Tema 4. DoE. Ejercicios obligatorios

Dra. Rosana Ferrero

Máxima Formación

Contents


```
library(tidyverse)
library(ggstatsplot)
library(rstatix)
library(ggpubr)
library(WRS2)
```

En el **TEMA 4** del curso DOE has aprendido a analizar respuestas numéricas de experimentos con un solo factor predictor, con el **ANOVA de una vía**, para diseños completamente aleatorizados (CRD). Has visto el caso de muestras independientes y muestras relacionadas. Utilizamos enfoques paramétricos, no paramétricos y robustos, así como pruebas de comparación múltiples post hoc para cada caso y gráficos finales en formato elegante.

Ejercicio 1.

Utiliza los datos coagulation del paquete faraway que contienen 24 tiempos de coagulación de sangre de un experimento donde 24 animales fueron aleatoriamente asignados a 4 dietas diferentes y las muestras se tomaron en orden aleatorio (Box, Hunter & Hunter, 1978). Indica si existen diferencias entre los tiempos de coagulación según la dieta del animal y, en tal caso, cómo son estas diferencias.

```
library(faraway)
data(coagulation)
head(coagulation)
```

```
##   coag diet
## 1   62    A
## 2   60    A
## 3   63    A
## 4   59    A
## 5   63    B
## 6   67    B
```

```
summary(coagulation)
```

```
##       coag       diet
##  Min.   :56.00   A:4
##  1st Qu.:61.75   B:6
##  Median :63.50   C:6
##  Mean   :64.00   D:8
##  3rd Qu.:67.00
```

```
##  Max.   :71.00
```

Tenemos una respuesta cuantitativa (tiempos de coagulación) y un predictor cualitativo con 4 niveles (la dieta, que son muestras independientes). Piensa entonces qué prueba de hipótesis tiene sentido realizar. Estudia los supuestos paramétricos y selecciona la versión adecuada de la prueba.

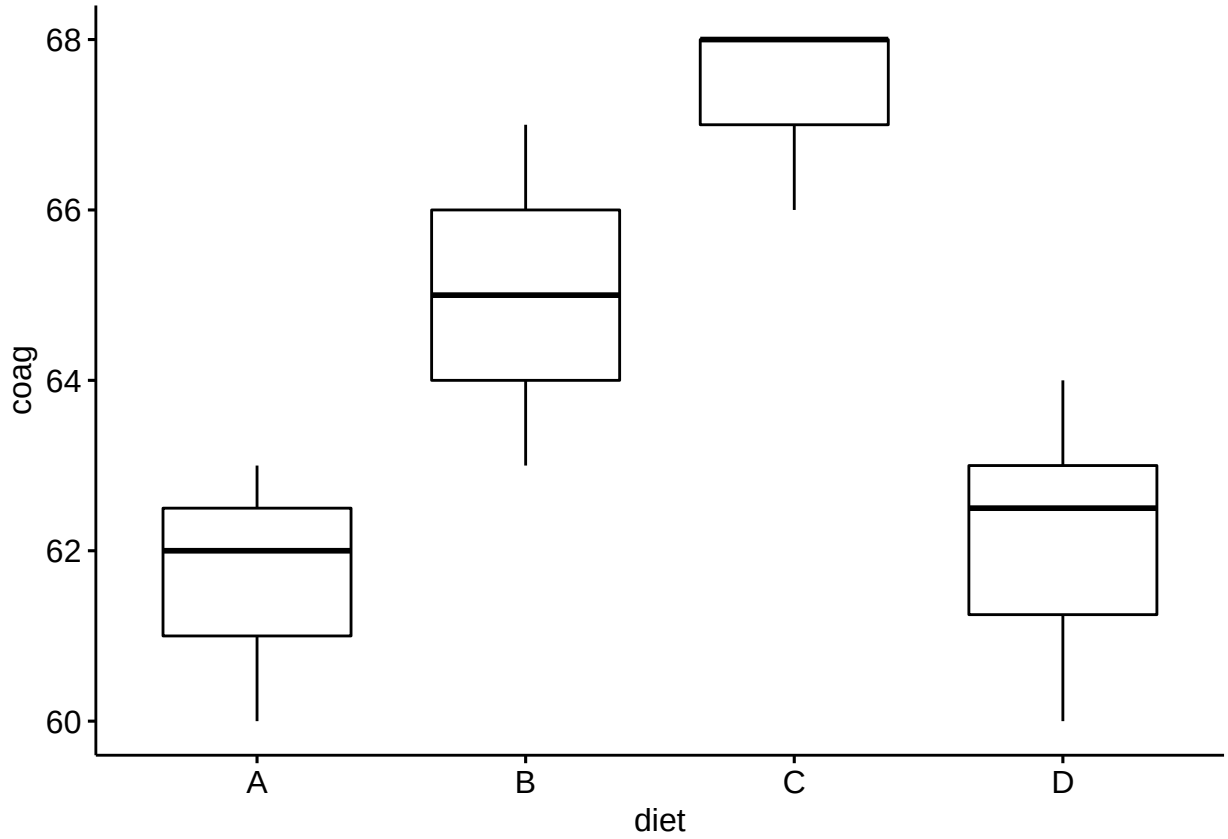
Respuesta: Puesto que tenemos cuatro categorias (dietas) y una variable contínua (coagulación) en muetras no relacionadas (cada valor corresponde a perros diferentes) optaremos por una anova de 1 vía para muestras no relacionadas. La hipótesis nula será que o existe diferencia etre las coagulaciones medias de cada tipo de dieta ($H_0 : \mu(d1) = \mu(d2) = \mu(d3) = \mu(d4)$), mientras que la hipótesis alternativa es que algunas de estas medias será diferente.

