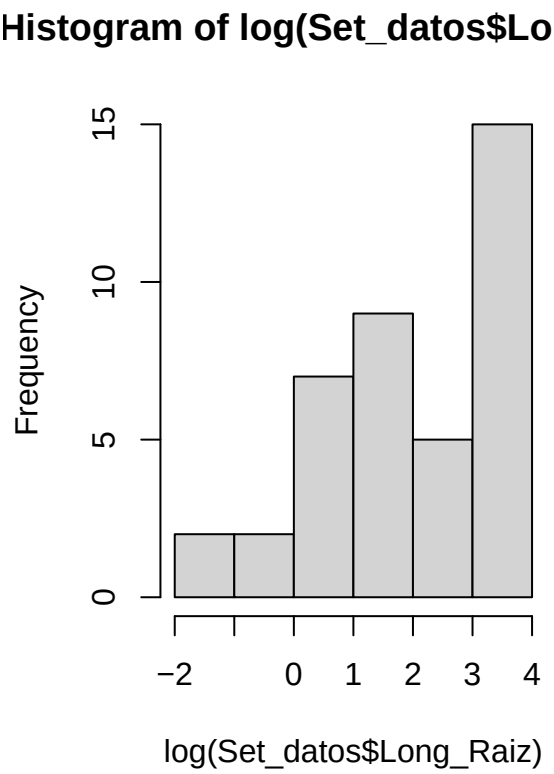


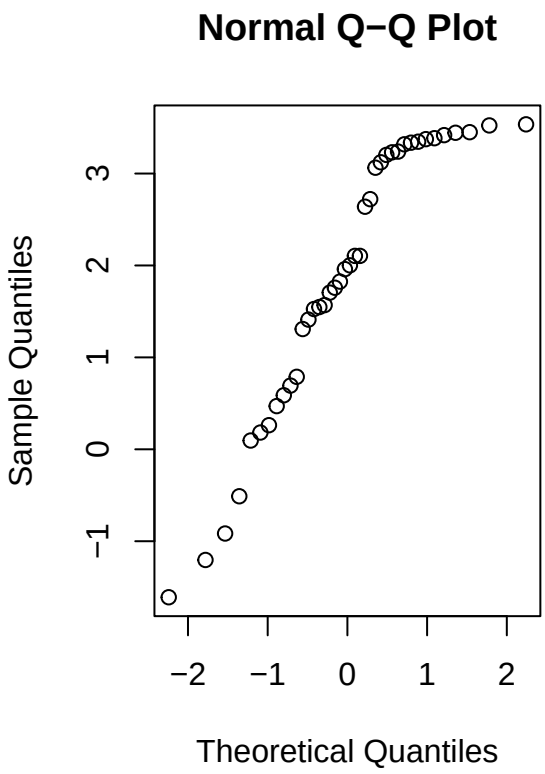
Al obtener un valor de p-value menor a 0.05 lo que procede es transformar los datos, en este caso, utilizaremos la funcion log ya que el histograma nos da como resultado una grafica con sesgo a la derecha.

```
shapiro.test(log(Set_datos$Long_Raiz))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(Set_datos$Long_Raiz)
## W = 0.90287, p-value = 0.00232
```

```
par(mfrow=c(1,2))
hist(log(Set_datos$Long_Raiz))
qqnorm(log(Set_datos$Long_Raiz))
```



