Analítica descriptiva e inferencial del Fifa 2017

Jesús González

23/4/2021

Contents

```
library(ggplot2)
library(dplyr)
library(gridExtra)
library(kableExtra)
```

1 Introducción

Esta actividad se centra en el análisis estadístico descriptivo e inferencial del conjunto de datos del videojuego Fifa 2017, así como estadísticas reales de jugadores de fútbol. El conjunto inicial contiene más de 17.500 registros repartidos en 53 variables.

Se parte de un fichero, fifa_clean.csv, en donde ya se ha realizado tareas de limpieza y consolidación del dato. En esta actividad, nos interesa investigar la puntuación del jugador (Variable Rating) y otras variables como el control de la pelota (variable Ball_Control) y la técnica del jugador (variable Dribbling). Se asume que los datos contenidos en el fichero, *constituyen una muestra representativa de los jugadores de la última década (población).*

2 Lectura del fichero

```
fifa.df <- read.csv(file="./Data/fifa_clean.csv", encoding = "UTF8",
                    stringsAsFactors = F, header = T)
paste0("Número de observaciones: ", nrow(fifa.df),". Número de variables: ",
       ncol(fifa.df))
```

```
## [1] "Número de observaciones: 17588. Número de variables: 54"
```

Verificamos si existen duplicados en el identificador.

```
if (length(fifa.df$ID) == length(unique(fifa.df$ID))) {
  print("No hay diferencias en el ID")} else {print("Existen diferencias en el ID")}
```

```
## [1] "No hay diferencias en el ID"
```

Convertimos a factor los literales.

```
column_name <- c('Nationality','National_Position', 'Club', 'Club_Position',
                 'Preffered_Foot', 'Preffered_Position', 'Work_Rate')
for (col in column_name) {
  fifa.df[,col] <- factor(fifa.df[,col])
}
```

Para el estudio, se va a investigar las variables: *Rating*, *Ball_Control* y *Dribbling*, por ello se va a realizar una subselección del dataset original, para quedarnos con las que necesitamos para nuestro estudio.

```
column_name <- c('ID', 'Name', 'Nationality', 'Preffered_Foot', 'Club_Position',
                 'Rating', 'Ball_Control', 'Dribbling' )
fifa.df.sub <- fifa.df[, column_name]
```

Con el fin de poder comprobar si los datos se han cargado correctamente, revisamos las primeras líneas de las observaciones cargadas.

```
head(fifa.df.sub, 5)
```

```
##   ID              Name Nationality Preffered_Foot Club_Position Rating
## 1  1 Cristiano Ronaldo    Portugal          Right            LW     94
## 2  2      Lionel Messi   Argentina           Left            RW     93
## 3  3            Neymar      Brazil          Right            LW     92
## 4  4       Luis Suárez     Uruguay          Right            ST     92
## 5  5      Manuel Neuer     Germany          Right            GK     92
##   Ball_Control Dribbling
## 1           93        92
## 2           95        97
## 3           95        96
```

```
## 4           91        86
## 5           48        30
```

A continuación, se expone una descripción de las variables cargadas:

- **Name**: Nombre del jugador. Literal.
- **Nationality**: Nacionalidad del jugador. Se ha convertido a factor.
- **Preffered_Foot**: Variable binaria que describe si el jugador es diestro (valor right) o zurdo (valor left).
- **Clup_Position**: Posición de juego del jugador en club.
- **Rating**: Valoración global de juego del jugador. Tiene un rango de valores de 0 a 100.
- **Ball_Control**: Valoración del control del balón del jugador. Tiene un rango de valores de 0 a 100.
- **Dribbling**: Valoración de la técnica del jugador.Tiene un rango de valores de 0 a 100.

Revisamos las estadísticas básicas.

```
summary(fifa.df.sub)
```

```
##        ID            Name              Nationality    Preffered_Foot
##  Min.   :    1   Length:17588       England  : 1618   Left : 4094
##  1st Qu.: 4398   Class :character   Argentina: 1097   Right:13494
##  Median : 8794   Mode  :character   Spain    : 1008
##  Mean   : 8794                      France   :  974
##  3rd Qu.:13191                      Brazil   :  921
##  Max.   :17588                      Italy    :  751
##                                     (Other)  :11219
##  Club_Position      Rating       Ball_Control     Dribbling
##  Sub    :7492   Min.   :45.00   Min.   : 5.00   Min.   : 4.0
##  Res    :3146   1st Qu.:62.00   1st Qu.:53.00   1st Qu.:47.0
##  RCB    : 633   Median :66.00   Median :63.00   Median :60.0
##  GK     : 632   Mean   :66.17   Mean   :57.97   Mean   :54.8
##  LCB    : 631   3rd Qu.:71.00   3rd Qu.:69.00   3rd Qu.:68.0
##  LB     : 549   Max.   :94.00   Max.   :95.00   Max.   :97.0
##  (Other):4505
```

En la variable Nationality podemos ver que los grupos de jugadores que dominan el dataset (mayor frecuencia) son de nacionalidad inglesa, argentina y española. Aunque en Otros, tenemos una agrupación mayor que el top 6 de nacionalidades que aparecen descritas, por lo que no hay relevancia estadística.

En cambio, Preffered_Foot si tenemos un balanceo hacia jugadores diestros en una proporción de 3 a 1 que se tiene que tener en cuenta respecto al sesgo que puede plantear.

