Módulo 4. Ejercicios obligatorios

Juan Manuel Cabrera Rodríguez

2023-09-10

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.

Análisis de datos exploratorio (EDA):

En primer lugar se realiza el análisis de datos exploratorios.

1. Datos: los datos son aumentar los impuestos o reducir los servicios, son datos nominales.
2. Objetivos: comparar la proporción de la población que quiere aumentar los impuestos con la que quiere reducir los servicios. Al ser un datos nominal podemos tratarla como proporción.
3. Muestras: hay una sola muestra de 1200 floridanos para una variable nominal. Se opta por una prueba de proporción.

Pregunta de investigación:

- ¿la proporción de floridanos a favor de subir los impuestos es igual que la proporción de floridianos de reducir los servicios?

A continuación se definen las variables del enunciado:

```
#Población
n = 1200

#Proporción aumentar los impuestos
p1 = 0.52

#Numero de floridanos a favor de subir los impuestos
x1 = 624

#proporción reducir los servicios (o de no aumentar los impuestos)
p2 = 0.48

#Numero de floridanos a favor de reducir los servicios
x2 = n-624
```

Resolución aplicando direcatamente la función **prop.test**.

Aplicamos la prueba de proporción.

```
prop.test(x = x1,
          n = n,
          conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  x1 out of n, null probability 0.5
## X-squared = 1.8408, df = 1, p-value = 0.1749
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4912980 0.5485725
## sample estimates:
##    p
## 0.52
```

El p-valor es 0.1749, superior al 5% de significación, por lo que **no aceptamos la hipótesis nula de igualdad**. Uno de las dos proporciones será mayor que la otra.

Graficamos

En primar lugar se crea un dataframe con los datos para luego poder poder graficar los datos con un gráfico tipo pie".

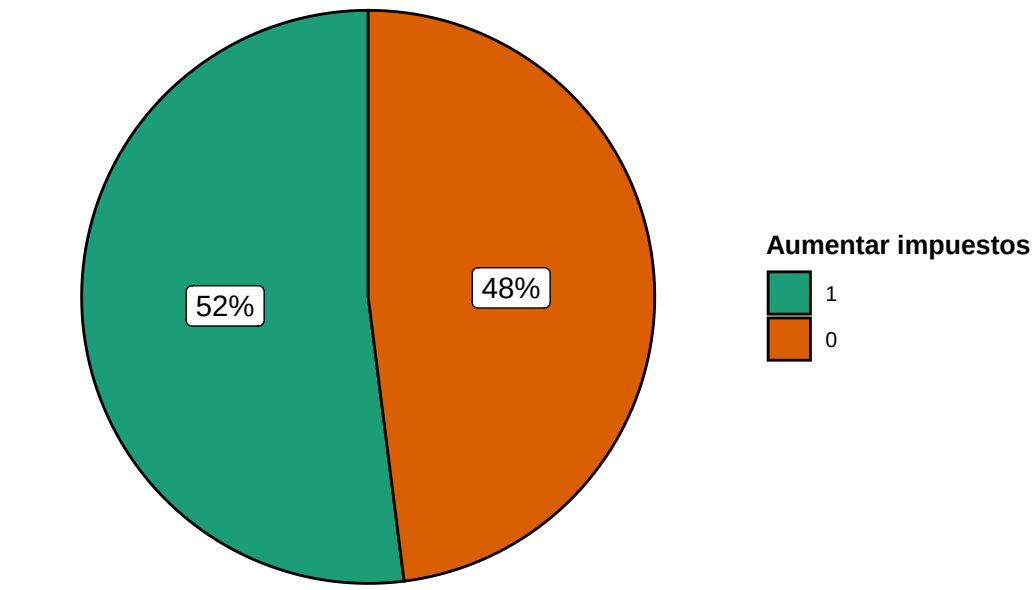
```
#Se genera el dataframe y se cambia el nombre de la columna
poblacion <- data.frame(c(rep(1,624), rep(0,576)))
colnames(poblacion) <-  "eleccion"
```

Se crea el gráfico.

```
library(ggstatsplot)
```

```
## You can cite this package as:
##      Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach.
##      Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167
```

```
ggpiestats(data = poblacion,
           x = eleccion,
           title = "Aumentar impuestos vs reducir servicios",
           legend.title = "Aumentar impuestos",
           bf.message = F)
```



Conclusión

El 52% de la población prefiere que aumenten los impuestos, mientras que el 48% de la población prefiere que se reduzcan los servicios o lo que es lo mismo, no aumenten los impuestos.

Resolución usando la tabla CrossTable

También se podría resorvel el ejercicio usando la función CrossTable.

```
library(gmodels)
CrossTable(poblacion$eleccion)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1200
##
##
##            |         0 |         1 |
##            |-----------|-----------|
##            |       576 |       624 |
```

```
##            |     0.480 |     0.520 |
##            |-----------|-----------|
##
##
##
##
```

Mediante la representación en tabla tenemos el mismo resultado dondel el 52% está a favor de aumentar los impuestos y el 48% en reducir los servicios.

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". Queremos contrastar si la proporción de condenados es mayor para los gemelos monocigóticos que para los dicigóticos.

La tabla de datos observados es la siguiente:

Gemelo	Condenado	No condenado
Dicigótico	2	15
Monocigotico	10	3

EDA:

1. Datos: para cada gemelo (monocigótico o dicigótico) se obtiene si este ha sido condenado o no.
2. Objetivo: comparar la proporción de gemelos condenados monocigótico y dicigótico.
3. Muestra: se quiere comparar 2 muestras independientes de dos variables nominales (proporción). Prueba de Fisher o Chi-cuadrado.

Pregunta de hipótesis:

- ¿La proporción de gemelos monocigóticos condenados esta relacionada con a proporción de gemelos dicigóticos condenados?

En primer lugar creamos la tabla de contingencia denominada *gemelos*.

```
gemelos <- matrix(c(2,10,15,3),
                  nrow = 2,
                  dimnames = list(c("dicigotico", "monocigotico"),
                                  c("condenado", "no.condenado")))
gemelos
```

```
##              condenado no.condenado
## dicigotico           2           15
## monocigotico        10            3
```

Comprobamos si debemos usar la **prueba de Fisher o Chi-squared**

```r
chisq.test(gemelos)$expected
```

```
##              condenado no.condenado
## dicigotico         6.8         10.2
## monocigotico       5.2          7.8
```

Como *NO* hay valores inferior a 5, optamos por la **prueba de Chi-squared**.

```r
chisq.test(gemelos, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  gemelos
## X-squared = 13.032, df = 1, p-value = 0.0003063
```

El valor **p-value** es menor al nivel del significación del 5%, por ello **rechazamos la hipótesis nula de igualdad de proporciones**, y por lo tanto, *no existe una relación significante entre los gemelos monocigóticos condenados y los gemelos dicigóticos condenados.*

También se podría resolver a través de la función CrossTable del paquete *gmodels*.

```r
library(gmodels)

CrossTable(gemelos,
           prop.r = F,
           prop.t = F,
           prop.chisq = F,
           chisq = T)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  30
##
##
##              |
##              |    condenado | no.condenado |    Row Total |
## -------------|--------------|--------------|--------------|
##   dicigotico |            2 |           15 |           17 |
##              |        0.167 |        0.833 |              |
## -------------|--------------|--------------|--------------|
## monocigotico |           10 |            3 |           13 |
##              |        0.833 |        0.167 |              |
## -------------|--------------|--------------|--------------|
## Column Total |           12 |           18 |           30 |
```

```
##              |        0.400 |        0.600 |              |
## -------------|--------------|--------------|--------------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  13.03167     d.f. =  1     p =  0.0003062666
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  10.45814     d.f. =  1     p =  0.001221099
##
##
```

Como era de esperar se obtiene el mismo resultado.

Pregunta de hipótesis:

- ¿La proporción de gemelos monocigóticos condenados es mayor que la proporción de gemelos dicigóticos condenados?

(PENDIENTE)

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 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 abo inducidos).

Se muestra la base de datos:

```r
library(datasets)
data <- infert

head(data)
```

```
##   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
```

EDA:

1. Datos: los datos son education e induced, ambas son variables nominales.

2. Objetivo: comparar si existe relación entre el nivel de educación y el número de abortos.
3. Muestra: existe una muestra de 248 personas, con dos variables nominales, el nivel educativo y el número de abortos. Debemos usar la prueba Fisher o Chi cuadrado.

