A2 - Analisis Estadistico I

Pablo A. Delgado

23 de April, 2021

Contents

Introduccion

En esta actividad se realizará un análisis estadístico descriptivo e inferencial de los datos procesados en la actividad 1. Recordamos que el conjunto de datos usado en la actividad previa consistía en el conjunto de datos Fifa.csv, que se encuentra disponible en la plataforma Kaggle: https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global.

Este conjunto de datos contiene el estilo de juego del videojuego de consola Fifa 2017, así como estadísticas reales de los jugadores de futbol. El conjunto de datos contiene más de 17,500 registros y 53 variables.

Las principales variables que se usarán en esta actividad son:

- `Name (Nombre del jugador)`
- `Nationality (Nacionalidad del jugador)`
- `National_Position (Posición de juego en equipo nacional).`
- `National_Kit (Número de equipación en equipo nacional)`
- `Club (Nombre del club)`
- `Club_Position (Posición de juego en club)`
- `Club_Kit (Número de equipación en club)`
- `Club_Joining (Fecha en la que empezó en el club)`
- `Contract_Expire (Año finalización del contrato)`
- `Rating (Valoración global del jugador, entre 0 y 100)`
- `Height (Altura)`
- `Weight (Peso)`
- `Preffered_Foot (Pie preferido)`
- `Birth_Date (Fecha de nacimiento)`
- `Age (Edad)`
- `Preffered_Position (Posición preferida)`
- `Work_Rate (valoración cualitativa en términos de ataque-defensa)`
- `Weak_foot (valoración de 1 a 5 de control y potencia de la pierna no preferida)`
- `Skill_Moves (valoración de 1 a 5 de la habilidad en movimientos del jugador)`
- `El resto de variables hacen referencia a atributos del jugador.`

La descripción de los atributos se puede consultar en https://www.fifplay.com/encyclopedia. La descripción de las abreviaturas de la posición del jugador en el campo se puede consultar en https://www.dtgre.com/2016/10/fifa-17-position-abbreviations-acronyms.html.

Puesto que el resultado del preprocesado de los datos puede ser ligeramente distinto entre las distintas soluciones que habéis aportado, os suministramos el fichero preprocesado. Esta actividad se realizará con el fichero que os suministramos, independientemente del proceso de preprocesado que hayáis realizado en la actividad anterior. El nombre del fichero es **fifa_clean.csv**.

En esta actividad realizaremos un **análisis descriptivo e inferencial**. En especial, nos interesa investigar la puntuación del jugador (Rating) y otras variables como el control de pelota (Ball_Control) y la técnica (Dribbling). Asumimos que este conjunto de datos es una muestra representativa de los jugadores de la última década (población).

Nota importante a tener en cuenta para entregar la actividad:

- Es necesario entregar el archivo Rmd y el fichero de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se debe respetar la misma numeración de los apartados que el enunciado.
- No se pueden realizar listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden usar las funciones **head** y **tail** que sólo muestran unas líneas del fichero de datos.
- Se valora la precisión de los términos utilizados (hay que usar de manera precisa la terminología de la estadística).
- Se valora también la concisión en la respuesta. No se trata de hacer explicaciones muy largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos de manera clara y concisa.
- No se puede compartir código entre compañeros ni copiar código de actividades anteriores. Cada estudiante debe encontrar su propia solución a las preguntas de la actividad.

1. Lectura del fichero

Leer el fichero fifa_clean.csv. Validar que los datos leídos son correctos. Si no es así, realizar las conversiones oportunas.

Como primer paso definimos que librerias estaremos usando. Para ellos generamos un vector con todas las posibles librerias que necesitaremos, instalamos las que no tengamos, para finalmente mediante el uso de lapply cargarlas.

```
packages <- c("ggplot2", "gridExtra", "kableExtra", "stats", "proxy")
new <- packages[!(packages %in% installed.packages()[,"Package"])]
if(length(new)) install.packages(new)
foo=lapply(packages, require, character.only=TRUE)
```

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

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Loading required package: gridExtra
```

```
## Warning: package 'gridExtra' was built under R version 4.0.5
```

```
## Loading required package: kableExtra
```

```
## Warning: package 'kableExtra' was built under R version 4.0.5
```

```
## Loading required package: proxy
```

```
## Warning: package 'proxy' was built under R version 4.0.5
```

```
##
## Attaching package: 'proxy'
```

```
## The following objects are masked from 'package:stats':
##
##     as.dist, dist
```

```
## The following object is masked from 'package:base':
##
##     as.matrix
```

Dado que el input de datos es un archivo csv, y en una inspeccion visual manual hemos visto que posee como separador de columna la coma, usaremos la funcion read.csv() para cargar los datos en un dataframe. Hemos hemos podido constatar que el csv cuenta con 54 columnas y 17590 lineas, la primera linea corresponde al header y la ultima una linea en blanco, siendo asi 17588 lineas con datos, mientras que la primer columna corresponde al ID de cada fila, quedandonos 53 variables "dato" tal como especifica el enunciado

Dicho esto carguemos el archivo y hagamos una verificacion rapida de las primeras y ultimas 5 filas del dataframe.

```
fifa2017 <- read.csv("fifa_clean.csv", stringsAsFactors = FALSE, header=TRUE)
cols = c('Name','Nationality','Club','Rating','Age','Height','Weight','Dribbling','Ball_Control')
head(fifa2017[,cols])
```

```
##                Name Nationality           Club Rating Age Height Weight
## 1 Cristiano Ronaldo    Portugal    Real Madrid     94  31    185     78
## 2      Lionel Messi   Argentina   FC Barcelona     93  29    179     72
## 3             Neymar      Brazil   FC Barcelona     92  24    174     68
## 4       Luis Suárez     Uruguay   FC Barcelona     92  29    182     85
## 5      Manuel Neuer     Germany      FC Bayern     92  30    193     85
## 6             De Gea       Spain Manchester Utd     90  26    186     82
##   Dribbling Ball_Control
## 1        92           93
## 2        97           95
## 3        96           95
## 4        86           91
## 5        30           48
## 6        13           31
```

```
tail(fifa2017[,cols])
```

```
##                   Name         Nationality           Club Rating Age Height
## 17583  Mark Mcelhinney Republic Of Ireland     Derry City     45  18    182
## 17584      Adam Dunbar Republic Of Ireland Wexford Youths     45  18    183
## 17585     Dylan Mcgoey Republic Of Ireland  Longford Town     45  19    185
## 17586  Tommy Ouldridge             England   Swindon Town     45  18    173
## 17587       Mark Foden            Scotland    Ross County     45  20    180
## 17588 Barry Richardson             England        Wycombe     45  47    185
```

```
##       Weight Dribbling Ball_Control
## 17583     76        14           15
## 17584     82        11           12
## 17585     80        11           13
## 17586     61        39           44
## 17587     80        13           17
## 17588     77        11           22
```

Veamos la cantidad de columnas, filas y como quedaron los tipos de datos en el dataframe para entender si hubo algun dato importado incorrectamente.

