Tema 3. DoE. Ejercicios obligatorios

Dra. Rosana Ferrero

Máxima Formación

Contents


```
library(tidyverse)
library(ggstatsplot)
library(rstatix)
```

En el **TEMA 3** del curso DOE has aprendido a realizar **pruebas de comparación para proporciones** (la prueba de independencia Chi-cuadrado, la prueba exacta de Fisher, la prueba de McNemar y la prueba Q de Cohran) y **pruebas de comparación para puntuaciones (medias)** (como la prueba t para la media de muestras independientes o relacionadas, y sus versiones no paramétricas y robustas).

Ejercicio 1

¿Reducir los servicios o aumentar los impuestos? En estos días, ya sea a nivel local, estatal o nacional, el gobierno a menudo enfrenta el problema de no tener suficiente dinero para pagar los diversos servicios que brinda. Una forma de abordar este problema es aumentar los impuestos. Otra forma es reducir los servicios. ¿Cual preferirías? Cuando la Encuesta de Florida preguntó recientemente a una muestra aleatoria de 1200 floridanos, el 52% (624 de los 1200) dijo que aumentaría los impuestos y el 48% dijo que reduciría los servicios. Determina si quienes están a favor de aumentar los impuestos en lugar de reducir los servicios son mayoría o minoría de la población.

Los datos son los siguientes:

```
datos <- data.frame(name=c("impuestos","servicios"), count=c(624, 1200-624))
datos
```

```
##        name count
## 1 impuestos   624
## 2 servicios   576
```

Piensa que aquí evalúas la opinión de los ciudadanos sobre qué prefieren ¿reducir los servicios o aumentar los impuestos? Tienes una única muestra de 1200 personas donde 624 prefieren aumentar los impuestos y el resto prefieren reducir los servicios. Para evaluar cuál es la opción preferida puedes analizar si la proporción de los que prefieren aumentar los impuestos es significativamente mayor al 50%.

Entonces tienes 1 muestra y con una respuesta categórica (proporciones), piensa qué prueba de hipótesis podrían utilizar.

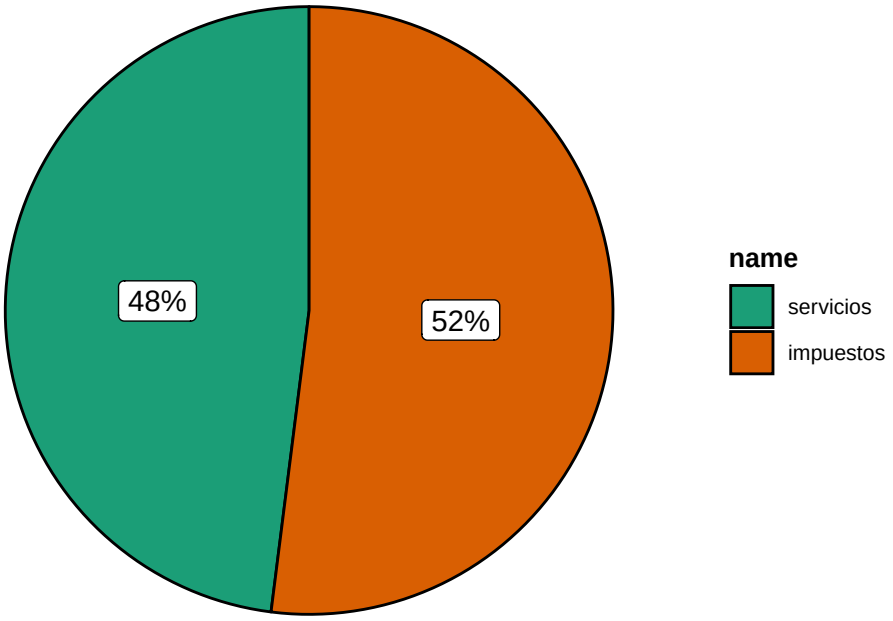
Respuesta: Prueba igualdad de proporciones Chi-cuadrado.

Aquí te enseño el gráfico para 1 variable categórica ggpiestats(), con la prueba de hipótesis correspondiente.

