Tarea extra: ANOVA desde cero

Diego Hernández Jiménez

Función principal

`ANOVA`: Realiza un ANOVA de uno o dos factores de efectos fijos.

- Argumentos:
  - `factor`: vector de caracteres de longitud 1 o 2 con el nombre de los predictores o factores. Deben hacer referencia a una variable de tipo factor de un dataframe. En caso de que el/los factores no sean de tipo factor, se llama a la función `check_factor` para transformar las variables del dataframe en en tipo factor.
  - `vd`: vector de caracteres de longitud 1 con el nombre de la variable criterio o dependiente. Debe hacer referencia a una variable numérica de un dataframe.
  - `data`: nombre del dataframe que contiene el factor o factores y la variable dependiente.
- Salida:

  En función del número de factores, devuelve la función `one_way` o la función `two_way`

```
ANOVA <- function(factor,vd,data){

  num_fact <- length(factor)

  data <- check_factor(factor,data)

  switch(num_fact,

         return(one_way(num_fact,factor,vd,data)),
         return(two_way(num_fact,factor,vd,data)))
}
```

Funciones auxiliares

`check_factor`: Comprueba que los predictores son factores. En caso de no serlo, transforma las variables en factores en el propio dataframe original y lo devuelve.

- Argumentos:
  - `vars`: vector de caracteres de longitud 1 ó 2 que contiene el nombre de los factores del ANOVA
  - `df`: dataframe que contiene los factores vars.
- Salida:
  - devuelve el mismo dataframe, pero con las variables transformadas en factores en caso de que no lo fueran.

```r
check_factor <- function(vars,df){

  for (var in vars){

    if (!(is.factor(df[,var])))

      warning(paste0('se convertira ',var,' en tipo factor '))
      df[,var] <- as.factor(df[,var])
  }
  return(df)
}
```

`anova_plot`: Devuelve gráfico de líneas con las medias de cada nivel (ANOVA de un factor) o de cada combinación de niveles (ANOVA de dos factores). Permite explorar el efecto de la interacción mediante inspección visual.

- Argumentos:
  - `num_fact`: Vector numérico de longitud 1. Longitud del vector que contiene los factores. Calculado en `ANOVA`.
  - `factor`: Igual que en `ANOVA`.
  - `vd`: Igual que en `ANOVA`.
  - `data`: Igual que en `ANOVA`.
- Salida:
  - devuelve gráfico de líneas

```r
anova_plot <- function(num_fact,factor,vd,data,medias=NULL,J=NULL){

  if(num_fact==1){

    labels <- levels(data[,factor])

    plot(1:J,medias,
         type='o', lwd=2,
         ylim=c(min(data[,vd]),max(data[,vd])),
         xlab=factor,ylab=vd,
         pch=19, col='#ff6b6b',
         xaxt='n')

    axis(1,1:J,labels=labels) # xticks
  }

  else{

    interaction.plot(x.factor = data[,factor[1]], trace.factor = data[,factor[2]],
                response = data[,vd], fun = mean,
                type = "b", legend = TRUE, lwd=2,
                ylim=c(min(data[,vd]),max(data[,vd])),
                xlab = factor[1], ylab=vd,trace.label=NULL,
                pch=c(1,19), col = c('#4ecdc4', '#ff6b6b'))
```

```
  }

  grid(col = "lightgray", lty = "dotted",
         lwd = par("lwd"), equilogs = TRUE) # rejilla para ver mejor

}
```

`one_way` : Realiza ANOVA de un factor de efectos fijos. El procedimiento está basado en el capítulo dedicado a ANOVA de un factor de Pardo y San Martín (2015). La función proporciona una lista con tres dataframes que contienen el valor del estadístico F, su nivel crítico y tres medidas del tamaño del efecto.