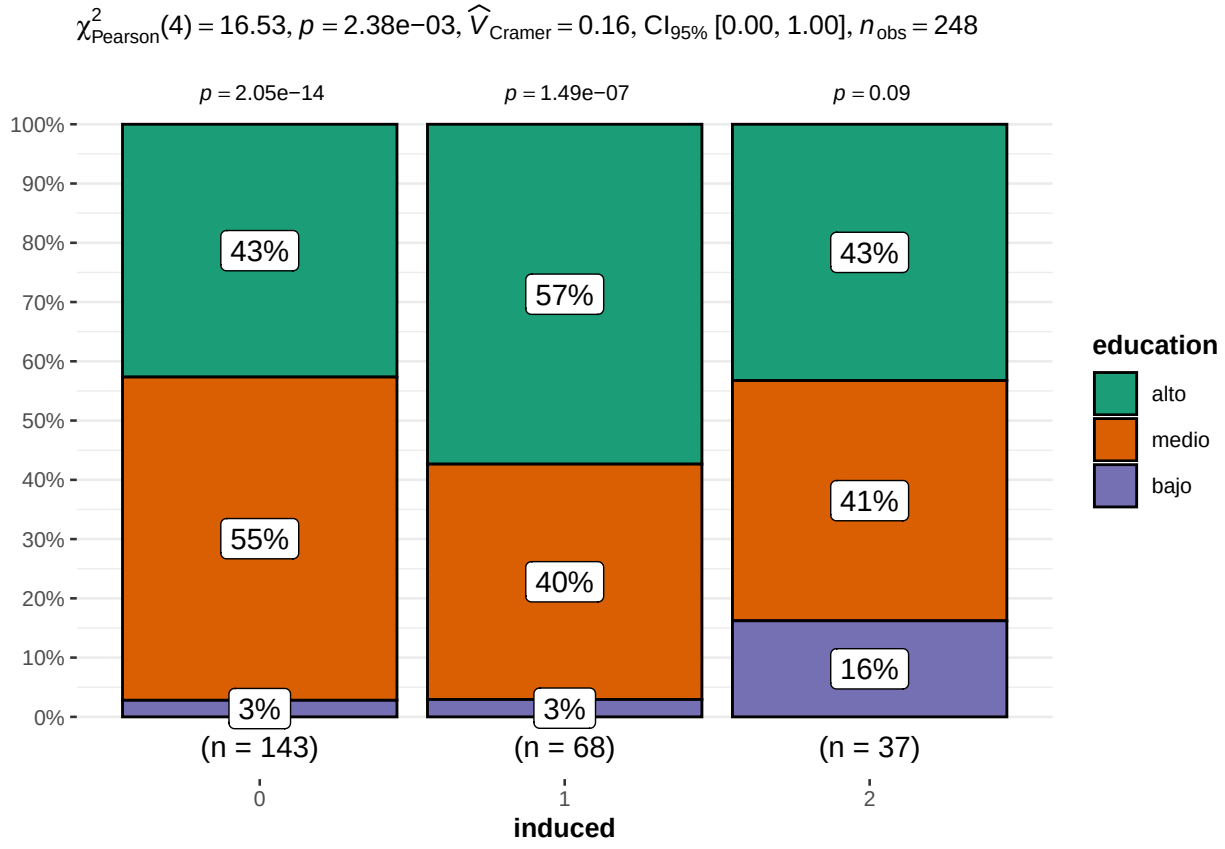
Para facilitar trabajar con estos datos, se modifica los niveles de la variable education, la cual es de tipo factor.

```
levels(infert$education) <- c("bajo", "medio", "alto")

head(infert)
```

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

Vamos a mostrar los datos en un gráfico de barras.

```
library(ggstatsplot)
ggbarstats(infert,
           education,
           induced,
           bf.message = F)
```



Se realiza la **prueba de independencia Chi-cuadrado**, donde vamos a usar unicamente 2 categorías de la variable *Induced*.

```
library(rstatix)
```

```
##
## Attaching package: 'rstatix'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
tab <- table(infert$education, infert$induced)

pairwise_prop_test(tab[ , c(1,3)],
                   p.adjust.method ="bonferroni")
```

```
## Warning in prop.test(x[c(i, j)], n[c(i, j)], ...): Chi-squared approximation
## may be incorrect
```

```
## Warning in prop.test(x[c(i, j)], n[c(i, j)], ...): Chi-squared approximation
## may be incorrect
```

```
## # A tibble: 3 x 5
##   group1 group2       p  p.adj p.adj.signif
## * <chr>  <chr>    <dbl>  <dbl> <chr>
## 1 bajo   medio  0.00425 0.0127 *
## 2 bajo   alto   0.0216  0.0647 ns
## 3 medio  alto   0.56    1      ns
```

Con la función *pairwise_prop_test* se va a evaluar la homogeneidad entre grupos.

Se observa que existe una diferencia significativa entre:

- el grupo de población con un nivel de educación bajo y un nivel de educación medio.

También se podría usar la función *fisher.multcomp* del paquete *RVAideMemoire* para comparar 2 grupos de 3 categorías cada grupo.

```
library(RVAideMemoire)
```

```
## *** Package RVAideMemoire v 0.9-83-2 ***
```

```
fisher.multcomp(tab, p.method = "bonferroni")
```

```
##
##         Pairwise comparisons using Fisher's exact test for count data
##
## data:  tab
##
```

```
##               0:1    0:2    1:2
## bajo:medio 1.0000 0.0404 0.5012
## bajo:alto  1.0000 0.1320 0.1592
## medio:alto 0.4671 1.0000 1.0000
##
## P value adjustment method: bonferroni
```

Aplicando la función de comparación de Fisher.

Unicamente se observa una diferencia significativa entre el grupo de nivel de educación bajo y medio para el grupo 0 y 2 abortos.

Se podría resumir que no existe una relación significativa entre el nivel de educación y el número de abortos.

Para faciliar la comprensión del ejercicio, a continuación vamos a estudiar la relación entre el nivel de educación y si ha habido aborto o no.

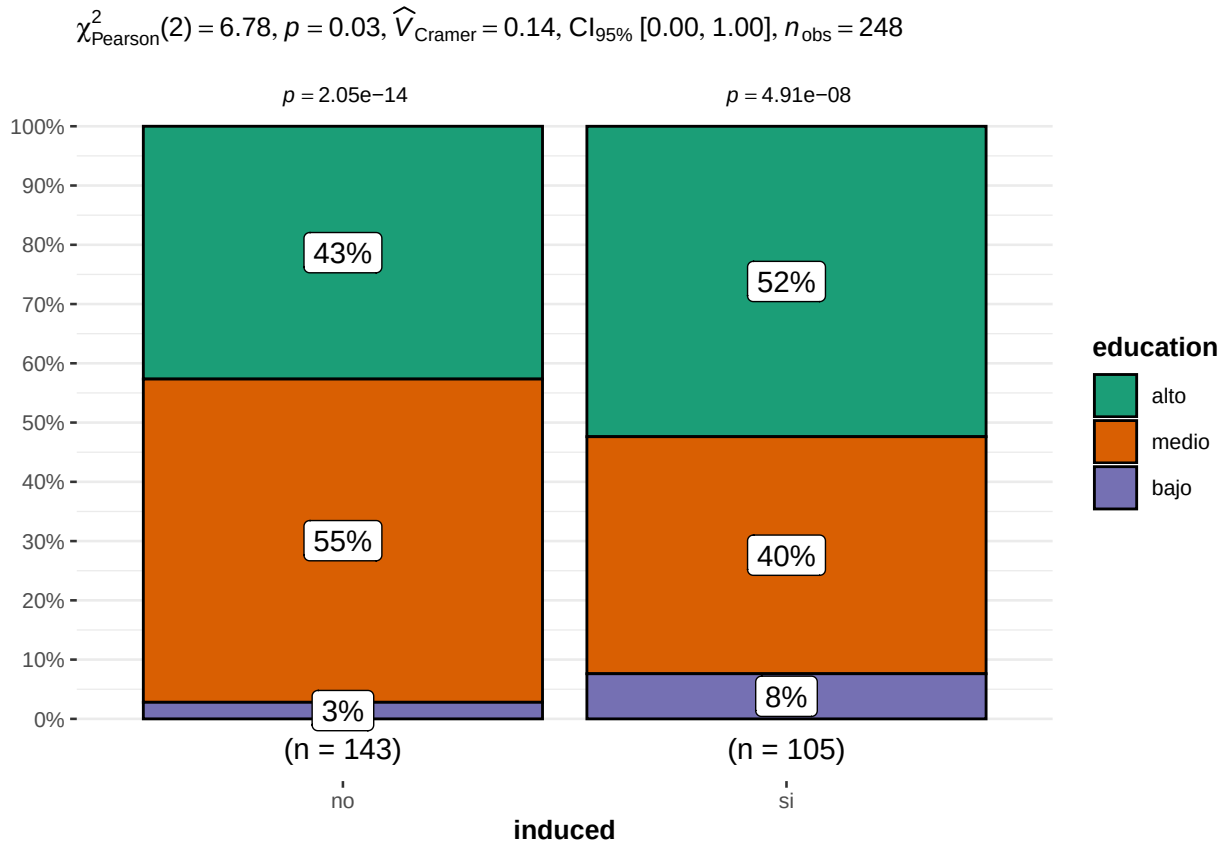
Comenzamos modificando transformando la variable *Induced* a factor y le asignamos 2 niveles ("no" para cuando no ha habido aborto y "si" para cuando ha habido aborto.

```
infert.long <- infert

infert.long$induced <- as.factor(infert.long$induced)

levels(infert.long$induced) <- c("no", "si", "si")

head(infert.long)
```

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

Mostramos el gráfico de barras.

```
ggbarstats(infert.long,
           education,
           induced,
           bf.message = F)
```



Aplicamos la función que compara entre proporciones.

```
pairwise_prop_test(table(infert.long$education, infert.long$induced), p.adjust.method ="bonferroni")
```

```
## Warning in prop.test(x[c(i, j)], n[c(i, j)], ...): Chi-squared approximation
## may be incorrect
```

```
## # A tibble: 3 x 5
##   group1 group2      p p.adj p.adj.signif
## * <chr>  <chr>   <dbl> <dbl> <chr>
## 1 bajo   medio  0.0652 0.196 ns
## 2 bajo   alto   0.334  1     ns
## 3 medio  alto   0.071  0.213 ns
```

Se observa que no existe una relación significativa entre el nivel de educación y si ha habido o no al menos un aborto inducido.