Fíjate que cuando tenemos los datos como data.fame, con una columna con el conteo, debemos ingresar la frecuencia en el argumento counts:

```
ggpiestats(datos, x=name, counts=count, bf.message=FALSE)
```

$$\chi^2_{gof}(1) = 1.92,\ p = 0.17,\ \widehat{C}_{Pearson} = 0.04,\ CI_{95\%}\ [0.00, 1.00],\ n_{obs} = 1{,}200$$



Plantea las hipótesis de la prueba e interpreta los resultados del gráfico. **¿Existen diferencias significativas entre el % de sujetos que desean reducir los servicios y aquellos que desean aumentar los impuestos? ¿Qué quiere la mayoría?**

Respuesta: Hipótesis nula H_0 : P(impuestos)=P(servicios)=50% (ambas opiniones tienen una proporción no significativament diferente de 50%). Hipótesis alternativa H_1 : P(impuestos)>50% (La proporción de personas favorables a pagar más impuestos es significativaemnte superior al 50%). La prueba de bondad de ajuste para la proporción de una muestra (que usa el chi-cuadrado) da un p-value de 0.17, y un tamaño de efecto de 0.04 por lo que podemos concluir que las diferencias no son significativamente diferentes del 50%, y por tanto no hay diferencias significaivas entre ambas preferencias. La mayoria encuestada aboga por subir los impuestos, pero este resultado no es extrapolable a la población general, al no haber una diferencia significativa en la muestra.

Ejercicio 2

Se quiere evaluar un estudio de gemelos del mismo sexo donde un gemelo había tenido una condena penal. Se recopiló la siguiente información: si el hermano también había tenido una condena penal y si los gemelos eran gemelos monocigóticos (idénticos) o dicigóticos (no idénticos). Los estudios de gemelos como este se han utilizado a menudo para investigar los efectos de la "naturaleza versus crianza".

La tabla de datos observados es la siguiente:

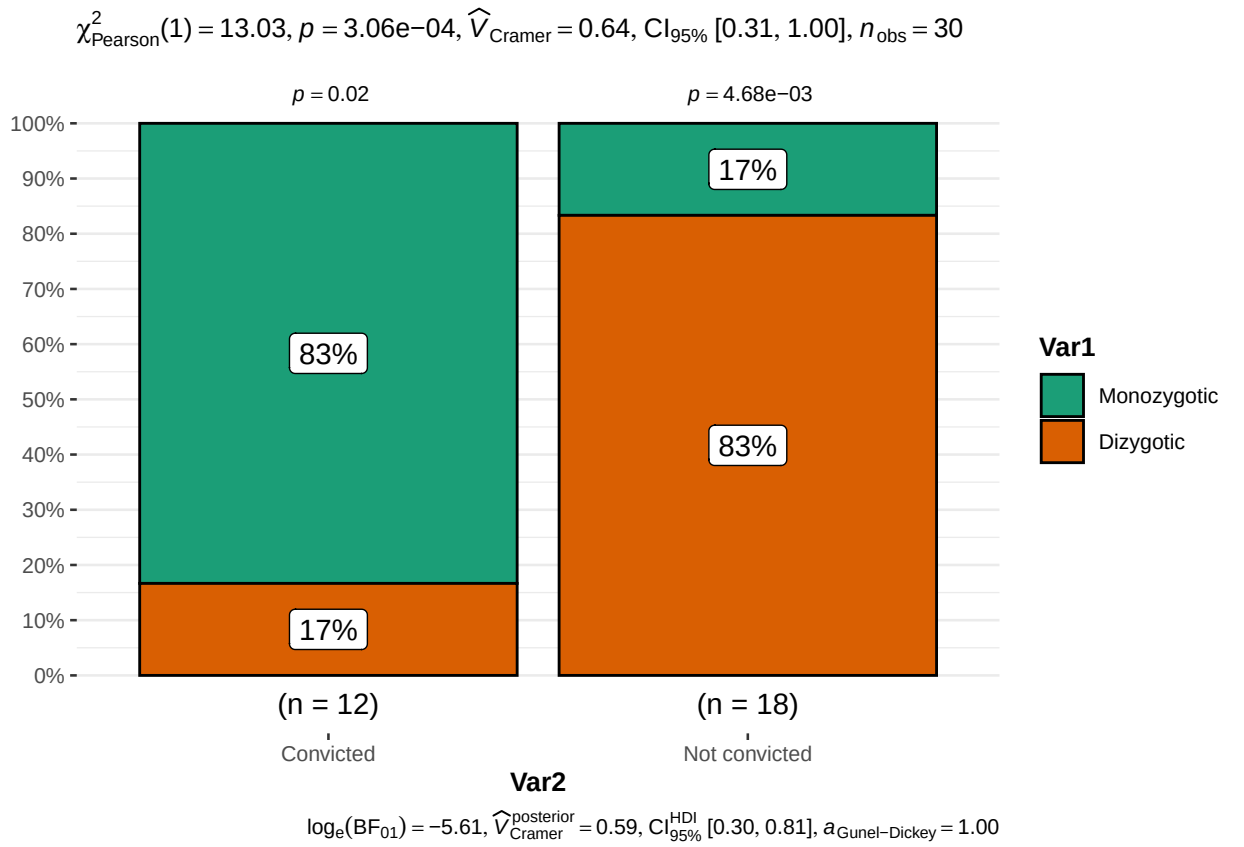
```
Convictions <-matrix(c(2, 10, 15, 3),
                     nrow = 2,
                     dimnames = list(c("Dizygotic", "Monozygotic"),
                                     c("Convicted", "Not convicted")))
```