- Argumentos:
  - `num_fact`: Vector numérico de longitud 1. Longitud del vector que contiene los factores. Calculado en `ANOVA`.
  - `factor`: Igual que en `ANOVA`.
  - `vd`: Igual que en `ANOVA`.
  - `data`: Igual que en `ANOVA`.
- Salida:
  - Lista con dataframes que resumen la información del ANOVA → Medias por nivel del factor, sumas cuadráticas, datos del contraste de significación y medidas del tamaño del efecto.

Fórmulas empleadas:

$$MC_E = \frac{\sum_j (n_j - 1) S_j^2}{N - J} \quad ; \quad MCI = MC_A = \frac{\sum_j n_j \, (\overline{Y}_j - \overline{Y})^2}{J - 1}$$

```
one_way <- function(num_fact,factor,vd,data){

  N <- nrow(data)                                    # Tamaño muestral
  J <- nlevels(data[,factor])                        # Número de niveles del factor A

 # ------------------------------------------------- Estimador variabilidad intra basado
 #                                                    en varianzas de cada nivel j del factor

  varianzas <- aggregate(data[vd],
                         by=list(data[[factor]]),
                         FUN=var)[[vd]]

  n <- aggregate(data[vd],
                 by=list(data[[factor]]),
                 FUN=length)[[vd]]

  SCE <- sum((n-1)*varianzas)
  glE <- N-J
  MCE <- SCE/glE

  # ------------------------------------------------- Estimador variabilidad inter
  #                                                    basado en las medias del factor
```

```
  medias_A_df <- aggregate(data[vd],
                           by=list(data[[factor]]),
                           FUN=mean)
  medias_A <- medias_A_df[[vd]]              # el df es para la salida

  media_tot <- mean(data[,vd])
  SCI <- sum( n * (medias_A-media_tot)^2  )
  glI <- J-1
  MCI <- SCI/glI

  F_stat <- MCI/MCE

  p_val <- 1-pf(F_stat,glI,glE)

  # ------------------------------------------------- Salida

  anova_plot(num_fact,factor,vd,data,medias_A,J)

 # -------------------------------------------------

  names(medias_A_df) <- c(factor,'medias')

  sumas_cuad <- data.frame('SCT'=SCI+SCE,
                           'SCI'=SCI,
                           'SCE'=SCE)
  contraste <- data.frame('Factores'=factor,
                          'F'=F_stat,
                          'nivel_crítico'=p_val)

  return(list('medias'=medias_A_df,
              'sumas_cuadráticas'=sumas_cuad,
              'contraste'=contraste,
              'tamaño_efecto'=effect_size(num_fact,N,
                                          glE,
                                          glA=glI,F_A=F_stat)))
}
```

`two_way` : Realiza ANOVA de dos factores de efectos fijos y asumiendo mismo número de participantes por combinación de niveles de los factores (diseño equilibrado). El procedimiento está basado en el capítulo dedicado a ANOVA de dos factores de Pardo y San Martín (2015). La función proporciona una lista con tres dataframes que contienen el valor del estadístico F para cada factor y para la interacción, los niveles críticos asociados y tres medidas del tamaño del efecto para cada uno.

- Argumentos:
  - `num_fact`: Vector numérico de longitud 1.Longitud del vector que contiene los factores. Calculado en `ANOVA`.
  - `factor`: Igual que en `ANOVA`.
  - `vd`: Igual que en `ANOVA`.
  - `data`: Igual que en `ANOVA`.
- Salida:

– Lista con dataframes que resumen la información del ANOVA → Medias por combinación de niveles de los factores, sumas cuadráticas, datos de los contrastes de significación y medidas del tamaño del efecto.