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). Para el grupo tratamiento, queremos comparar las edades de los pacientes que no mostraron mejoría con los que sí tuvieron una marcada mejoría.

```r
library(vcd)
```

```
## Loading required package: grid
```

```r
artritis <- Arthritis

head(artritis)
```

```
##   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
```

EDA:

1. Datos: la edad de los pacientes (Age) y la mejoría (Improved). La variable edad es numérica y la de mejoria es nominal.
2. Objetivo: para el grupo que recibió tratamiento se quiere comparar las edades que no mostraron mejoría con los que sí tuvieron uan mejoria.
3. Muestra: existe una sola muestra, con dos variables independientes, una variable numérica y otra nominal. Se podrá usará la prueba la prueba t-Student, Yuen o Mann-Whitney.

En primer lugar filtramos los pacientes que se sometieron al tratamiento y solo aquellos que obtuvieron una mejoría marcada (Marked) o ninguna mejoría (none).

```r
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
```

```r
artritis.long <- artritis %>%
  filter(Treatment == "Treated" & (Improved == "None" | Improved == "Marked"))

head(artritis.long)
```

```
##   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
```

Aplicando la siguiente función podemos determinar la edad media para cada categoría (none y marked).

```
artritis.long %>%
  group_by(Improved) %>%
  get_summary_stats(Age, type = "mean_sd")
```

```
## # A tibble: 2 x 5
##   Improved variable     n  mean    sd
##   <ord>    <fct>    <dbl> <dbl> <dbl>
## 1 None     Age         13  49.8 16.8
## 2 Marked   Age         21  56.8  9.02
```

Se observa que la edad media para el grupo Marked es 56.81 años, mientras que para el grupo None es de 49.85 años.

Podemos observar estos resultados en un gráfico de cajas.

```
library(ggpubr)
```

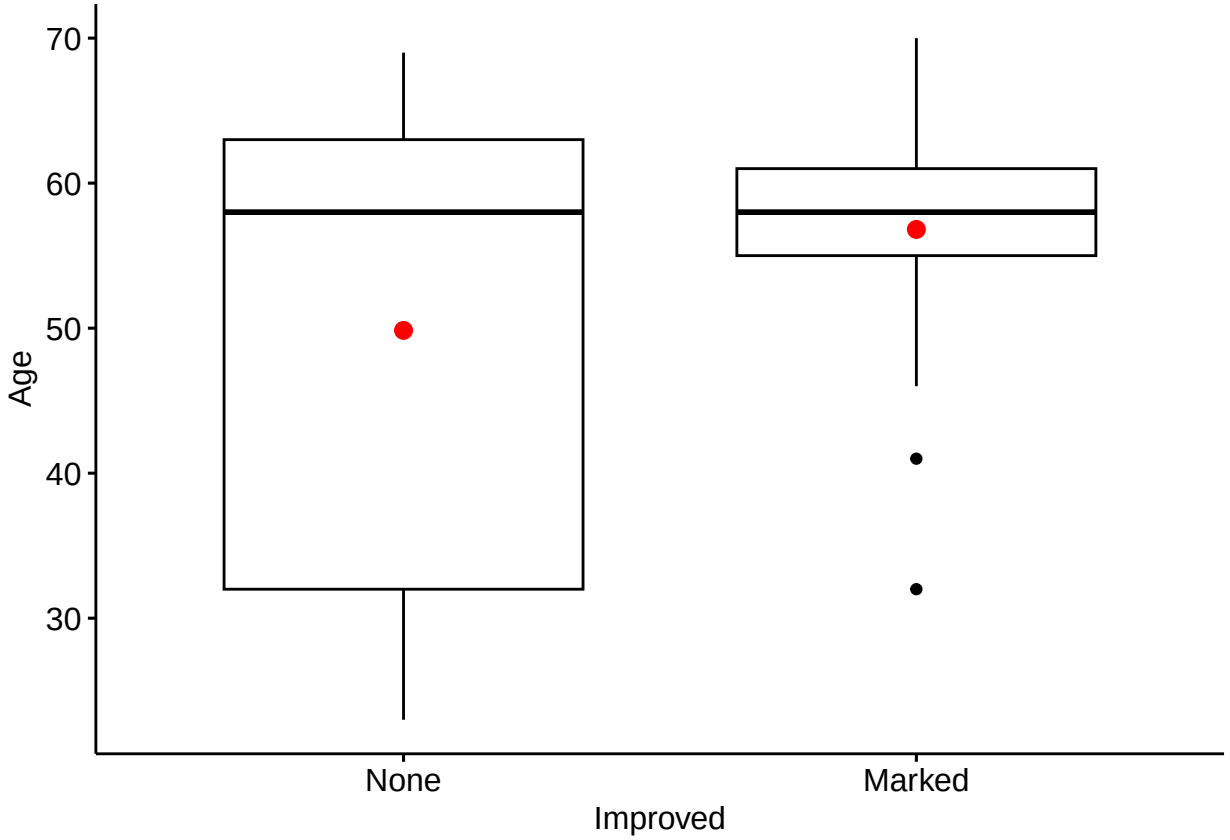
```
## Loading required package: ggplot2
```

```
ggboxplot(x="Improved",
          y="Age",
          data = artritis.long,
          add=c("mean"),
          add.params=list(color='red'))
```

```
## Warning: The `fun.y` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
## i Please use the `fun` argument instead.
## i The deprecated feature was likely used in the ggpubr package.
##   Please report the issue at <https://github.com/kassambara/ggpubr/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The `fun.ymin` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
## i Please use the `fun.min` argument instead.
## i The deprecated feature was likely used in the ggpubr package.
##   Please report the issue at <https://github.com/kassambara/ggpubr/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The `fun.ymax` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
## i Please use the `fun.max` argument instead.
## i The deprecated feature was likely used in the ggpubr package.
##   Please report the issue at <https://github.com/kassambara/ggpubr/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Se objerva valores atípicos en el grupo Marked. Por ello descrtamos la prueba t-Student y vamos a usar la prueba robusta de Yuen.

Aplicamos la función *identify_outliers* para determinar la cantidad de valores típicos y extremos.

```
artritis.long %>%
  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
```

Tenemos 2 valores atípicos y 1 valor extremo.

A continuación volvemos a determinar la edad media de cada categoría pero recortando un 20% los datos.

```
artritis.long %>%
  group_by(Improved) %>%
  filter(between(Age,
                 quantile(Age, 0.1),
                 quantile(Age, 0.9))) %>%
  get_summary_stats(Age, type="mean_sd")
```

```
## # A tibble: 2 x 5
##   Improved variable     n  mean    sd
##   <ord>    <fct>    <dbl> <dbl> <dbl>
## 1 None     Age          9  50.9 13.6
## 2 Marked   Age         17  57.8  5.27
```

Se observa que la edad media para el grupo Marked es 57.765 años (antes 56.81 años), mientras que para el grupo None es de 50.889 años (antes 49.85 años).

Volvemos a obervar el número de valores atípicos/extremos para los datos recortados.