Tenemos 2 variables categóricas (tipo de gemelo y delincuencia) con lo cual podemos utilizar el gráfico ggbarstats para visualizar los datos.

```
#Para realizar el gráfico necesitamos los datos en formato dataframe
as.data.frame(as.table(Convictions))
```

```
##          Var1          Var2 Freq
## 1   Dizygotic     Convicted    2
## 2 Monozygotic     Convicted   10
## 3   Dizygotic Not convicted   15
## 4 Monozygotic Not convicted    3
```

```
ggbarstats( data = as.data.frame(as.table(Convictions)),
            x = Var1,
            y = Var2,
            counts = Freq)
```



Queremos evaluar si existe o no una relación entre el tipo de gemelo y la delincuencia pero tenemos una baja frecuencia de observaciones, por lo que tenemos que utilizar la prueba exacta de Fisher. Pero fíjate que no tienes la opción de seleccionar la prueba de Fisher en el gráfico ggbarstats.

Así que vamos a realizar la prueba de Fisher con la función fisher_test. Como en este caso queremos

contrastar si la proporción de condenados es menor para los gemelos dicigóticos que para los gemelos monocigóticos, planteamos las siguientes hipótesis:

$H_0 : p_d \geq p_m$ (proporciones similares) $H_1 : p_d < p_m$ (la proporción de condenados es menor para los gemelos dicigóticos que para los monocigóticos, es decir, hay una influencia genética en la delincuencia)

NOTA: **R considera la primer categoría según la primera fila de datos**, en las hipótesis debemos escribir los gemelos dicigóticos en primer lugar.

Realizamos una prueba unilateral con **alternative=less** para contastar si la proporción de condenados es menor para los gemelos dicigóticos que para los monocigóticos.

```
# para realizar la prueba necesitamos los datos en formato tabla
fisher_test(Convictions, alternative = "less")
```

```
## # A tibble: 1 x 3
##       n        p p.signif
## * <dbl>    <dbl> <chr>
## 1    30 0.000465 ***
```

Según este estudio, ¿La delincuencia tiene un componente genético? ¿el tipo de gemelo y la delincuencia está relacionado? en caso afirmativo ¿cómo?

Respuesta: El test de Fisher da un p-value de 4.6E-4, es decir, muy por debajo del nivel e significación de 0.05, por lo que podemos dar por válida la hipótesis alternativa. Atendiendo a los datos de este experimento vemos una relación entre la conducta criminal y el tipo de gemelo, pues la proporción de gemelos dizigoticos no convictos es significativamente menor que la de monocigoticos. Así, se podría entender que este experimento encuentra un componente genético en la delincuencia (auqnue el experimento debería ser replicado múltiples veces antes de generalizar el resultado).

Ejercicio 3

Vamos a evaluar si existe una relación entre el nivel educativo y el número abortos inducidos. La base de datos infert corresponde a un estudio de caso-control pareado donde la variable "Education" está formada por 3 categorías (0 = 0-5 años, 1 = 6-11 años, 2 = 12+ años); y la variable "number of prior induced abortions" también (0 = 0, 1 = 1, 2 = 2 o más abortos inducidos).