Para hacer este test paramétrico tenemos que ver si se cumplen los supuestos: independencia (si), no outliers, normalidad y homogeneidad de varianzas. Veamos cómo.

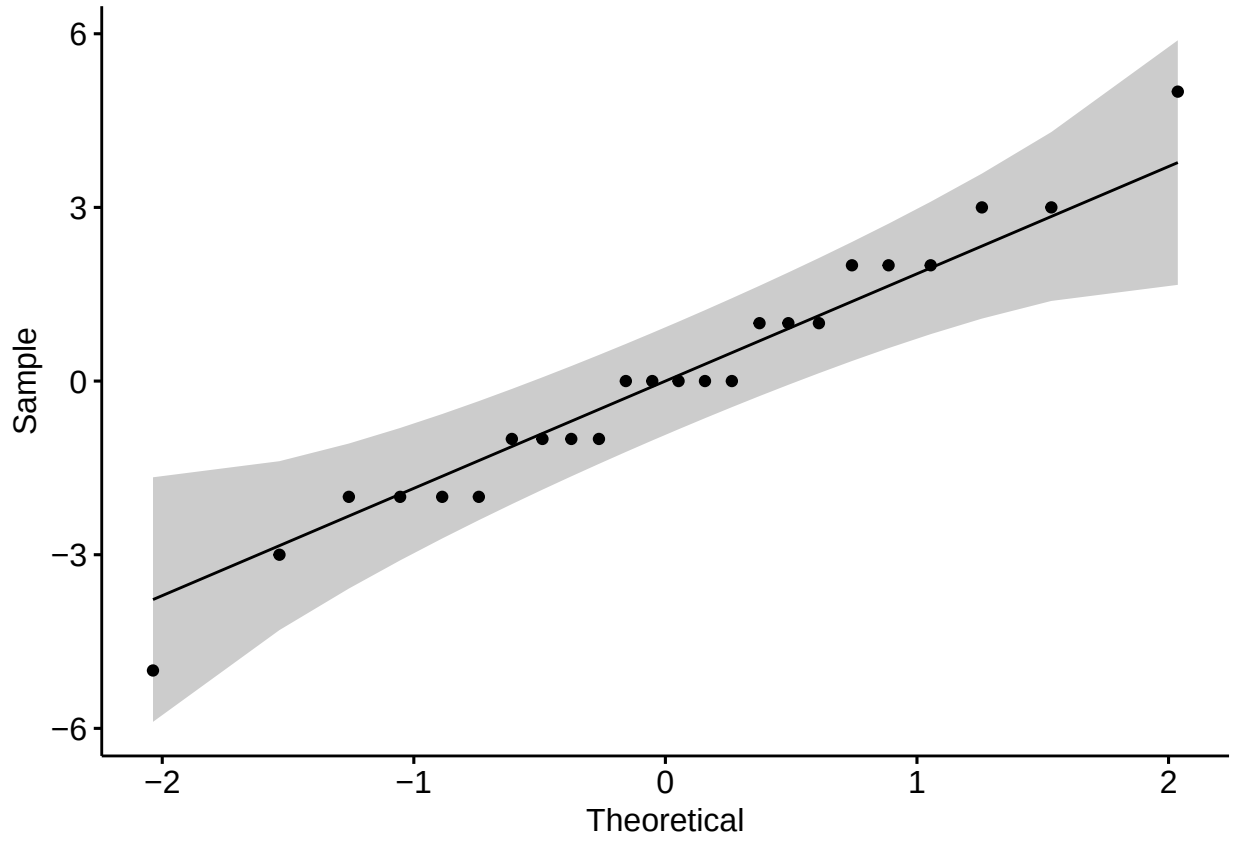
```
c_df <- coagulation
#outliers
c_df %>%  group_by(diet) %>% identify_outliers(coag) #encuentra 3 outliers, optaremos por pruebas robus
```

```
## # A tibble: 3 x 4
##   diet   coag is.outlier is.extreme
##   <fct> <dbl> <lgl>      <lgl>
## 1 B        71 TRUE       FALSE
## 2 C        66 TRUE       FALSE
## 3 C        71 TRUE       TRUE
```

```
c_df %>%  filter(between(coag,quantile(coag,0.1),quantile(coag,0.9))) %>% ggboxplot(x="diet", y="coag")
```