```
str(fifa2017)
```

```
## 'data.frame':    17588 obs. of  54 variables:
##  $ ID                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name              : chr  "Cristiano Ronaldo" "Lionel Messi" "Neymar" "Luis Suárez" ...
##  $ Nationality       : chr  "Portugal" "Argentina" "Brazil" "Uruguay" ...
##  $ National_Position : chr  "LS" "RW" "LW" "LS" ...
##  $ National_Kit      : int  7 10 10 9 1 1 9 11 NA 1 ...
##  $ Club              : chr  "Real Madrid" "FC Barcelona" "FC Barcelona" "FC Barcelona" ...
##  $ Club_Position     : chr  "LW" "RW" "LW" "ST" ...
##  $ Club_Kit          : int  7 10 11 9 1 1 9 11 9 13 ...
##  $ Club_Joining      : chr  "07/01/2009" "07/01/2004" "07/01/2013" "07/11/2014" ...
##  $ Contract_Expiry   : int  2021 2018 2021 2021 2021 2019 2021 2022 2017 2019 ...
##  $ Rating            : int  94 93 92 92 92 90 90 90 90 89 ...
##  $ Height            : int  185 179 174 182 193 186 185 183 196 199 ...
##  $ Weight            : int  78 72 68 85 85 82 78 74 95 91 ...
##  $ Preffered_Foot    : chr  "Right" "Left" "Right" "Right" ...
##  $ Birth_Date        : chr  "02/05/1985" "06/24/1987" "02/05/1992" "01/24/1987" ...
##  $ Age               : int  31 29 24 29 30 26 28 27 35 24 ...
##  $ Preffered_Position: chr  "LW/ST" "RW" "LW" "ST" ...
##  $ Work_Rate         : chr  "High / Low" "Medium / Medium" "High / Medium" "High / Medium" ...
##  $ Weak_foot         : int  4 4 5 4 4 3 4 3 4 3 ...
##  $ Skill_Moves       : int  5 4 5 4 1 1 3 4 4 1 ...
##  $ Ball_Control      : int  93 95 95 91 48 31 87 88 90 23 ...
##  $ Dribbling         : int  92 97 96 86 30 13 85 89 87 13 ...
##  $ Marking           : int  22 13 21 30 10 13 25 51 15 11 ...
##  $ Sliding_Tackle    : int  23 26 33 38 11 13 19 52 27 16 ...
##  $ Standing_Tackle   : int  31 28 24 45 10 21 42 55 41 18 ...
##  $ Aggression        : int  63 48 56 78 29 38 80 65 84 23 ...
##  $ Reactions         : int  96 95 88 93 85 88 88 87 85 81 ...
##  $ Attacking_Position: int  94 93 90 92 12 12 89 86 86 13 ...
##  $ Interceptions     : int  29 22 36 41 30 30 39 59 20 15 ...
##  $ Vision            : int  85 90 80 84 70 68 78 79 83 44 ...
##  $ Composure         : int  86 94 80 83 70 60 87 85 91 52 ...
##  $ Crossing          : int  84 77 75 77 15 17 62 87 76 14 ...
##  $ Short_Pass        : int  83 88 81 83 55 31 83 86 84 32 ...
##  $ Long_Pass         : int  77 87 75 64 59 32 65 80 76 31 ...
##  $ Acceleration      : int  91 92 93 88 58 56 79 93 69 46 ...
##  $ Speed             : int  92 87 90 77 61 56 82 95 74 52 ...
##  $ Stamina           : int  92 74 79 89 44 25 79 78 75 38 ...
##  $ Strength          : int  80 59 49 76 83 64 84 80 93 70 ...
##  $ Balance           : int  63 95 82 60 35 43 79 65 41 45 ...
```

```
##  $ Agility           : int  90 90 96 86 52 57 78 77 86 61 ...
##  $ Jumping           : int  95 68 61 69 78 67 84 85 72 68 ...
##  $ Heading           : int  85 71 62 77 25 21 85 86 80 13 ...
##  $ Shot_Power        : int  92 85 78 87 25 31 86 91 93 36 ...
##  $ Finishing         : int  93 95 89 94 13 13 91 87 90 14 ...
##  $ Long_Shots        : int  90 88 77 86 16 12 82 90 88 17 ...
##  $ Curve             : int  81 89 79 86 14 21 77 86 82 19 ...
##  $ Freekick_Accuracy : int  76 90 84 84 11 19 76 85 82 11 ...
##  $ Penalties         : int  85 74 81 85 47 40 81 76 91 27 ...
##  $ Volleys           : int  88 85 83 88 11 13 86 76 93 12 ...
##  $ GK_Positioning    : int  14 14 15 33 91 86 8 5 9 86 ...
##  $ GK_Diving         : int  7 6 9 27 89 88 15 15 13 84 ...
##  $ GK_Kicking        : int  15 15 15 31 95 87 12 11 10 69 ...
##  $ GK_Handling       : int  11 11 9 25 90 85 6 15 15 91 ...
##  $ GK_Reflexes       : int  11 8 11 37 89 90 10 6 12 89 ...
```