Para acceder a los datos escribe en la consola de R:

```
data(infert, package = "datasets")
head(infert)
```

```
##   education age parity induced case spontaneous stratum pooled.stratum
## 1    0-5yrs  26      6       1    1           2       1              3
## 2    0-5yrs  42      1       1    1           0       2              1
## 3    0-5yrs  39      6       2    1           0       3              4
## 4    0-5yrs  34      4       2    1           0       4              2
## 5   6-11yrs  35      3       1    1           1       5             32
## 6   6-11yrs  36      4       2    1           1       6             36
```

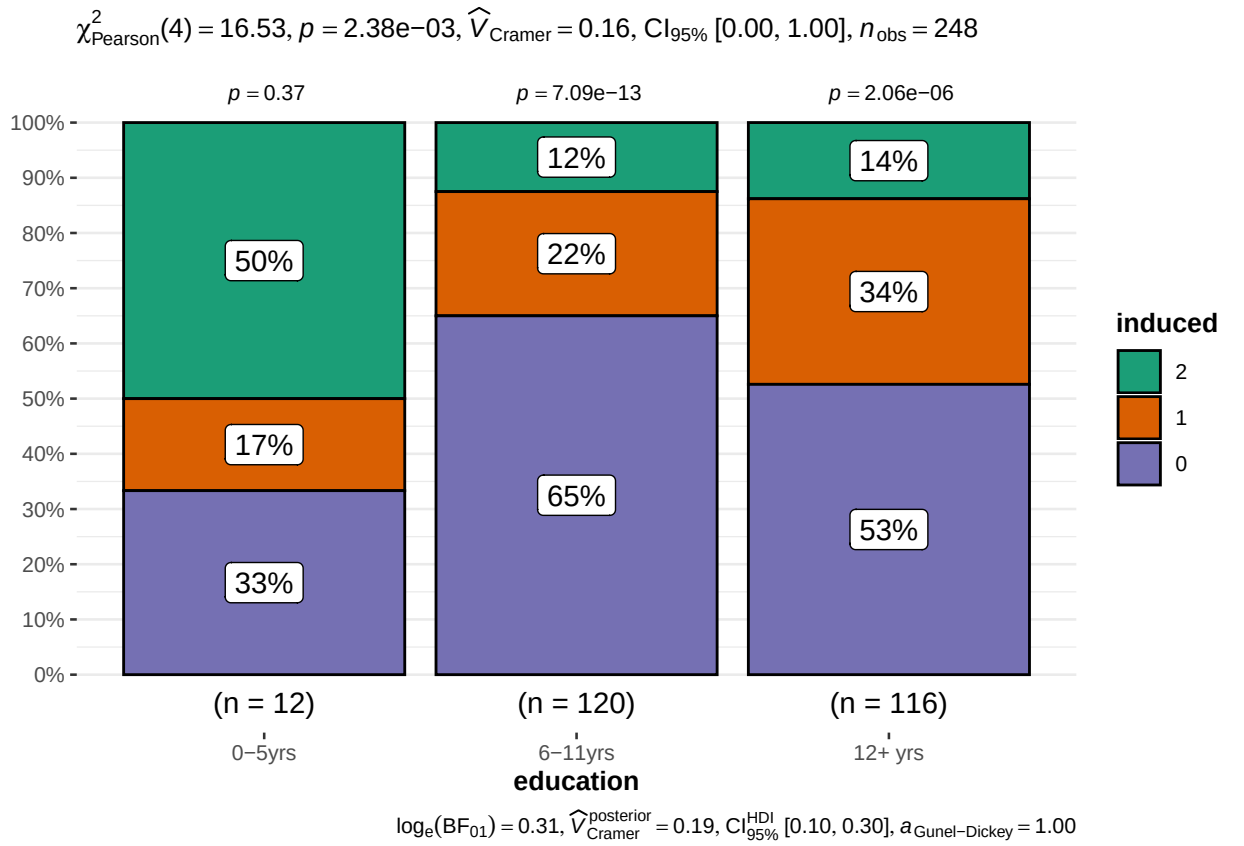
```
table(infert$education, infert$induced)
```

```
##
##            0  1  2
##   0-5yrs   4  2  6
##   6-11yrs 78 27 15
##   12+ yrs 61 39 16
```

Estamos trabajando con 2 variables cualitativas (con 3 muestras cada una) y queremos evaluar si son independientes o no (si existe o no relación entre ellas), y si es necesario luego evaluar cómo es la relación. Para

analizar dos variables cualitativas usamos la función ggbarstats():

```
ggbarstats(data = infert, x = induced, y = education)
```



Plantea las hipótesis, interpreta el gráfico y la prueba realizada.

En el gráfico se enseña, además de la prueba global de independencia Chi-cuadrado, pruebas en cada nivel educativo comparando el % de abortos inducidos (0, 1 o 2).

Plantea las hipótesis, interpreta los resultados del a prueba y el gráfico.

