Comparación entre dos Grupos

Juan C. Correa

3/11/2021

En este tutorial, vamos a explicar cómo hacer algunos análisis de comparación entre dos grupos exclusivamente. Comencemos usando los datos que aparecen a continuación.

```
ipak <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)
}

# uso
packages <- c("readxl","apa","dplyr","tidyBF","BayesFactor","ggstatsplot")
ipak(packages)
```

```
## Loading required package: readxl

## Loading required package: apa

## Loading required package: 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

## Loading required package: tidyBF

## This package is no longer being maintained and might be removed from CRAN in future.
##     All its functionality has now moved to `statsExpressions` package.
##     Please see: https://indrajeetpatil.github.io/statsExpressions/

## Loading required package: BayesFactor
```

```
## Loading required package: coda
```

```
## Loading required package: Matrix
```

```
## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Morey (richarddmorey
##
## Type BFManual() to open the manual.
## ************
```

```
## Loading required package: ggstatsplot
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## In case you would like cite this package, cite it as:
##      Patil, I. (2018). ggstatsplot: "ggplot2" Based Plots with Statistical Details. CRAN.
##      Retrieved from https://cran.r-project.org/web/packages/ggstatsplot/index.html
```

```
##      readxl         apa       dplyr      tidyBF BayesFactor ggstatsplot
##        TRUE        TRUE        TRUE        TRUE        TRUE        TRUE
```

```
setwd('~/Documents/GitHub/Pantaleon')
df <- read_excel("PantaleonDatos.xls")
```

```
## New names:
## * TOTAL -> TOTAL...24
## * CONTAR -> CONTAR...25
## * edad -> edad...34
## * edad -> edad...40
## * TOTAL -> TOTAL...89
## * ...
```

```
head(df)
```

```
## # A tibble: 6 x 104
##   `id-com` Finca Gerente   Descripcion       Lote    TC  Área   TCH Estacion
##   <chr>    <dbl> <chr>     <chr>            <dbl> <dbl> <dbl> <dbl> <chr>
## 1 10166206 10166 Gerente A SANTA CRUZ CHIPO   206  800. 11.0   72.6 Cocales
## 2 10112401 10112 Gerente B AGRICOLA           401 2392. 24     99.7 Tehuantepeq
## 3 10112601 10112 Gerente A AGRICOLA           601 1626. 25     65.1 Tehuantepeq
## 4 10655202 10655 Gerente B LA PRESA Baul      202 2463. 27.9   88.3 Tehuantepeq
## 5 10012301 10012 Gerente A LA CUCHILLA        301  418.  6.40  65.3 Bouganvilia
## 6 10123210 10123 Gerente B LIMONES S.A.       210 1790. 11.0  162.  Tehuantepeq
## # ... with 95 more variables: Estación_Actualizada <chr>, rendimiento <dbl>,
## #   TAH <dbl>, fecha_corte <dttm>, Fecha ultima <dttm>, dias <dbl>,
## #   FI_E1 <dttm>, FF_E1 <dttm>, FI_EII <dttm>, FF_EII <dttm>, Cuadrante <chr>,
```

```
## #   Sum rad I <dbl>, Sum radII <dbl>, totalrad <dbl>, TOTAL...24 <dbl>,
## #   CONTAR...25 <dbl>, Rad <dbl>, Temperatura x <lgl>, round t° 0.5 <lgl>,
## #   round t° 1 <lgl>, total <lgl>, <35 <lgl>, cant_dias <lgl>, 15_dia <lgl>,
## #   edad...34 <dbl>, round <dbl>, edad_ran <chr>, Cortes <dbl>,
## #   Cortes_agrup <chr>, t_cosecha <chr>, edad...40 <dbl>, Region <chr>,
## #   variedad_cortada <chr>, grup_var <chr>, Status <chr>,
## #   KG_azucar_core_ant <dbl>, Kg_azucar_Industrial_ant <dbl>,
## #   KG_Azucar_Core_real <dbl>, KG_Azucar_industrial_real <dbl>,
## #   Recuperación <dbl>, prequema <chr>, area_cos <dbl>, tons_ant <dbl>,
## #   TCH_ant <dbl>, altitud <chr>, TERCIO <chr>, chinche <chr>, trash <chr>,
## #   r_trash <dbl>, caña_seca <chr>, sistema de Riego <chr>, REGION_NE <chr>,
## #   CLEANER <chr>, Pureza_Core <dbl>, Total pur <dbl>, Humedad_Core <dbl>,
## #   Round_h <dbl>, round_2H <chr>, Madurante <chr>, Premadurante <chr>,
## #   Tipo <chr>, Formulacion <chr>, formulacion1 <chr>, foliar_1 <chr>,
## #   foliar_2 <chr>, complemento <chr>, Quemas <chr>, quema_ran <chr>,
## #   Psuelo <chr>, Brix <dbl>, Pol <dbl>, Delta <dbl>, jugo <dbl>, ph <dbl>,
## #   fibra <dbl>, Conductividad <dbl>, Ingenio <chr>, Semana <dbl>, Month <dbl>,
## #   TOTAL...89 <dbl>, CONTAR...90 <dbl>, T°Max <dbl>, TOTAL...92 <dbl>,
## #   CONTAR...93 <dbl>, Amplitud <dbl>, TOTAL...95 <dbl>, CONTAR...96 <dbl>,
## #   T°Min <dbl>, TCHp <dbl>, KG <dbl>, mes cosehca <dbl>, round tch <dbl>,
## #   round rad <dbl>, round rad% <dbl>, round2 rad% <dbl>
```