```
nrow(fifa2017)
```

```
## [1] 17588
```

```
ncol(fifa2017)
```

```
## [1] 54
```

En principio como vemos, 17588 obs. of 54 variables, la cantidad de filas y columnas en el dataframe coinciden con la inspecion visual que hemos hecho sobre el archivo csv. Y los tipos de datos de cada variable corresponden a los esperados.

2. Rating de los jugadores

Nos interesa investigar los valores que toma la variable Rating en la población. Para ello, realizad un primer análisis visual de esta variable a partir de la muestra. Posteriormente, calculad el intervalo de confianza de la variable Rating de los jugadores. Seguid los pasos que se indican a continuación.

2.1. Análisis visual

Mostrad visualmente la distribución de la variable Rating. Usad el gráfico o gráficos que creáis más oportunos. Describid brevemente lo que se observa en los gráficos que representáis.

Verifiquemos estadisticas basicas de la variable rating de forma numerica:

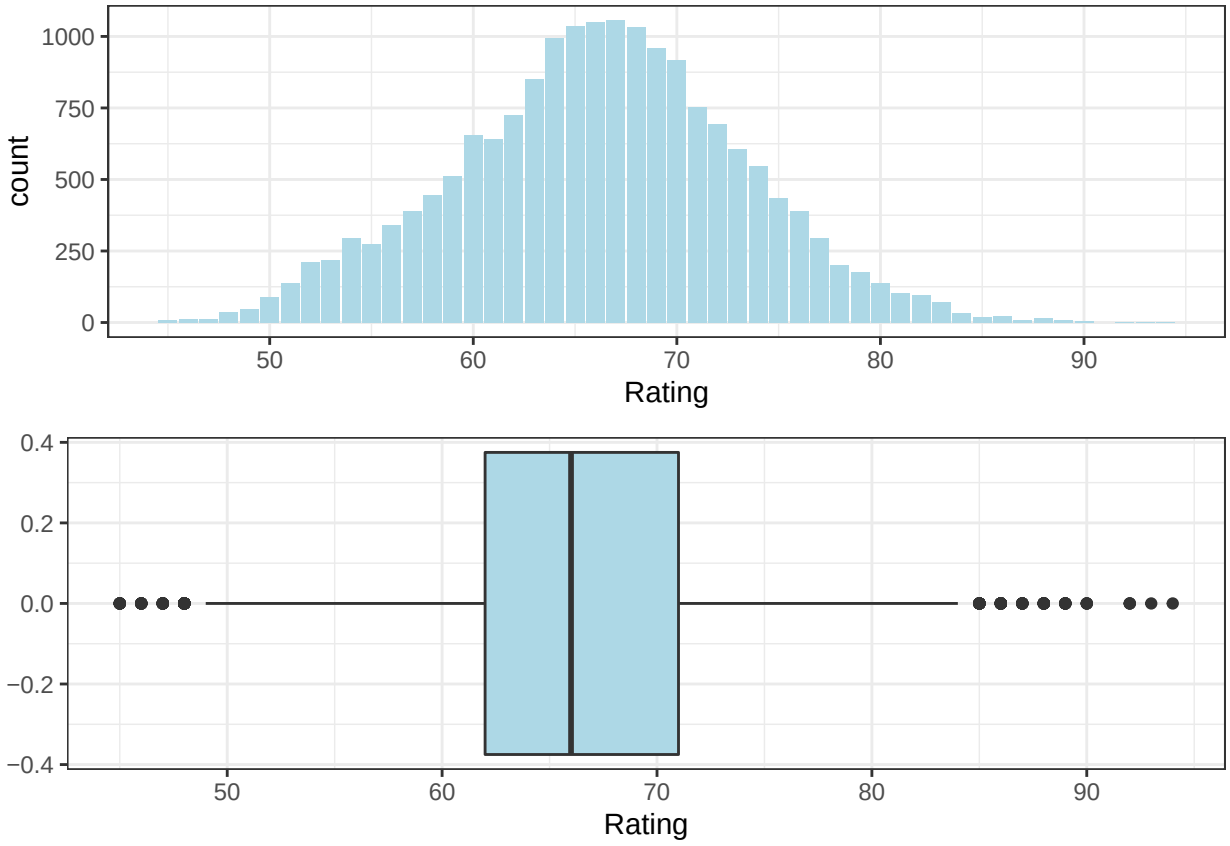
```
summary(fifa2017$Rating)
```

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

Tenemos una media de 66.17 y el IQR esta entre 62.0 y 71.0.

Analizemosla ahora visualmente:

```
gbp=ggplot(fifa2017,aes(x=Rating,fill=Rating))+geom_boxplot(fill="lightblue")+theme_bw()
gb=ggplot(fifa2017,aes(x=Rating,fill=Rating))+geom_bar(fill="lightblue")+theme_bw()
grid.arrange(grobs = list(gb, gbp ) , nrow = 2, ncol = 1)
```



Con el barplot puede verse una distribucion casi normal, si es verdad con varios outliers a la derecha. Situacion que se hace mas notoria al usar un boxplot, donde se ven muchos mas outliers por la derecha que por la izquierda. Sin embargo, no vemos un left o righ skew exagerado, como dijimos estamos ante un distribucion normal de la variable rating por lo que podremos aplicar test parametricos sin problema en las siguientes secciones al menos sobre esta variable.

2.2. Intervalo de confianza

Calculad el intervalo de confianza de la variable Rating. A continuación, explicad el resultado y cómo se debe interpretar el resultado obtenido.

Nota: Los cálculos se deben realizar manualmente. No se pueden usar funciones de R que calculen directamente el intervalo de confianza. En cambio, sí se pueden usar funciones como **mean, sd, qnorm, pnorm, qt** y **pt**.

Asumiendo un intervalo de confianza del 95% y siendo que no tenemos previamente calculada la varianza de la poblacion y debemos estimarla a partir de la desviacion de la muestra, nuestra variable sigue de esta forma una distribucion t de Student con n-1 grados de libertad.

Por lo que calcularemos el intervalo de confianza para la media de la variable Rating de los jugadores cuando la varianza es desconocida previamente siguiendo estos calculos:

```
alfa <- 1-0.95
sd <- sd(fifa2017$Rating)
n <- nrow(fifa2017)
SE <- sd / sqrt(n)
# Para obtener las probabilidades o cuantiles de la distribución t
# se usan las funciones pt y qt de R, que son análogas a las
# funciones pnorm y qnorm de la distribución normal
z <- qt( alfa/2, df=n-1, lower.tail=FALSE )
peso_medio  = mean(fifa2017$Rating)
L <- peso_medio - z*SE
U <- peso_medio + z*SE
peso_medio
```

```
## [1] 66.16619
```

```
c(L,U)
```

```
## [1] 66.06151 66.27088
```

Luego aplicamos t.test solo a modo de chequeo de los calculos realizados:

```
t.test( fifa2017$Rating, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  fifa2017$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
```

Por lo tanto, el intervalo de confianza del 95% del rating de jugadores es: [66.06151 , 66.27088].

Dado que no se conocía la varianza de la población y se ha estimado a partir de la muestra, hemos usado la distribución t de Student en lugar de la distribución normal. La consecuencia de esto es que el intervalo de confianza calculado con la distribución t es más ancho que el equivalente calculado con distribución normal. Por lo tanto, para un mismo nivel de confianza, hay más incertidumbre en el valor del parámetro de la población cuando se usa la distribución t. Pero tambien es cierto que, para tamaños de muestra grandes, la distribución t de Student se aproxima a la distribución normal.

De hecho solo a modo de chequeo si hubiesemos aplicado qnorm, vemos que el intervalo de confianza es casi identico, ya que como mencionamos para muestras grandes t student tiende a una distribucion normal.