Fórmulas empleadas:

$$SC_T = \sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y})^2 \quad ; \quad MC_E = \frac{\sum_j (n_j - 1) S_{jk}^2}{N - JK}$$

$$MC_A = \frac{nK \sum_j (\overline{Y}_{j+} - \overline{Y})^2}{J - 1} \quad ; \quad MC_B = \frac{nJ \sum_k (\overline{Y}_{+k} - \overline{Y})^2}{k - 1} \quad ; \quad MC_{AB} = \frac{SC_T - (SC_A + SC_B)}{(J-1)(K-1)}$$

```
two_way <- function(num_fact,factor,vd,data){

  N <- nrow(data)                                    # Tamaño muestral
  J <- nlevels(data[,factor[1]])                     # Número de niveles del factor A
  K <- nlevels(data[,factor[2]])                     # Número de niveles del factor B
  n <- N/(J*K)                                       # Número de sujetos por combinación jk
                                                     # (se asume anova con diseño equilibrado)

  # ------------------------------------------------

  media_tot <- mean(data[,vd])

  SCT <- sum((data[vd]-media_tot)^2)
  glT <- N-1
 # ------------------------------------------------ Estimador variabilidad intra
 #                                                  basado en varianzas de cada combinación jk

  varianzas <- aggregate(data[vd],
                         by=list(data[[factor[1]]],data[[factor[2]]]),
                         FUN=var)[[vd]]
  SCE <- sum((n-1)*varianzas)
  glE <- N-(J*K)
  MCE <- SCE/glE

 # ------------------------------------------------ Estimador variabilidad inter
 #                                                  basado en medias del factor A
  medias_A <- aggregate(data[vd],
                        by=list(data[[factor[1]]]),
                        FUN=mean)[[vd]]

  SCa <- n*K*sum( (medias_A-media_tot)^2 )
  glA <- J-1
  MCa <- SCa/glA
  F_stat_A <- MCa/MCE

  p_val_A <- 1-pf(F_stat_A,glA,glE)

 # ------------------------------------------------ Estimador variabilidad inter
 #                                                  basado en medias del factor B
```

```r
medias_B <- aggregate(data[vd],
                      by=list(data[[factor[2]]]),
                      FUN=mean)[[vd]]

SCb <- n*J*sum( (medias_B-media_tot)^2 )
glB <- K-1
MCb <- SCb/glB
F_stat_B <- MCb/MCE

p_val_B <- 1-pf(F_stat_B,glB,glE)

# ------------------------------------------------ Estimador variabilidad inter
#                                                  basado en medias del factor interacción

medias_inter_df <- aggregate(data[vd],
                      by=list(data[[factor[1]]],data[[factor[2]]]),
                      FUN=mean)
medias_inter <- medias_inter_df[[vd]]              # el df es para la salida

SCab <- SCT-(SCa+SCb+SCE)
glAB <- (J-1)*(K-1)
MCab <- SCab/glAB
F_stat_AB <- MCab/MCE

p_val_AB <- 1-pf(F_stat_AB,glAB,glE)

# ------------------------------------------------ Ajuste condicional. Compara
#                                                  modelo aditivo e interactivo

aditiv_vs_inter <- ajuste_condicional(factor,vd,
                                      SCT,SCa,SCb,SCE,
                                      glT,glA,glB,glE)

# ------------------------------------------------ Salida

anova_plot(num_fact,factor,vd,data)

# ------------------------------------------------


names(medias_inter_df) <- c(factor,'medias') # nombrar adecuadamente df con medias

sumas_cuad <- data.frame('SCT'=SCT,
                         'SCa'=SCa,
                         'SCb'=SCb,
                         'SCab'=SCab,
                         'SCE'=SCE)
contraste <- data.frame('Factores'=c(factor,paste0(factor[1],'*',factor[2])),
                        'F'=c(F_stat_A,F_stat_B,F_stat_AB),
                        'nivel_crítico'=c(p_val_A,p_val_B,p_val_AB))

tamano_efecto <- data.frame(contraste[1],
                                    effect_size(num_fact,N,
```

```
                                                    glE,
                                                    glA,F_A=F_stat_A,
                                                    glB,F_B=F_stat_B,
                                                    glAB,F_AB=F_stat_AB))
  return(list('ajuste_condicional'=aditiv_vs_inter,
              'medias'=medias_inter_df,
              'sumas_cuadráticas'=sumas_cuad,
              'contraste'=contraste,
              'tamaño_efecto'=tamano_efecto))
}
```