```
artritis.long %>%
  group_by(Improved) %>%
  filter(between(Age,
                 quantile(Age, 0.1),
                 quantile(Age, 0.9))) %>%
  identify_outliers(Age)
```

```
## # A tibble: 3 x 7
##   Improved    ID Treatment Sex      Age is.outlier is.extreme
##   <ord>    <int> <fct>     <fct>  <int> <lgl>      <lgl>
## 1 Marked      36 Treated   Male      46 TRUE       FALSE
## 2 Marked      82 Treated   Female    48 TRUE       FALSE
## 3 Marked      13 Treated   Female    67 TRUE       FALSE
```

Recortando el 20% eliminamos los valores extremos pero seguimos teniendo valores atípicos, en este caso 3 valores (1 nuevo).

A continuación vamos a aplicar la prueba de Shapiro-Wilks para comprobar si existe normalidad en la distribución de los datos.

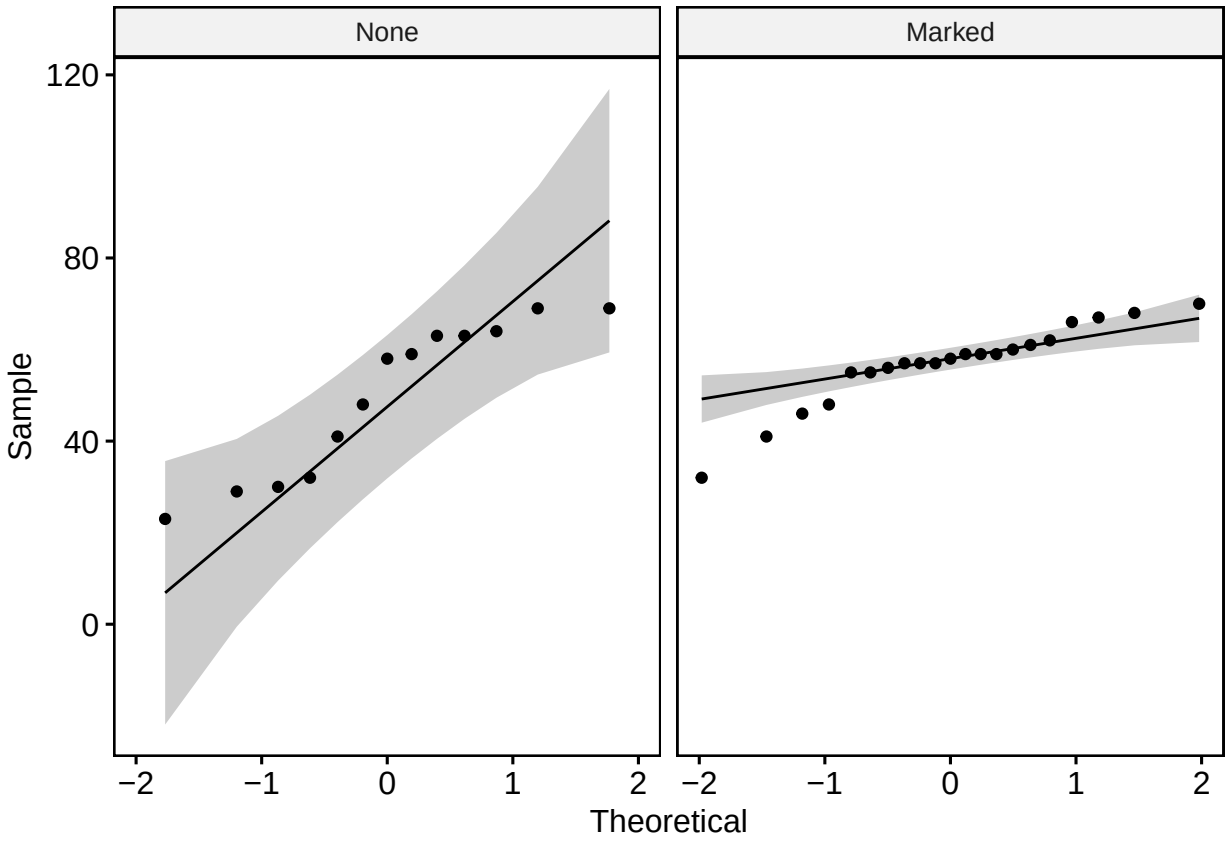
```
artritis.long %>%
  group_by(Improved) %>%
  shapiro_test(Age)
```

```
## # A tibble: 2 x 4
##   Improved variable statistic      p
##   <ord>    <chr>        <dbl>  <dbl>
## 1 None     Age          0.877 0.0659
## 2 Marked   Age          0.903 0.0395
```

La prueba de Shapiro-Wilks, para la variable Marked da como resultado un p-valor < 0.05, no aceptamos la hipótesis nula de normalidad.

Con la función Q-Q plots

```
ggqqplot(artritis.long,
         x="Age",
         facet.by = "Improved")
```



En el gráfico se comprueba que los datos de la variable Marked no siguien una distribución normal por lo que descartamos la prueba paramétrica de Yuen.

Vamos a usar una **prueba no paramétrica de Mann-Whitney** ya que los datos no cumplen con el supuesto de normalidad.

```
library(rstatix)
#wilcox_test(data = artritis.long,
#            formula = Age ~ Improved)
```

Al aplicar la prueba wilcox_test se obtiene el siguiente error:

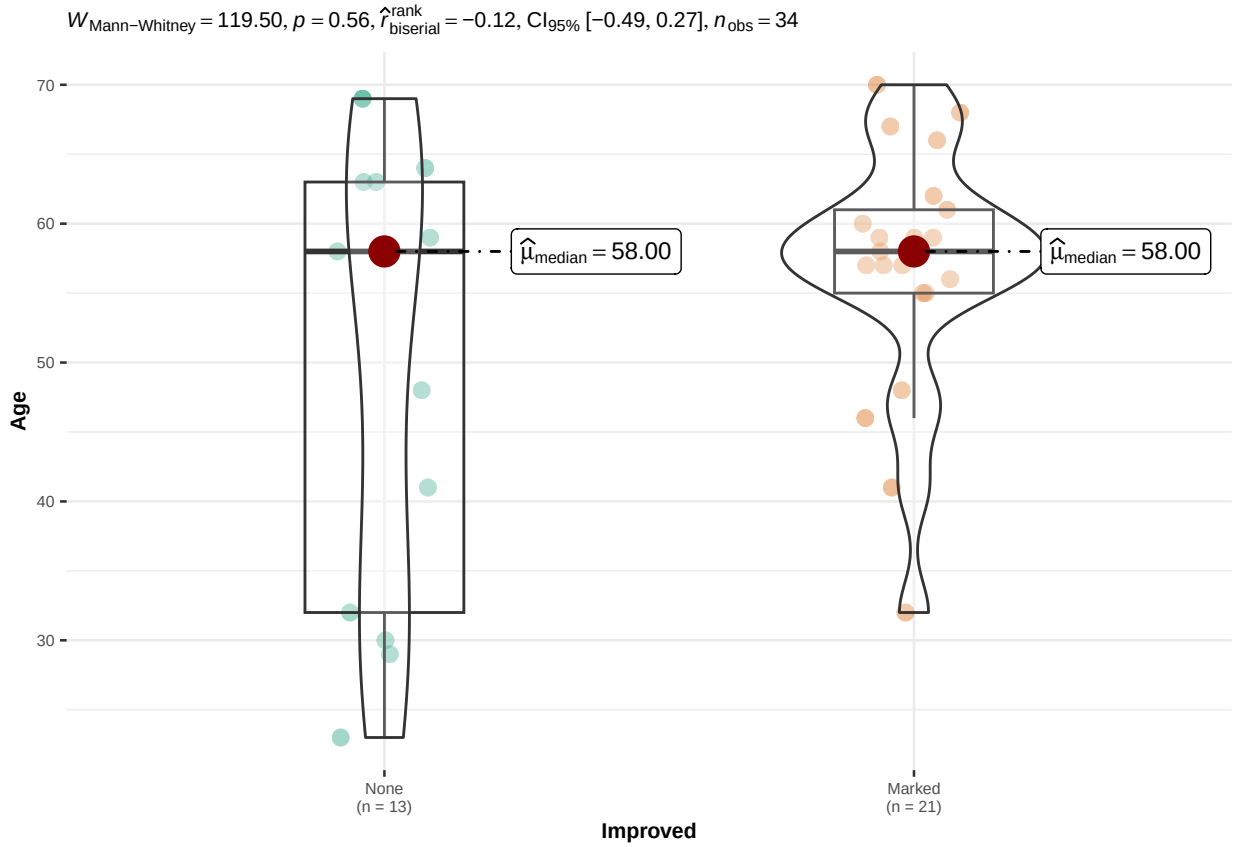
"Error in wilcox.test.default(x = c(29L, 30L, 59L, 63L, 63L, 64L, 69L, : not enough 'y' observations"

Como no he podido solventarlo, continuo el ejercicio.

Aplico la función ggbetweenstats de tipo 'np' no parametrico.

```
library(ggstatsplot)

ggbetweenstats(x = Improved,
               y = Age,
               data = artritis.long,
```

```
              type = 'np',
              bf.message = F) +
  theme(text = element_text(size=8),
        plot.subtitle = element_text(size=8))
```



Se observa que el tratamiento funciona princpialmente con una edad media de 58 años.

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.Comprueba mediante pruebas paramétricas, no paramétricas y robustas si han cambiado los valores medios de la producción de cebada. Interpreta y compara los resultados.

Se cargan los datos.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## The following object is masked from 'package:rstatix':
##
##     select
```

```
cebada <- immer

head(cebada)
```

```
##   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
```

Las variables de los datos son:

- **Loc**: localización
- **Var**: Variedad de cebada
- **Y1**: cocecha de 1931
- **Y2**: cocecha de 1932

ADE

- Datos: tenemos dos variables Y1 e Y1 que corresponde al rendimiendo en la recolección de cebada. Ambas variables son numéricas.
- objetivo: comprobar la media de la producción de cebada.
- muestra: comparar 2 muestras independiente para una variable numérica.

Prueba paramétrica

Para realizar la prueba paramétrica usaremos la **t de Student**.