Para Club_Position, el valor Sub es mayoritario.

En cuanto a las tres variables cuantitativas, Rating, Ball_control y Dribbling, la primera de ellas, presenta un buen centrado de la población respecto a la media, cosa que no ocurre en las otras dos, que se encuentra desbalanceadas hacia la izquierda.

Vamos a estudiar estas variables con más atención.

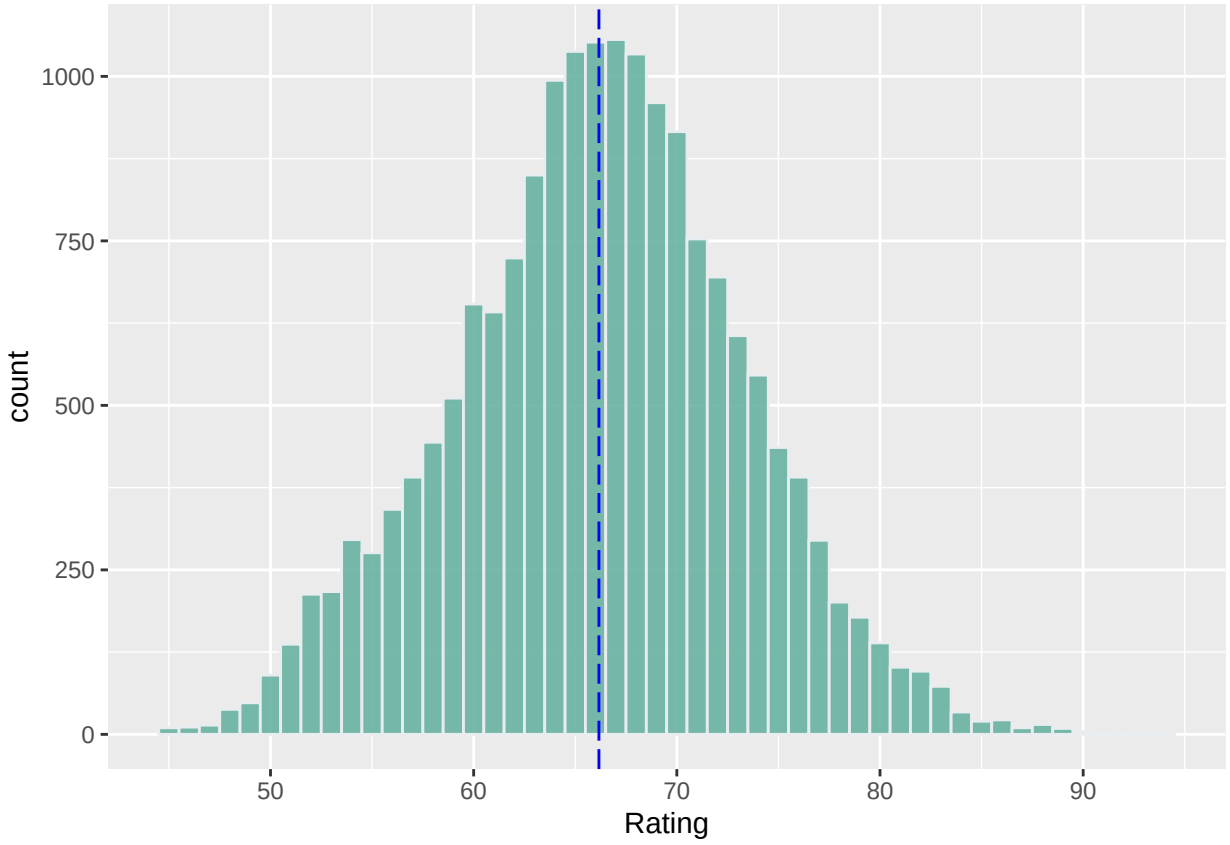
3 Rating de los jugadores

En el primer apartado de este estudio, interesa investigar la variable Rating, que muestra la puntuación del jugador.

3.1 Análisis visual

A continuación se va a mostrar visualmente que distribución sigue esta variable.

```
ggplot(fifa.df.sub, aes(x= Rating)) +
  geom_histogram(binwidth = 1, fill="#69b3a2", color="#e9ecef", alpha=0.9) +
  geom_vline(xintercept = mean(fifa.df.sub$Rating), colour="blue", linetype = "longdash")
```



De la observación de la salida del histograma, podemos comprobar como la variable puntuación sigue una distribución muy similar a la normal y prácticamente simétrica. Se muestra con la línea azúl de referencia que indica la media de la serie. Visualmente se comprueba que la distribución se encuentra casi centrada en la media.