`ajuste_condicional`: Compara el ajuste del modelo simple que solo incluye efectos principales (modelo aditivo) frente al del modelo más complejo, que incluye el componente de interacción (modelo interactivo). Un resultado no estadísticamente significativo indica que no hay evidencia de que el modelo interactivo produzca una reducción significativa de los errores de predicción en comparación con el modelo más simple. En este caso se da un aviso al usuario, indicando que modelo ANOVA generado puede no ser el óptimo. El procedimiento ha sido copiado de Ato y Vallejo (2015) (p.138).

- Argumentos:
  - `factor`: Igual que en `ANOVA`.
  - `vd`: Igual que en `ANOVA`.
  - `SCT,SCa,SCb,SCab`: Vectores numéricos de longitud 1. Suma cuadrática total, del factor A, del factor B y de la interacción AB respectivamente.
  - `glT,glA,glB,glAB`: Vectores numéricos de longitud 1. Grados de libertad totales, del factor A, del factor B y de la interacción AB respectivamente.
- Salida:
  - dataframe con información que resume el contraste de significación del ajuste condicional (modelos enfrentados, valor estadístico F y nivel crítico).

Fórmulas empleadas:

$$F = \frac{\frac{SCE_{aditivo} - SCE_{inter}}{glE_{aditivo} - glE_{inter}}}{\frac{SCE_{inter}}{glE_{inter}}}$$

```
ajuste_condicional <- function(factor,vd,
                               SCT,SCa,SCb,SCE,
                               glT,glA,glB,glE){

  SCE_aditivo <- SCT-(SCa+SCb)
  glE_aditivo <- glT-(glA+glB)

  SCE_inter <- SCE            # SCE de modelo inter = SCT-(SCa+SCb+SCab)
  glE_inter <- glE            # glE de modelo inter = glT-(glA+glB+glAB)

  F_stat <- ( (SCE_aditivo-SCE_inter)/(glE_aditivo-glE_inter) )/(SCE_inter/glE_inter)
  p_val <- 1-pf(F_stat,(glE_aditivo-glE_inter),(glE_inter))
```

```
 adit <-  paste0(vd,'=',factor[1],'+',factor[2])
 inter <- paste0(vd,'=',factor[1],'+',factor[2],'+',factor[1],'*',factor[2])

 comparacion <- data.frame('modelo_aditivo'=adit,
                           'modelo_interactivo'=inter,
                           'F'=F_stat,
                           'nivel_crítico'=p_val)

  if (p_val>0.05) warning('Diferencia no significativa en grado de ajuste con
                          alfa=0,05. El modelo interactivo podría no ser apropiado')

  return(data.frame('modelos'=c(adit,inter),
                    'F'=c(NA,F_stat),
                    'nivel_crítico'=c(NA,p_val)))
}
```

`effect_size` : Calcula tres medidas del tamaño del efecto para todos los factores involucrados en el ANOVA. En función del tipo de ANOVA (con 1 factor o con 2), devuelve unas fórmulas u otras. Todas las fórmulas empleadas se han copiado directamente de los capítulos dedicados a ANOVA de Pardo y San Martín (2015). Se asume que los factores son de efectos fijos. Problema: al comparar resultados con los que se obtienen con software JASP, se observa que omega cuadrado de la función es mayor, siendo la discrepancia bastante grande en el caso del ANOVA de dos factores.

- Argumentos:
  - `num_fact`: Igual que en `one_way` y `two_way`.
  - `N`: Vector numérico de longitud 1 correspondiente al tamaño total de la muestra. Se calcula en `one_way` o `two_way`
  - `F_A,F_B,F_AB`: Vectores numéricos de longitud 1. Estadístico F del factor A, del factor B y de la interacción AB respectivamente.
  - `glE,glA,glB,glAB`: Vectores numéricos de longitud 1. Grados de libertad de los errores, del factor A, del factor B y de la interacción AB respectivamente.
- Salida:
  - dataframe con información que resume la magnitud de los diferentes efectos estudiados.