Hasta este momento, puede observarse que hemos definido una función llamada "ipak" que nos permite definir cuáles son los paquetes o librerías que vamos a usar y si no lo tenemos instalado, pedirle a R que me los instale de manera automática y me los cargue antes de proceder con el análisis.

Ahora que tenemos los datos abiertos, vamos a hacer algunos cálculos. Empecemos mirando que a los datos de Pantaleon, le hemos creado una variable dicotómica llamada "Gerente". Desde luego esta es una variable ficticia, pero hace referencia a dos estilos de gerenciar la producción. A la mitad de la producción se le ha puesto bajo la dirección del Gerente A, y al resto se le ha puesto bajo la dirección del Gerente B. Entonces queremos ver si hay o no una asociación entre el estilo de gerencia y el rendimiento de las fincas.

t de Student (Prueba Paramétrica)

```
t.test(df$rendimiento ~ df$Gerente, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  df$rendimiento by df$Gerente
## t = 0.6366, df = 1528, p-value = 0.5245
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7484278  1.4676448
## sample estimates:
## mean in group Gerente A mean in group Gerente B
##                131.7172                131.3576
```

El resultado de aplicar esta sintaxis nos muestra un estadístico (en este caso t) cuyo valor sugiere que no existe una relación o asociación significativa entre el rendimiento y el estilo de gerencia ($t = 0.636$, $df = 1528$, $p = 0.5245$). La diferencia en el rendimiento de los lotes bajo la dirección del Gerente A (131.7172) no es estadísticamente superior al rendimiento de los lotes bajo la dirección del Gerente B (131.3576).

Prueba de Wilcoxon (Prueba no paramétrica)

Veamos ahora, cómo serían los resultados si corremos una prueba de tipo no paramétrico.

```
wilcox.test(rendimiento ~ Gerente, data = df, exact = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  rendimiento by Gerente
## W = 296842, p-value = 0.6246
## alternative hypothesis: true location shift is not equal to 0
```

Ahora vemos como resultado otro estadístico (en este caso W) cuyo valor sugiere que no existe una relación o asociación significativa entre el rendimiento y el estilo de gerencia (W = 296842, p = 0.6246). En este ejemplo, se observa que la conclusión que se obtiene por ambos métodos coinciden en la interpretación de no haber relación estadísticamente significativa entre el estilo de gerencia y el rendimiento de los lotes.

#t test bayesiano con paquete BayesFactor

```
ttestBF(formula = rendimiento ~ Gerente, data = df)
```

```
## Warning: data coerced from tibble to data frame
```

```
## Bayes factor analysis
## --------------
## [1] Alt., r=0.707 : 0.07015332 ±0.11%
##
## Against denominator:
##   Null, mu1-mu2 = 0
## ---
## Bayes factor type: BFindepSample, JZS
```

```
veamos <- bf_ttest(data = df, x = Gerente, y = rendimiento)
variable.names(veamos)
```

```
##  [1] "term"               "estimate"           "conf.level"
##  [4] "conf.low"           "conf.high"          "pd"
##  [7] "rope.percentage"    "prior.distribution" "prior.location"
## [10] "prior.scale"        "bf10"               "method"
## [13] "log_e_bf10"
```

