Proyecto

Mariana Corte Rodriguez

2023-10-26

Introducción

La lechuga *(Lactuca sativa L.)* es una planta herbacea anual perteneciente a la familia Compositae. Es una planta que presenta un elevado contenido de agua del 90 al 95% y su valor nutricional se resalta por el alto contenido de minerales y vitaminas.

Figure 1: Plantulas de lechuga

Liga con informacion extra sobre la germinacion para las semillas de lechuga

Se realizó un estudio para medir la efectividad de un bioestimulante en la germinación de las plántulas de lechuga y se comparo con el crecimiento con el uso de un fertilizante a dos temperaturas distintas para determinar el mejor tratamiento y la temperatura optima mediante un analisis estadistico.

```
# 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(readr)
library(tidyverse)
```

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

## -- 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
```

```
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
```

```
library(tinytex)
library(devtools)
```

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

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

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

```
# Set de datos que se utilizara y que se cargo al ambiente utilizando la libreria readr:

Set_datos <- read.csv("~/CursoR/CursoRgit/Proyectos/Set_datos_nuevo.csv")
Estanque_plantas <- read.csv("~/CursoR/CursoRgit/Proyectos/Estanque_plantas.csv")
```

Analisis estadistico

Antes de comenzar a realizar el analisis estadistico del set de datos es necesario seguir los siguientes pasos para asegurarnos que los datos cumplan con todas las suposiciones:

- Normalización
- Transformación
- Balancear
- Homogeneidad de varianzas

1. Normalización y Transformacion: Los datos a analizar deben seguir una distribución normal por lo que se realizará el test de shapiro, un histograma y un qqplot. Se utilizan los siguientes codigos:

- shapiro.test()
- hist()
- qqnorm()

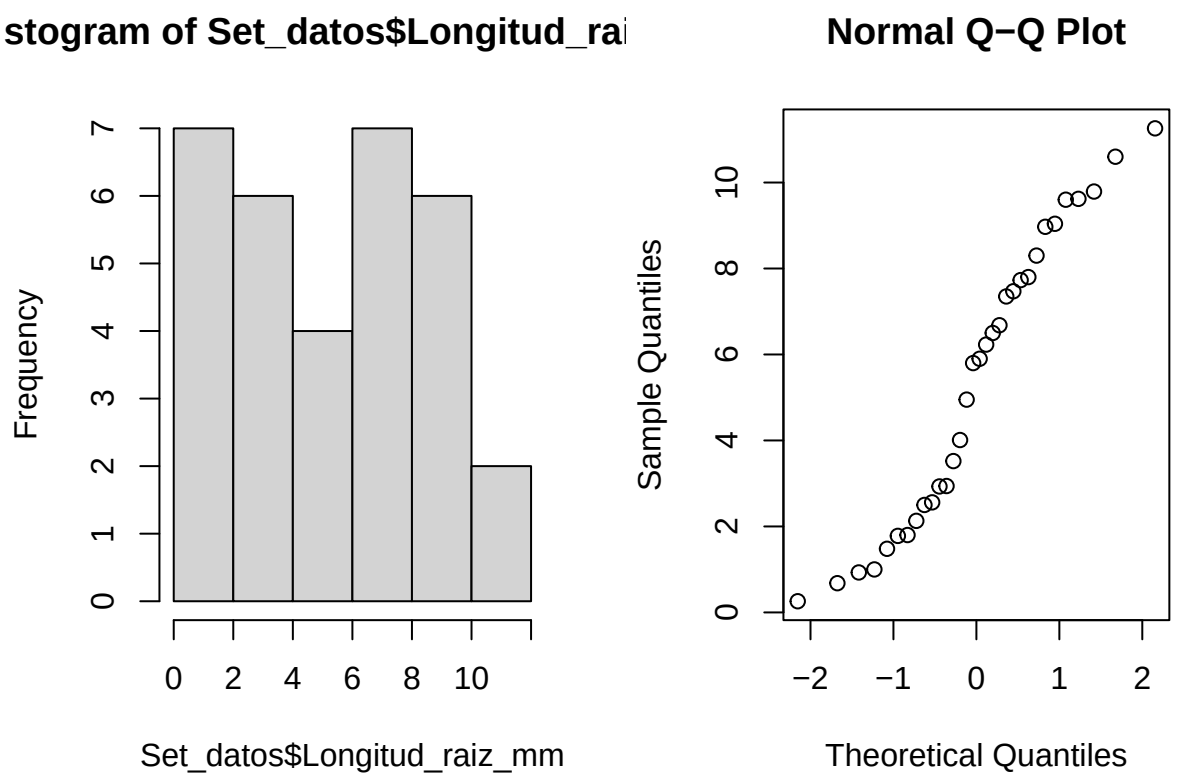
Si los datos de las variables a analizar no estan normalizados se deberán transformar siguiendo las normas dependiendo de como esten distribuidos los datos.