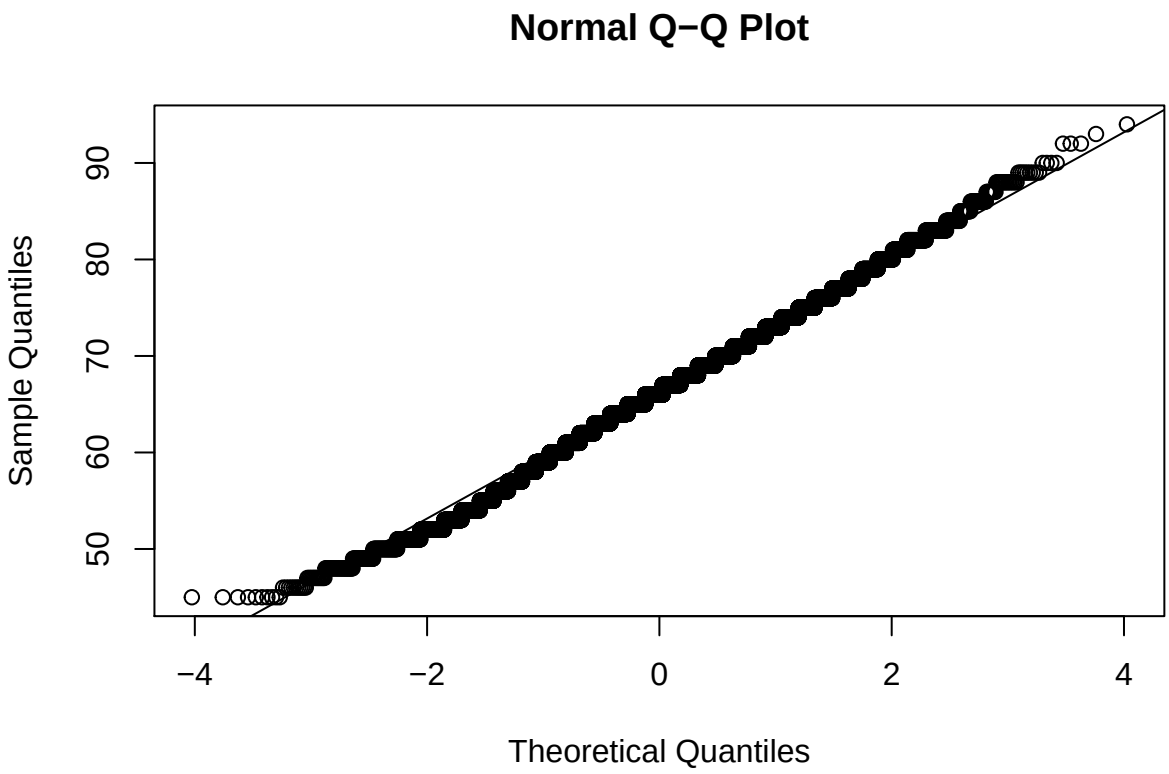
```
summary(fifa.df.sub$Rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   45.00   62.00   66.00   66.17   71.00   94.00
```

Efectivamente, esta simetría la podemos comprobar numéricamente al fijarnos en los valores de media y mediana, podemos ver que se encuentran prácticamente solapadas.

Podemos apoyar aún más nuestra afirmación realizando un gráfico Q-Q para esta variable.

```
qqnorm(fifa.df.sub$Rating)
qqline(fifa.df.sub$Rating)
```



En donde vemos que el error de los puntos de observación con respecto a la recta es mínimo y localizados en los extremos.

3.2 Intervalo de confianza

Vamos a establecer el intervalo de confianza de esta variable a partir de su media, y calculando su desviación estándar.

```
NC <- 0.95
n <- length(fifa.df.sub$Rating)
print(paste('Número de observaciones: ', n))
```

```
## [1] "Número de observaciones:  17588"
```

```
# Cálculo de la desviación estándar
sd <- sd(fifa.df.sub$Rating)

# Cálculo de z_alpha/2
#z_alpha_1_2 <- qnorm(0.025)

alpha <- 1-NC
SE <- sd / sqrt(n)
z <- qnorm(alpha/2, lower.tail = F)
L <- mean(fifa.df.sub$Rating) - z*SE
U <- mean(fifa.df.sub$Rating) + z*SE
cat(paste('Los niveles de confianza se encuentran definidos \n',
            'en el rango [',round(L, 2),', ', round(U, 2), '] \n',
            'para un nivel de confianza del ', NC*100, '%'))
```

```
## Los niveles de confianza se encuentran definidos
##  en el rango [ 66.06 ,  66.27 ]
##  para un nivel de confianza del  95 %
```

Para acabar de confirmar la normalidad de la variable, a más a más de las observaciones de la serie, muy superior en número a los 30 valores, nos podemos apoyar en el **teorema del límite centrar** (TLC). Por lo tanto, y debido a que la distribución sigue una normal, por simetría, el nivel crítico se calcula con la mitad de alfa para un nivel de confianza (NC) del 95%.

4 Diferencias entre jugadores

En este apartado, se va a comprobar si los jugadores zurdos tienen mejor control de la pelota (variable Ball_Control), valoración (Rating) y mejor Dribbling que lo diestros. Como paso previo vamos a diferenciar en dos subsets a los jugadores zurdos de los diestros, eliminando a los porteros (Código GK de la variable Club_Position).

```
left_gamers <- fifa.df.sub %>%
  filter(Club_Position!='GK' & Preffered_Foot=='Left') %>%
  select('ID', 'Rating', 'Ball_Control', 'Dribbling')

right_gamers <- fifa.df.sub %>%
  filter(Club_Position!='GK' & Preffered_Foot=='Right') %>%
  select('ID', 'Rating', 'Ball_Control', 'Dribbling')
```

4.1 Pregunta de investigación

La pregunta que se quiere responder es si los valores de las tres variables a estudiar son **significativamente mayores** entre los jugadores zurdos y los diestros. Estamos ante un caso de contraste unilateral de dos muestras poblacionales independientes.

Por tanto, definiendo la hipótesis nula como que no hay diferencias por el hecho de ser diestros o zurdos, tendríamos que la hipótesis alternativa es aquella en la que los zurdos consiguen mejor puntuación en las tres variables.