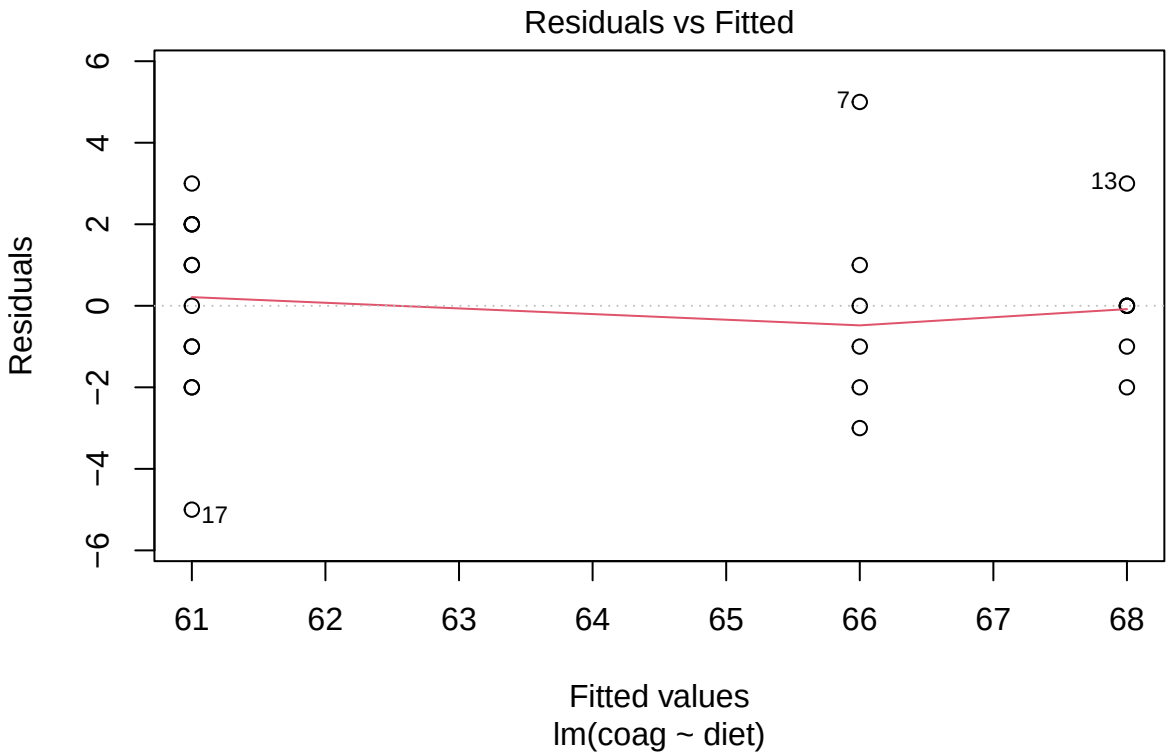
```r
#normalidad
fit <- lm(coag ~ diet, data=c_df)
ggqqplot(residuals(fit))
```




```r
shapiro_test(residuals(fit)) #normalidad ok
```

```
## # A tibble: 1 x 3
##   variable       statistic p.value
##   <chr>              <dbl>   <dbl>
## 1 residuals(fit)     0.978   0.863
```

```r
#homogeneidad de varianza
plot(fit,1) #los residuos de cada grupo parecen homogeneos
```




```
c_df %>% levene_test(coag ~ diet)#no es significativamente diferente por grupos
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     3    20     0.649 0.593
```