```
# Test de Shapiro

for (i in 3:ncol(Set_datos)) {
  # quiere decir que: "para la celda (i) a partir de la columna 3 de la tabla"
  shapiro <- shapiro.test(Set_datos[,i]) # va a realizar el test de shapiro
  normal <- ifelse(shapiro[["p.value"]]>0.05, "YES", "NO")
  # si la respuesta es verdadera va a imprimir YES y si es falsa es NO
  print(c(i,normal))
}
```

```
## [1] "3"   "YES"
## [1] "4"  "NO"
## [1] "5"  "NO"
```

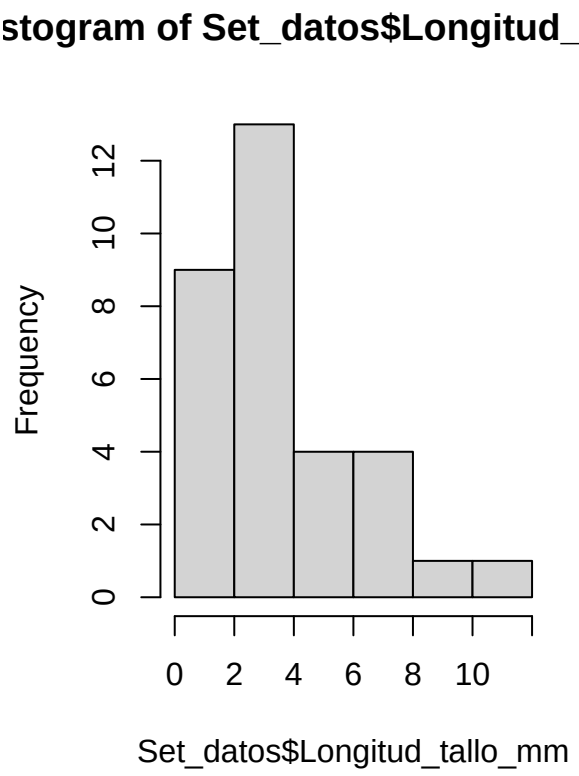
Se utilizo la funcion “FOR LOOP IF ELSE” para resolver de una sola vez cuales de las tres variables a analizar si cumplen con el test de shapiro.Se obtuvo como resultado que para la variable de longitud de raiz si cumple con el test por que el valor de pvalue si es mayor a 0.05 y la respuesta que arrojo fue “YES”, mientras que para la longitud de tallo y el numero de hojas no cumplen.

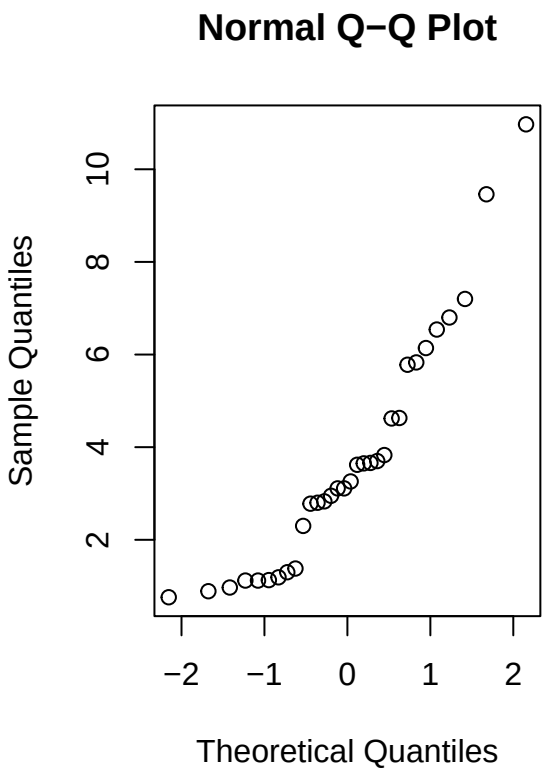
```
# Histograma y qqplot para la variable Longitud de Raiz

