A2 - Analítica descriptiva e inferencial

Nicolás Caruso

Marzo 2021

Índice

1. Lectura del fichero **2**

2. Rating de los jugadores **4**

2.1. Análisis visual 4

2.2. Intervalo de confianza 5

3. Diferencia entre jugadores **6**

3.1. Pregunta de investigación 6

3.2. Representación visual 6

3.3. Hipótesis nula y alternativa 10

3.3.1. Rating 10

3.3.2. Dribbling 10

3.3.3. Ball control 11

3.4. Método 11

3.5. Cálculo 12

3.6. Tabla de resultados 14

3.7. Interpretación 14

4. Comparación por pares **15**

4.1. Jugador más similar 15

4.2. Muestras 16

4.3. Hipótesis nula y alternativa 16

4.4. Método 16

4.5. Cálculos 16

4.6. Interpretación 17

4.7. Reflexión 17

5. Comparación entre clubes **17**

5.1. Hipótesis nula y alternativa 18

5.2. Método 18

5.3. Cálculos 18

5.4. Resultados e interpretación 19

6. Resumen ejecutivo **19**

---

1. Lectura del fichero

Para la lectura del archivo voy a autilizar la función read.csv del paquete utils que viene por *default* instalado en R. Luego voy a mostrar las primeras filas de la tabla, el nombre de las columnas y un resumen del tipo de datos que aloja cada columna. Finalmente mostraré un *summary* de cada una de ellas.

```
# leo el archivo
df <- read.csv("./fifa_clean.csv", header=T, dec='.')

# muestro las primeras filas
head(df)
```

```
##   ID              Name Nationality National_Position National_Kit
## 1  1 Cristiano Ronaldo    Portugal                LS            7
## 2  2      Lionel Messi   Argentina                RW           10
## 3  3            Neymar      Brazil                LW           10
## 4  4    Luis Su\xe1rez     Uruguay                LS            9
## 5  5      Manuel Neuer     Germany                GK            1
## 6  6            De Gea       Spain                GK            1
##             Club Club_Position Club_Kit Club_Joining Contract_Expiry Rating
## 1    Real Madrid            LW        7   07/01/2009            2021     94
## 2   FC Barcelona            RW       10   07/01/2004            2018     93
## 3   FC Barcelona            LW       11   07/01/2013            2021     92
## 4   FC Barcelona            ST        9   07/11/2014            2021     92
## 5      FC Bayern            GK        1   07/01/2011            2021     92
## 6 Manchester Utd            GK        1   07/01/2011            2019     90
##   Height Weight Preffered_Foot Birth_Date Age Preffered_Position
## 1    185     78          Right 02/05/1985  31              LW/ST
## 2    179     72           Left 06/24/1987  29                 RW
## 3    174     68          Right 02/05/1992  24                 LW
## 4    182     85          Right 01/24/1987  29                 ST
## 5    193     85          Right 03/27/1986  30                 GK
## 6    186     82          Right 11/07/1990  26                 GK
##         Work_Rate Weak_foot Skill_Moves Ball_Control Dribbling Marking
## 1      High / Low         4           5           93        92      22
## 2 Medium / Medium         4           4           95        97      13
## 3   High / Medium         5           5           95        96      21
```

```
## 4   High / Medium         4           4           91        86      30
## 5 Medium / Medium         4           1           48        30      10
## 6 Medium / Medium         3           1           31        13      13
##   Sliding_Tackle Standing_Tackle Aggression Reactions Attacking_Position
## 1             23              31         63        96                 94
## 2             26              28         48        95                 93
## 3             33              24         56        88                 90
## 4             38              45         78        93                 92
## 5             11              10         29        85                 12
## 6             13              21         38        88                 12
##   Interceptions Vision Composure Crossing Short_Pass Long_Pass Acceleration
## 1            29     85        86       84         83        77           91
## 2            22     90        94       77         88        87           92
## 3            36     80        80       75         81        75           93
## 4            41     84        83       77         83        64           88
## 5            30     70        70       15         55        59           58
## 6            30     68        60       17         31        32           56
##   Speed Stamina Strength Balance Agility Jumping Heading Shot_Power Finishing
## 1    92      92       80      63      90      95      85         92        93
## 2    87      74       59      95      90      68      71         85        95
## 3    90      79       49      82      96      61      62         78        89
## 4    77      89       76      60      86      69      77         87        94
## 5    61      44       83      35      52      78      25         25        13
## 6    56      25       64      43      57      67      21         31        13
##   Long_Shots Curve Freekick_Accuracy Penalties Volleys GK_Positioning GK_Diving
## 1         90    81                76        85      88             14         7
## 2         88    89                90        74      85             14         6
## 3         77    79                84        81      83             15         9
## 4         86    86                84        85      88             33        27
## 5         16    14                11        47      11             91        89
## 6         12    21                19        40      13             86        88
##   GK_Kicking GK_Handling GK_Reflexes
## 1         15          11          11
## 2         15          11           8
## 3         15           9          11
## 4         31          25          37
## 5         95          90          89
## 6         87          85          90
```

```
# muestro el nombre de las columnas
colnames(df)
```

```
##  [1] "ID"                 "Name"               "Nationality"
##  [4] "National_Position"  "National_Kit"       "Club"
##  [7] "Club_Position"      "Club_Kit"           "Club_Joining"
## [10] "Contract_Expiry"    "Rating"             "Height"
## [13] "Weight"             "Preffered_Foot"     "Birth_Date"
## [16] "Age"                "Preffered_Position" "Work_Rate"
## [19] "Weak_foot"          "Skill_Moves"        "Ball_Control"
## [22] "Dribbling"          "Marking"            "Sliding_Tackle"
## [25] "Standing_Tackle"    "Aggression"         "Reactions"
## [28] "Attacking_Position" "Interceptions"      "Vision"
## [31] "Composure"          "Crossing"           "Short_Pass"
## [34] "Long_Pass"          "Acceleration"       "Speed"
```

```
## [37] "Stamina"            "Strength"           "Balance"
## [40] "Agility"            "Jumping"            "Heading"
## [43] "Shot_Power"         "Finishing"          "Long_Shots"
## [46] "Curve"              "Freekick_Accuracy"  "Penalties"
## [49] "Volleys"            "GK_Positioning"     "GK_Diving"
## [52] "GK_Kicking"         "GK_Handling"        "GK_Reflexes"
```