Como hemos encontrado outliers pero los datos cumplen los supuestos de normalidad e igualdad de varianzas podemos hacer una anova robusta de 1 vía.

```
#anova test
#prueba
t1way(coag ~ diet, data=c_df )# sale significativo
```

```
## Call:
## t1way(formula = coag ~ diet, data = c_df)
##
## Test statistic: F = 24.4582
## Degrees of freedom 1: 3
## Degrees of freedom 2: 6.81
## p-value: 5e-04
##
## Explanatory measure of effect size: 0.85
## Bootstrap CI: [0.7; 1.1]
```

```
#para saber entre los grupos que hay diferencia
lincon(coag ~ diet, data=c_df) #encuentra diferencias
```

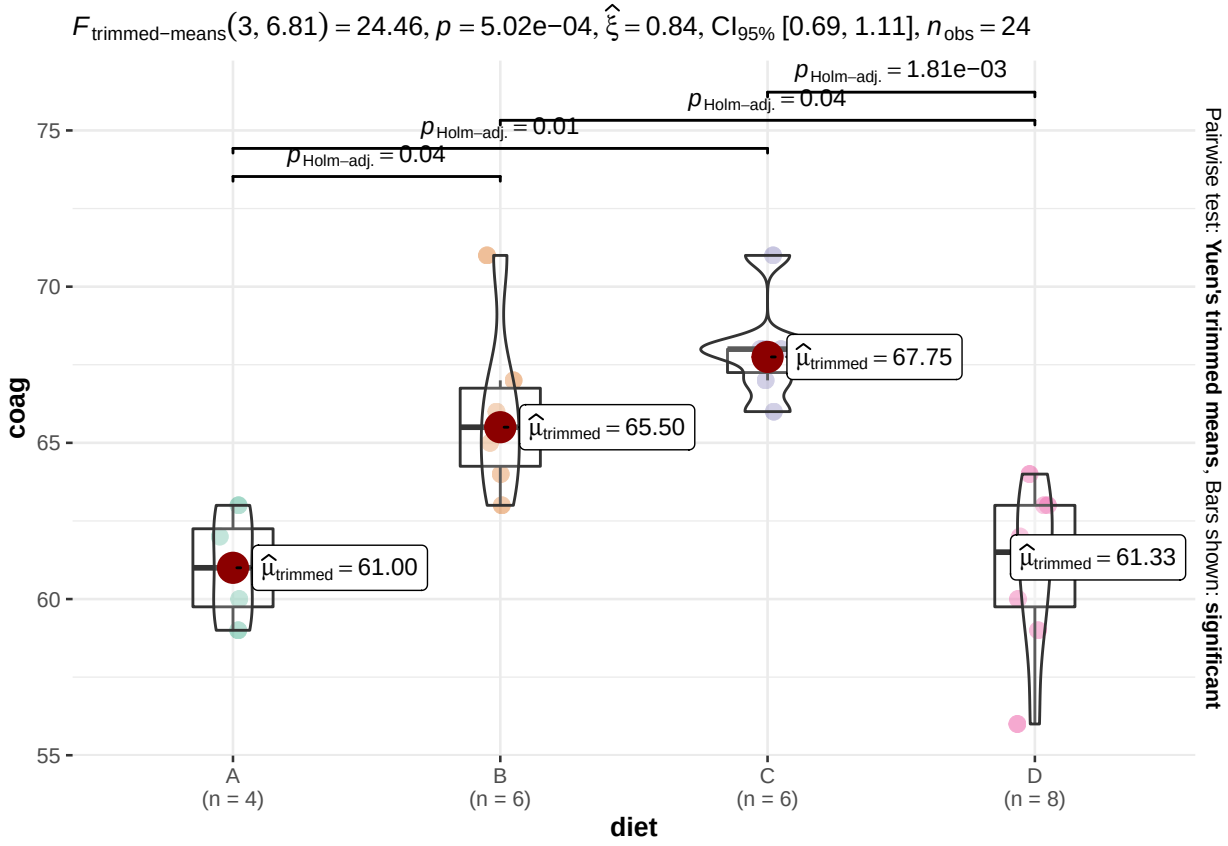
```
## Call:
## lincon(formula = coag ~ diet, data = c_df)
##
##           psihat  ci.lower ci.upper p.value
## A vs. B -4.50000  -9.16673  0.16673 0.03714
## A vs. C -6.75000 -11.13858 -2.36142 0.01380
## A vs. D -0.33333  -4.64436  3.97769 0.79642
## B vs. C -2.25000  -6.51373  2.01373 0.15987
## B vs. D  4.16667  -0.06698  8.40032 0.03714
## C vs. D  6.41667   3.15654  9.67679 0.00181
```

```
#resultado
```

Esto mismo podríamos haberlo obteido directamente como veremos a contiuación con el plot de ggbetweenstats.