```
z_qnorm <- qnorm( alfa/2, lower.tail=FALSE )
L_qnorm <- peso_medio - z_qnorm*SE
U_qnorm <- peso_medio + z_qnorm*SE
c(L_qnorm,U_qnorm)
```

```
## [1] 66.06151 66.27087
```

3. Diferencias entre jugadores

Existe una creencia que los jugadores zurdos tienen mejor control de la pelota que los diestros. Vamos a comprobar qué dicen los datos al respecto. Nos preguntamos si los jugadores zurdos tienen mejor control de pelota (**Ball_Control**), valoración (**Rating**) y mejor **Dribbling** que los diestros. Para ello, primero seleccionad los jugadores que no son porteros (los porteros tienen el valor **GK** -Goal Keeper- en Club_Position). Entonces, debéis obtener dos muestras. La primera muestra contiene todos los jugadores de campo (no porteros) zurdos (**Preffered_Foot** igual a Left). La segunda muestra contiene todos los jugadores de campo (no porteros) diestros (**Preffered_Foot** Right). Usad un nivel de confianza del 95 %.

Aspectos a tener en cuenta para resolver este ejercicio:

- Se deben realizar los cálculos manualmente. No se pueden usar funciones de R que calculen directamente el contraste como 't.test' o similar. Sí se pueden usar funciones como 'mean', 'sd', 'qnorm', 'pnorm', 'qt' y 'pt'. Sí podéis usar 'var.test' si lo necesitáis.
- Debido a que se preguntan las diferencias en tres variables, es aconsejable estructurar el código con una función, a la que se pasa como parámetro la variable a analizar. No deberías escribir el mismo código tres veces.

Seguid los pasos que se especifican a continuacion.

3.1. Pregunta de investigación

La pregunta de investigacion en este caso es: los jugadores zurdos son mejores que los diestros en relacion del rating, control y dribling? o sea, puntuan mejores en estas 3 variables?

Podemos tambien plantear la pregunta para cada variable:

- El rating de los jugadores zurdos es mejor que el de los diestros?
- El Ball_Control de los jugadores zurdos es mejor que el de los diestros?
- El Dribbling de los jugadores zurdos es mejor que el de los diestros?

3.2. Representación visual

Representad visualmente, mediante el gráfico que sea más apropiado el valor de estas variables en jugadores de campo (no porteros) diestros y zurdos. Se deben mostrar los valores de forma comparativa entre zurdos y diestros. Interpretad los gráficos.

Como primer paso creamos una funcion que nos permite obtener facilmente los jugadores de campo.

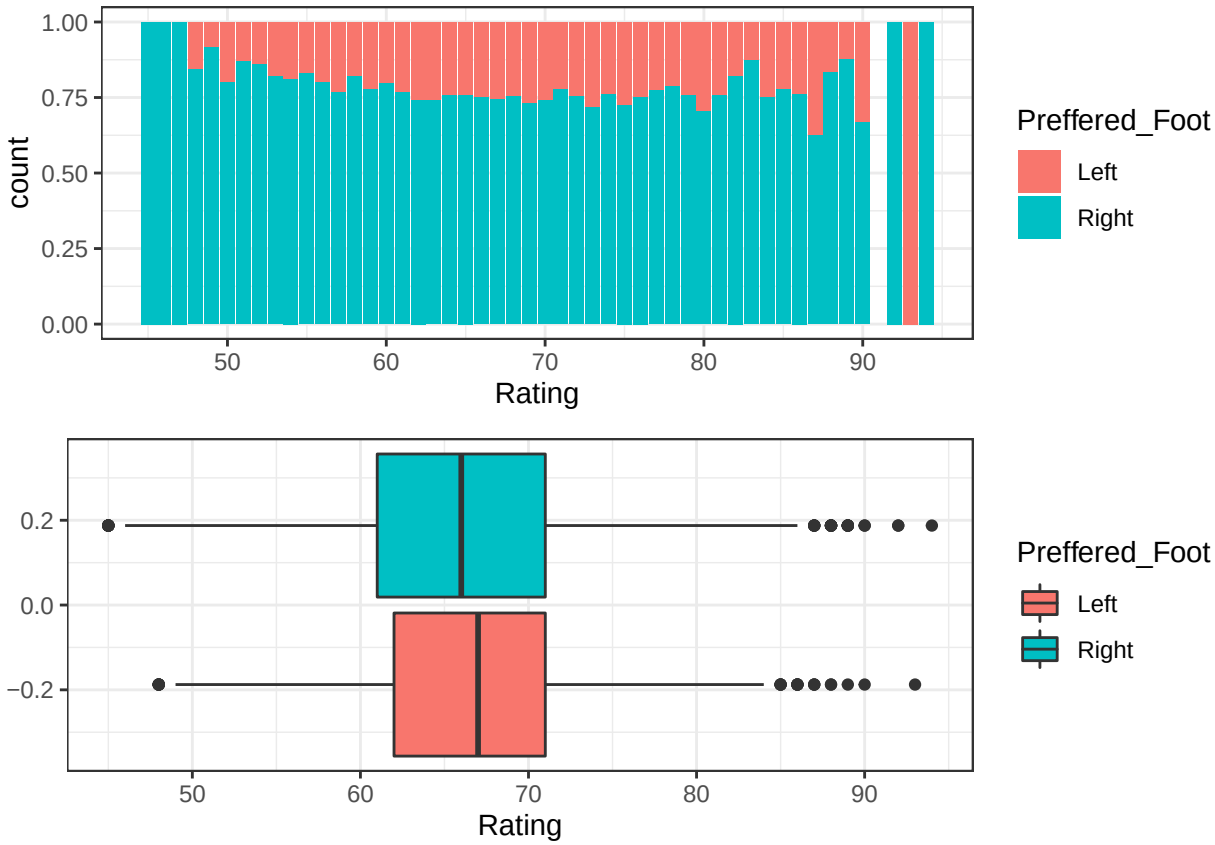
```
# Esta funcion devuelve un dataframe con todos los jugadores de campos y
# sus caractaeristicas
# Si se llama sin parametros devuelve todos los jugadores de campo por default
# mientras que devuelve solo los zurdos o diestros segun el parametro que se le pase
# Ejemplos:
# Devuelve los jugadores de campo zurdos   -> jugador_de_campo('Left')
# Devuelve los jugadores de campo diestros -> jugador_de_campo('Right')
# Devuelve todos los jugadores de campo    -> jugador_de_campo()

jugador_de_campo <- function(foot=c('Left','Right')) {
fifa2017[fifa2017$Preffered_Foot %in% foot & !fifa2017$Club_Position=='GK',]
```