Fórmulas empleadas:

$$\hat{\eta}^2_{factor} = \frac{gl_{factor}\ F_{factor}}{gl_{factor}\ F_{factor} + N - JK} \quad ; \quad \hat{\omega}^2_{factor} = \frac{gl_{factor}\ (F_{factor} - 1)}{gl_{factor}\ (F_{factor} - 1) + N} \quad ; \quad \hat{\delta}_{factor} = \sqrt{\frac{\hat{\omega}^2_{factor}}{1 - \hat{\omega}^2_{factor}}}$$

```
effect_size <- function(num_fact,N,
                        glE,
                        glA,F_A,
                        glB=NULL,F_B=NULL,
                        glAB=NULL,F_AB=NULL){
  etas <- double()
  omegas <- double()
  deltas <- double()
```

```
  for (gl_F in list(A=c(glA,F_A), B=c(glB,F_B), AB=c(glAB,F_AB))){

      eta <- (gl_F[1]*gl_F[2])/(gl_F[1]*gl_F[2]+glE)
      omega <- (gl_F[1]*(gl_F[2]-1))/(gl_F[1]*(gl_F[2]-1)+N)
      delta <- sqrt( omega/(1-omega) )
      etas <- c(etas,eta); omegas <- c(omegas,omega); deltas <- c(deltas,delta)
  }

  efs <- data.frame('delta_Cohen' = deltas,
                    'eta_cuadrado' = etas,
                    'omega_cuadrado' = omegas)

  return(efs)
}
```

EJEMPLO 1: ANOVA DE UN FACTOR

Este archivo de datos, "Amigos de Facebook", proporciona clasificaciones de preferencia para los perfiles de Facebook. Cinco grupos juzgan cinco perfiles de Facebook idénticos, salvo en un aspecto: el número de amigos para ese perfil. ¿Tiene algún efecto sobre la valoración el número de amigos que muestre el perfil?

Variables:

Friends - Grupo experimental - número de amigos (el número indica el número de cuentas de amigos con el perfil de la maqueta).

Participant - Número de participante.

Score - Calificación de atractivo social del perfil de la maqueta (1 = puntuación más baja, 7 = puntuación más alta).