Respuesta: Queremos saber si existe una relación entre el nivel educativo de las mujeres y el número de abortos provocados. La hipotésis nula es que los porcentajes no van a ser diferentes ($H_0 = \mu(0-5) = \mu(6-11) = \mu(12+)$, mientras que la alternativa H_1 es que el número de abortos provocados si depende del nivel educativo de la mujer. Mirando el gráfico podemos ver que usando la Chi-cuadrado podemos rechazar la hipótesis nula (p=0.002) y que existe una relación entre el nivel educativo y número de abortos provocados, con las mujeres con menor nivel educativo mostrando una mayor proporcion de abortos. El tamaño del efecto, sin embargo, es pequeño, con un valor de 0.16. Sin emabrgo, dado que 2 frecuencias en el nivel educativo 0 están por debajo de 5, quizás sería más conveniente usar la prueba exacta de Fisher en lugar de la aproximación de la Chi-cuadrado.

Ejercicio 4

Utiliza los datos "Arthritis", del paquete "vcd", sobre un ensayo clínico de doble ciego que investiga un nuevo tratamiento para la artritis reumatoide.Tenemos información de 84 observaciones de 5 variables: la identificación del paciente (ID), el tratamiento (Treatment: Placebo, Treated), el sexo (Sex: Female, Male), la edad (Age) y la mejoría (Improved: None, Some, Marked).

```
library(vcd)
data(Arthritis)
head(Arthritis)
```

```
##   ID Treatment  Sex Age Improved
## 1 57   Treated Male  27     Some
## 2 46   Treated Male  29     None
## 3 77   Treated Male  30     None
## 4 17   Treated Male  32   Marked
## 5 36   Treated Male  46   Marked
## 6 23   Treated Male  58   Marked
```

```
table(Arthritis[which(Arthritis$Treatment=="Treated"),5])
```

```
##
##   None   Some Marked
##     13      7     21
```

Para el grupo tratamiento, queremos comparar las edades de los pacientes que no mostraron mejoría con los que sí tuvieron una marcada mejoría. Entonces, primero selecciona los pacientes tratados y solo aquellos sin ninguna mejoría o marcada mejoría.

```
datos <- Arthritis %>%
    filter(Treatment == "Treated" & Improved != "Some") %>%
    mutate(Improved = droplevels(Improved)) #borra la categoría fantasma

head(datos)
```

```
##   ID Treatment  Sex Age Improved
## 2 46   Treated Male  29     None
## 3 77   Treated Male  30     None
## 4 17   Treated Male  32   Marked
## 5 36   Treated Male  46   Marked
## 6 23   Treated Male  58   Marked
## 7 75   Treated Male  59     None
```

```
summary(datos)
```

```
##        ID          Treatment      Sex          Age          Improved
##  Min.   : 2.00   Placebo: 0   Female:22   Min.   :23.00   None  :13
##  1st Qu.:27.25   Treated:34   Male  :12   1st Qu.:48.00   Marked:21
##  Median :43.00                            Median :58.00
##  Mean   :46.00                            Mean   :54.15
##  3rd Qu.:64.50                            3rd Qu.:62.75
##  Max.   :84.00                            Max.   :70.00
```

Ahora piensa que estamos comparando las edades (variable cuantitativa) entre 2 muestras independientes (sin ninguna mejoría o marcada mejoría). ¿Qué prueba de hipótesis deberías utilizar? ¿Qué supuestos debería cumplir? Evalúa los supuestos paramétricos.

Graficamos los datos con el paquete ggstatsplot para 2 muestras independientes. Selecciona el tipo de prueba que corresponde según el análisis de supuestos e indícalo en el argumento:

Respuesta: Tenemos dos categorias (no mejora, mejora claramente) y queremos comparar la edad media de los pacientes en cada una de las categorias, por lo que la primera opción sería un t-test, en el supuesto de que los datos cumplan que no hay outliers, normalidad e igualdad de varianza en ambas muestras. Vamos a ver si a) hay outliers, b) ambas varianzas son similares y c) se cumple el supuesto de normalidad.

```
library(DescTools)
library(WRS2)

#outliers
datos %>% group_by(Improved) %>% identify_outliers(Age)
```

```
## # A tibble: 2 x 7
##   Improved    ID Treatment Sex      Age is.outlier is.extreme
##   <ord>    <int> <fct>     <fct>  <int> <lgl>      <lgl>
## 1 Marked      17 Treated   Male      32 TRUE       TRUE
## 2 Marked      72 Treated   Female    41 TRUE       FALSE
```

```
#puesto que hay outlayers haremos una prueba robusta, para lo que vamos a filtar el 20% extremo
#levene test
datos %>% filter(between(Age,quantile(Age,0.1),quantile(Age,0.9))) %>% levene_test(Age ~ Improved) #no
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     1    25      1.12 0.299
```

```
#shapiro
datos %>% group_by(Improved) %>%  filter(between(Age,
                                                 quantile(Age, 0.1),
                                                 quantile(Age, 0.9)))  %>% shapiro_test(Age) # no signi
```

```
## # A tibble: 2 x 4
##   Improved variable statistic      p
##   <ord>    <chr>        <dbl>  <dbl>
## 1 None     Age          0.850 0.0740
## 2 Marked   Age          0.930 0.219
```

```
#yo diria que lo mejor es un test de Yuen, que es parametrica pero sin outliers
#se hace asi:
YuenTTest(Age ~ Improved, data=datos) #esta te recorta por defecto el 20% #DescTools# resultado no sign
```