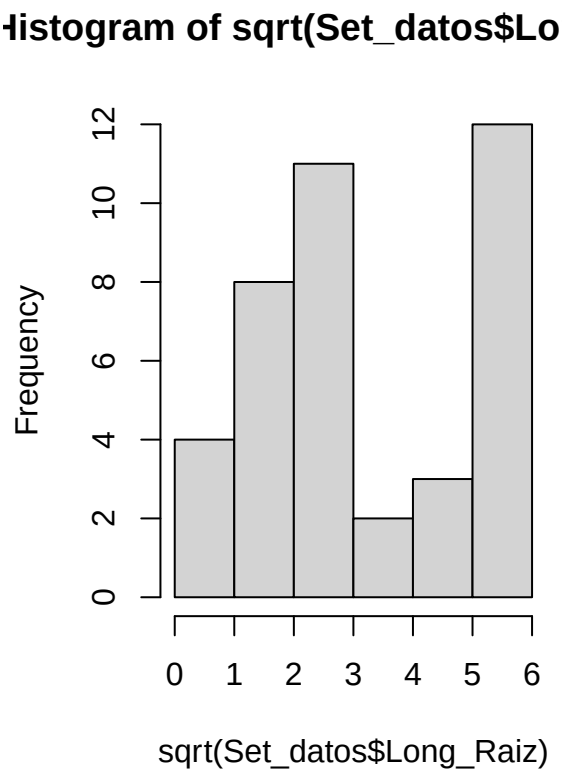


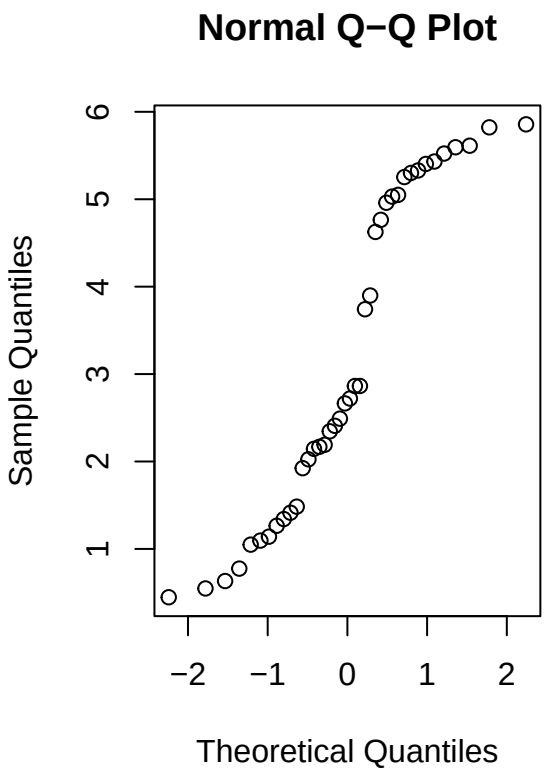
Al trasformar con funcion logaritmica nos damos cuenta de que si hay un arreglo en nuestros datos el cual es positivo pero no del todo bueno, ya que nuestro p-value es menor que nuestro nivel de significancia, por ende, decidi hacer una modificación y transformar con la funcion de raiz cuadrada ya que esta tambien nos ayuda cuando tenemos un histograma con sesgo a la derecha.

```
shapiro.test(sqrt(Set_datos$Long_Raiz))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sqrt(Set_datos$Long_Raiz)
## W = 0.89294, p-value = 0.001199
```

```
par(mfrow=c(1,2))
hist(sqrt(Set_datos$Long_Raiz))
qqnorm(sqrt(Set_datos$Long_Raiz))
```





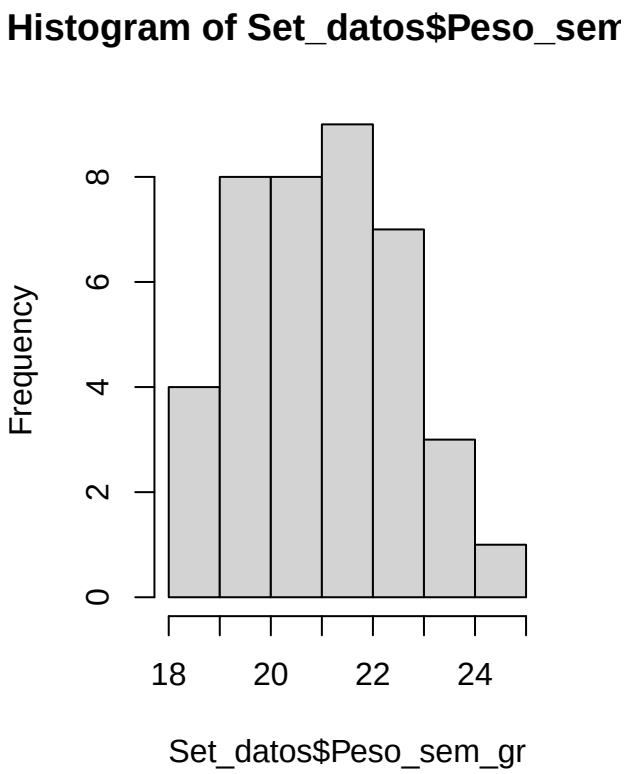
Al utilizar la funcion raiz cuadrada podemos observar que tambien obtenemos un arreglo positivo, pero si comparamos el resultado de shapiro (p-value) y los histogramas en cada funcion (logaritmica y raiz cuadrada), nos percatamos de que la mejor opción sería utilizar la función logaritmica para transformar nuestros datos. Sin pasar por alto que, aún asi nuestro p-value es menor al nivel de significancia y sabiedo que shapiro es una prueba muy exigente y se debería rechazar e intentar con pruebas no parametricas, en esta ocasion como aun no revisamos ese tema no la rechazaremos y seguiremos adelante...