Como estamos comparando muestras independientes, utlizamos el gráfico ggbetweenstats para visualizar los datos. Selecciona el tipo de prueba que corresponde según el análisis de supuestos e indícalo en el argumento:

```
ggbetweenstats(data = coagulation,
               x = diet,
               y = coag,
               type = "r",    #INDICA
               tr=0.2,
               var.equal = TRUE)   #INDICA
```



Plantea las hipótesis, interpreta los resultados y el gráfico. **¿Existen diferencias significativas en los tiempos de coagulación según la dieta suministrada?, ¿qué dieta se relaciona con menores**

tiempos de coagulación?.

Respuesta: La prueba de anova robusta nos indica que podemos rechazar la hipótesis nula (p=5E-4) y que por tanto exise un efecto de la dieta sobre el tiempo de coagulación. El test post-hoc nos muestra que existen diferencias significativas entre todas las dietas menos entre la B y la C, y la A y la D. El tamaño del efecto además es grande (0.84). Es decor, las dietas A y D tienen tiempos de coagulación menores que las dietas B y C.

Ejercicio 2.

Utiliza los datos selfesteem del paquete datarium para evaluar cómo varía la puntuación de autoestima de 10 personas en tres momentos durante una dieta específica para determinar si su autoestima mejoró. Indica si hay diferencias en el nivel de autoestima a lo largo del tiempo y, en tal caso, cómo son estas diferencias.

```
library(datarium)
data(selfesteem)
head(selfesteem)
```

```
## # A tibble: 6 x 4
##      id    t1    t2    t3
##   <int> <dbl> <dbl> <dbl>
## 1     1  4.01  5.18  7.11
## 2     2  2.56  6.91  6.31
## 3     3  3.24  4.44  9.78
## 4     4  3.42  4.71  8.35
## 5     5  2.87  3.91  6.46
## 6     6  2.05  5.34  6.65
```

```
summary(selfesteem)
```

```
##        id              t1              t2              t3
##  Min.   : 1.00   Min.   :2.046   Min.   :3.908   Min.   :6.308
##  1st Qu.: 3.25   1st Qu.:2.914   1st Qu.:4.411   1st Qu.:6.700
##  Median : 5.50   Median :3.212   Median :4.601   Median :7.463
##  Mean   : 5.50   Mean   :3.140   Mean   :4.934   Mean   :7.636
##  3rd Qu.: 7.75   3rd Qu.:3.486   3rd Qu.:5.301   3rd Qu.:8.440
##  Max.   :10.00   Max.   :4.005   Max.   :6.913   Max.   :9.778
```

Tenemos una respuesta cuantitativa (puntuación de autoestima) y un predictor cualitativo con 3 niveles (el tiempo, muestras relacionadas). Piensa entonces qué prueba de hipótesis tiene sentido realizar. Estudia los supuestos paramétricos y selecciona la versión adecuada de la prueba.

Respueta: En este caso tendremos que hacer una prueba anova para variables relacionadas (medimos la autoestima de un mismo sujeto tres veces). Veamos si nuestros datos cumple los supuestos necesarios para hacer una prueba paramétrica, o si por el contrario tendremos que optar por otra.