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



Ambos graficos se observan normalizados por lo que no es necesario transformar los datos.

```
# Histograma, qqplot y test de shapiro para la variable Longitud de Tallo

par(mfrow=c(1,2)) # se acomodan los graficos en 1 fila y 2 columnas para antes de transformar
hist(Set_datos$Longitud_tallo_mm)
qqnorm(Set_datos$Longitud_tallo_mm)
```



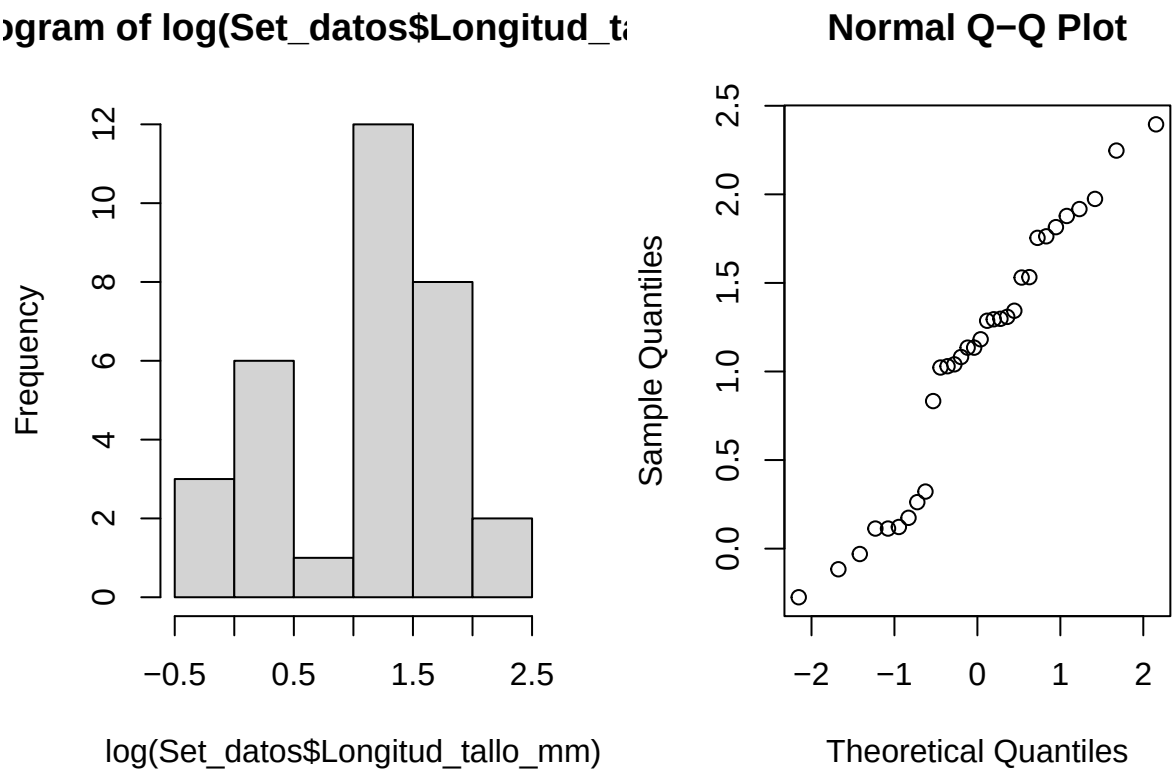



```
shapiro.test(Set_datos$Longitud_tallo_mm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Set_datos$Longitud_tallo_mm
## W = 0.89863, p-value = 0.005716
```

Se deben transformar los datos para la variable Longitud_tallo_mm debido a que los datos se observan con sesgo a la derecha y segun las normas de normalizacion se debera usar “log” para transformar los datos.

```
par(mfrow=c(1,2)) # se acomodan los graficos en 1 fila y 2 columnas para despues de transformar
hist(log(Set_datos$Longitud_tallo_mm))
qqnorm(log(Set_datos$Longitud_tallo_mm))
```




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

```
##
##  Shapiro-Wilk normality test
##
## data:  log(Set_datos$Longitud_tallo_mm)
## W = 0.94596, p-value = 0.1106
```

Una vez transformados los datos se obtiene un pvalue mayor a 0.05 para el test de shapiro, el histograma se observa simetrico y el qqplot esta un poco más linearizado, por lo que los datos se pueden considerar normalizados.