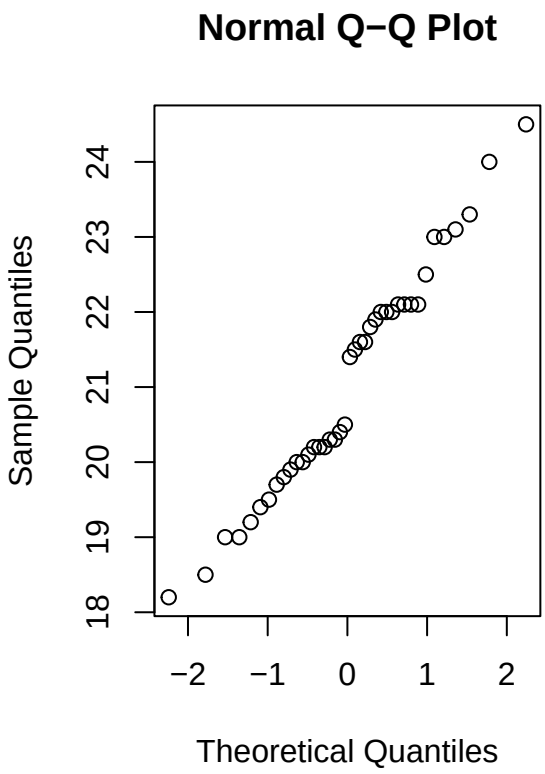
```
# Por otro lado, contamos con otra variable que a continuacion revisaremos...

shapiro.test(Set_datos$Peso_sem_gr)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Set_datos$Peso_sem_gr
## W = 0.96697, p-value = 0.2874
```

```
par(mfrow=c(1,2))
hist(Set_datos$Peso_sem_gr)
qqnorm(Set_datos$Peso_sem_gr)
```





Obtenemos como resultado que nuestro p-value es mayor al nivel de significancia que es 0.05 por lo cual, para esta variable, no sería necesario realizar una transformacion de datos ya que nuestros datos entrarian como datos normales, por lo tanto, no tenemos evidencia para rechazar.

Para continuar, una vez que contamos con nuestro set de datos, se debe revisar si este se encuentra *balanceado*.

```
# Balanceo
Set_datos %>%
  group_by(Tratamiento,Dia) %>%
  summarise(n()) #tomando en cuenta las variables independientes, obtenemos como resultado...
```

```
## `summarise()` has grouped output by 'Tratamiento'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 4 x 3
## # Groups:   Tratamiento [2]
##   Tratamiento Dia   `n()`
##   <chr>       <chr> <int>
## 1 "Control "  Dia 2    10
## 2 "Control "  Dia 7    10
## 3 "ECAP15"    Dia 2    10
## 4 "ECAP15"    Dia 7    10
```

```
#...que nuestro set se encuentra balanceado.
```

Ahora, revisaremos la *homogeneidad* de varianza de nuestros datos

```
leveneTest(log(Long_Raiz)~Tratamiento*Dia, data = Set_datos)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.8357 0.4832
##       36
```

Nuestra prueba de levene para la variable transformada de *longitud de raiz* nos da como resultado un valor mayor a 0.05, por lo tanto, nos dice que no existe una diferencia significativa y que los datos de nuestra variable son estadisticamente iguales.

```
leveneTest(Peso_sem_gr~Tratamiento*Dia, data = Set_datos)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.5038 0.6821
##       36
```

Para la variable de peso de semillas en gr la cual no fue necesaria transformar, nos da como resultado un valor mayor a 0.05, por ende, nos dice que tampoco existe diferencia significativa entre sus datos.