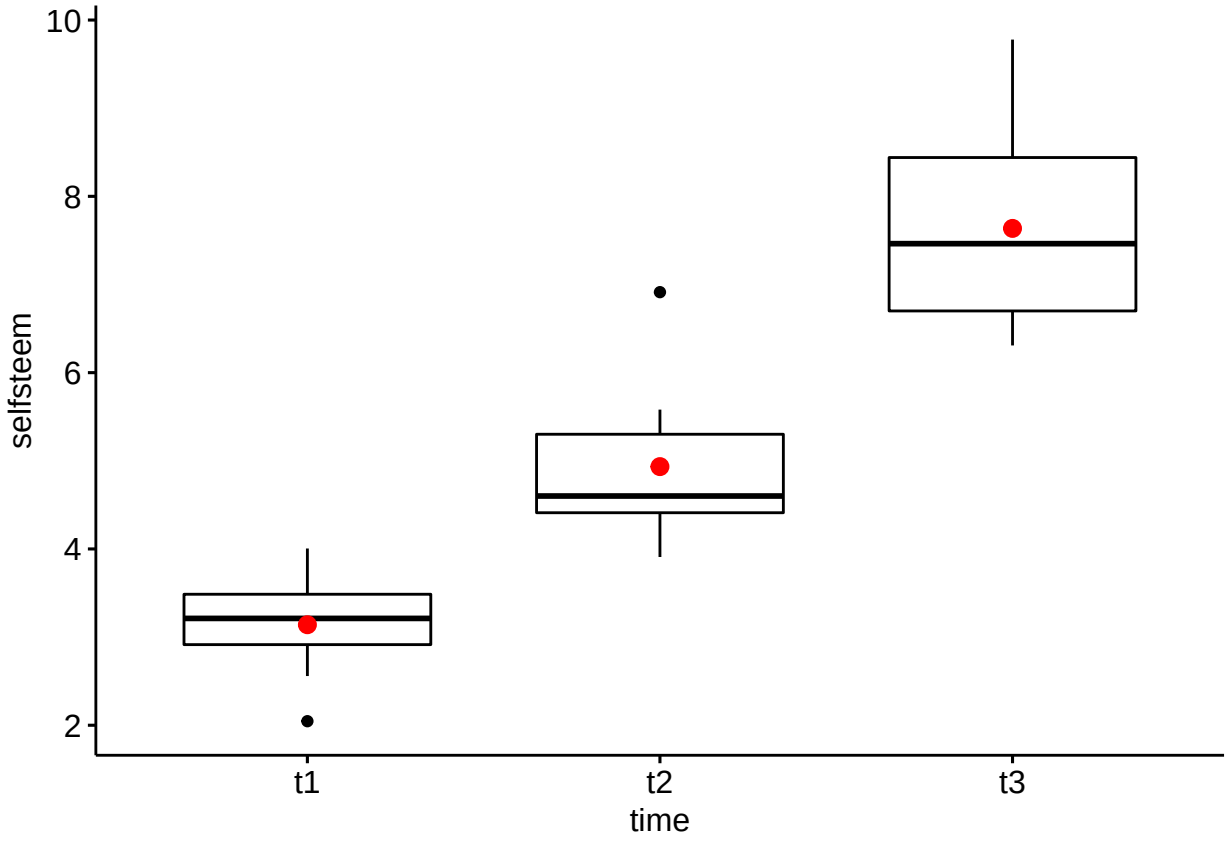
```
s_df <- selfesteem
#primero tenemos que pasarlo a estructura tidy
s_long <- s_df %>% pivot_longer(c(t1,t2,t3), names_to = "time", values_to= "selfsteem")

#explore and outliers detection
s_df %>% dplyr::select(t1,t2,t3) %>% get_summary_stats(type="mean_sd")
```

```
## # A tibble: 3 x 4
##   variable     n  mean    sd
##   <chr>    <dbl> <dbl> <dbl>
## 1 t1          10  3.14 0.552
## 2 t2          10  4.93 0.863
```

```
## 3 t3          10  7.64 1.14
```

```
ggboxplot(x="time", y="selfsteem", data=s_long, add=c("mean"), add.params = list(color="red")) # vemos
```