enlace a base de datos

```
str(data)
```

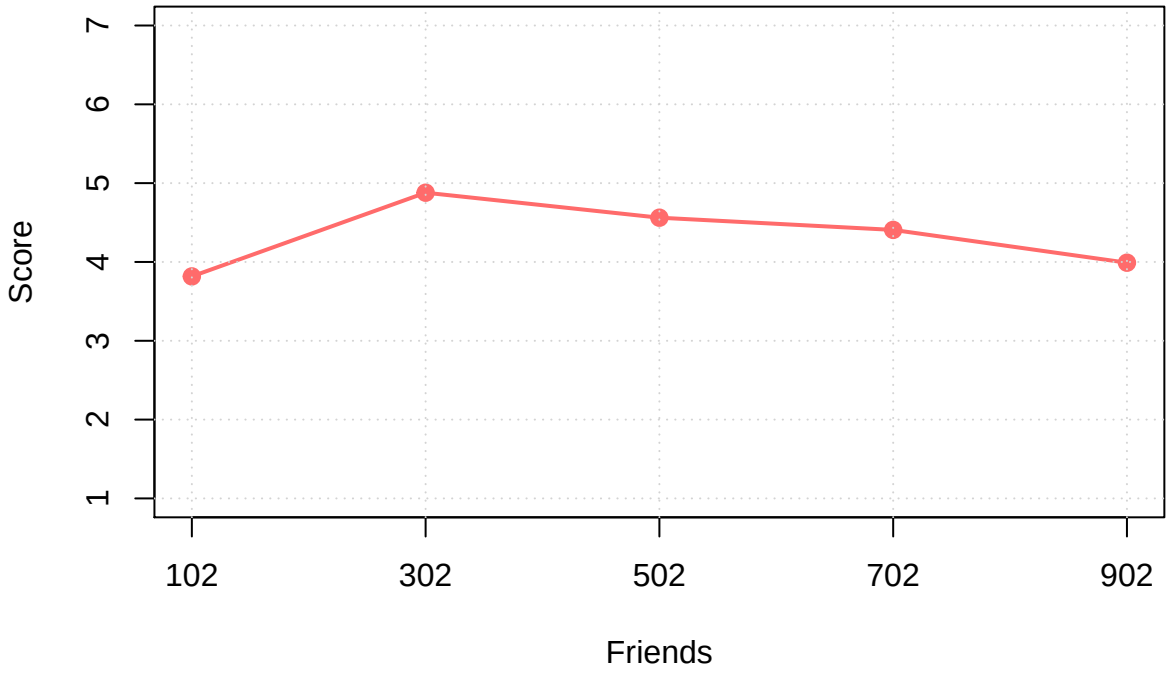
```
## 'data.frame':    134 obs. of  3 variables:
##  $ Friends    : int  102 102 102 102 102 102 102 102 102 102 ...
##  $ Participant: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Score      : num  3.8 3.6 3.2 2.4 4.8 3 4.2 3.6 3.2 3 ...
```

```
head(data)
```

```
##   Friends Participant Score
## 1     102           1   3.8
## 2     102           2   3.6
## 3     102           3   3.2
## 4     102           4   2.4
## 5     102           5   4.8
## 6     102           6   3.0
```

```r
ANOVA(factor='Friends',vd='Score',data=data)
```

```
## Warning in check_factor(factor, data): se convertira Friends en tipo factor
```




```
## $medias
##   Friends   medias
## 1     102 3.816667
## 2     302 4.878788
## 3     502 4.561538
## 4     702 4.406667
## 5     902 3.990476
##
## $sumas_cuadráticas
##       SCT      SCI      SCE
## 1 174.757 19.89023 154.8668
##
## $contraste
##   Factores        F nivel_crítico
## 1  Friends 4.142011   0.003439561
##
## $tamaño_efecto
##   delta_Cohen eta_cuadrado omega_cuadrado
## 1   0.3062538    0.1138165     0.08574887
```

EJEMPLO 2: ANOVA DE DOS FACTORES

Este arichivo de datos, "Ritmo cardíaco", proporciona los ritmos cardíacos de corredores y corredoras y, en general, de participantes sedentarios después de 6 minutos de ejercicio. ¿Afecta el género al ritmo cardíaco promedio? ¿Tiene un efecto sobre el ritmo cardíaco el estilo de vida (corredores y no corredores)? ¿Existe un efecto interactivo del estilo de vida y el género sobre la tasa cardíaca?

Variables:

Gender - Género del participante (Female, Male).

Group - Grupo de "Runners" (con un promedio de más de 15 millas por semana) y grupo de "Control" (generalmente participante sedentario).

Heart.Rate - Ritmo cardíaco después de seis minutos de ejercicio.