Seguimos con el *ANOVA...*

```
Long_anova <- aov(log(Long_Raiz)~Tratamiento*Dia, data = Set_datos)
Anova(Long_anova)
```

```
## Anova Table (Type II tests)
##
## Response: log(Long_Raiz)
##                 Sum Sq Df F value    Pr(>F)
## Tratamiento      4.952  1  4.3921    0.0432 *
## Dia             37.589  1 33.3367 1.394e-06 ***
## Tratamiento:Dia  1.275  1  1.1304    0.2948
## Residuals       40.592 36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para la variable de longitud de raiz encontramos que hay diferencias entre los valores de tratamiento y de dia ya que los valores del p son menores al nivel de significancia el cual es 0.05, lo que nos interpreta que es poco probable que la H0 sea cierta.

```r
Peso_anova <- aov(Peso_sem_gr~Tratamiento*Dia, data = Set_datos)
Anova(Peso_anova)
```

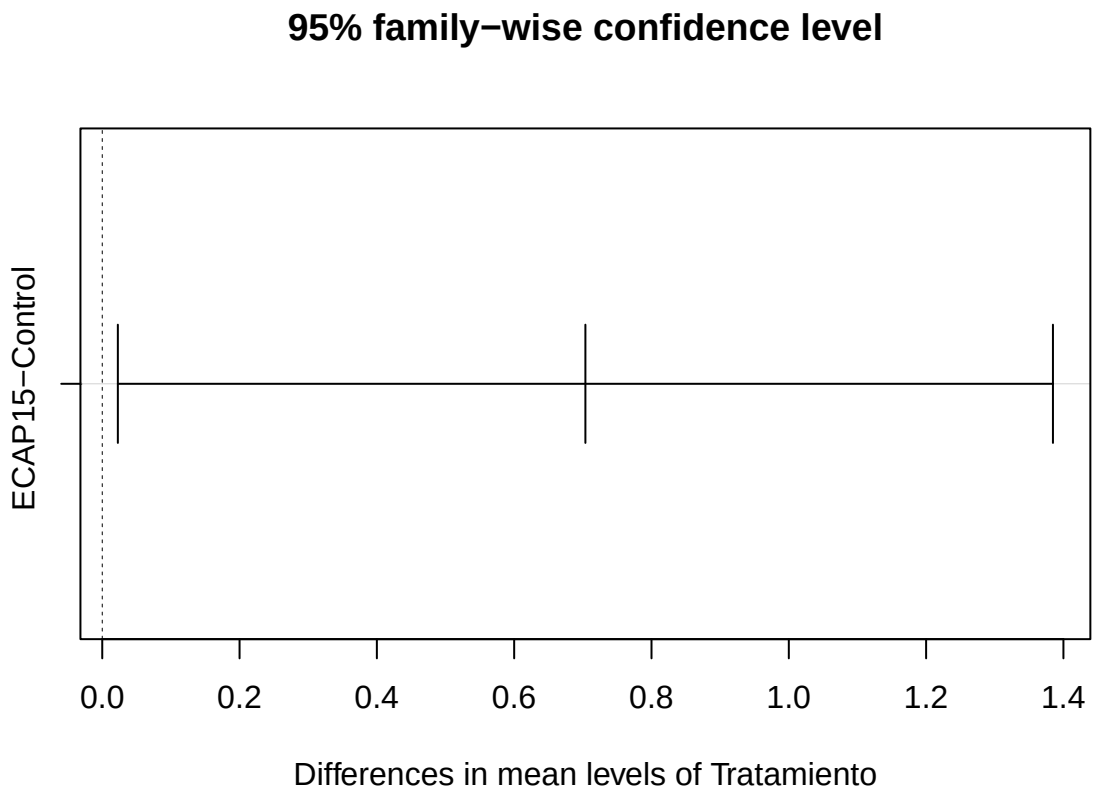
```
## Anova Table (Type II tests)
##
## Response: Peso_sem_gr
##                 Sum Sq Df F value    Pr(>F)
## Tratamiento     26.896  1 15.0598 0.0004264 ***
## Dia              0.841  1  0.4709 0.4969699
## Tratamiento:Dia  0.289  1  0.1618 0.6898658
## Residuals       64.294 36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