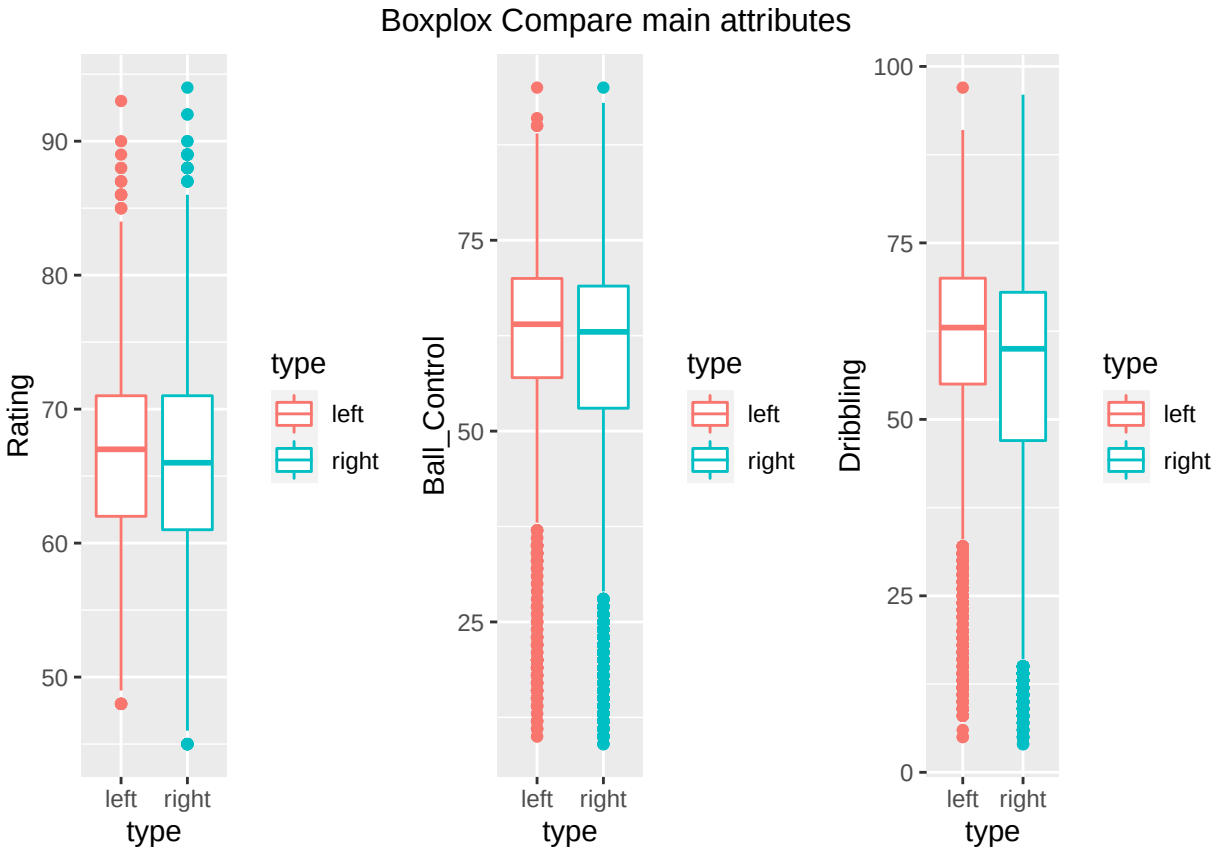
Distribución unilateral por la derecha.

4.2 Representación visual

Veamos el comportamiento de las tres variables del estudio en función de si los jugadores son zurdos o diestros

```
left_gamers$type <- 'left'
right_gamers$type <- 'right'
columns_names <- c('Rating', 'Ball_Control', 'Dribbling', 'type')
data <- merge(x=left_gamers[, columns_names], y =right_gamers[,columns_names], all=T )
```

```
p1 <- ggplot(data=data,aes(x=type, y=Rating,color=type))+geom_boxplot()
p2 <- ggplot(data=data,aes(x=type, y=Ball_Control,color=type))+geom_boxplot()
p3 <- ggplot(data=data,aes(x=type, y=Dribbling,color=type))+geom_boxplot()
grid.arrange(p1, p2, p3, ncol=3, top= 'Boxplox Compare main attributes')
```



A nivel general, vemos como en el caso de los zurdos, las medianas de las tres variables están por encima en puntuación que la de los diestros (valor superior indica mejor puntuación). Es destacable también, la presencia de valores extremos inferiores para el control de la pelota y del juego, sobretodo en jugadores zurdos. Para estas dos variables, tenemos también valores extremos superiores para los jugadores zurdos, lo que indica que hay jugadores zurdos con muy buenas puntuaciones en control del balón y juego. En cuanto al rating, los valores extremos superiores e inferiores se encuentran equilibrados para los dos tipos de jugadores.

4.3 Hipótesis nula y alternativa

A partir de la pregunta realizada definimos la hipótesis nula como que no hay diferencias por el hecho de ser diestros o zurdos. En contraposición, tendríamos que la hipótesis alternativa es aquella en la que los zurdos consiguen mejor puntuación en las tres variables.