En primer lugar vamos a modificar los datos

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## v readr     2.1.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks rstatix::filter(), stats::filter()
## x dplyr::lag()    masks stats::lag()
## x MASS::select()  masks dplyr::select(), rstatix::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(dplyr)

cebada.long <- cebada %>%
  pivot_longer(c(Y1, Y2),
               names_to = "year",
               values_to = "score") %>%
  arrange(year)

head(cebada.long)
```

```
## # A tibble: 6 x 4
##   Loc   Var   year  score
##   <fct> <fct> <chr> <dbl>
## 1 UF    M     Y1     81
## 2 UF    S     Y1    105.
## 3 UF    V     Y1    120.
## 4 UF    T     Y1    110.
## 5 UF    P     Y1     98.3
## 6 W     M     Y1    147.
```

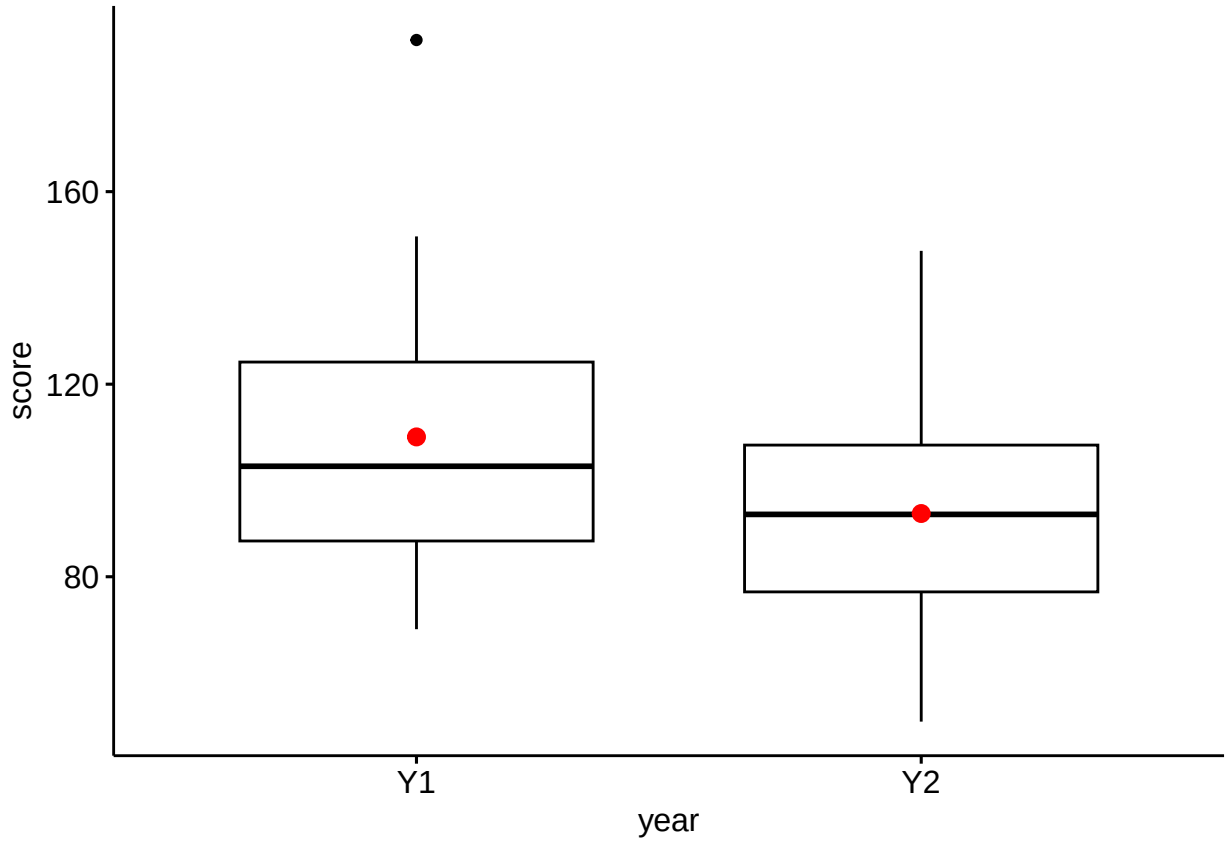
Obtenemos la media.

```r
cebada.long %>%
  group_by(year) %>%
  get_summary_stats(score, type = "mean")
```

```
## # A tibble: 2 x 4
##   year  variable     n  mean
##   <chr> <fct>    <dbl> <dbl>
## 1 Y1    score       30 109.
## 2 Y2    score       30  93.1
```

La media de producción para Y1 es de 109.05, mientras que la media par Y2 es de 93.13. En un comienzo se observa que en el año Y1 (1931) la producción fue mayor.

```r
ggboxplot(x="year",
          y="score",
          data = cebada.long,
          add=c("mean"),
          add.params=list(color='red'))
```



En el gráfico box se observa que existe un valor atípico para Y1.

Vamos a comprobar si este valor atípico es significativo.

```
cebada.long %>%
  group_by(year) %>%
  identify_outliers(score)
```

```
## # A tibble: 1 x 6
##   year  Loc   Var   score is.outlier is.extreme
##   <chr> <fct> <fct> <dbl> <lgl>      <lgl>
## 1 Y1    W     T      192. TRUE       FALSE
```

Tal como se muestra en el gráfico tipo box, se comprueba que existe un valor atípico pero no extremo.

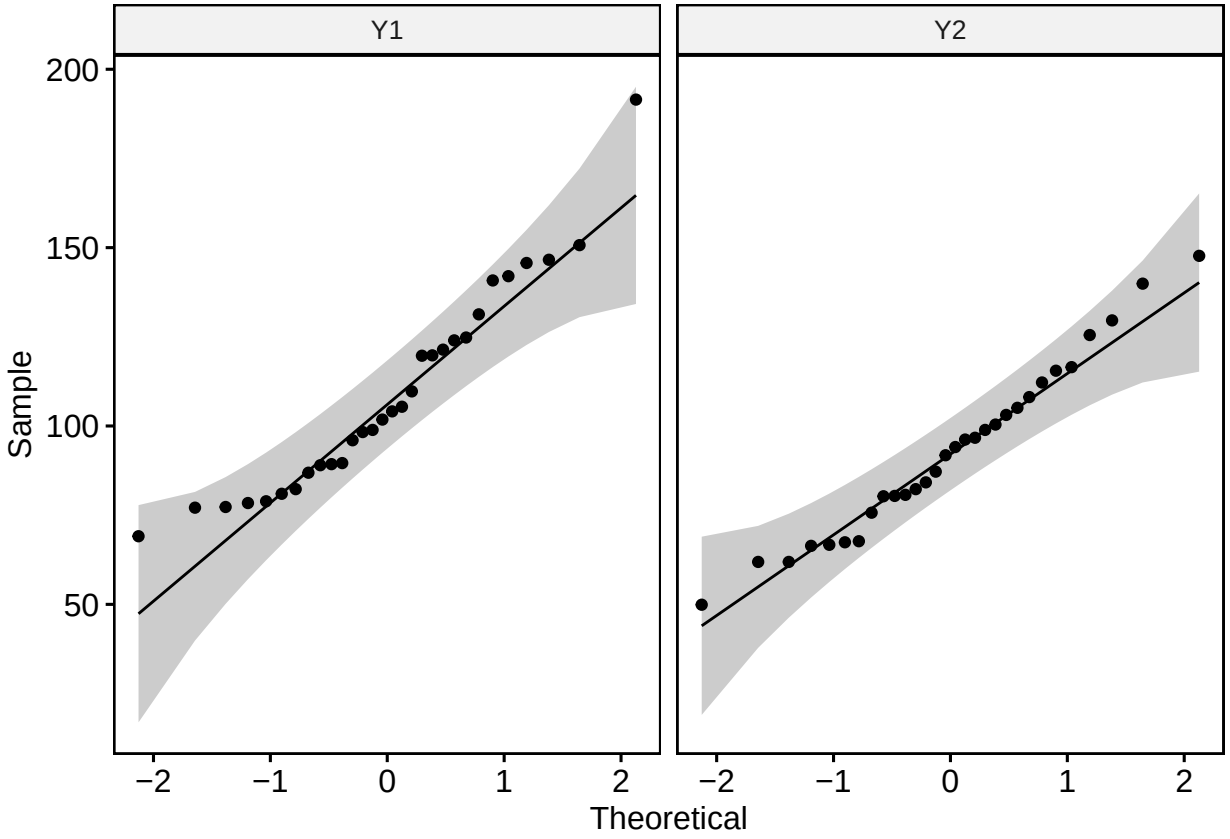
Comprobamos si cumple con la normalidad.

```
cebada.long %>%
  group_by(year) %>%
  shapiro_test(score)
```

```
## # A tibble: 2 x 4
##   year  variable statistic      p
##   <chr> <chr>        <dbl>  <dbl>
## 1 Y1    score        0.930 0.0483
## 2 Y2    score        0.976 0.703
```

Como p-value es < 0.05 para Y1, no aceptamos H0 de normalidad, lo que indica que esta variable no siguen una distribución normal.