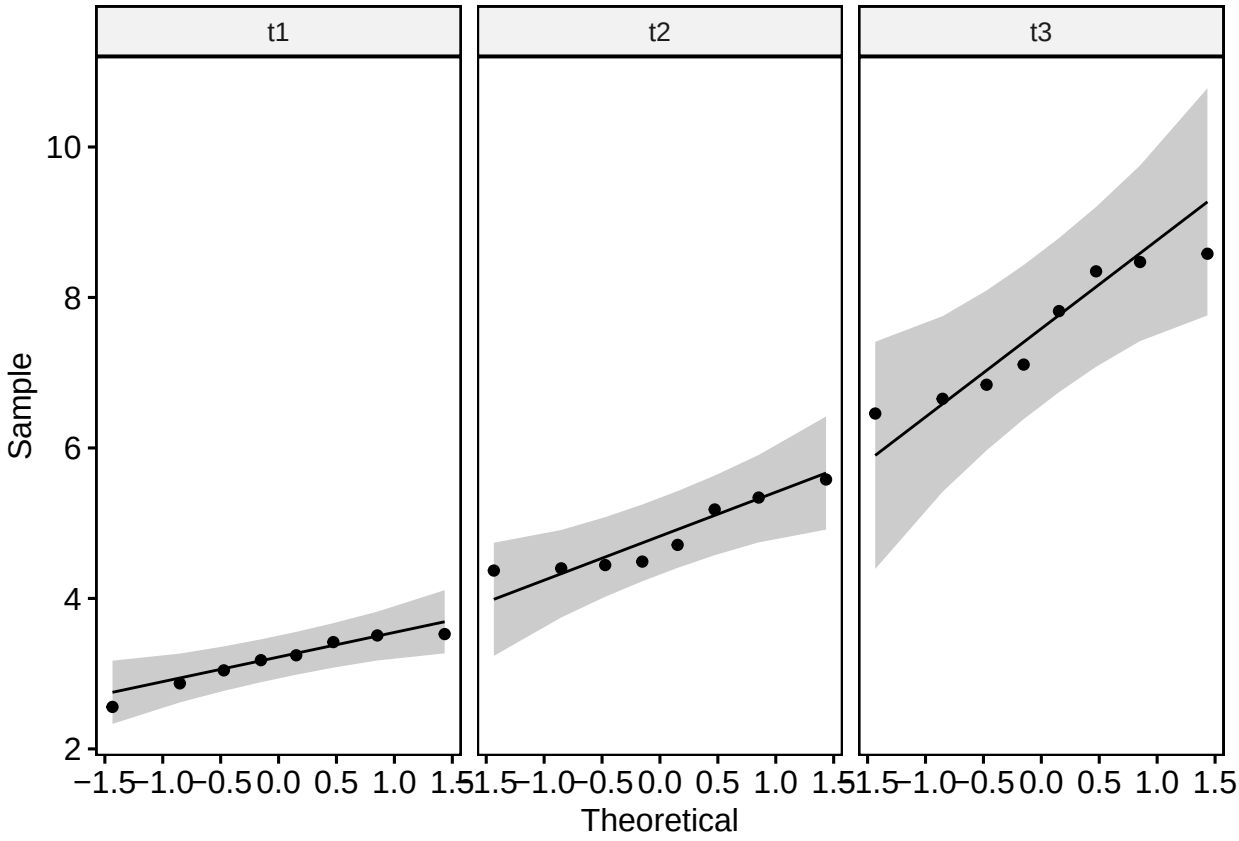

```
#outliers
s_long %>%  group_by(time) %>%  identify_outliers(selfsteem) #hay 2, auqnue no sean extremos optaremos
```

```
## # A tibble: 2 x 5
##   time     id selfsteem is.outlier is.extreme
##   <chr> <int>     <dbl> <lgl>      <lgl>
## 1 t1        6      2.05 TRUE       FALSE
## 2 t2        2      6.91 TRUE       FALSE
```

```
#esto significa que tenemos que chequar si cumplen normalidad
# supuesto de normalidad
s_long %>% group_by(time) %>% filter(between(selfsteem,quantile(selfsteem,.1),quantile(selfsteem,.9)))
```

```
## # A tibble: 3 x 4
##   time  variable  statistic      p
##   <chr> <chr>         <dbl>  <dbl>
## 1 t1    selfsteem     0.930 0.520
## 2 t2    selfsteem     0.847 0.0878
## 3 t3    selfsteem     0.881 0.193
```

```
s_long %>% group_by(time) %>% filter(between(selfsteem,quantile(selfsteem,.1),quantile(selfsteem,.9)))
```



Puesto que encontramos outliers pero los datos cumplen los supuestos de normalidad optaremos por una prueba paramétrica robusta:

```
#el test rmanova del WRS2
rmanova(y=s_long$selfsteem, groups = s_long$time,blocks = s_long$id) #sale  significativo
```

```
## Call:
## rmanova(y = s_long$selfsteem, groups = s_long$time, blocks = s_long$id)
##
## Test statistic: F = 42.1534
## Degrees of freedom 1: 1.16
## Degrees of freedom 2: 5.78
## p-value: 0.00063
```

```
#prubas post hoc
rmmcp(y=s_long$selfsteem, groups = s_long$time ,blocks = s_long$id) #todas significativas
```

```
## Call:
## rmmcp(y = s_long$selfsteem, groups = s_long$time, blocks = s_long$id)
##
##             psihat ci.lower ci.upper p.value p.crit  sig
## t1 vs. t2 -1.39328 -2.14470 -0.64186 0.00124 0.0250 TRUE
## t1 vs. t3 -4.41229 -5.61008 -3.21450 0.00005 0.0169 TRUE
## t2 vs. t3 -2.82372 -5.09852 -0.54892 0.00711 0.0500 TRUE
```

Vemos que aparecen diferencias significativas entre los tres tiempos. Esto mismo podríamos haberlo hecho directament con el gráfico de ggwithinstats.