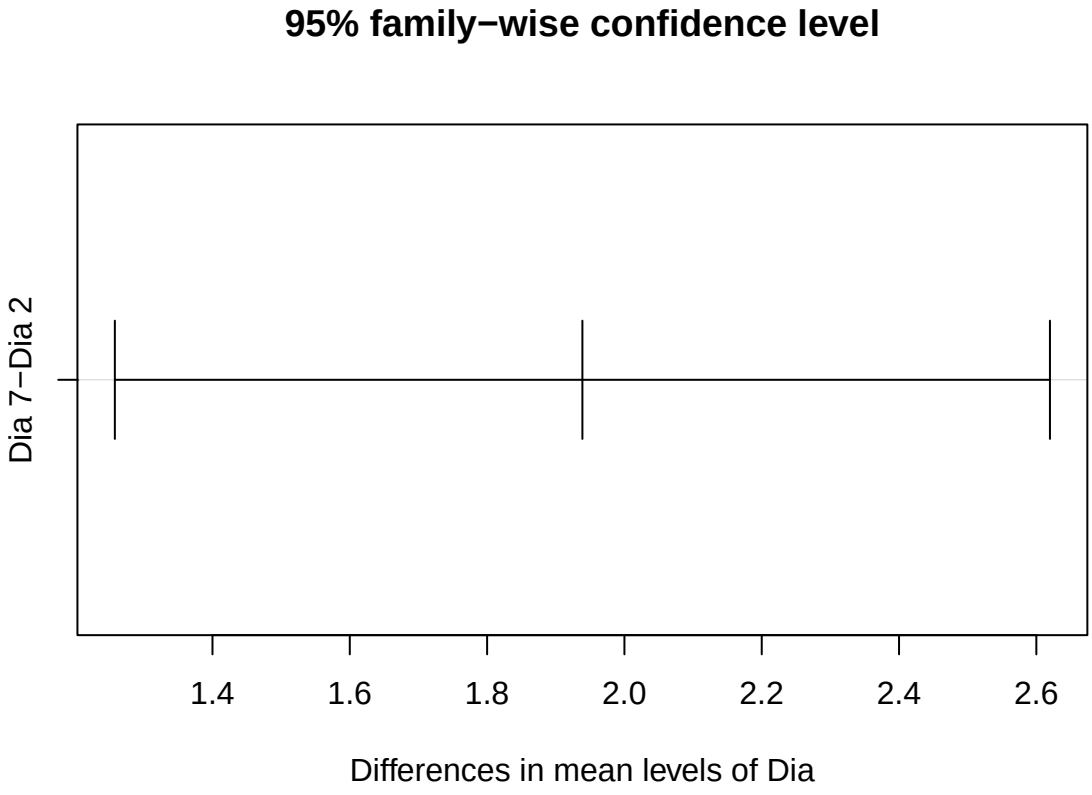
Mientras tanto, para la variable de peso de la semilla en gr encontramos que solo en el tratamiento hay diferencia ya que el p-value es menor al nivel de significancia, que de igual manera nos interpreta que es poco probable que la hipotesis nula sea verdadera.