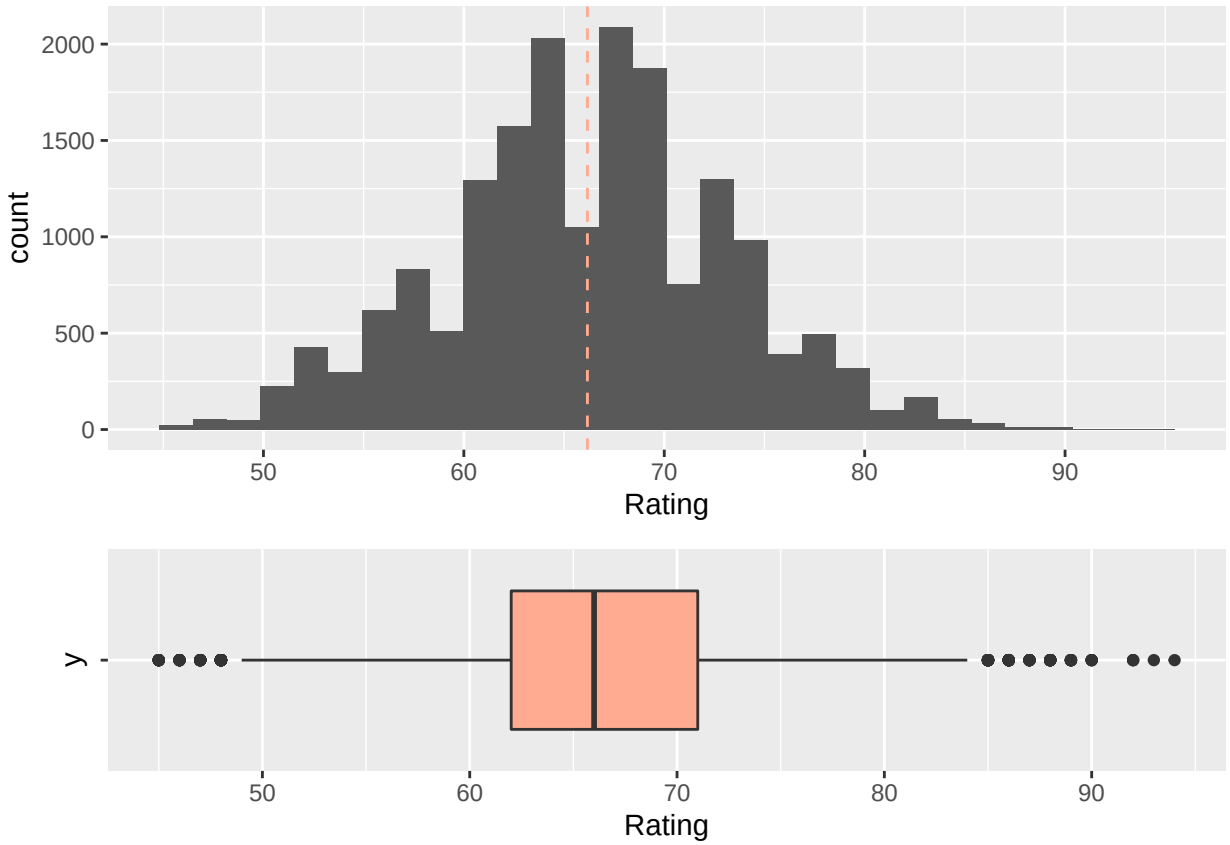
---

2. Rating de los jugadores

2.1. Análisis visual

Vamos a mostrar la distribución de la variable rating. Usaremos un histograma y un *boxplot*.

```
library(ggplot2)
library(gridExtra)
library(egg)

#histograma con el valor de la media
histp<-ggplot(df, aes(x=Rating)) +
      geom_histogram()+
      geom_vline(aes(xintercept=mean(Rating)), col='#FFAB91', linetype='dashed')

# boxplot
boxp<-ggplot(df, aes(y='',x=Rating))+
  geom_boxplot(fill='#FFAB91')

egg::ggarrange(histp, boxp, heights = 2:1)
```



Se decidió realizar una composición de gráficos formada por un histograma y un *boxplt*. Ambos muestran la distribución de valores de la variable. Podemos ver que la variable *Rating* parecería tener una forma bastante similar a la distribución normal con una media en torno a los 65. El *boxplot* confirma esto y muestra que la proporción de valores extremos es relativamente baja.