```
}
table(jugador_de_campo()$Preffered_Foot)
```

```
##
##  Left Right
##  4022 12934
```

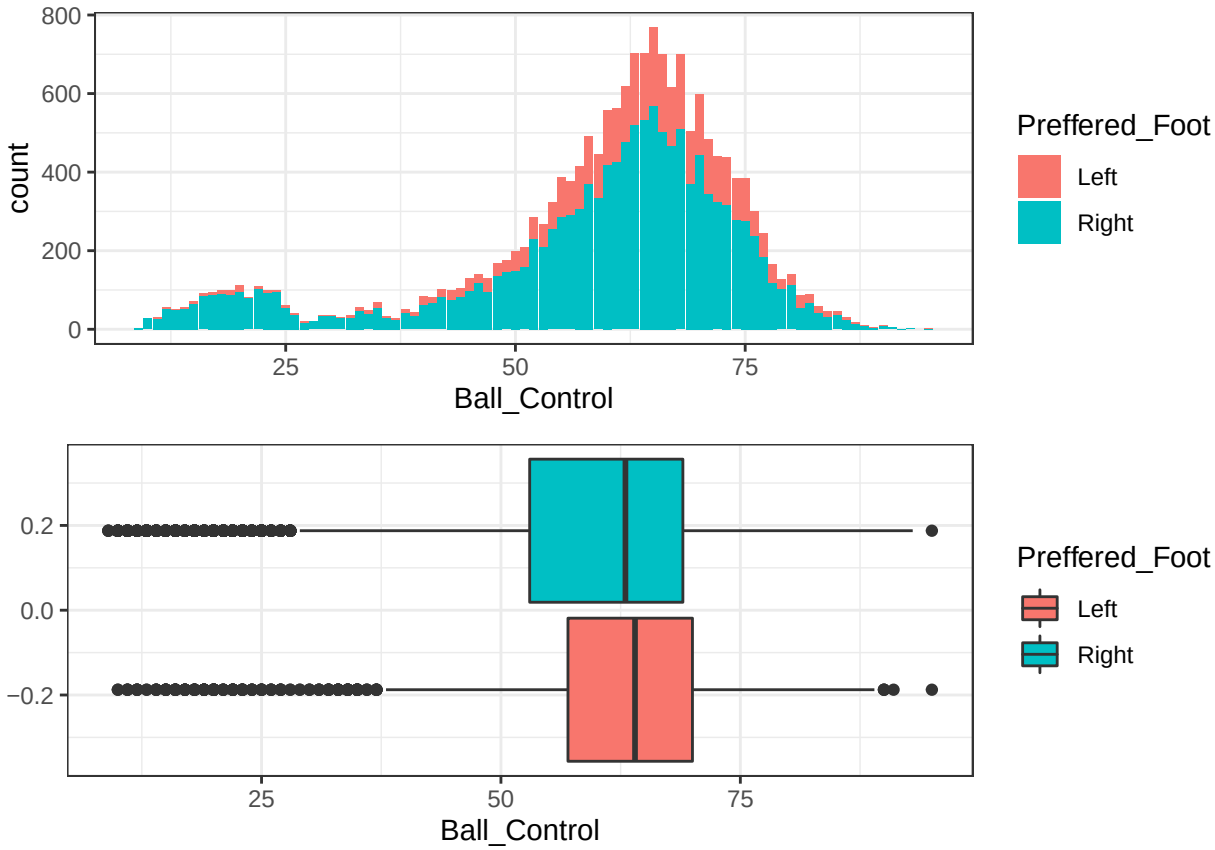
Realicemos los primeros analisis de las distintas variables para ambas muestras de jugadores, comenzando por el Rating

```
gb=ggplot(jugador_de_campo(),aes(x=Rating,fill=Preffered_Foot)) +
  geom_bar(position="fill") +
  theme_bw()

gbp=ggplot(jugador_de_campo(),aes(x=Rating,fill=Preffered_Foot)) +
  geom_boxplot() +
  theme_bw()

grid.arrange(grobs = list(gb, gbp ) , nrow = 2, ncol = 1)
```

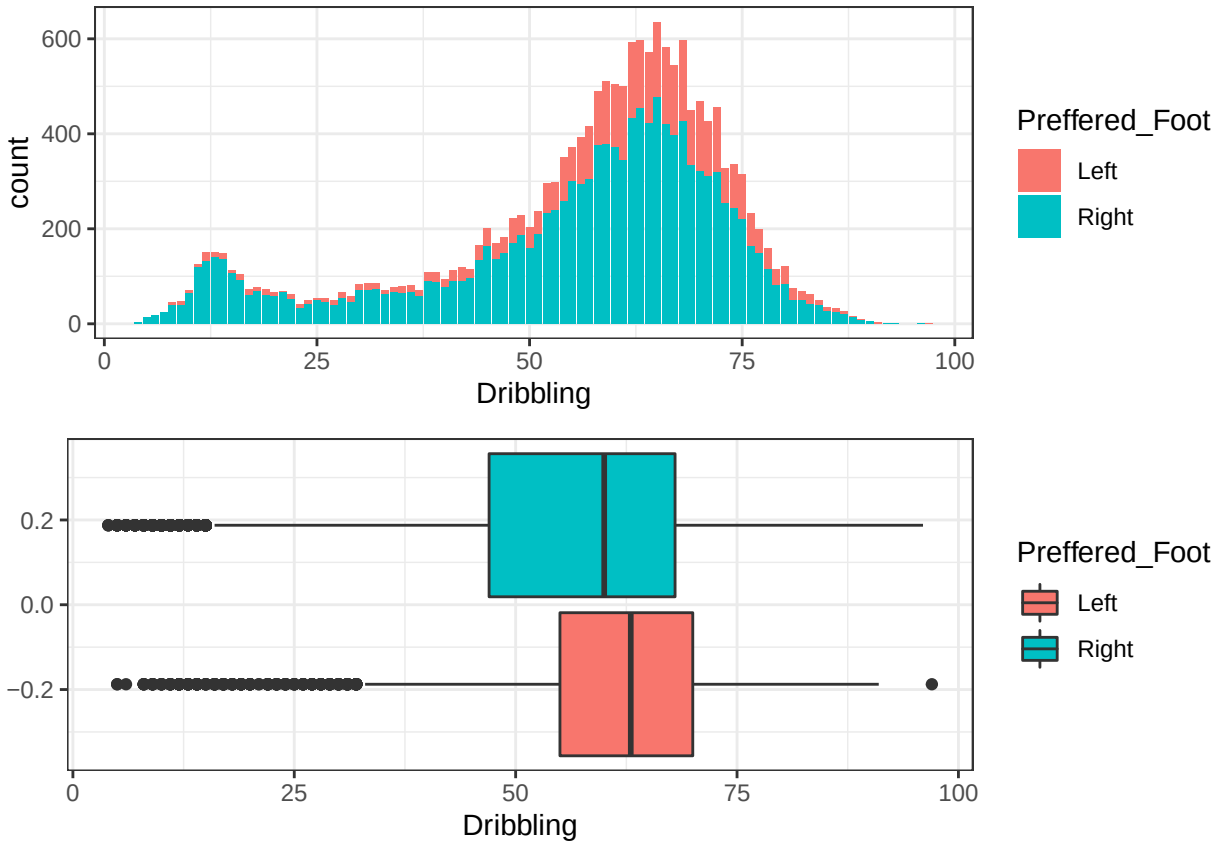



```
gb=ggplot(jugador_de_campo(),aes(x=Ball_Control,fill=Preffered_Foot)) +
  geom_bar() +
  theme_bw()
```