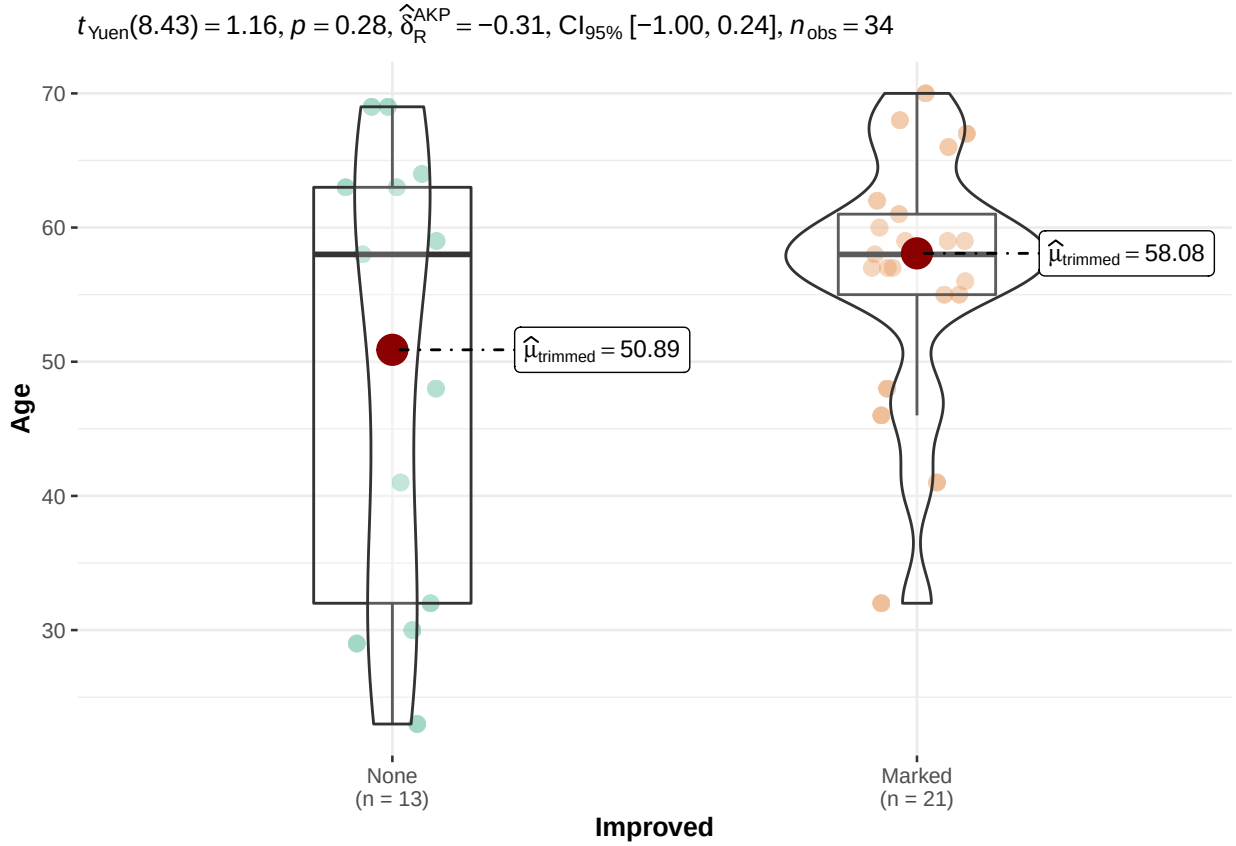
```
##
##  Yuen Two Sample t-test
##
## data:  Age by Improved
## t = -1.1907, df = 8.4527, trim = 0.2000, p-value = 0.2662
## alternative hypothesis: true difference in trimmed means is not equal to 0
## 95 percent confidence interval:
##  -20.980775   6.604706
## sample estimates:
##   trimmed mean in group None trimmed mean in group Marked
##                     50.88889                     58.07692
```

```
#ggbetweenstats(x=Improved,y=Age, data=datos, tr=0.2, type="r", bf.message = FALSE)
```