2.2. Intervalo de confianza

Calcularemos el intervalo de confianza para la variable Rating

```
#establezco el nivel de significación con el nivel de significación
alfa <- 1-0.95

#calculo el desvío estandar
sd <- sd(df$Rating)

#obtengo en numero de muestras
n <- nrow(df)

#calculo el error estandar
SE <- sd / sqrt(n)

#calculo el estadístico
z <- qt( alfa/2, df=n-1, lower.tail=FALSE )

#calculo el limite inferior del intervalo
L <- mean(df$Rating) - z*SE
```

```r
#calculo el límite superior del intervalo
U <- mean(df$Rating) + z*SE

#muestro el resultado del intervalo redondeado a 2 decimales
round( c(L,U), 2)
```

```
## [1] 66.06 66.27
```

```r
#puedo comprobar que esté bien el cálculo usando la función t.test
t.test(df$Rating, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  df$Rating
## t = 1238.9, df = 17587, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  66.06151 66.27088
## sample estimates:
## mean of x
##  66.16619
```

Así pues, el intervalo de confianza del 95 % de la variable *rating* es: [66.06, 66.27]. Podríamos afirmar que si obtenemos infinitas muestras de los jugadores, el 95 % de los intervalos de confianza calculados a partir de estas muestras contendrían el valor medio de la variable *rating*.

---

3. Diferencia entre jugadores

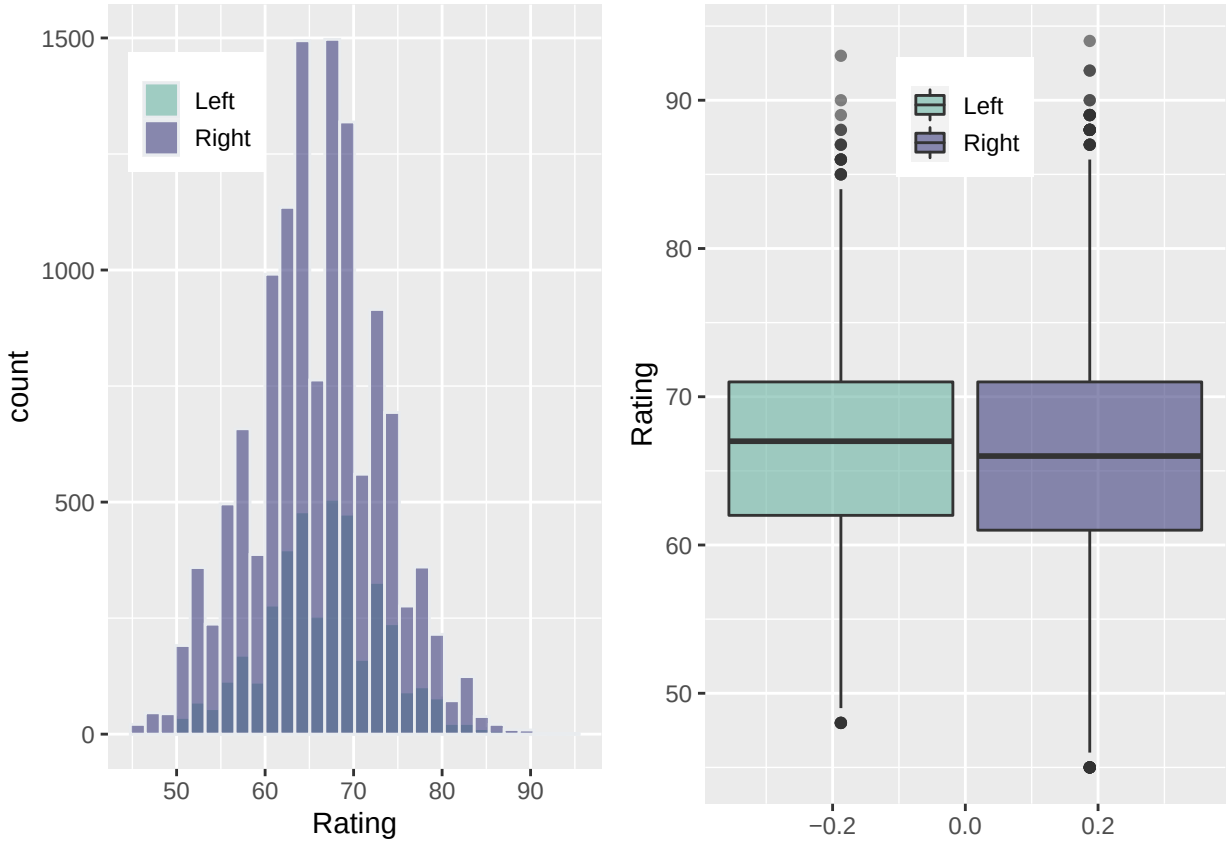
3.1. Pregunta de investigación

Queremos investigar si existen diferencias entre los jugadores zurdos y los diestros en cuanto al nivel de performance en el juego, teniendo en cuenta tres variables: *rating*, *dribbling* y *ball control*. En terminos de pregunta podrías formular las siguientes: 1. ¿Los jugadores zurdos tienen valores, en promedio, significativamente mayores de *rating* que los diestros? 2. ¿Los jugadores zurdos tienen valores, en promedio, significativamente mayores de *dribbling* que los diestros? 3. ¿Los jugadores zurdos tienen valores, en promedio, significativamente mayores de *ball control* rating que los diestros?