```
gbp=ggplot(jugador_de_campo(),aes(x=Ball_Control,fill=Preffered_Foot)) +
  geom_boxplot() +
  theme_bw()

grid.arrange(grobs = list(gb, gbp ) , nrow = 2, ncol = 1)
```




```
gb=ggplot(jugador_de_campo(),aes(x=Dribbling,fill=Preffered_Foot)) +
  geom_bar() +
  theme_bw()

gbp=ggplot(jugador_de_campo(),aes(x=Dribbling,fill=Preffered_Foot))+
  geom_boxplot() +
  theme_bw()

grid.arrange(grobs = list(gb, gbp ) , nrow = 2, ncol = 1)
```



A nivel visual se ve que tiende haber mayor "acumulacion" de jugadores zurdos en los valores mas grades de rating en el grafico de barras. Ademas de que hay mayor concentracion de jugadores zurdos alrededor de la mediana (IQR mas angosto), lo que podria implicar que los zurdos en general tienen a ser medianamente mas parejos y cerca de esos valores. Mientras que los diestros estan mas acumulados en los sectores mas bajos y su IQR es mas amplio.

Situacion similar ocurre con el Ball Control y el Dribling, entre zurdos y diestros.

Sin olvidarnos que la mediana de los zurdos es mayor para las 3 variables respecto a la de los diestros.

Como opcion se podrian quitar los outliers para no sesgar los valores que analizamos anteriormente, pero en general la mediana al ser un estimador robusto, podria no ser el caso.

3.3. Hipótesis nula y alternativa

Planteemos ahora las hipotesis para la primer pregunta de investigacion:

El rating de los jugadores zurdos es mejor que el de los diestros?

Osea aqui queremos validar que:

rating_zurdos > *rating_diestros*

Por lo que las hipotesis nula (H0) y alternativa (H1) sera:

H0: *rating_zurdos* <= *rating_diestros*

H1: *rating_zurdos* > *rating_diestros*

entonces segun lo que comprobemos en las siguientes secciones podremos llegar a decir:

- Rechazo la H0 a favor de la H1 y por tanto si que hay evidencias que demuestran que los zurdos son mejores que los diestros. Y por tanto la respuesta a la pregunta de investigacion es SI, con el 95 de confianza

o

- No hay evidencia que permita rechazar la hipotesis nula por lo que no puede afirmarse que los zurdos tengan mejor puntuacion que lo diestros para la muestra seleccionada.

Con la misma logica seguida arriba podemos representar lo mismo para las dos otra preguntas de investigacion (para Ball_Control y Dribbling)

3.4. Método

En función de las características de la muestra, decidid qué método aplicar para validar la hipótesis planteada.

Para ello, debéis especificar como mínimo:

a. si es un contraste de una muestra o de dos muestras (en caso de dos muestras, si éstas son independientes o están relacionadas),

   En este caso se trata de un Contraste de hipotesis de dos muestras independientes sobre la media (zurdos y diestros). Son Muestras Independientes, tenemos dos poblaciones, de diestros y de zurdos, no estan relacionados entre ellos.

b. si podéis asumir normalidad y por qué,

   Se podria evaluar la normalidad para cada variable, pero tambien podemos basarnos en el Teorema del Limite Central, por lo que asumimos normalidad por el TLC, dado que el tamaño de la muestra es suficientemente grande (mayor a 30). De hecho la muestra de zurdos tiene un tamaño de 4022 y la de diestros tiene uno de 12934, lo cual se cumple en nuestro ejemplo. Entonces por el TLC, podemos asumir que las dos medias de cada muestra siguen distribuciones normales.

c) si el test es paramétrico o no paramétrico,

   Es un test parametrico, porque contamos con distribuciones donde podemos asumir la normalidad de datos, porque aplica el Teorema del Limite Central al tener muestras de mas de 30 elementos y porque estamos aplicando contrastes de hipotesis sobre variables o datos numericos.

d) si el test es bilateral o unilateral,

   Para determinar si es bilateral o unilateral se debe evaluar la H1, si es mayor o menor entonces es unilateral, mientras que si es por igual entonces bilateral En nuestro caso como queremos evaluar si los zurdos son mejores que los diestros tenemos algo como: H1: zurdos $>$ diestros Entonces si despejamos a la izquieda obtenemos: zurdos - diestros > 0 Por lo que podemos concluir que se trata de un test unilateral por la derecha. Dicho esto siendo que usamos un nivel de confianza del 95% el el nivel de significancia (alfa) sera de 0.05.

e) si se puede asumir homocedasticidad o heterocedasticidad.

> Como se sabe la homocedasticidad y la heterocedasticidad., tiene que ver con la variabiblidad de las muestras, es decir, los zurdos se pueden parecer muchos mientras que los diestros no, o viceversa. Si las varianzas entre las dos muestras son similares se podra aplicar una formula, mientras que si son distintas se debe usar otra formula. Dicho todo esto y...

Dados que asumimos que no se conocen las varianzas de la población y, por lo tanto, hay que estimarlas a partir de las muestras, para aplicar el estadístico adecuado, hay que comprobar si las varianzas de las dos poblaciones son iguales o diferentes. Para ello, aplicamos primero el test de igualdad de varianzas, tambien denomidado test de homoscedasticidad.

Primero crearemos una funcion para reutilizar el test de homocedasticidad para cada variable

```
homoscedasticidad <- function(muestra1, muestra2, variable) {
  alfa <- 0.05
  zurdos <- muestra1[,variable]
  diestros <- muestra2[,variable]
  mean1 <- mean(zurdos); n1 <- length(zurdos); s1 <- sd(zurdos)
  mean2 <- mean(diestros); n2 <- length(diestros); s2 <- sd(diestros)

  fobs<-s1^2 / s2^2
  fcritL <- qf(alfa/2, df1=n1-1, df2=n2-2 )
  fcritU <- qf(1-alfa/2, df1=n1-1, df2=n2-2)
  pvalue <- min(pf( fobs, df1=n1-1, df2=n2-2, lower.tail=FALSE ),
                pf( fobs, df1=n1-1, df2=n2-2)) *2
  return ( data.frame(fobs, fcritL, fcritU, pvalue) )
}
```

Ahora apliquemos las funciones a las muestras para cada variable.