2. Checar si los datos estan balanceados: Una vez que los datos esten normalizados se procede a revisar que las variables independientes del set de datos esten balanceadas para decidir que tipo de ANOVA se utilizara.

```
Set_datos %>%
  group_by(Tratamiento,Temperatura)%>%
  summarise(n())
```

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

## # A tibble: 4 x 3
## # Groups:   Tratamiento [2]
```

```
##   Tratamiento      Temperatura `n()`
##   <chr>            <chr>       <int>
## 1 "Bioestimulante" 15 grados C     8
## 2 "Bioestimulante" 20 grados C     8
## 3 "Fertilizante "  15 grados C     8
## 4 "Fertilizante "  20 grados C     8
```

```
# Si estan balanceados: hay 8 muestras para c/tratamiento y temperatura,
# por lo que se utilizara la funcion de ANOVA TIPO I.
```

3. Homogeneidad de varianzas: Se caracteriza por poder comparar 2 o más poblaciones y esto es importante a la hora de contrastar la homogeneidad de varianzas para determinar si los grupos se distribuyen de forma normal o no. Si los datos son homogeneos se puede proceder a realizar el analisis de ANOVA y Tukey. Se utiliza el siguiente codigo:

- leveneTest(varable dep ~ variable indep, datos =)

```
# Variable Longitud de raiz

leveneTest(Longitud_raiz_mm ~ Tratamiento*Temperatura, data = Set_datos)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  1.3368 0.2823
##       28
```

```
# p-value > 0.05 significa que si existe homogeneidad
```

Para la variable Longitud_raiz_mm da como resultado un p-value de 0.28, por lo que es mayor a 0.05, y por lo tanto se puede continuar con el analisis de ANOVA porque los datos son homogeneos entre si.

```
# Variable Longitud de tallo

leveneTest(log(Longitud_tallo_mm) ~ Tratamiento*Temperatura, data = Set_datos)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  3  2.5636 0.07474 .
##       28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# p-value > 0.05 significa que si existe homogeneidad
```

Para la variable Longitud_tallo_mm da como resultado un p-value de 0.074, de igual manera es mayor a 0.05, y por lo tanto se puede continuar con el analisis de ANOVA, ya que si existe homogeneidad.

4. ANOVA: Es una fórmula estadística que se utiliza para comparar las varianzas entre las medias de mas de dos grupos. Existe el ANOVA de una sola vía en donde se compara una variable independiente sobre una variable dependiente. Así mismo, existe el ANOVA de dos vías en donde se compara dos variables independientes sobre una variable dependiente. Codigo:

- aov(var dep ~ var indep, data =)
- Anova(datos)
- Anova(datos, type = 2)
- Anova(datos, type = 3)

```
# Variable Longitud de raiz

Est_anova <- aov(Longitud_raiz_mm ~ Tratamiento*Temperatura, data = Set_datos)
# para la interaccion entre dos variables se recomienda utilizar el signo "*"

Anova(Est_anova) # p-value < 0.05 indica que si hay diferencias significativas
```

```
## Anova Table (Type II tests)
##
## Response: Longitud_raiz_mm
##                          Sum Sq Df F value    Pr(>F)
## Tratamiento             237.784  1 80.0137 1.062e-09 ***
## Temperatura              24.448  1  8.2265  0.007762 **
## Tratamiento:Temperatura   2.732  1  0.9193  0.345866
## Residuals                83.210 28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como el valor de p-value es menor a 0.05 entonces si hay diferencias significativas en la longitud de raiz para los tratamientos y la temperatura.

```
# Variable Longitud de tallo

Est_anova2 <- aov(log(Longitud_tallo_mm) ~ Tratamiento*Temperatura, data = Set_datos)
Anova(Est_anova2)
```