Al tener outliers y tener unas distribuciones normales (p-value entre 0.07 y 0.21 en el shapiro test) podemos hacer una prueba de Yuen, que es paramétrica pero no considera a los autliers.

```
ggbetweenstats(data = datos,
               x = Improved,
               y = Age,
               type = "r",   #INDICAR
```

```
                var.equal = TRUE)   #INDICAR
```



Plantea las hipótesis, interpreta los resultados y el gráfico.

Respuesta: Queremos saber si la edad de los pacientes afecta a su mejoría. La hipóteis nula es que no habrá efecto de la edad en los dos grupos, es decir que la edad media de los que han mejorado no va a ser significativamente diferente de los que si han mejorado (H_0=μ(None)=μ(marked)). Despué de hacer el test de Yuen vemos que la diferencia no es significativa (p=0.28), por lo que no podemos rechazar la hipótesis nula: no hemos detectado efecto de la edad en la mejoría.

Ejercicio 5

Utiliza los datos "immer", del paquete "MASS", sobre el rendimiento de la cebada en los años 1931 y 1932 en un mismo campo de recolección.

```
library(MASS)
head(immer)
```

```
##   Loc Var    Y1    Y2
## 1  UF   M  81.0  80.7
## 2  UF   S 105.4  82.3
## 3  UF   V 119.7  80.4
## 4  UF   T 109.7  87.2
## 5  UF   P  98.3  84.2
## 6   W   M 146.6 100.4
```

Evalúa mediante pruebas paramétricas, no paramétricas y robustas si han cambiado los valores medios del rendimiento de cebada. Interpreta y compara los resultados.

Estamos comparando el rendimiento (variable cuantitativa) entre 2 muestras relacionadas en el tiempo. ¿Qué prueba de hipótesis deberías utilizar? ¿Qué supuestos debería cumplir? Evalúa los supuestos.

Para 2 muestras relacionadas realizamos el gráfico con la función ggwithinstats del paquete ggstatsplot. Fíjate que debemos transformar los datos de formato ancho a largo para el gráfico. Selecciona el tipo de prueba que corresponde según el análisis de supuestos e indícalo en el argumento:

```
immer_largo <- immer %>%
    pivot_longer(Y1:Y2, names_to = "variable", values_to = "valor") %>%
  mutate(variable = as.factor(variable))

head(immer_largo)
```

```
## # A tibble: 6 x 4
##   Loc   Var   variable valor
##   <fct> <fct> <fct>    <dbl>
## 1 UF    M     Y1        81
## 2 UF    M     Y2        80.7
## 3 UF    S     Y1       105.
## 4 UF    S     Y2        82.3
## 5 UF    V     Y1       120.
## 6 UF    V     Y2        80.4
```

```
summary(immer_largo)
```

```
##  Loc     Var    variable     valor
##  C :10   M:12   Y1:30    Min.   : 49.90
##  D :10   P:12   Y2:30    1st Qu.: 80.62
##  GR:10   S:12            Median : 97.50
##  M :10   T:12            Mean   :101.09
##  UF:10   V:12            3rd Qu.:119.72
##  W :10                   Max.   :191.50
```

```
#los supuestos que debemos chequear son: outliers, similitud de varianzas, normalidad
# outliers
immer <- immer %>% mutate(differences=Y2-Y1)
immer %>% identify_outliers(differences) #no hay
```

```
## [1] Loc         Var         Y1          Y2          differences is.outlier
## [7] is.extreme
## <0 rows> (or 0-length row.names)
```

```
#levene: supuestamente no es necesario al darse por sentado q al ser del mismo sujeto son similares
immer_largo %>% levene_test(valor ~ variable)#efectivamente son similares p=0.43
```