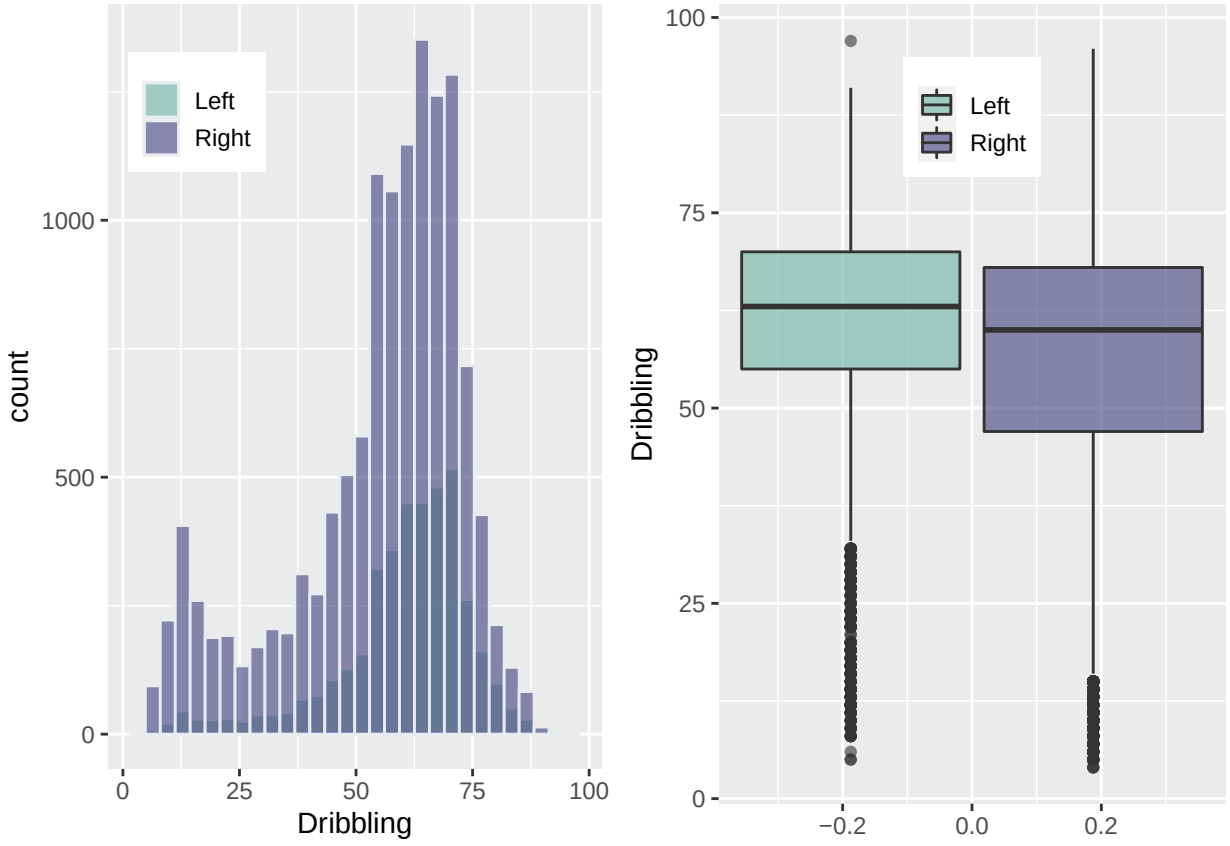
3.2. Representación visual

Vamos a preparar el set de datos que necesitamos para responder la pregunta planteada. Para eso primero vamos a generar dos set de datos: uno con los jugadores zurdos y otro con los jugadores diestros (siempre excluyendo los porteros).