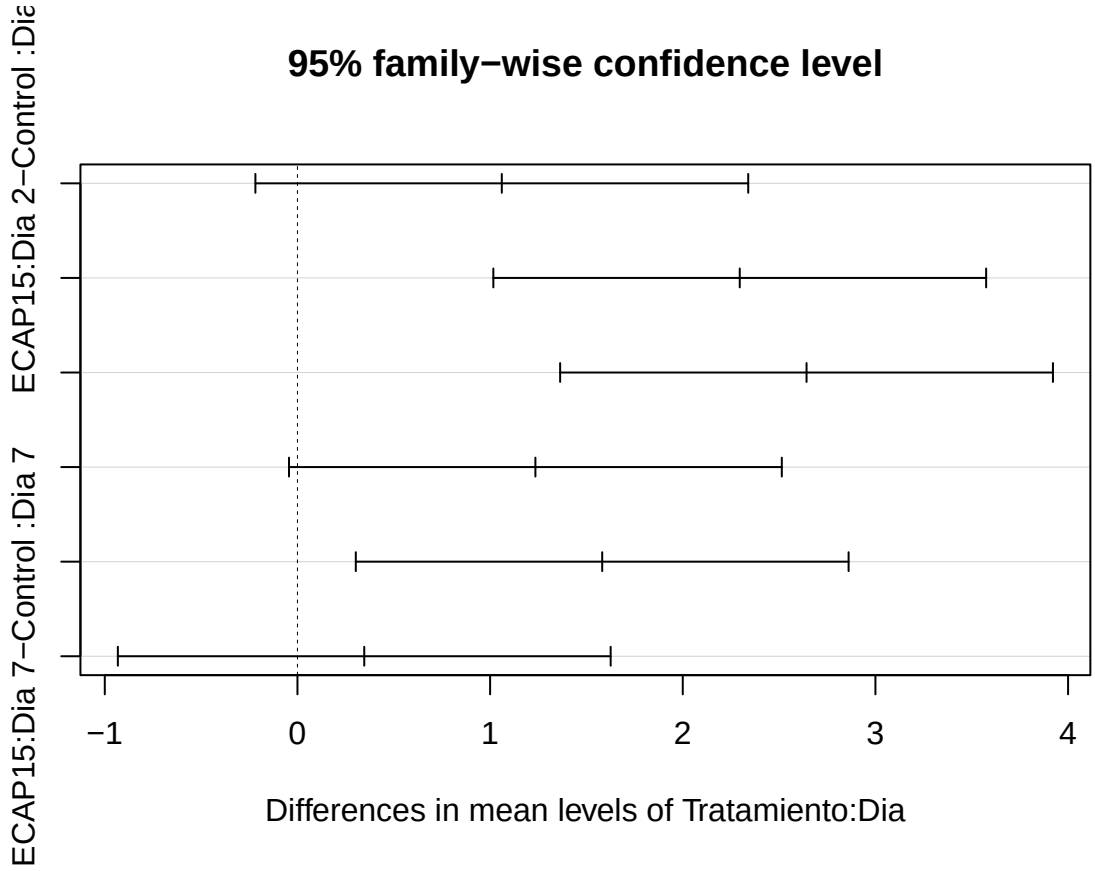
Por ultimo, terminamos con la prueba de *Tukey*...

```r
# Para la variable de longitud de raiz tenemos que:

Long_Tukey <- TukeyHSD(Long_anova)
plot(Long_Tukey)
```




95% family−wise confidence level
Dia 7−Dia 2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
Differences in mean levels of Dia