```
zurdos = jugador_de_campo('Left')
diestros = jugador_de_campo('Right')
homoscedasticidad(zurdos, diestros, 'Rating')
```

```
##        fobs    fcritL   fcritU       pvalue
## 1 0.8456888 0.9508611 1.050977 1.037705e-10
```

```
homoscedasticidad(zurdos, diestros, 'Ball_Control')
```

```
##        fobs    fcritL   fcritU       pvalue
## 1 0.5909475 0.9508611 1.050977 1.122406e-85
```

```
homoscedasticidad(zurdos, diestros, 'Dribbling')
```

```
##        fobs    fcritL   fcritU       pvalue
## 1 0.6269791 0.9508611 1.050977 1.243947e-68
```

Como se puede observar, las funciones qf y pf devuelven el cuantil y la probabilidad de la distribución F respectivamente.

Como vemos aqui el valor observado (para las 3 variables) cae fuera de la zona de aceptacion de la hipotesis nula, y de hecho es menor que el limite superior (fcritU) por ende se puede puede rechazar la hipotesis nula

(H0=igualdad de varianzas). Sucede lo mismo si analizamos el valor p, basandonos en el, tambien podemos rechazar H0. Todo esto implica que estamos antes varianzas distintas.

Tambien se puede aplicar simplemente el var.test en para evaluar estos conceptos o sea si podemos asumir varianzas parecidas o no. y segun eso aplicar una formula u otra.

En R, la función var.test calcula el test de igualdad de varianzas:

```
var.test(zurdos$Rating, diestros$Rating,   conf.level = 0.95 )
```

```
##
##  F test to compare two variances
##
## data:  zurdos$Rating and diestros$Rating
## 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
```

```
var.test(zurdos$Ball_Control, diestros$Ball_Control,  conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  zurdos$Ball_Control and diestros$Ball_Control
## 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
```

```
var.test(zurdos$Dribbling, diestros$Dribbling,conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  zurdos$Dribbling and diestros$Dribbling
## 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
```

Una vez que podemos asumir que las varianzas son distintas, aplicaremos el metodo correspondiente a la media de dos poblaciones independientes con varianza desconocida distinta (referencia: “5.2.2 Varianzas desconocidas diferentes”) en la siguiente seccion.

3.5. Cálculos

Calcularemos aqui los siguientes puntos:

- *estadístico de contraste(t)*, o tambien denominado valor observado. Cuya formula depende del tipo de contraste, tipo de muestra, etc… En nuestro caso aplicaremos las formulas correspondiente a Contrates de hipotesis de dos muestras independientes con varianzas desconocidas distintas.
- *el valor crítico*, es el valor contra el cual comparar para decidir si la H0 es rechazada o no. En este caso como es unilateral a la derecha deberemos encontrar un valor positivo suficientemente grande vs el valor critico, que se obtiene a partir del nivel de confianza (alfa es el area debajo de la curva), que corresponde al error maximo permitido.

Por ejemplo supongamos tener distribucion normal si buscamos una confianza de 95% entonces el nivel de significancia alfa es 0.05 (error maximo), y el valor critico por tabla es 1.64 tendremos que testear: Si valor observado > valor critico entonces rechazo H0 a favor de H1. Ya que ese valor observado estaria fuera de la zona de aceptacion de la H0.

- *el valor p*, se calcula a partir de la curva normal y el valor observado, ej si el valor critico superior fuera 1.64, entonces lo que se hace es calcular el area a la derecha de 1.64 con la funcion pnorm (o pt segun la distribucion), osea la P(X>1.64)

Hay dos formas de rechazar la H0, como vimos arriba con el valor critico o evaluando valor P (que es el error que estaria cometiendo) contra alfa osea si valor P < alpha, osea si el error es menor que el error maximo permitido, si esto se cumple se puede rechazar la H0.

Dada este breve introduccion calculemos cada uno de los valores y evaluemos los resultados.

```
contraste_5.2.2 <- function(muestra1, muestra2, variable) {
  alfa <- 0.05
  zurdos <- muestra1[,variable]
  diestros <- muestra2[,variable]
  mean1 <- mean(zurdos); n1 <- length(zurdos); s1 <- sd(zurdos)
  mean2 <- mean(diestros); n2 <- length(diestros); s2 <- sd(diestros)

  tobs <- (mean1-mean2) / sqrt((s1**2)/n1 + (s2**2)/n2)
  numerador  <- ((s1**2)/n1 + (s2**2)/n2)**2
  denominador  <- ((s1**2)/n1)**2/(n1-1) + ((s2**2)/n2)**2/(n2-1)
  grados_libertad  <- numerador / denominador
  tcritL <- "-INF"
  tcritU <- qt( 1-alfa, df=grados_libertad)
  pvalue <-pt( tobs, df=grados_libertad, lower.tail=FALSE)

  return (
    data.frame(mean_Left=mean1,mean_Right=mean2, n_Left=n1,
               n_Right=n2, tobs, tcritL, tcritU, df=grados_libertad, pvalue)
        )
}
```

```
zurdos = jugador_de_campo('Left')
diestros = jugador_de_campo('Right')
rating_result = contraste_5.2.2(zurdos, diestros, 'Rating')
rating_result[,c('tobs', 'tcritL','tcritU','df','pvalue')]
```

```
##       tobs tcritL   tcritU       df       pvalue
## 1 5.933765   -INF 1.645065 7218.306 1.548773e-09
```

Por lo que vemos aqui el valor observado (tobs: t observado) esta por fuera de la zona de aceptacion, mayor al limite superior (t critico). Y esto sumado a que el pvalue es mucho menor que el nivel de significancia podemos rechazar la hipotesis nula. Esto implica que rechazamos la H0 a favor de aceptar la Hipotesis alternativa.

Lo mismo sucede para las otras dos variables:

```
Ball_Control_result = contraste_5.2.2(zurdos, diestros, 'Ball_Control')
Dribbling_result = contraste_5.2.2(zurdos, diestros, 'Dribbling')
Ball_Control_result[,c('tobs', 'tcritL','tcritU','df','pvalue')]
```

```
##       tobs tcritL  tcritU       df       pvalue
## 1 15.18192   -INF 1.64503 8623.776 1.071938e-51
```

```
Dribbling_result[,c('tobs', 'tcritL','tcritU','df','pvalue')]
```

```
##       tobs tcritL   tcritU       df       pvalue
## 1 18.13756   -INF 1.645036 8359.387 1.923087e-72
```

Reconfirmemos estos calculos con la ayuda de la funcion t.test:

```
t.test(zurdos$Rating,diestros$Rating, alternative="greater", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  zurdos$Rating and diestros$Rating
## t = 5.9338, df = 7218.3, p-value = 1.549e-09
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.5228087       Inf
## sample estimates:
## mean of x mean of y
##  66.58155  65.85820
```

En ambos casos, el valor p obtenido es menor que el nivel de significancia y, por lo tanto, se puede rechazar la hipótesis nula a favor de la alternativa, por lo que se reconfirma con el t.test que los zurdos tiene mejor Rating que los diestros.

Lo mismo sucede con el control del balon y el regate como vemos aqui:

```
t.test(zurdos$Ball_Control,diestros$Ball_Control,alternative="greater", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  zurdos$Ball_Control and diestros$Ball_Control
## t = 15.182, df = 8623.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
```

```
## 95 percent confidence interval:
##  3.286679      Inf
## sample estimates:
## mean of x mean of y
##  62.16335  58.47727
```

```
t.test(zurdos$Dribbling,diestros$Dribbling,alternative="greater", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  zurdos$Dribbling and diestros$Dribbling
## t = 18.138, df = 8359.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  4.597236      Inf
## sample estimates:
## mean of x mean of y
##  60.15266  55.09688
```

Si el p-value es menor que el valor de seleccionado, existen evidencias suficientes para rechazar H0 en favor de H1, osea:

H0: *rating_zurdos* <= *rating_diestros* => RECHAZADA

H1: *rating_zurdos* > *rating_diestros* => ACEPTADA

Dado que el p-value es menor que alpha, se dispone de evidencia suficiente para considerar que los zurdos puntuan mejor en las 3 variables analizadas que los diestros con una confianza del 95%.

3.6. Tabla de resultados

Incorporad una tabla de resultados con el formato que se indica a continuación. Adjuntamos el fragmento de código que hemos usado para generar la tabla para que podáis usar este mismo formato. Necesitáis usar la librería kableExtra (si es necesario, debéis instalarla previamente a su uso).