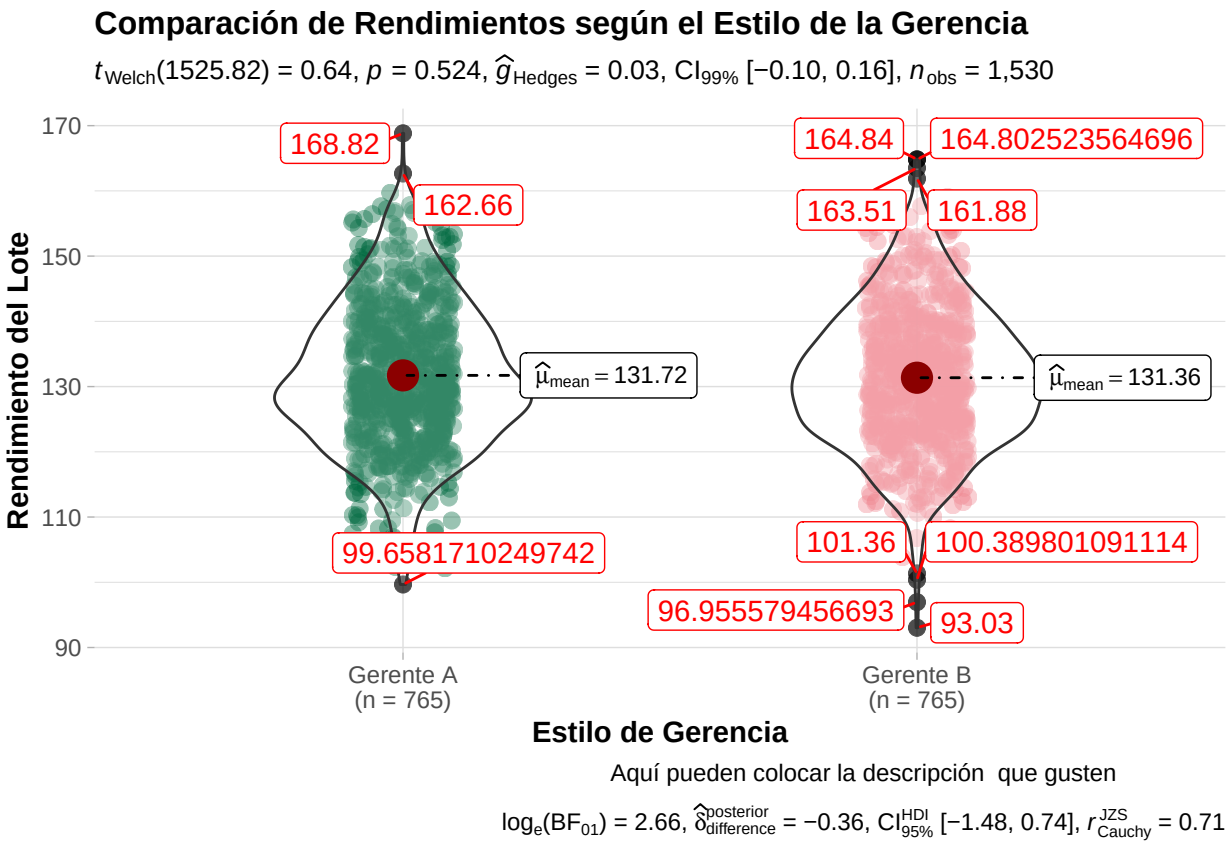
```
veamos[11]
```

```
## # A tibble: 2 x 1
##     bf10
##    <dbl>
## 1 0.0702
## 2 0.0702
```

Al correr una t test bayesiana, nos interesa saber el factor de bayes que se encuentra en la columna 11 de la tabla que hemos llamado "veamos". El resultado es 0.0702 que al ser casi cero, se interpreta como una asociación no importante desde el punto de vista estadístico.

Algo que quizás no se ha mencionado explíctamente hasta ahora es que las pruebas estadísticas que nos permiten estimar relaciones entre variables pueden complementarse con técnicas de graficación o visualización. Por ejemplo, en el siguiente gráfico estamos representando el comportamiento estadístico del rendimiento de los lotes en función del Gerente.

```
ggstatsplot::ggbetweenstats(
  data = df,
  x = Gerente, # Variable de agrupación/independiente
  y = rendimiento, # Variable dependiente
  xlab = "Estilo de Gerencia", # etiqueta para el eje X
  ylab = "Rendimiento del Lote", # etiqueta para el eje y
  type = "p",# "p" de parametrico, "np" no-parametrico, "r" de robusto, "bf" de Bayesiano.
  effsize.type = "g", # se refiere al método de cálculo de efecto
  conf.level = 0.99,
  plot.type = "violin", # dos opciones "boxplot" o "violin"
  outlier.tagging = TRUE, # Te indica cuales son outliers
  outlier.coef = 1.5, # coeficiente de Tukey para detección de outliers
  outlier.label.args = list(color = "red"), # la etiqueta de los outlairs en rojo
  ggtheme = ggplot2::theme_light(), # cambiar el fondo de la gráfica todos los temas en https://ggplot2
  package = "yarrr", # El paquete asociado a la paleta de colores.
  palette = "info2", # Elegir la paletta dentro del paquete
  title = "Comparación de Rendimientos según el Estilo de la Gerencia",
  caption = "Aquí pueden colocar la descripción  que gusten"
)
```



Otras pruebas de comparación

Por su gran aplicabilidad, el análisis de varianza (y su versión no-paramétrica llamada prueba H de Kruskal-Wallis) funcionan de manera muy similar a las pruebas de comparación entre dos grupos que se han documentado acá. La diferencia entre el ANOVA o la prueba Kruskal-Wallis y las pruebas mostradas acá es básicamente el número de grupos a comparar. Así pues, en la prueba t de Student (o la prueba Wilcoxon)