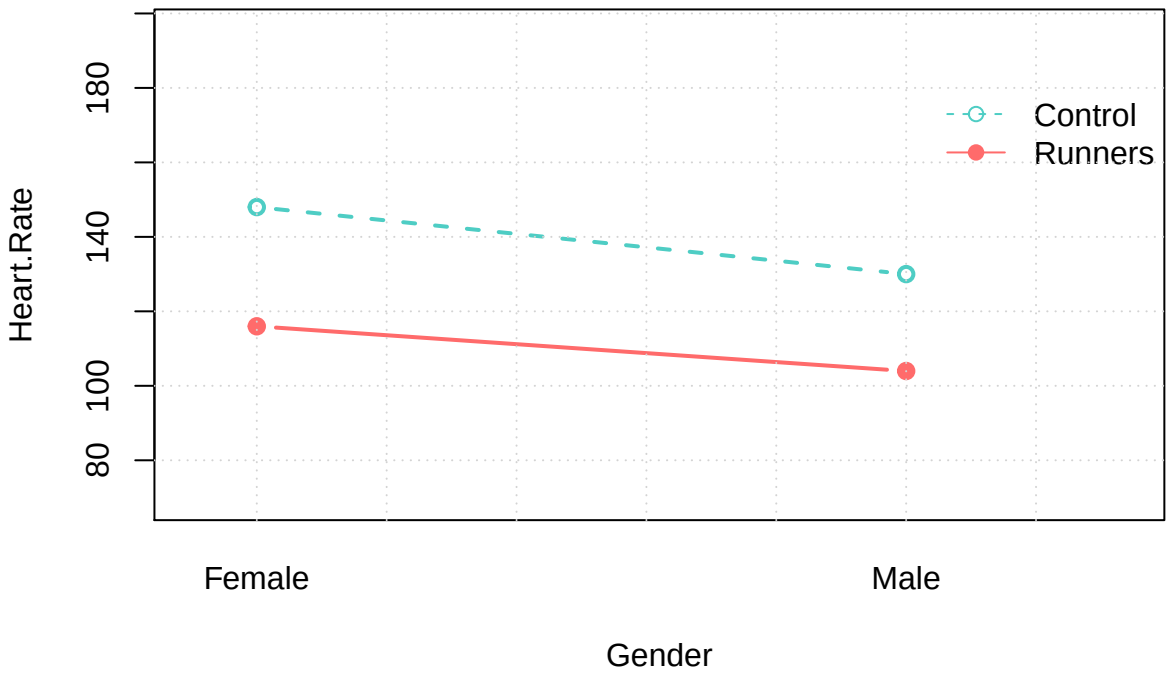
enlace a base de datos

```
str(data2)
```

```
## 'data.frame':    800 obs. of  3 variables:
##  $ Gender    : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Group     : Factor w/ 2 levels "Control","Runners": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Heart.Rate: int  119 84 89 119 127 111 115 109 111 120 ...
```

```
head(data2)
```

```
##   Gender   Group Heart.Rate
## 1 Female Runners        119
## 2 Female Runners         84
## 3 Female Runners         89
## 4 Female Runners        119
## 5 Female Runners        127
## 6 Female Runners        111
```




```
## $ajuste_condicional
##                                 modelos        F nivel_crítico
## 1              Heart.Rate=Gender+Group       NA            NA
## 2 Heart.Rate=Gender+Group+Gender*Group 7.409481   0.006629953
##
## $medias
##   Gender   Group  medias
## 1 Female Control 148.000
## 2   Male Control 130.000
## 3 Female Runners 115.985
## 4   Male Runners 103.975
##
## $sumas_cuadráticas
##        SCT      SCa      SCb     SCab      SCE
## 1 407985.9 45030.01 168432.1 1794.005 192729.8
##
## $contraste
##       Factores          F nivel_crítico
## 1       Gender 185.979949   0.000000000
## 2        Group 695.647040   0.000000000
## 3 Gender*Group   7.409481   0.006629953
##
## $tamaño_efecto
##       Factores delta_Cohen eta_cuadrado omega_cuadrado
## 1       Gender  0.48085854  0.189392817    0.187800726
## 2        Group  0.93183089  0.466361694    0.464756575
```

```
## 3 Gender*Group  0.08950894  0.009222546    0.007948171
```

```
## pdf
##   2
```

Referencias

Pardo, A., y San Martín, R. (2015). Análisis de datos en ciencias sociales y de salud II. Editorial Síntesis.

Ato, M. y Vallejo, G. (2015). Diseños de investigación en Psicología. Ediciones Pirámide.