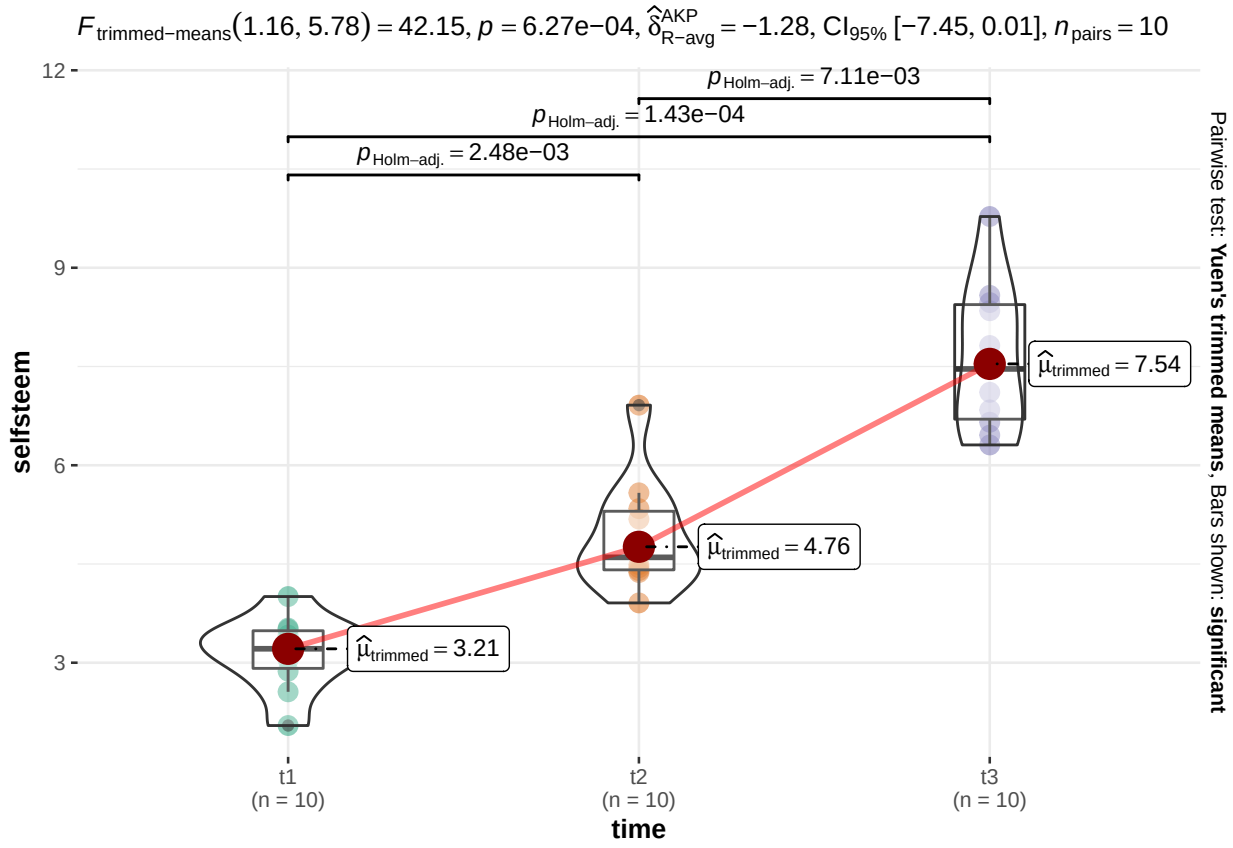
Como estamos comparando muestras relacionadas utlizamos el gráfico ggwithinstats para visualizar los datos.

Recuerda que para utilizar la función ggwithinstats necesitamos ingresar los datos en formato largo, por lo que previamente debes transformarlos. Selecciona el tipo de prueba que corresponde según el análisis de supuestos e indícalo en el argumento:

```
#self_largo <- selfesteem %>%
#    pivot_longer(t1:t3, names_to = "variable", values_to = "valor") %>%
#  mutate(variable = as.factor(variable))

#head(self_largo)
#summary(self_largo)

ggwithinstats(data = s_long,
              x = time,
              y = selfsteem,
              type = "r",    #INDICA
              var.equal = TRUE) #INDICA
```



Plantea las hipótesis, interpreta los resultados y el gráfico. **¿Existen diferencias significativas en los niveles de autoestima en el tiempo?, ¿Qué tiempo se relaciona con mayores valores de autoestima?**.

Respouesta: La hipótesis nula será que no existen diferencias significativas de la autoestima en los tres tiempos ($H_0 : \mu(t1) = \mu(t2) = \mu(t3)$)*mientrasquelahiptesisalternativaserquealgunadeestasmediasdeautoestimaesdiferente.El*
$6E{-}4$)*porloquepodemosdescartarlahiptesisnula.Elgrficomuestraqueexisteuncrecimientoestadisticamentesignificativode*
Aunque en la clase no se ha especificado como interpretar el tamaño del efecto interpretando el gráfico entiendo que debe ser moderado (solamente se han enseñado muestras sin diferencias significativas y no parece que se mencione esto).