```
ggqqplot(cebada.long,
         x="score",
         facet.by = "year")
```



En el gráfico se observa que los datos se aproximan a la normalidad.

```
library(rstatix)
t_test(score ~ year,
       data = cebada.long,
       paired = F,
       alternative = "two.sided",
       mu = 0,
       conf.level = 0.95)
```

```
## # A tibble: 1 x 8
##   .y.   group1 group2    n1    n2 statistic    df     p
## * <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl> <dbl>
## 1 score Y1     Y2        30    30      2.32  56.5 0.024
```

La **prueba t_test** describe la diferencia entre dos medias.

Como p-value es menor al 5% de confianza, rechazamos H0 de igualdad, el valor medio es distintos entre los años Y1 e Y2.

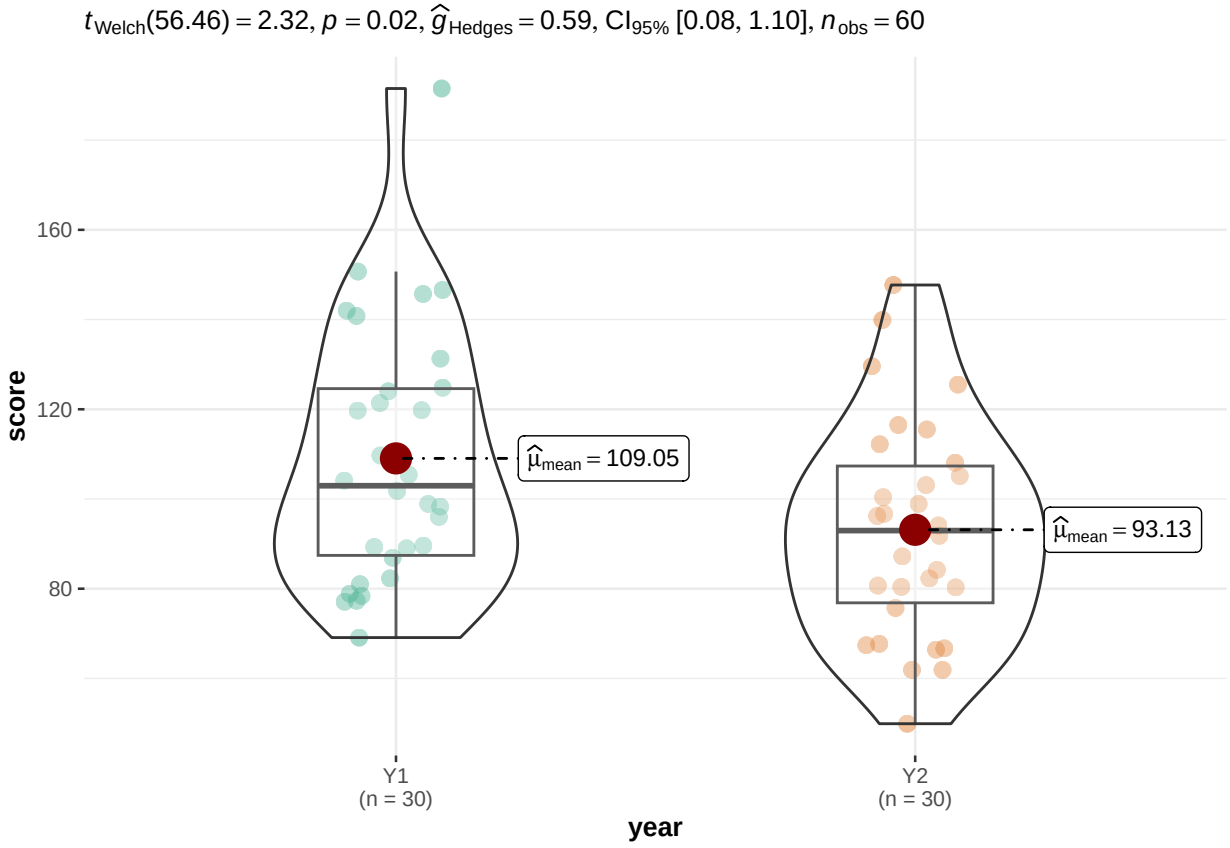
Tamaño del efecto

```
cohens_d(score ~ year,
         data = cebada.long)
```

```
## # A tibble: 1 x 7
##   .y.   group1 group2 effsize    n1    n2 magnitude
## * <chr> <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 score Y1     Y2       0.599    30    30 moderate
```

El tamaño del efecto es moderado, es decir, hay una diferencia significativa entre la producción media de cada año.

```
library(ggstatsplot)
ggbetweenstats(x = year,
               y = score,
               data = cebada.long,
               bf.message = F)
```



En el caso de Y1 se observa una distribución bimodal, con una media en la producción de 109.05, mientra que en el caso de Y2 se observa que los datos siguen una distribución normal con una media en la producción de 93.13.

La prueba t de Welch para las dos muestras, señala que la diferencia es estadísticamente significativa (t(56.46) = p < 0.02, d = 0.59)

Prueba robusta

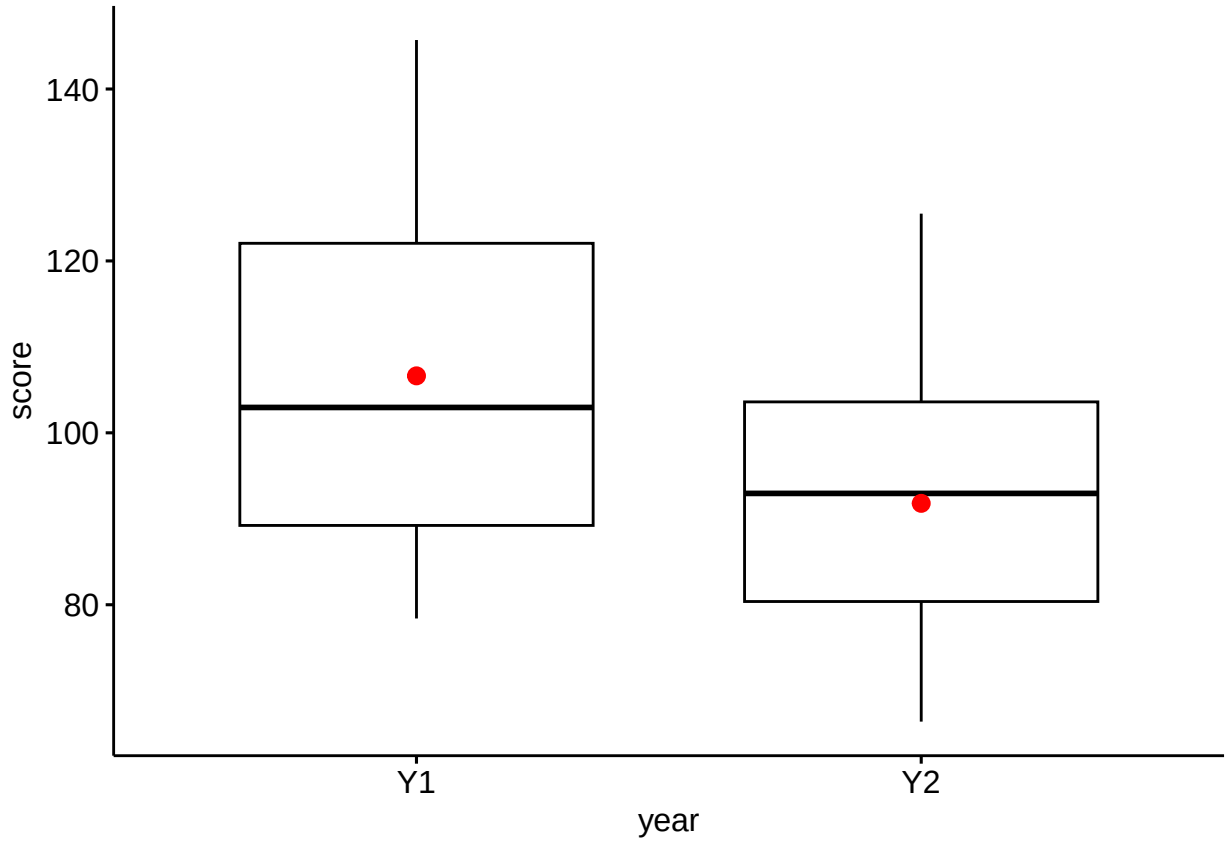
Repetimos el ejercicio pero usando una prueba robusta. Para ello se recortan el 20% de los datos (10% por cada lado)

```
cebada.long %>%
  group_by(year) %>%
  filter(between (score,
                  quantile(score, 0.1),
                  quantile(score, 0.9))) %>%
  get_summary_stats(score, type = "mean")
```

```
## # A tibble: 2 x 4
##   year  variable     n  mean
##   <chr> <fct>    <dbl> <dbl>
## 1 Y1    score       24 107.
## 2 Y2    score       24  91.8
```

La media de producción para Y1 es de 106.63, mientras que la media par Y2 es de 91.8.