Planteado de un modo más matemático:

$$\begin{cases} H_0: & \mu_{left} <= \mu_{right} \\ H_1: & \mu_{left} > \mu_{right} \end{cases}$$

Para cada una de las tres variables observadas, Rating, Ball control y Dribbling, y dentro del intervalo de confianza (NC) del 95 %.

4.4 Método

En este apartado vamos a responder una serie de cuestiones que nos ayudarán a escoger el método adecuado para validar la hipótesis planteada.

- **Tipo de contraste**. En este estudio estamos valorando dos muestras independientes de tamaños diferentes. Para ello realizaremos un **contraste sobre la media**.

- **Normalidad de la muestra**: El número de muestras es suficientemente grande como para poder aplicar el teorema del límite central (TLC) que dice que la media poblacional **se comporta** como una normal si el número de muestras es suficientemente grande, estimándose como suficiente para la gran mayoría de los casos cuando se superan las 30 observaciones.

Propiedades del test a aplicar:

- **Test paramétrico o No paramétrico.** La normalidad de la muestra justificadas en el punto anterior nos permite afirmar que podemos aplicar un test paramétrico.
- **Test bilateral o Unilateral:** Si examinamos la hipótesis alternativa (H_1), estamos hablando de un contraste unilateral por la derecha.
- **Homocedasticidad o heterocedasticidad:** Para resolver este parámetro nos fijaremos en la variabilidad de las muestras aplicando un F test.

```
columns_names <- c('Rating', 'Ball_Control', 'Dribbling')
var_table <- data.frame(matrix(ncol=5, nrow = 0))
column_names_table <- c('Variable', 'F test', 'p valor', 'NC1', 'NC2')
for (col in columns_names){
  out <- var.test(x=left_gamers[,col], y=right_gamers[,col])
  var_table <- rbind(var_table, c(col, round(out$statistic,4),
                                  out$p.value, round(out$conf.int[1],4),
                                  round(out$conf.int[2],4)))
}
colnames(var_table) <- column_names_table
var_table
```

```
##       Variable F test              p valor    NC1    NC2
## 1       Rating 0.8457 1.03733283647206e-10 0.8047 0.8894
## 2 Ball_Control 0.5909  1.1193822394462e-85 0.5623 0.6215
## 3    Dribbling  0.627 1.24119263217829e-68 0.5966 0.6594
```

El test **F de Snedecor** establece como hipótesis nula la homogeneidad de las varianzas. Los p-valores encontrados nos indican que podemos rechazar la H_0, encontrándonos con que las muestras de las tres variables observadas presentan **heterocedasticidad**.

4.5 Cálculos

Ahora procederemos a calcular el estadístico de contraste, el valor crítico y el valor de p, teniendo en cuenta todo las características de la muestra que hemos visto anteriormente (**contraste de dos muestras independientes sobre la media con varianzas desconocidas**).

```
t_2_sample_unknow_var <- function(df, column_name, NC){
  # El intervalo de confianza es dado (95% por defecto)

  alpha <- 1-NC
  # Cálculo del valor crítico
  vc <- qnorm(alpha)
  # Cálculo de las medias
  mean_left <- mean(df[df$type == 'left', column_name])
  mean_right <- mean(df[df$type == 'right', column_name])
  # Desviación estándar
  sd_left <- sd(df[df$type == 'left', column_name])
  sd_right <- sd(df[df$type == 'right', column_name])
  # Número de muestras
  n_left <- length(df[df$type == 'left', column_name])
  n_right <- length(df[df$type == 'right', column_name])
```

Variable	t-valor	valor_critico
Rating	5.93376544583843	1.64485362695147
Ball_Control	15.181923101687	1.64485362695147
Dribbling	18.1375623401995	1.64485362695147

```
  # Cálculo del estimador
  t <- (mean_left - mean_right) / sqrt((sd_left^2/n_left) + (sd_right^2/n_right))
  # Resultado
  return(c(column_name, t, abs(vc)))
}
column_names <- c('Rating', 'Ball_Control', 'Dribbling')
t2_sample_table <- data.frame(matrix(ncol=3, nrow = 0))
for (col in column_names){
  out <- t_2_sample_unknow_var(data, col, 0.95)
  t2_sample_table <- rbind(t2_sample_table, out)
}
names(t2_sample_table) <- c("Variable", "t-valor", "valor_critico")
```

4.6 Tabla de resultados

Mostramos los resultados para las tres variables de la muestra de jugadores

```
t2_sample_table %>% kable() %>% kable_styling()
```

4.7 Interpretación

Como podemos observar de los resultados de la tabla anterior, en las tres variables estudiadas el valor observado (t-valor) es mayor que el valor crítico

```
t valor > valor crítico
```

lo que permite rechazar la hipótesis nula aceptando que los jugadores zurdos obtienen mejores puntuaciones que los diestros para estas variables con un nivel de confianza del 95%.

5 Comparación por pares

Nos preguntamos si obtendríamos el mismo resultado si **comparásemos** los jugadores de campo zurdos con aquellos jugadores de campo diestros que tienen un peso, altura y edad con **características similares**. Para dar respuesta a esta pregunta, realizaremos un proceso similar al denominado `propensity score matching`, aunque un poco simplificado.

El procedimiento se basa en calcular la **distancia euclídea** entre los vectores de jugadores zurdos y diestros para cada una de las variables. El resultado serán dos muestras, una primera con los jugadores zurdos, y por cada posición de jugador, en la segunda muestra encontraremos el jugador diestro más cercano, existiendo una relación 1 a 1 por cada jugador de cada muestra.