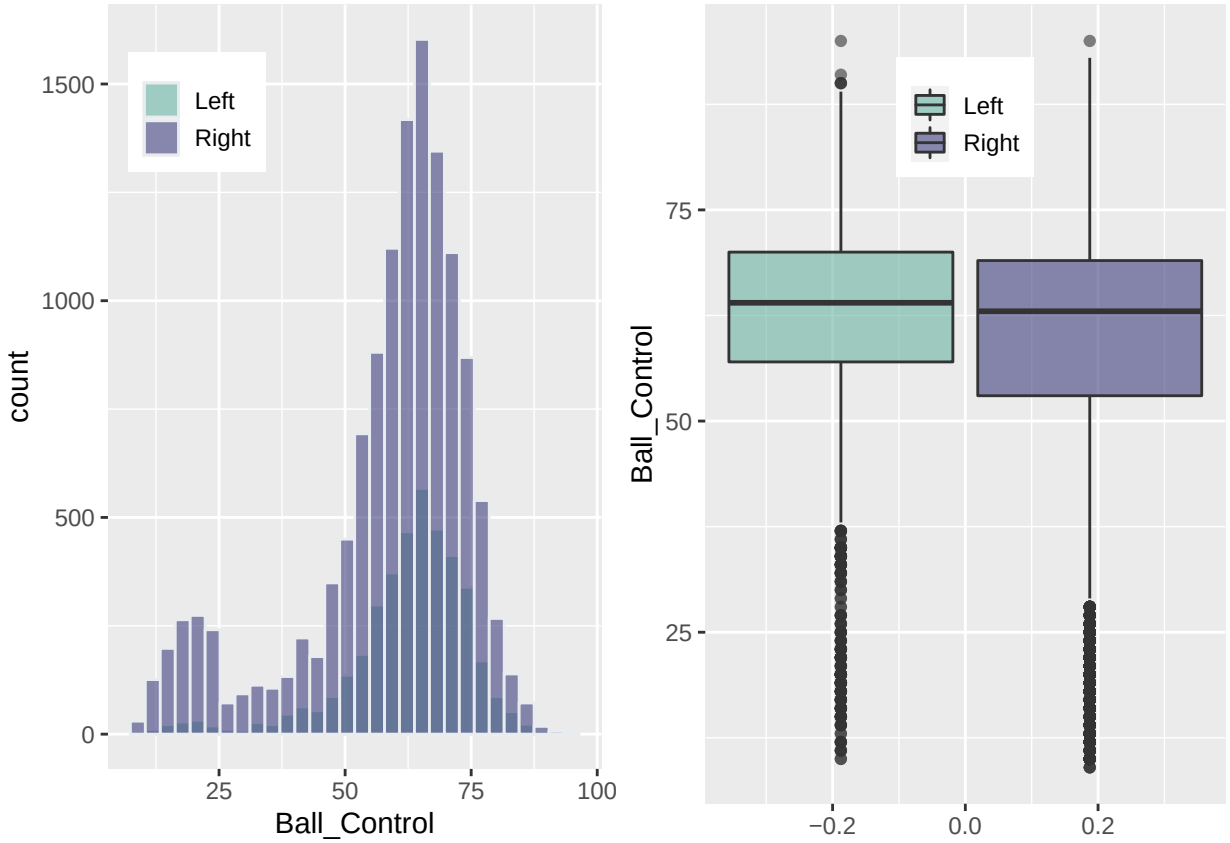
```r
#elimino los porteres
df_sinPort<-subset(df, df$Club_Position != 'GK')

# Creo una función para graficar un histograma con un parametro que indique
# la variable a graficar
plotHist<-function(data, variable){

  a<-ggplot(data, aes_string(x=variable, fill=data$Preffered_Foot)) +
    geom_histogram(color="#e9ecef", alpha=0.6, position = 'identity')+
    scale_fill_manual(values=c("#69b3a2", "#404080"))+
    theme(legend.position = c(0.18, 0.86), legend.key.size = unit(0.5, 'cm'),
          legend.title = element_blank())

  return(a)
}

#idem anterior pero boxplot
plotBox<-function(data, variable){

  a<-ggplot(data, aes_string(y=variable, fill=data$Preffered_Foot)) +
    geom_boxplot(alpha=0.6)+
    scale_fill_manual(values=c("#69b3a2", "#404080"))+
    theme(legend.position = c(0.50, 0.86), legend.key.size = unit(0.5, 'cm'),
          legend.title = element_blank())
    guides(fill=guide_legend(title="Preffered foot"))

  return(a)
}

#plot de Rating
h1=plotHist(df_sinPort, variable = "Rating")
b1=plotBox(df_sinPort,variable = "Rating")

grid.arrange(h1, b1, nrow=1)
```




```
#plot de Dribbling
h1=plotHist(df_sinPort, variable = "Dribbling")
b1=plotBox(df_sinPort,variable = "Dribbling")

grid.arrange(h1, b1, nrow=1)
```




```
#plot de Ball control
h1=plotHist(df_sinPort, variable = "Ball_Control")
b1=plotBox(df_sinPort,variable = "Ball_Control")

grid.arrange(h1, b1, nrow=1)
```



Los gráficos no muestran una diferencia notable entre zurdos y derechos para ninguna de las variables. La distribución parece ser similar, diferenciándose más en la variable *dribbling* (como muestra el *boxplot*). La variable *rating* parece mostrar una distribución bastante similar a la normal, mientras que las otras dos parecen estar un poco más alejada de la normal. De todas maneras, por el tamaño de la muestra y por el Teorema Central del Límite, podemos asumir normalidad. Las variables *dribbling* y *ball control* son las que muestran mayor número de valores extremos, hacia los valores más chicos de la variable (como puede verse en los *boxplots*).

3.3. Hipótesis nula y alternativa

3.3.1. Rating

De manera coloquial:
Hipótesis nula: Los jugadores zurdos tienen menor o igual *Rating* que los derechos.
Hipótesis alternativa: Los jugadores zurdos tienen mayor *Rating* que los derechos.

O también:
H_0: $\mu_{\text{zurdos}} <= \mu_{\text{derechos}}$
H_1: $\mu_{\text{zurdos}} > \mu_{\text{derechos}}$

3.3.2. Dribbling

De manera coloquial:
Hipótesis nula: Los jugadores zurdos tienen menor o igual *Dribbling* que los derechos.

Hipótesis alternativa: Los jugadores zurdos tienen mayor *Dribbling* que los derechos.

O también:
H_0: $\mu_{\text{zurdos}} <= \mu_{\text{derechos}}$
H_1: $\mu_{\text{zurdos}} > \mu_{\text{derechos}}$

3.3.3. Ball control

De manera coloquial:
Hipótesis nula: Los jugadores zurdos tienen menor o igual *Ball control* que los derechos.
Hipótesis alternativa: Los jugadores zurdos tienen mayor *Ball control* que los derechos.