```
cebada.long %>%
  group_by(year) %>%
  filter(between(score,
                 quantile(score,0.1),
                 quantile(score,0.9))) %>%
  ggboxplot(x="year",
            y="score",
            add=c("mean"),
            add.params=list(color='red'))
```



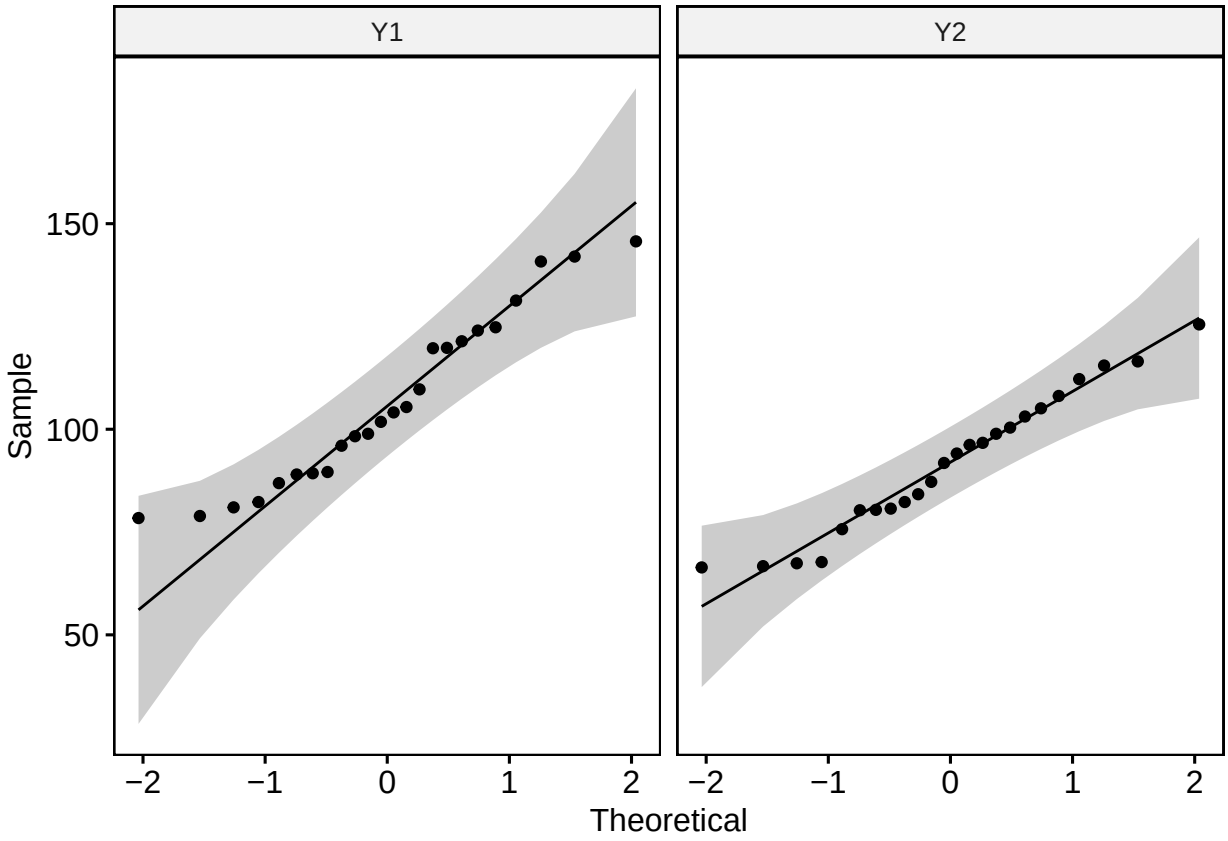
Al recortar el 20% de los datos nos desprendemos del valor atípico.

```
cebada.long %>%
  group_by(year) %>%
  filter(between(score,
                 quantile(score,0.1),
                 quantile(score,0.9))) %>%
   shapiro_test(score)
```

```
## # A tibble: 2 x 4
##   year  variable statistic     p
##   <chr> <chr>        <dbl> <dbl>
## 1 Y1    score        0.936 0.135
## 2 Y2    score        0.962 0.485
```

Ahora como p $>$ 0.05, no podemos rechazar la hipótesis nula de normalidad.

```
cebada.long %>%
  group_by(year) %>%
  filter(between(score,
                 quantile(score,0.1),
                 quantile(score,0.9))) %>%
   ggqqplot(x="score",
            facet.by = "year")
```




```
library(DescTools)
```

```
## Registered S3 method overwritten by 'DescTools':
##   method         from
##   reorder.factor gdata
```

```
YuenTTest(score ~ year,
          data = cebada.long)
```

```
##
##  Yuen Two Sample t-test
##
## data:  score by year
## t = 1.6771, df = 33.227, trim = 0.200, p-value = 0.1029
## alternative hypothesis: true difference in trimmed means is not equal to 0
## 95 percent confidence interval:
##  -2.922596 30.389263
## sample estimates:
## trimmed mean in group Y1 trimmed mean in group Y2
##                105.12778                91.39444
```

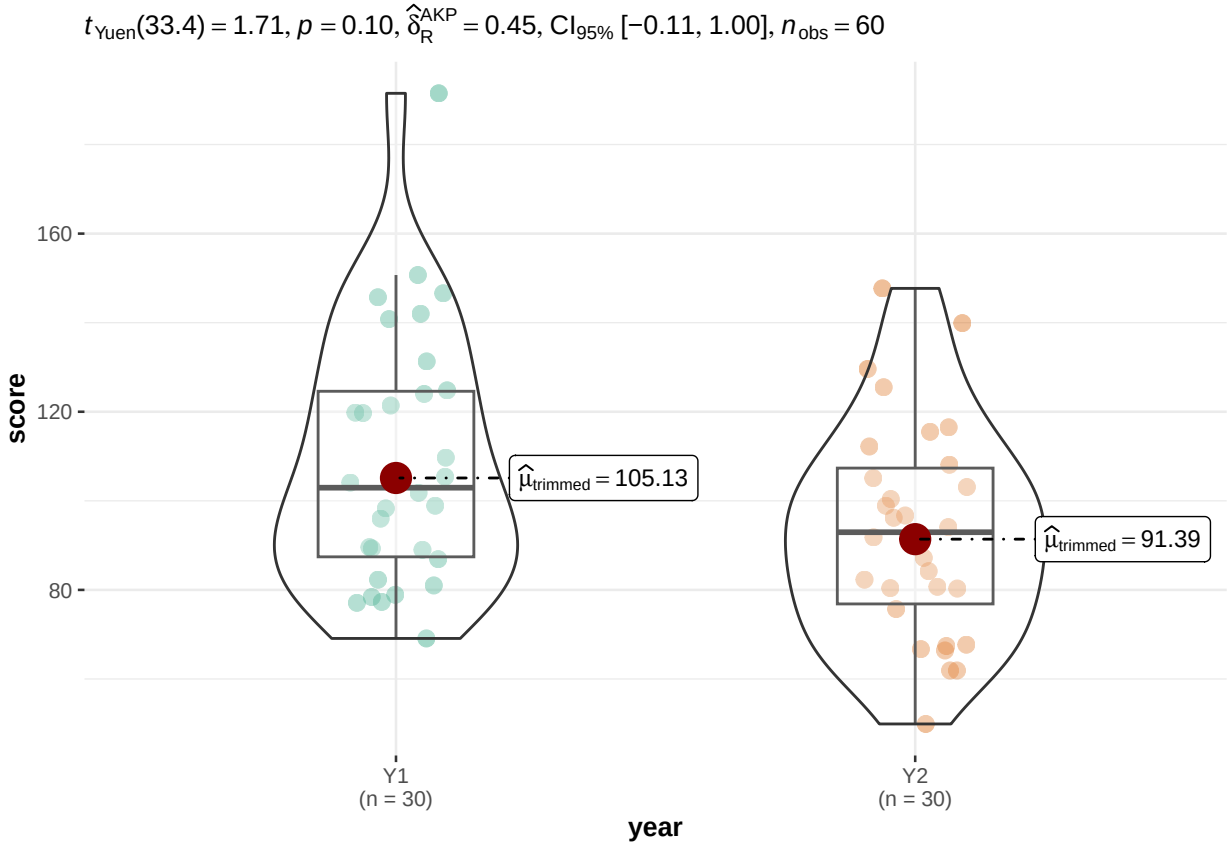
El p-valor es > 0.05, por lo que no podemos rechazar H0 de igualdad.

```
library(WRS2)
yuen.effect.ci(score ~ year,
               data = cebada.long,
               tr = 0.2,
               alpha = 0.05,
               nboot = 10)
```

```
## $effsize
## [1] 0.3510129
##
## $alpha
## [1] 0.05
##
## $CI
## [1] 0.0000000 0.5650439
```

El tamaño del efecto es 0.35, existe una diferencia moderada en las medias de las producciones.