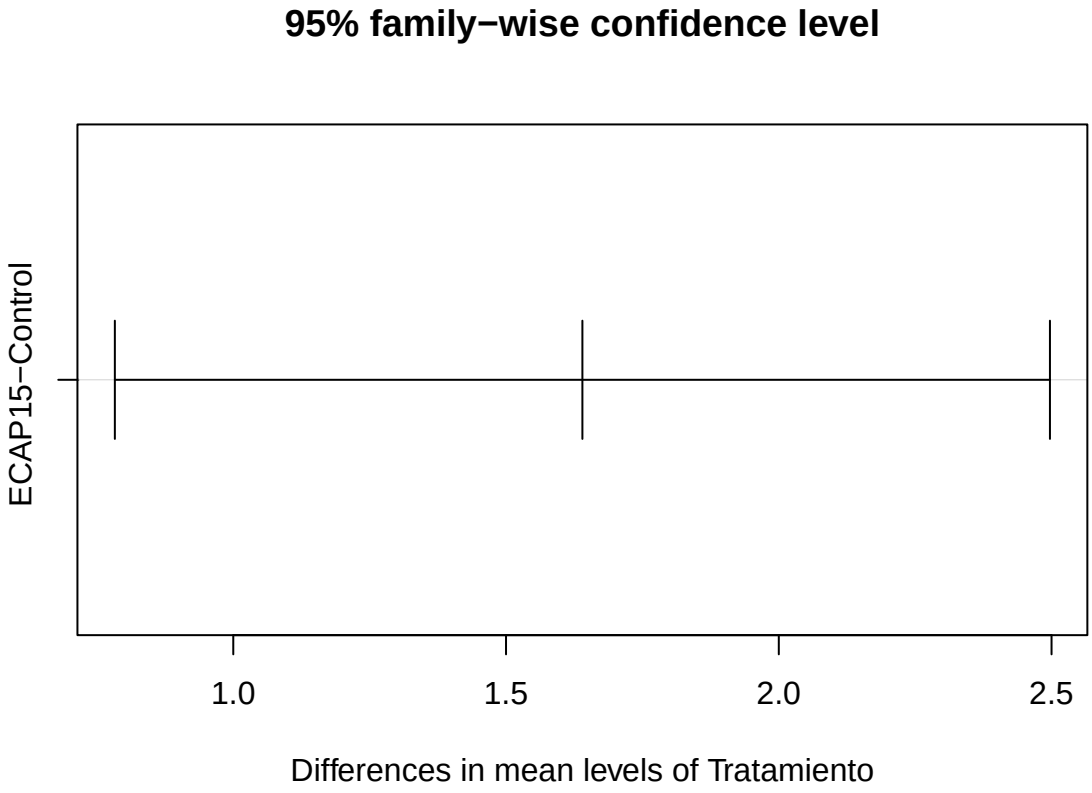


Interpretacion de resultados en prueba Tukey para la variable de longitud de raiz

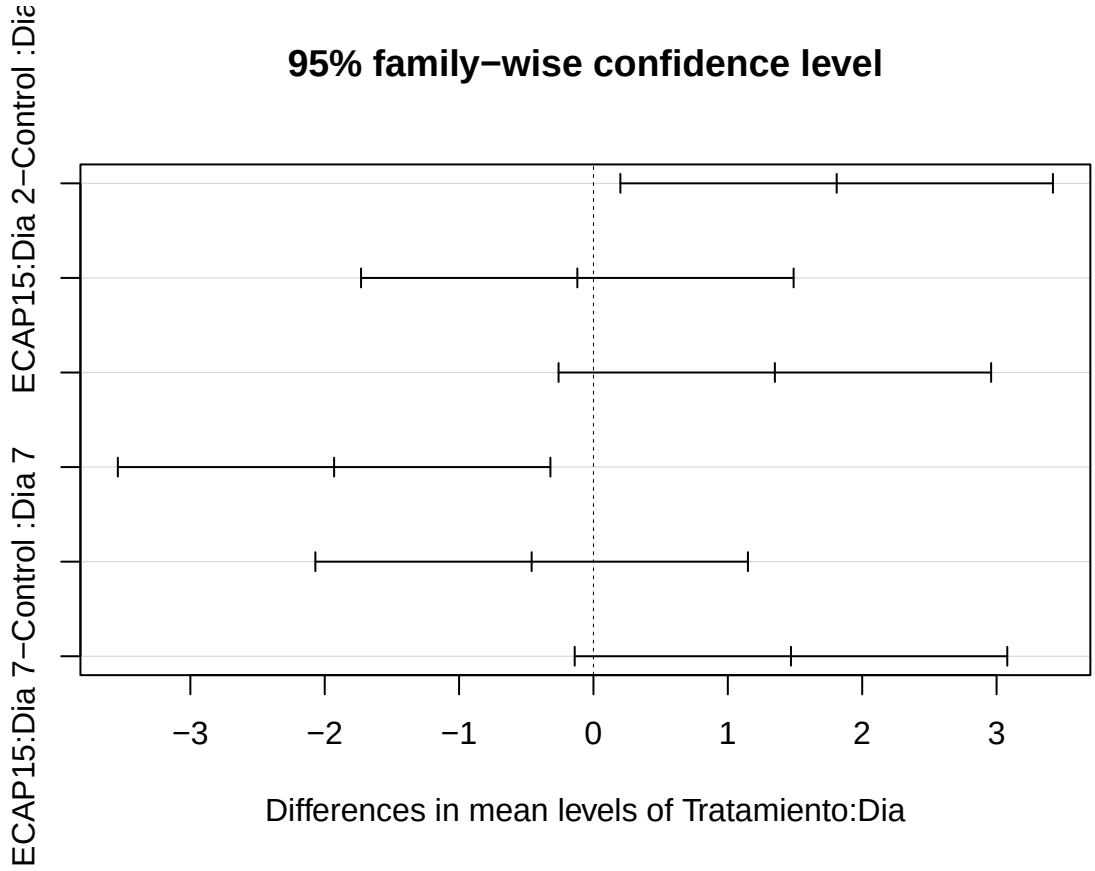
- **ECAP15: Dia2 - Control: Dia2** nos da como resultado que este intervalo de confianza no tiene diferencia significativa con respecto a los tratamientos antes mencionados.
- **Control: Dia7 - Control: Dia2** nos dice que si existe diferencia significativa entre estos tratamientos ya que su intervalo de confianza no entra en 0.
- **ECAP15: Dia7 - Control: Dia2** nos muestra que de igual manera si existe diferencia entre estos tratamientos y de nueva cuenta su intervalo de confianza no entra en el 0.
- **Control: Dia7 - ECAP15: Dia2** nos muestra que su intervalo de confianza si entra dentro del 0 por lo que no existe diferencia.
- **ECAP15: Dia7 - ECAP15: Dia2** da como resultado que tiene diferencia entre los tratamientos ya que su intervalo de confianza no esta dentro del 0.
- **ECAP15: Dia7 - Control: Dia7** por ultimo, tenemos que nuestro intervalo de confianza perteneciente a estos tratamientos no cuentan con diferencia significativa.

```
# Para la variable de peso de semilla tenemos que:

Peso_Tukey <- TukeyHSD(Peso_anova)
plot(Peso_Tukey)
```