O también:
H_0: $\mu_{\text{zurdos}} <= \mu_{\text{derechos}}$
H_1: $\mu_{\text{zurdos}} > \mu_{\text{derechos}}$

3.4. Método

a) El método que aplicaremos para responder las preguntas planteadas en este ejerircicio será un contraste de hipótesis de dos muestras, dado que se quiere comparar el valor de ciertas variables (*raitng*, *dribbling* y *ball control*) para dos poblacines (jugadores zurdos y jugadres diestros). Ambas, se tratan de muestras independientes ya que no existe relación (desde el punto de vista matemático) entre los jugadores zurdos y los derechos.

b) Dado que el tamaño de la muestra es grande (4022 jugadres zurdos y 12936 diestros) podemos aplicar el Teorema Central del Límite y asumir normalidad en los datos.

c) Consecuentemente, el test que aplicaremos es de tipo paramétrico (haremos un contraste de medias). Dado que queremos saber si el parámetro a comparar es *mayor* en los zurdos que en los diestrs, el contraste que realizaremos es unilateral.

d) Para decidir si podemos asumir homocedasticidad o no, podemos aplicar un test de igualdad de varianzas. Como podemos ver en los resultados, para ninguna de las tres variables podemos asumir hmocedasticidad.

```
#Rating
left<-df_sinPort$Rating[df_sinPort$Preffered_Foot=='Left']
right<-df_sinPort$Rating[df_sinPort$Preffered_Foot=='Right']

var.test(left, right)
```

```
##
##  F test to compare two variances
##
## data:  left and right
## F = 0.84569, num df = 4021, denom df = 12933, p-value = 1.037e-10
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8046699 0.8893922
## sample estimates:
```

```
## ratio of variances
##          0.8456888
```

```
#Dribbling
left<-df_sinPort$Dribbling[df_sinPort$Preffered_Foot=='Left']
right<-df_sinPort$Dribbling[df_sinPort$Preffered_Foot=='Right']

var.test(left, right)
```

```
##
##  F test to compare two variances
##
## data:  left and right
## F = 0.62698, num df = 4021, denom df = 12933, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5965684 0.6593800
## sample estimates:
## ratio of variances
##          0.6269791
```

```
#Ball control
left<-df_sinPort$Ball_Control[df_sinPort$Preffered_Foot=='Left']
right<-df_sinPort$Ball_Control[df_sinPort$Preffered_Foot=='Right']

var.test(left, right)
```

```
##
##  F test to compare two variances
##
## data:  left and right
## F = 0.59095, num df = 4021, denom df = 12933, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5622844 0.6214864
## sample estimates:
## ratio of variances
##          0.5909475
```

3.5. Cálculo

Voy a crear una función que calcule el contraste y devuelve el estadístico, el valor crítico y el valor de P. Hacerlo mediante una función me permitirá aplicarla a las diferentes varables.

```
testT<-function(data, variable, alpha=0.05){

  # filtro la variable de interes
  left<-subset(data[data$Preffered_Foot=="Left",], select = variable)
  right<-subset(data[data$Preffered_Foot=="Right",], select = variable)

  # Calculo la media, el n y la varianza de cada muestra
  mean1 <- mean(left[,1]); n1 <- dim(left)[1]; s1 <- sd(left[,1])
```

```
  mean2 <- mean(right[,1]); n2 <- dim(right)[1]; s2 <- sd(right[,1])

  # calculo el estadistico
  tobs<-(mean1-mean2) / (sqrt((s1^2/n1)+(s2^2/n2)))

  #calculo los grados de libertad
  df<-(((s1^2/n1)+(s2^2/n2))^2) / (((s1^2/n1)^2)/(n1-1)+((s2^2/n2)^2)/(n2-1))

  #calculo el valor critico
  tcrit <- qt(1-alfa, df)

  #calculo el p-value
  pvalue <-pt(abs(tobs), df=df, lower.tail=FALSE)

  return(c(mean1, mean2, n1, n2, tobs, tcrit, pvalue))
}
```

Aplico la función para cada variable y muestro los resultados:

```
# Rating
resR<-testT(df_sinPort, "Rating")

a<-paste("Para la variable Rating los resultados son:")
b<-paste("El estadístico es: ", round(resR[5],3))
c<-paste("El valor crítico es: ", round(resR[6],3))
d<-paste("El p-value es: ", resR[7])

cat(paste(a,b,c,d, sep='\n'))
```

```
## Para la variable Rating los resultados son:
## El estadístico es:  5.934
## El valor crítico es:  1.645
## El p-value es:  1.54877273162864e-09
```

```
cat(paste("\n\n"))
```

```
# Dribbling
resD<-testT(df_sinPort, "Dribbling")

a<-paste("Para la variable Dribbling los resultados son:")
b<-paste("El estadístico es: ", round(resD[5],3))
c<-paste("El valor crítico es: ", round(resD[6],3))
d<-paste("El p-value es: ", resD[7])

cat(paste(a,b,c,d, sep='\n'))
```

```
## Para la variable Dribbling los resultados son:
## El estadístico es:  18.138
## El valor crítico es:  1.645
## El p-value es:  1.92308680577809e-72
```

```r
cat(paste("\n\n"))
```

```r
# Ball control
resBC<-testT(df_sinPort, "Ball_Control")

a<-paste("Para la variable Ball Control los resultados son:")
b<-paste("El estadístico es: ", round(resBC[5],3))
c<-paste("El valor crítico es: ", round(resBC[6],3))
d<-paste("El p-value es: ", resBC[7])

cat(paste(a,b,c,d, sep='\n'))
```

```
## Para la variable Ball Control los resultados son:
## El estadístico es:  15.182
## El valor crítico es:  1.645
## El p-value es:  1.07193804956118e-51
```