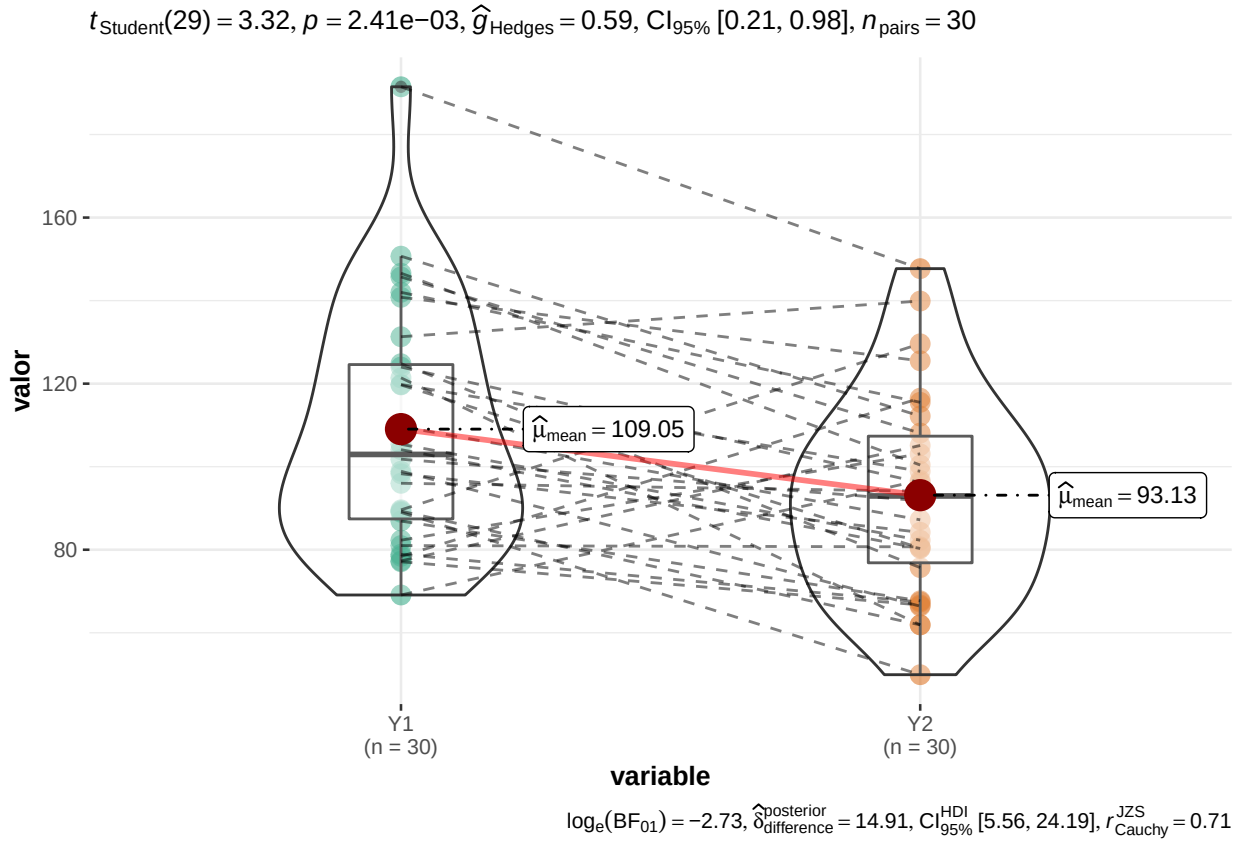
```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     1    58     0.624 0.433
```

```
#shaphiro
immer %>% shapiro_test(differences) #no son diferentes a la normalidad
```

```
## # A tibble: 1 x 3
##   variable    statistic      p
##   <chr>           <dbl>  <dbl>
## 1 differences     0.938 0.0796
```

```
#por lo tanto optaremos por un t-test ggwithinstats(x=test,y=score, data=hsb2_long, bf.message = FALSE)

ggwithinstats(data = immer_largo,
              x = variable,
              y = valor,
              type = "p")  #INDICA
```



Plantea las hipótesis, interpreta los resultados y el gráfico.

Respuesta: Queremos comprobar si hay un cambio significativo entre la media de la productividad entre la medida Y1 y la Y2. La hipótesis nula es que no encontraremos diferencias significativas entre las medias en los dos tiempos ($H_0 : \mu(Y_1) = \mu(Y_2)$), mientras que la hipótesis alternativa es que las medias de producción no van a ser iguales ($H_1 : \mu(Y_1)! = \mu(Y_2)$). El t-test arroja un p-value de 0.002, bastante menor a 0.05, por lo que existe una diferencia significativa. El tamaño del efecto es de 0.59, es decir moderado. Así que podemos concluir que ha habido una reducción significativa, de tamaño moderado, en la productividad de los campos.