95% family−wise confidence level
ECAP15−Control
1.0
1.5
2.0
2.5
Differences in mean levels of Tratamiento


95% family−wise confidence level
Dia 7−Dia 2
−1.0
−0.5
0.0
0.5
Differences in mean levels of Dia



Interpretacion de resultados en prueba Tukey para la variable de peso de semilla en gr

- **ECAP15: Dia2 - Control: Dia2** nos da como resultado que este intervalo de confianza si tiene diferencia significativa con respecto a los tratamientos.
- **Control: Dia7 - Control: Dia2** nos dice que no existe diferencia significativa entre estos tratamientos ya que su intervalo de confianza si entra en 0.
- **ECAP15: Dia7 - Control: Dia2** nos muestra que de igual manera no existe diferencia entre estos tratamientos y de nueva cuenta su intervalo de confianza si entra en el 0.
- **Control: Dia7 - ECAP15: Dia2** nos muestra que su intervalo de confianza no entra dentro del 0 por lo que si existe diferencia.
- **ECAP15: Dia7 - ECAP15: Dia2** da como resultado que no tiene diferencia entre los tratamientos ya que su intervalo de confianza esta dentro del 0.
- **ECAP15: Dia7 - Control: Dia7** por ultimo, tenemos que nuestro intervalo de confianza perteneciente a estos tratamientos no cuentan con diferencia significativa.

for loop aplicado a nuestro proyecto...

En esta ocacion, aplicaremos un for loop en donde modificaremos en la variable de peso de semilla un cambio en la unidad de medida. Tenemos que nuestro peso se encuentra en gr pero en este caso vamos a realizar una modificacion y poner el peso en mg.

```
for (i in 5 : nrow(Set_datos)) {
  Peso_mg <- Set_datos$Peso_sem_gr[i]*1000
  print(Peso_mg)
}
```

```
## [1] 20500
## [1] 20200
## [1] 20200
## [1] 21500
## [1] 22100
## [1] 22100
## [1] 20000
## [1] 19000
## [1] 19400
## [1] 20400
## [1] 19200
## [1] 19700
## [1] 19000
## [1] 22000
## [1] 21600
## [1] 21400
## [1] 21800
## [1] 23100
## [1] 22100
## [1] 23300
## [1] 23000
## [1] 23000
## [1] 22500
## [1] 21900
## [1] 20000
## [1] 20300
## [1] 22100
## [1] 22000
## [1] 22000
## [1] 21600
## [1] 24000
## [1] 24500
## [1] 19900
## [1] 19800
## [1] 20300
## [1] 20200
```

Si bien se transformo la unidad de peso de gr a mg, observamos que solo se puede observar y que en realidad la tabla no se modifico ya que estos resultados solo se imprimieron y no se guardaron ya que no creamos un objeto.

Link para más información sobre funciones que puedes utilizar en RStudio

Figure 1: Imagen