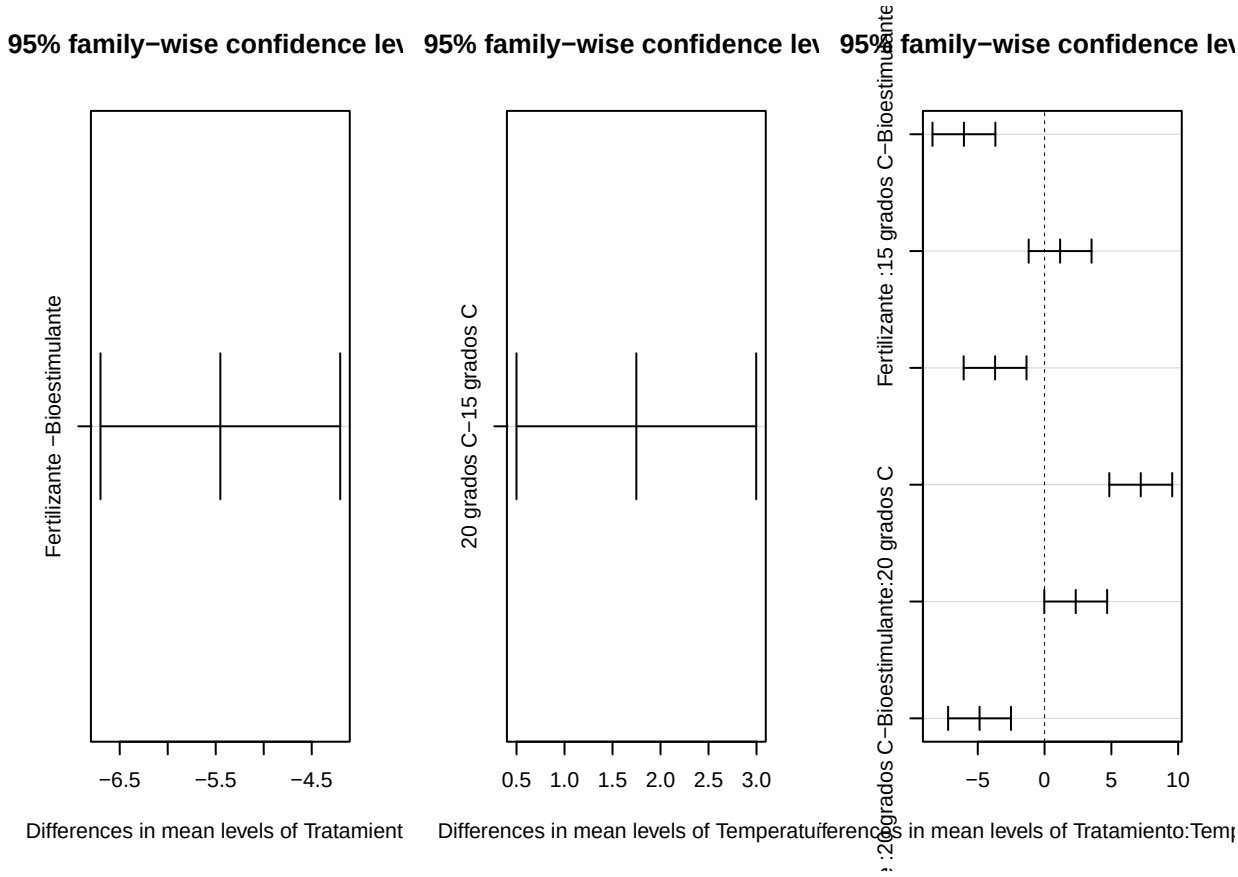
```
## Anova Table (Type II tests)
##
## Response: log(Longitud_tallo_mm)
##                          Sum Sq Df F value    Pr(>F)
## Tratamiento              9.7503  1 39.0290 9.432e-07 ***
## Temperatura              0.0426  1  0.1704    0.6829
## Tratamiento:Temperatura 0.0006  1  0.0022    0.9627
## Residuals                6.9950 28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los resultados indican que para la variable de longitud de tallo los tratamientos que se probaron en el ensayo si presentaron diferencias significativas debido a que el p-value es menor a 0.05.

5. Tukey: En el caso que en el ANOVA existan diferencias significativas entre los tratamientos, el siguiente paso es realizar un test de Tukey en el cual nos indique cuales de los tratamientos son diferentes entre si. Codigo:

- TukeyHSD()

```r
# Variable Longitud de raiz

Est_Tukey <- TukeyHSD(Est_anova) # se creo un objeto para guardar los resultados de  Tukey