Como medida de comparación de la calidad del juego entre jugadores, utilizaremos la variable Rating, también empleada en el anterior artículo.

5.1 Jugador más similar

Implantamos la función euclídea que nos mostrará la distancia entre los vectores.

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

```
my.nn <- function(x, sample){
  max_row <- nrow(sample)
  #max_row <-200
  n_col <- 3
  i <- max_row
  # pos es un vector que guarda la posición y la media calculada
  # La distancia más pequeña es 0.
  # Se inicializa a un valor elevado.
  pos <- c(0, 10000)

  while (i > 0){
    eu <- euclidean(x, sample[i, ])
    if (eu < pos[2]){
      pos <- c(i, eu)
    }
    i <- i -1
  }
  # Return the position of the second dataframe
  return(pos[1])
}
# For testing purpose
  #my.nn(left_gamers[1, 2:4], right_gamers[, 2:4])
```

```
my.nn.sample <- function( sample1, sample2, id_vector){
   max_row <- nrow(sample1)

  # Right.paired continene el listado de las posiciones de los jugadores diestros
   # más similares a los zurdos.
  right.vector <-c()
  i <- max_row
  while (i > 0){
    pos <- c()
    if (i%%10 == 0){
      print(paste('# de elementos restantes: ', i))
    }
    # Escogemos el vector de sample1 y buscamos el más cercano en sample2
    pos <- my.nn(sample1[i,], sample2[, ])
    # Con la posición, buscamos el identificador del jugador
    right.vector <-c(right.vector, id_vector[pos])
    i <- i -1
  }
  print('Proceso Finalizado')
  return(right.vector)
}
```

Vamos a preparar las muestras

```
column_names <- c('ID', 'Weight', 'Height', 'Age')
left_gamers <- fifa.df %>%
  filter(Club_Position!='GK' & Preffered_Foot=='Left') %>%
  select(all_of(column_names))
```

```
right_gamers <- fifa.df %>%
  filter(Club_Position!='GK' & Preffered_Foot=='Right') %>%
  select(all_of(column_names))
print(paste('# jugadores zurdos: ', nrow(left_gamers)))
```

```
## [1] "# jugadores zurdos:  4022"
```

```
print(paste('# jugadores diestros: ', nrow(right_gamers)))
```

```
## [1] "# jugadores diestros:  12934"
```

Vemos que hay tres veces más jugadores diestros que zurdos.

Lanzamos el script que nos va a permitir encontrar los jugadores diestros más similares en cuanto a morfología.

```
# right.paired es un vector que contiene los identificadores de los jugadores
# diestros más cercanos a los zurdos.
number_left <- 100
number_right <- 500
right.paired <- c()
right.paired  <- my.nn.sample(left_gamers[1:number_left, 2:4],
                              right_gamers[1:number_right, 2:4],
                              right_gamers[, 'ID'])
```

```
## [1] "# de elementos restantes:  100"
## [1] "# de elementos restantes:  90"
## [1] "# de elementos restantes:  80"
## [1] "# de elementos restantes:  70"
## [1] "# de elementos restantes:  60"
## [1] "# de elementos restantes:  50"
## [1] "# de elementos restantes:  40"
## [1] "# de elementos restantes:  30"
## [1] "# de elementos restantes:  20"
## [1] "# de elementos restantes:  10"
## [1] "Proceso Finalizado"
```

Para acortar el tiempo de cálculo se ha escogido una muestra de 100 jugadores zurdos y de 500 dentro de la población de jugadores diestros.

La dos muestras tienen que tener el mismo tamaño, dado que queremos tener **muestras apareadas**. Vamos a comprobarlo.

```
nrow(left_gamers[1:number_left,]) == length(right.paired)
```

```
## [1] TRUE
```

5.2 Muestras

Una vez finalizado el proceso de cálculo, vamos a montar el dataframe para poder comparar los jugadores zurdos y diestros según la salida del algoritmo.

```
right.paired.df <- as.data.frame(right.paired)
right.paired.df <- right.paired.df %>% mutate(line_number = row_number())

# Recuperamos el Rating de los jugadores diestros
right.paired.df <- dplyr::inner_join(right.paired.df, fifa.df[,c('ID', 'Rating')],
                                     by=c('right.paired'='ID'))
```

ID.l	Weight.l	Height.l	Age.l	Rating.l	ID.r	Rating.r	Weight.r	Height.r	Age.r
2	72	179	29	93	55	86	85	188	28
8	74	183	27	90	324	81	76	180	27
14	76	180	28	89	489	80	87	190	28
20	67	176	25	88	265	82	61	170	30
28	84	187	32	88	520	80	82	187	30
35	75	180	25	87	73	85	75	175	28

```
paired.df <- left_gamers[1:number_left,] %>%
  mutate(line_number = row_number())

# Recuperamos el Rating de los jugadores zurdos
paired.df <- dplyr::inner_join(paired.df, fifa.df[,c('ID', 'Rating')],
                               by=c('ID'='ID'))

# Join de los dos datasets
paired.df <- dplyr::inner_join(paired.df, right.paired.df, by= 'line_number')
paired.df <- dplyr::inner_join(paired.df, right_gamers,
                               by= c("right.paired" = 'ID'))
# Ya podemos eliminar el campo de la posición
paired.df$line_number <- NULL

names(paired.df) <-c('ID.l', 'Weight.l', 'Height.l', 'Age.l', 'Rating.l',
                     'ID.r', 'Rating.r', 'Weight.r', 'Height.r', 'Age.r')

head(paired.df) %>% kable() %>% kable_styling()
```

Se muestra la tabla con los primeros resultados de los jugadores zurdos y diestros. Se ha colocado el sufijo '.l' para los zurdos y el '.r' para los diestros.