```
library(ggstatsplot)
ggbetweenstats(x = year,
               y = score,
               data = cebada.long,
               type = "robust",
               bf.message = F)
```



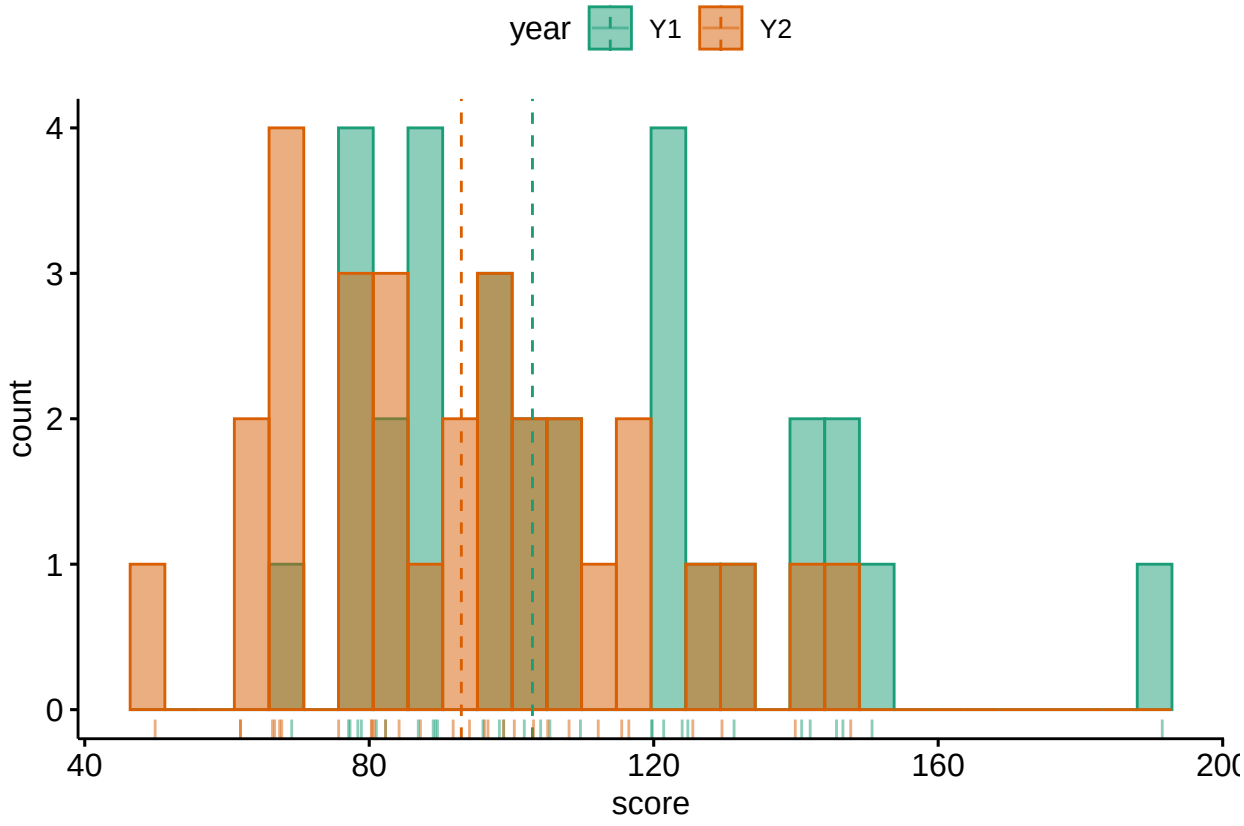
En el caso de Y1 se observa una distribución bimodal, con una media en la producción de 105.13, mientra que en el caso de Y2 se observa que los datos siguen una distribución normal con una media en la producción de 91.39.

La prueba t de Yen para las dos muestras, señala que la diferencia no es estadísticamente significativa (t(33.4) = p = 0.10, $ = 0.59$)

Prueba no paramétrica

La prueba no paramétrica para dos variables independientes es la **prueba U de Mann-Whitney**.

```
library(ggpubr)
gghistogram(cebada.long,
            x = "score",
            add = "median",
            rug = T,
            bins = 30,
            color = "year",
            fill = "year",
            palette = "Dark2")
```



En el histograma se observa el valor atípico.

```
library(rstatix)
wilcox_test(score ~ year,
            data = cebada.long)
```

```
## # A tibble: 1 x 7
##   .y.   group1 group2    n1    n2 statistic      p
## * <chr> <chr>  <chr>  <int> <int>     <dbl>  <dbl>
## 1 score Y1     Y2        30    30       589 0.0406
```

El p-valor es inferior a 0.05, por lo que rechazamos la hipótesis nula de igualdad, es decir, no podemos aceptar la hipótesis de que la media entre la producción en ambos años es la misma.

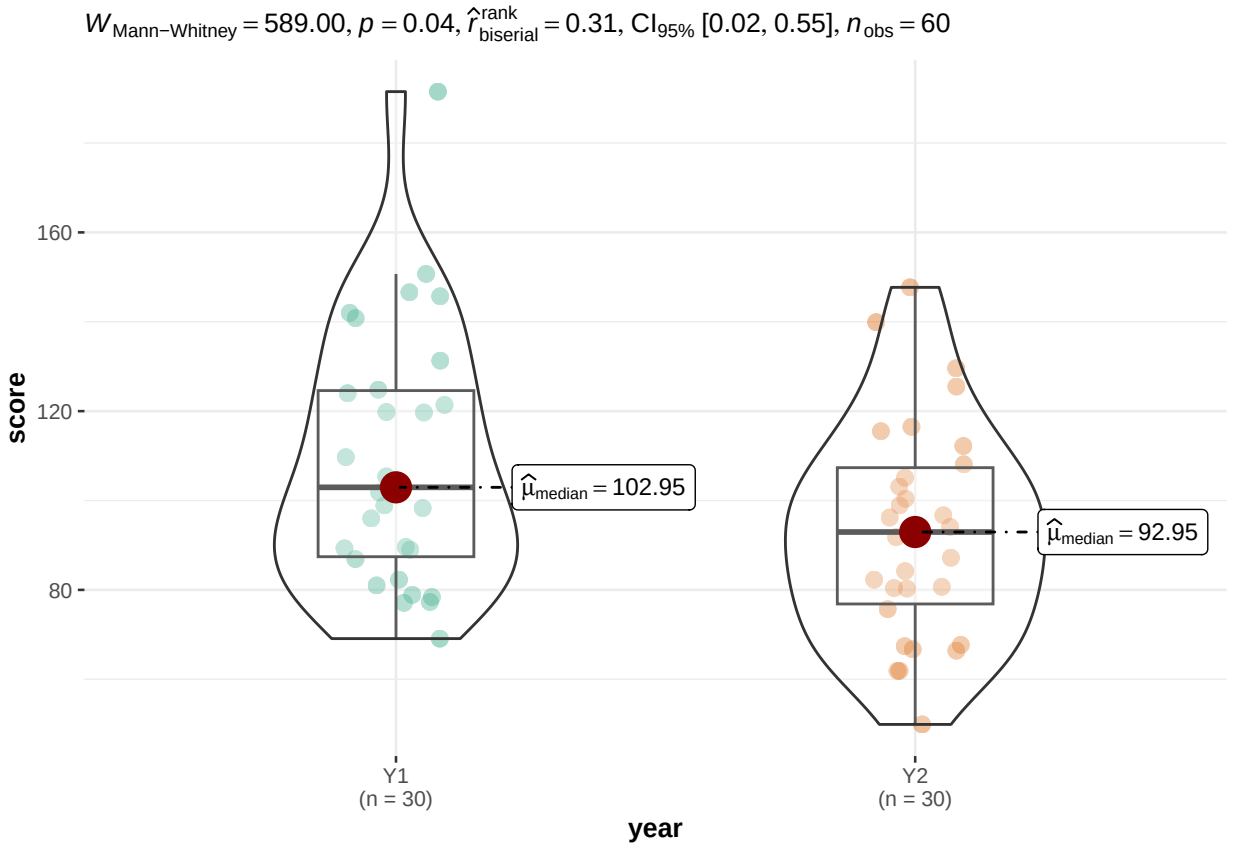
```
wilcox_effsize(score ~ year,
               data = cebada.long)
```

```
## # A tibble: 1 x 7
##   .y.   group1 group2 effsize    n1    n2 magnitude
## * <chr> <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 score Y1     Y2       0.265    30    30 small
```

El tamaño del efecto es de 0.265.

La diferencia entre las medias de Y1 e Y2 es de magnitud pequeña.

```
library(ggstatsplot)
ggbetweenstats(x = year,
               y = score,
               data = cebada.long,
               type = "np",
               bf.message = F)
```



La media de producción para el año Y1 es de 102.95 mientras que para el año Y2 es de 92.95.

La prueba de Wilcoxon monstró que la diferencia fue significativa (W = 589, p = 0.0406, r = 0.31)

Concluciones

Dado que los datos tiene valores atípicos y la distribución de los datos recortados siguen una distribución normal, podemos optar por la prueba robusta de Yuen como prueba definitiva.