3.6. Tabla de resultados

Mostramos los resultados en la tabla como solicita el enunciado. El parámetro "HOLD_position" deja la tabla en el lugar donde se la llama, si no se usa la tabla pasa al encabezado de la página.

```r
library(kableExtra)

out <- data.frame( var=c("Rating", "BallControl", "Dribbling"),
                   mean_Left=c(resR[1],resD[1],resBC[1]),
                   mean_Right=c(resR[2],resD[2],resBC[2]),
                   n_Left=c(resR[3],resD[3],resBC[3]),
                   n_Right=c(resR[4],resD[4],resBC[4]),
                   obs_value=c(resR[5],resD[5],resBC[5]),
                   critical=c(resR[6],resD[6],resBC[6]),
                   pvalue=c(resR[7],resD[7],resBC[7]))

out %>% kable() %>% kable_styling(latex_options = 'HOLD_position')
```

var	mean_Left	mean_Right	n_Left	n_Right	obs_value	critical	pvalue
Rating	66.58155	65.85820	4022	12934	5.933765	1.645065	0
BallControl	60.15266	55.09688	4022	12934	18.137562	1.645036	0
Dribbling	62.16335	58.47727	4022	12934	15.181923	1.645030	0

3.7. Interpretación

Para las tres variables analizadas y dado el p-value inferior al nivel de significancia fijado (0.05), rechazamos las hipótesis nulas. Por lo tanto podemos afirmar, con un nivel de significación del 95 %, que en promedio los valores de *Rating*, *Dribbling* y *Ball_Control* son mayores en jugadores zurdos que en derechos.

4. Comparación por pares

4.1. Jugador más similar

Primero definimos la función que calcula la distancia euclídea entre dos vectores.

```
# definimos la funcion
euclidean <- function(x1, x2){
  return (sqrt(sum((x1-x2)^2)))
}

a<-c(2,3,4)
b<-c(33,4,5)

euclidean(a,b)
```

```
## [1] 31.03224
```

Definimos la función "my.nn" que encuentra el jugador Y más similar a X, buscando en un *data frame*.

```
my.nn <- function(x, sample){
  #vector para guardar los valores de distancia euclidea
  euc<-c()

  #vector para guardar los indices
  ind<-c()

  #iteracion
  for (i in 1:dim(sample)[1]) {
    euc[i]<-euclidean(x, sample[i, c("Age","Weight", "Height")])
    ind[i]<-i
  }

  #creo un data frame
  a<-data.frame(euc,ind)

  #devuelvo el indice para el menor valor de distancia euclidea
  return(a[order(a$euc),][1,]$ind)
}
```

Una vez creada la función "my.nn" podemos generar la funcion "my.nn.sample" para trabajar con dos set de datos (zurdos y diestros). Voy a trabajar con un subset de datos como sugiere el enunciad ya que mi computadora no es muy buena en recursos.

```
# trabajo con un subset de 100 dats de jugadores zurdos y 200 de derechs
left<-subset(df_sinPort[df_sinPort$Preffered_Foot=="Left",])[0:100,]
right<-subset(df_sinPort[df_sinPort$Preffered_Foot=="Right",])[0:200,]

#construyo la funcion
my.nn.sample<-function(sample1, sample2){
```

```
  #vector para alojar los resultados de cada my.nn
  res<-c()
  for (i in 1:dim(sample1)[1]) {
    #cada uno de las triadas edad, peso, altura con la que comparar en el set
    p<-sample1[i, c("Age","Weight", "Height")]
    #resultados de my.nn
    res[i]<-my.nn(p,sample2)
  }

  #devuelvo un set de 100 diestros
  return(sample2[res,])
}
```

4.2. Muestras

Obtenemos las muestras finales con las que trabajaremos, aplicando las funciones antes construidas.

```
Left.sample<-left
Right.sample<-my.nn.sample(left, right)
```

4.3. Hipótesis nula y alternativa

Las hipótesis son las mismas que en el ejercicio anterior, pero acá las muestras son diferentes. Ahora comparamos jugadores que son más parecidos entre sí (ver reflexión final).

De manera coloquial:
Hipótesis nula: Los jugadores zurdos tienen menor o igual *Rating* que los derechos.
Hipótesis alternativa: Los jugadores zurdos tienen mayor *Rating* que los derechos.

O también:
H_0: $\mu_{\text{zurdos}} <= \mu_{\text{derechos}}$
H_1: $\mu_{\text{zurdos}} > \mu_{\text{derechos}}$

4.4. Método

Para este ejercicio aplicaremos un contraste de medias para muestras emparejadas. Hemos escojido este método ya que las muestras ahora no son independiente, debido a la manera en la que se obtuvieron, y mantienen cierto grado de dependencia entre ellas. Dado el tamaño de las muestras, podemos asumir normalidad.

4.5. Cálculos

```
#establecemos el alpha
alpha<-0.05

#obtenemos el vectr de diferencias sobre el que aplicaremos el test
d<-Left.sample$Rating - Right.sample$Rating
```

```r
#calculamos la media de las diferencias
mean <- mean(d)

#su desviación estandar
sd <- sd(d)

#el N
n <- length( d )

#Estadístico de contraste (comparamos con una mu=0)
tobs <- (mean-0)/(sd/sqrt(n))

#Región de aceptación
tcrit <- qt(1-alfa, df=n-1)

#Cálculo del valor p
pvalue <- pt(tobs, lower.tail=FALSE, df=n-1)
```

Muestro los resultados

```r
a<-paste("El estadístico es: ", round(tobs,3))
b<-paste("El valor crítico es: ", round(tcrit,3))
c<-paste("El p-value es: ", pvalue)

cat(paste(a,b,c, sep='\n'))
```

```
## El estadístico es:  -6.312
## El valor crítico es:  1.66
## El p-value es:  0.999999996087869
```

4.6. Interpretación

En este caso el p-value es mayor que el alpha establecido, por lo que no se puede rechazar la hipótesis nula. Consecuentemente, no podemos afirmar que los jugadores zurdos tengan mejor *Rating* que los derechos cuando son comparados jugadores similares en edad, peso y altura.