# pvalue < 0.05 indica que hay diferencias significativas

par(mfrow=c(1,3)) # se van acomodar los graficos en 1 fila y 3 columnas
plot(Est_Tukey)
```

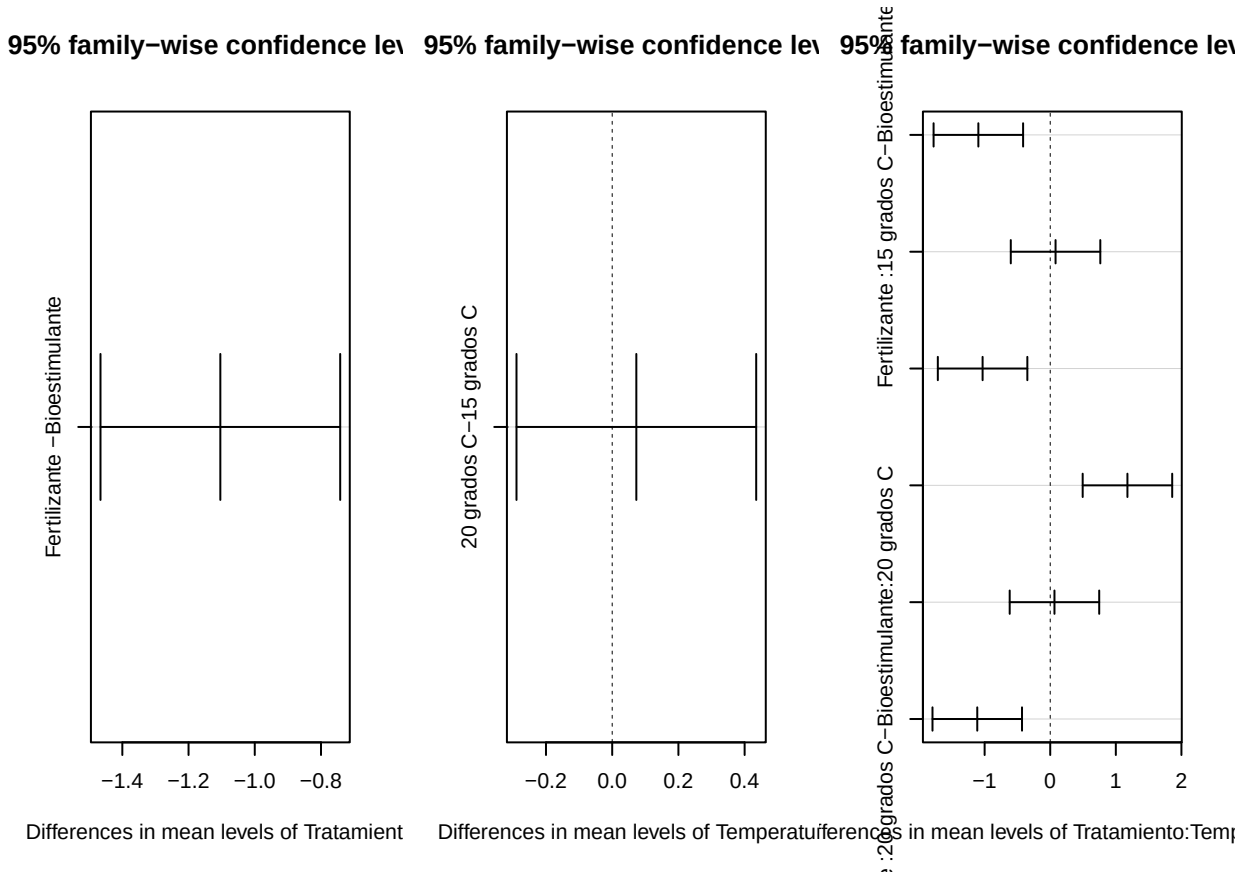


Se obtuvieron los siguientes resultados despues del analisis estadistico por Tukey:

- **p-value = 7.48e-7:** Las semillas a las que se les aplico fertilizante y que crecieron a 15°C presentaron diferencias significativas sobre aquellas semillas que se les aplico bioestimulante a 15°C. Esto indica que si hay diferencias sobre el tratamiento que se utilizo en cuanto a la longitud de raiz de las plantas.
- **p-value = 5.39e-1:** Las semillas a las que se les aplico bioestimulante y germinaron a 20°C no presentaron diferencias significativas sobre aquellas semillas que se les aplico bioestimulante a 15°C. Esto indica que no hay diferencias en cuanto a la temperatura de germinación para un mismo tratamiento, ya que las plantas crecieron casi iguales en cuanto a la longitud de raiz.
- **p-value = 1.02e-3:** Las semillas a las que se les aplico fertilizante y germinaron a 20°C presentaron diferencias significativas sobre aquellas semillas que se les aplico bioestimulante a 15°C. Esto nos indica que si hay diferencias en la longitud de raiz en cuanto a la temperatura y el tratamiento a utilizar.
- **p-value = 2.54e-8:** Las semillas a las que se les aplico bioestimulante y germinaron a 20°C presentaron diferencias significativas sobre aquellas semillas que se les aplico fertilizante a 15°C. Esto nos indica que si hay diferencias en la longitud de raiz en cuanto a la temperatura y tratamiento a utilizar.

- **p-value = 5.27e-2:** Las semillas a las que se les aplico fertilizante y germinaron a 20°C no presentaron diferencias significativas sobre aquellas semillas que se les aplico fertilizante a 15°C. Esto nos indica que no hay diferencias en la longitud de raiz en cuanto a la variacion de temperatura de germinacion, sin embargo, estan justo en el rango.
- **p-value = 2.70e-5:** Las semillas a las que se les aplico fertilizante y germinaron a 20°C presentaron diferencias significativas sobre aquellas semillas que se les aplico bioestimulante a 20°C. Esto nos indica que si hay diferencias en la longitud de raiz en cuanto al tratamiento a utilizar.