Con la visualización de la tabla se muestra como las variables fisiológicas de los jugadores zurdos y diestros están dentro de un orden similar, con alturas entre los 180 y los 190 cm, pesos similares (en la fila cuarta, los pesos de los jugadores son de 67 y 61 kg), y sin grandes diferencias de edad.

Comprobado este punto, es hora de dejar limpio el dataset con sólo las variables necesarias para el estudio.

```
paired.df[, c('ID.l', 'Weight.l', 'Height.l', 'Age.l',
              'ID.r',  'Weight.r', 'Height.r', 'Age.r')] <- NULL
head(paired.df)
```

```
##   Rating.l Rating.r
## 1       93       86
## 2       90       81
## 3       89       80
## 4       88       82
## 5       88       80
## 6       87       85
```

5.3 Hipótesis nula y alternativa

Como conservamos la misma pregunta, que en el ejercicio anterior, exponemos la hipótesis nula como que no hay diferencias por el hecho de ser diestros o zurdos y la hipótesis alternativa como aquella en la que los zurdos consiguen mejor puntuación en las tres variables.

Matemáticamente:

$$\begin{cases} H_0: & \mu_{left} = \mu_{right} \\ H_1: & \mu_{left} > \mu_{right} \end{cases}$$

A continuación, vamos a estudiar la variable Rating de los jugadores zurdos y diestros, y dentro del intervalo de confianza (NC) del 95 %, pero teniendo en cuenta que trabajaremos con las diferencias entre estas variables, tal y como se explica a continuación (**contraste sobre la media unilateral por la derecha**).

5.4 Método

Tal y como hemos ido viendo, vamos a realizar una **comparación de dos muestra apareadas sobre la media**.

5.5 Cálculos

En este tipo de contraste, en el que las muestras están vinculadas 1 a 1, calculamos las diferencias entre ellas, reduciendo el problema a un contraste de una muestra con las diferencias de los dos jugadores.

```
paired.df.sub <- paired.df %>% mutate(Rating =Rating.l - Rating.r)  %>% select(Rating)
head(paired.df.sub)
```

```
##   Rating
## 1      7
## 2      9
## 3      9
## 4      6
## 5      8
## 6      2
```

La asunción de la normalidad ya la demostramos en anteriormente. Así que podemos aplicar directamente el estadístico.

```
t_2_sample_by_dif <- function(df, column_name, NC){
  # El intervalo de confianza es dado (95% por defecto)

  # Al tratarse de un estudio unilateral, alpha no se divide.
  alpha <- 1-NC

  # Cálculo del valor crítico
  vc <- qnorm(alpha)

  # Cálculo de la media de las diferencias
  mean_dif <- mean(df[, column_name])

  # Cálculo de la Desviación estándar
  sd_dif <- sd(df[, column_name])

  # Número de muestras
  n <- length(df[, column_name])

  # Cálculo del estimador
  t <- mean_dif / (sd_dif / sqrt(n))

  # Resultado
  return(c(column_name, t, abs(vc)))
}
```

Variable	t-valor	valor_critico
Rating	4.06650420602581	1.64485362695147

```
# Parámetros
NC <- 0.95
column_names <- c('Rating')

t2_sample_table <- data.frame(matrix(ncol=3, nrow = 0))
for (col in column_names){
  out <- t_2_sample_by_dif(paired.df.sub, col, NC)
  t2_sample_table <- rbind(t2_sample_table, out)
}
names(t2_sample_table) <- c("Variable", "t-valor", "valor_critico")

#t2_sample_table
```

5.6 Interpretación

Mostremos el resultado del estimador

```
t2_sample_table %>% kable() %>% kable_styling()
```

Si comparamos la salida de la anterior tabla, para la variable estudiada el valor observado (t-valor) es mayor que el valor crítico

```
t valor > valor crítico
```

lo que permite rechazar la hipótesis nula, y por tanto se puede apreciar diferencia estadística entre los jugadores zurdos y los diestros para estas variables, con un nivel de confianza del 95%.

5.7 Reflexión

En este estudio hemos comparado las muestras entre jugadores 1 a 1, buscando entre los jugadores aquellos con características similares en las variables de peso, altura y edad. Posteriormente, hemos complementado el estudio confrontando este par de jugadores zurdos y diestros, con identidades morfológicas parecidas, con los valores de sus variables de Rating, medida que resume la calidad de juego de estos jugadores. El resultado obtenido ha sido que hay evidencia estadística para poder afirmar que los jugadores zurdos son mejores que los diestros dentro del nivel de confianza del 95%.

6 Comparación entre clubes

En este capítulo del estudio, buscamos calcular si el porcentaje de jugadores con un Rating superior a 90 es **diferente** entre los clubes de Barcelona y Madrid, con un nivel de confianza del 97%.