4.7. Reflexión

Es interesante la diferenciación entre el contraste realizado en el ejercicio 3 con el realizado en este ejercicio, ya que se arriba a conclusiones opuestas. Desde mi punto de vista, la comparación realizada aquí tiene más sentido ya que se están comparando jugadores zurdos y derechos, pero similares en edad, altura y peso. En el contraste del ejercicio 3 no había diferenciación en ests aspectos. Entonces, parece tener más sentido la conclusión de que el *Rating* de los jugadores no difiere entre zurdos y derechos, cuando se compara jugadores similares en aspectos físicos y etarios. Desde mi punto de viste esta contraste es más instructivo y pertinente que el realizado en el ejereicio anterior.

---

5. Comparación entre clubes

5.1. Hipótesis nula y alternativa

De manera coloquial:
Hipótesis nula: La proporción de jugadores con *Rating* mayor a 90 es igual en los equipos de Madrid que en los de Barcelona.
Hipótesis alternativa: La proporción de jugadores con *Rating* mayor a 90 es distinta en los equipos de Madrid que en los de Barcelona.

O también:
H_0: $p_{\text{Barcelona}} = p_{\text{Madrid}}$
H_1: $p_{\text{Barcelna}} \neq p_{\text{Madrid}}$

5.2. Método

En este ejericicio se busca comparar dos muestras independientes (jugadores de Barcelona y jugadores de Madrid) en terminos de la *proporción* de jugadores con *Rating* mayor a 90. Dado que se trata de un contraste de proporciones, utilizaremos un contraste de muestras sobre la proporción. Nuevamente, dado el tamaño de las muestras podemos asumir normalidad. Como indica el enunciado, el contraste solo busca ver si hay diferencias en las proporcines, por lo que aplicaremos un test a dos colas. El enunciado no es claro sobre si incorporar o no los porteros, así que tomaré el criterio usado hasta ahora que trabajar con la base de dats sin los porteros.

5.3. Cálculos

```
# establezco el alpha
alpha<-0.03

# muestras de Barcelona y Madrid
madrid<-subset(df_sinPort, Club=="Atl\xe9tico Madrid" | Club=="Real Madrid")
barcelona<-subset(df_sinPort, Club=="FC Barcelona" | Club=="RCD Espanyol")

# separo las que tienen rating >90
x1 <- barcelona[barcelona$Rating>90,]
x2 <- madrid[madrid$Rating>90,]

# calculo las proporciones
p1 <- dim(x1)[1]/dim(barcelona)[1]; p1
```

```
## [1] 0.05084746
```

```
p2 <- dim(x2)[1]/dim(madrid)[1]; p2
```

```
## [1] 0.01612903
```

```
# N
n1<-dim(barcelona)[1]
n2<-dim(madrid)[1]
```

```r
#calculo p
p<-(dim(barcelona)[1]*p1 + dim(madrid)[1]*p2) / (dim(barcelona)[1]+dim(madrid)[1])

#calculo el estadistico
zobs <- (p1-p2)/( sqrt(p*(1-p)*(1/n1+1/n2)) )

#calculo las zonas criticas
zcritL <- qnorm(alfa/2, lower.tail = F)
zcritU <- qnorm(1-alfa/2, lower.tail = F)

#calculo el p-value
pvalue <-pnorm(zobs,lower.tail=FALSE)*2
```

5.4. Resultados e interpretación

Muestro los resultados

```r
a<-paste("El estadístico es: ", round(zobs,3))
b<-paste("El valor crítico inferior es: ", round(zcritL,3))
c<-paste("El valor crítico superior es: ", round(zcritU,3))
d<-paste("El p-value es: ", pvalue)

cat(paste(a,b,c,d, sep='\n'))
```

```
## El estadístico es:  1.068
## El valor crítico inferior es:  1.96
## El valor crítico superior es:  -1.96
## El p-value es:  0.285653211277346
```

El p-value es mayor que el alpha establecido por lo que no podemos rechazar la hipótesis nula. Consecuentemente no podemos afirmar que la proporción de jugadores con *rating* mayor a 90 sea distinto en los clubes de Barcelona que en los de Madrid.

---

6. Resumen ejecutivo

En el primer ejercicio hemos realizado tres contrastes de hipótesis para ver si los jugadores zurdos mostraban valores mayores que los diestros para las variables *rating*, *dribbling* y *ball control*. Para las tres variables, nuestros análisis mostraron que los jugadores zurdos tenían valores significativamente mayores, trabajando con un nivel de significación de 95 %.
En el segundo ejercicio testeamos la misma hipótesis para la variable *rating* pero en este caso comparando jugadores similares en cuanto a edad, peso y altura. En este caso, no se pudo rechazar la hipótesis nula; por lo que no se puede afirmar que ls jugadores zurdos tengan mejor *performance* en cuanto a *rating* que los diestros cuando se compara jugadores similares en edad y físico. Para este caso también se trabajó con un nivel de significación del 95 %.
Por último hemos realizad un contraste para saber si la proporción de jugadores con *rating* mayor a 90 es igual o diferente entre los equipos de Barcelona y Madrid. No pudimos rechazar la hipótesis nula, por lo cual no se puede afirmar que las proporciones difieran entre equipos. En este ejerciccio se trabajó con un nivel de significación del 98 %.