```
# Variable Longitud de tallo

Est_Tukey2 <- TukeyHSD(Est_anova2)

par(mfrow=c(1,3))
plot(Est_Tukey2)
```



Se obtuvieron los siguientes resultados despues del analisis estadistico por Tukey:

- **p-value = 0.0008:** Las semillas a las que se les aplico fertilizante y que crecieron a 15°C presentaron diferencias significativas sobre aquellas semillas que se les aplico bioestimulante a 15°C. Esto indica que si hay diferencias sobre el tratamiento que se utilizo en cuanto a la longitud de tallo de las plantas.
- **p-value = 0.9878:** Las semillas a las que se les aplico bioestimulante y crecieron a 20°C no presentaron diferencias significativas sobre aquellas semillas que se les aplico bioestimulante a 15°C. Esto nos indica que no hay diferencias en el crecimiento del tallo de las plantas al variar la temperatura, en ambos casos las plantas crecieron bien al utilizar el tratamiento del bioestimulante.

- **p-value = 0.0016:** Las plantas a las que se les aplico fertilizante a 20°C presentaron diferencias significativas sobre aquellas semillas que se les aplico bioestimulante a 15°C. Esto nos indica que si hay diferencias en el crecimiento del tallo de las plantas al variar el tratamiento.
- **p-value = 0.0003:** Las plantas a las que se les aplico bioestimulante a 20°C presentaron diferencias significativas sobre aquellas semillas que se les aplico fertilizante a 15°C. Esto nos indica que si hay diferencias en el crecimiento del tallo de las plantas al variar la temperatura y el tratamiento.
- **p-value = 0.9938:** Las semillas a las que se les aplico fertilizante a 20°C no presentaron diferencias significativas sobre aquellas semillas que se les aplico fertilizante a 15°C. Por lo que no hay diferencias en el crecimiento del tallo de las plantas al variar la temperatura.
- **p-value = 0.0006:** Las semillas a las que se les aplico fertilizante a 20°C presentaron diferencias significativas sobre aquellas semillas que se les aplico bioestimulante a 20°C. Esto nos indica que si hay diferencias en el crecimiento del tallo de las plantas al variar el tratamiento.

Ejercicio de correlación

Usando los datos "modernos" de la tabla Estanques_plantas, determinar si existe una correlación entre la biomasa de dos especies acuáticas de plantas en los estanques de Alaska: Carex y Arctophila.

1. Revisar si los datos cumplen todas las suposiciones de una correlación.

2. Reportar el coeficiente de correlación y su p-value

3. Explicar que significan estos valores y denle una interpretación a los resultados

```
Estanque_filt <- Estanque_plantas[which(Estanque_plantas$Era == "Modern"),]
# Se filtraron los datos utilizando el comando which para seleccionar solo los datos
# modernos de la columna "Era".

cor.test(Estanque_filt$Artophila,Estanque_filt$Carex,
         method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Estanque_filt$Artophila and Estanque_filt$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
```

- **Coeficiente de correlación:** 0.6625
- **p-value:** 0.0019
- **Resultados:** El valor del coeficiente de correlación nos indica que los datos de la biomasa de las dos especies acuáticas de plantas *(Carex y Arctophilaque)* del estanque de Alaska, si estan relacionadas entre si. Asi mismo, al tener un p-value menor a 0.05 nos indica que esta relacion entre las especies si es significativa.