Comenzamos por preparar los datos.

```
rating_bcn <- fifa.df[fifa.df$Club == 'FC Barcelona', 'Rating']
rating_md <- fifa.df[fifa.df$Club == 'Real Madrid', 'Rating']
print(paste('# Casos del FC Barcelona: ', length(rating_bcn)))
```

```
## [1] "# Casos del FC Barcelona:  33"
```

```
print(paste('# Casos del Real Madrid: ', length(rating_md)))
```

```
## [1] "# Casos del Real Madrid:  33"
```

Vamos a calcular las proporciones según los parámetros de enunciado.

```
limit <- 90
p_bcn <- length(rating_bcn[rating_bcn > limit] ) / length(rating_bcn)
p_md <- length(rating_md[rating_md > limit] ) / length(rating_md)
print(paste('Rating del FC Barcelona: ', p_bcn))
```

```
## [1] "Rating del FC Barcelona:  0.0909090909090909"
```

```
print(paste('Rating del Real Madrid: ', p_md))
```

```
## [1] "Rating del Real Madrid:  0.0303030303030303"
```

6.1 Hipótesis nula y alternativa

La pregunta inicial, busca comprobar si existe diferencia entre la proporción de jugadores de ambos clubes cuando el valor de la variable Rating > 90. Estamos ante un contraste de Proporciones bilateral, en el que definimos la hipótesis nula H_{0} como que las proporciones son iguales, y la hipótesis alternativa H_{1} con el caso contrario, todo ello dentro del intervalo de confianza (NC) del 97 %..

Esquemáticamente es:

$$\begin{cases} H_0: & p_{bcn} = p_{md} \\ H_1: & p_{bcn} \neq p_{md} \end{cases}$$

6.2 Método

Ya lo hemos comentado antes, pero nos encontramos ante un **contraste de Proporciones bilateral**. La normalidad queda asegurada por el Teorema del Límite Central al tener un número de observaciones ligeramente superior a 30.

6.3 Cálculos

Realizamos los cálculos del estimador.

```
t_2_ind_prop <- function(p1, n1, p2, n2, NC){

  # Simetría, implica la mitad de alpha.
  alpha <- (1-NC)/2

  # Cálculo del valor crítico
  vc <- qnorm(alpha)

  # Cálculo de p
  p <- (n1*p1 + n2*p2) / (n1+n2)

  # Cálculo del estimador
  z <- (p1-p2) / sqrt(p*(1-p) * (1/n1 + 1/n2))

  # Cálulo del p_value
  p_value <- pnorm(z, lower.tail = T)

  return(c(z, vc, p_value, alpha))
}

# Fijamos el Nivel crítico del test.
NC <- 0.97
```

z_value	vc	p_value	alpha
1.031754	-2.17009	0.8489063	0.015

```
t2_out <- t_2_ind_prop(p_bcn, length(rating_bcn), p_md, length(rating_md), NC)
t2_out <- data.frame('z_value' = t2_out[1], 'vc' = t2_out[2],
                     'p_value' = t2_out[3], 'alpha' = t2_out[4])
```

6.4 Resultados e interpretación

Veamos los resultados obtenidos.

```
t2_out %>% kable() %>% kable_styling()
```

Con los valores encontrados no podemos rechazar la hipótesis nula debido a que el valor del estimador es inferior al valor crítico (en términos absolutos), y el p valor no es inferior al valor de alfa.

Por tanto, no podemos determinar que haya diferencia a nivel de Rating entre los jugadores de los dos clubes, Barcelona y Madrid, con un nivel de confianza del 97%.

7 Resumen ejecutivo

En este estudio hemos trabajado en un análisis descriptivo e inferencial sobre los datos de los jugadores del juego Fifa 2017, como muestra representativa de los jugadores de fútbol.

Primeramente hemos investigado la variable Rating de los jugadores, realizando un análisis visual de la misma y encontrando sus **intervalos de confianza**.

A continuación, nos hemos centrado en un estudio sobre las diferencias entre los jugadores zurdos y diestros. Para ello, hemos trabajado sobre tres de sus variables que determinan la calidad del juego, formulando las preguntas de investigación (una por cada variable), y planteando la hipótesis nula y alternativa. El estadístico utilizado, se basa en un **contraste unilateral de las dos muestras poblacionales independientes**, que nos ha llevado a la conclusión de que los jugadores zurdos obtienen mejores puntuaciones que los diestros para las variables analizadas, con un nivel de confianza del 95%.

En el siguiente capítulo de nuestro estudio, hemos continuado la comparación de la población de jugadores zurdos respecto a los diestros. Para ello, hemos emparejando a cada jugador zurdo con otro diestro de similares características morfológicas, reduciendo el estudio a una de las variables que determina la calidad del juego. Por lo tanto, se ha realizado un **contraste de dos muestras poblaciones emparejadas sobre la media**. El resultado obtenido, ha mostrado que existe evidencia estadística para afirmar que los jugadores zurdos obtienen mejor puntuación, con un nivel de confianza fijado al 95%.

Finalmente, en el último capítulo hemos realizado la comparación respecto a los jugadores de dos clubes, El Barcelona y el Madrid. Para ello, se ha realizado una selección de los mejores jugadores cuyo Ranking es superior a 90, respecto al total de los jugadores de la plantilla de ambos equipos. Por lo tanto, se ha realizado **un contraste de proporciones de dos muestras bilateral**, el resultado de la cual nos dice que no se ha apreciado evidencia estadística para afirmar que un equipo es suprior al otro, con un 97% de nivel de confianza.