```
var=c("Rating", "BallControl", "Dribbling")
mean_Left=c(rating_result$mean_Left,Ball_Control_result$mean_Left,Dribbling_result$mean_Left)
mean_Right=c(rating_result$mean_Right,Ball_Control_result$mean_Right,Dribbling_result$mean_Right)
n_Left=c(rating_result$n_Right,Ball_Control_result$n_Right,Dribbling_result$n_Right)
n_Right=c(rating_result$n_Left,Ball_Control_result$n_Left,Dribbling_result$n_Left)
obs_value=c(rating_result$tobs,Ball_Control_result$tobs,Dribbling_result$tobs)
critical=c(rating_result$tcritU,Ball_Control_result$tcritU,Dribbling_result$tcritU)
pvalue=format(c(rating_result$pvalue,Ball_Control_result$pvalue,Dribbling_result$pvalue),
digits = 3)

data.frame(var, mean_Left,mean_Right, n_Left, n_Right,obs_value,critical, pvalue) %>%
kable() %>%
kable_styling()
```

var	mean_Left	mean_Right	n_Left	n_Right	obs_value	critical	pvalue
Rating	66.58155	65.85820	12934	4022	5.933765	1.645065	1.55e-09
BallControl	62.16335	58.47727	12934	4022	15.181923	1.645030	1.07e-51
Dribbling	60.15266	55.09688	12934	4022	18.137562	1.645036	1.92e-72

3.7. Interpretación

Como hemos comentado en la seccion 3.5, si el p-value es menor que el valor de seleccionado, existen evidencias suficientes para rechazar H0 en favor de H1, osea:

H0: *rating_zurdos* <= *rating_diestros* => RECHAZADA

H1: *rating_zurdos* > *rating_diestros* => ACEPTADA

Dado que el p-value es menor que alpha y que el valor observado esta fuera de la zona de aceptacion de la hipotesis nula, podemos afirmar que se dispone de evidencia suficiente para considerar que los zurdos puntuan mejor en las 3 variables analizadas que los diestros con un nivel confianza del 95%.

4. 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 similar**. Para dar respuesta a esta pregunta, realizaremos un proceso similar al denominado propensity score matching, aunque un poco simplificado. Realizaremos lo siguiente:

- Para cada jugador de campo zurdo, localizaremos el jugador de campo diestro más similar, en cuanto a peso, altura y edad.
- Para realizar esta búsqueda, debemos implementar un algoritmo del tipo vecino más cercano.
- Para calcular la similitud entre dos jugadores, nos basaremos en la función de distancia euclídea.

La función de distancia euclídea entre dos vectores es:

```
euclidean <- function( x1, x2 ){

  return ( sqrt( sum ( (x1-x2)^2 ) ) )

}
```

El resultado de este proceso de matching serán dos muestras:

- La muestra original está compuesta por el conjunto de jugadores de campo zurdos.
- La segunda muestra tendrá el mismo tamaño que la primera. En esta muestra hay, para cada jugador zurdo de la primera muestra, el jugador diestro con características físicas similares. Es decir, existe una correspondencia entre el jugador 1 de la muestra de zurdos, con el jugador 1 de la muestra de diestros, y así para todos los jugadores.

A partir de estas muestras, nos preguntaremos si los jugadores zurdos son mejores en Rating que los jugadores diestros relacionados.

Para poder realizar este análisis, calcularemos primero el jugador diestro más similar a cada jugador zurdo.

4.1. Jugador más similar

En primer lugar, implementad una función **my.nn** que, dado un jugador calcule el jugador más similar en términos de edad, peso y altura. La función debe ser:

```
my.nn <- function( x, sample ){

}
```

donde x es el jugador zurdo, de tipo vector que guarda la edad, peso y altura. Y **sample** es la muestra de diestros con valores en edad, peso y altura. Para calcular el ejemplo de **sample** más similar a x, usad la función de distancia euclídea que os hemos suministrado. La función **my.nn** devuelve el índice (posición) del jugador de la muestra **sample** que se parece más a x.

Una vez realizada esta función, podemos usarla para calcular, para cada jugador de la muestra de zurdos, el jugador diestro más similar. Para ello, implementad la función siguiente:

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

}
```

donde **sample1** es la muestra de jugadores zurdos y **sample2** es la muestra de jugadores diestros. Esta función devuelve la muestra Right.paired, que contiene el listado de jugadores diestros más similares a los jugadores zurdos de la muestra 1. En definitiva, esta función realiza una iteración sobre la muestra de zurdos. Para cada jugador de la muestra de zurdos, llama a la función **my.nn**.

Recomendamos que probéis la función con una muestra pequeña de datos para validar que el código es correcto.

Cambiamos un poco la estrategia definida para generar una funcion que en lugar de una lista de zurdos y otra lista de diestros, genere como salida un dataframe con dos columnas, la primera contiene el index del dataframe de zurdos y la segunda columna el index del dataframe de diestros, donde este ultimo index corresponde al index del jugador diestro mas similar al zurdo.

Se utiliza la funcion dist del paquete proxy que nos permite comparar una observacion (con todas sus variables) contra un dataframe completo y nos devuelve un vector con las distancias. Vector sobre el cual buscamos el minimo para esa observacion. Lo hacemos de esta manera porque estamos aprovechando una funcion de R que es mas optimo que realizar for anidados, algo poco optimo en cualquier lenguaje de programacion. Esto nos permite ademas usar la muestra completa de jugadores, no solo un sampleo de la muestra.

```
zurdos_some = zurdos[,c("Age", "Weight", "Height")]
diestros_some = diestros[,c("Age", "Weight", "Height")]

my.nn = function(df1, df2) {
  near_rows= data.frame(matrix(ncol = 2, nrow = 0))
  names(near_rows) = c("index_left", "index_right")
  for(k in 1:nrow(df1)) {
    df.dist<-dist(df1[k,],df2)
    closest.row<-row.names(df2)[which.min(df.dist)]
    near_rows[k,] = cbind(row.names(df1[k,]), closest.row)
  }
  return (near_rows)
}

jugadores_similares = my.nn(zurdos_some, diestros_some)
```

Aqui vemos un sample del resultado:

```
head(jugadores_similares)
```

```
##   index_left index_right
## 1          2        4179
## 2          8        7929
## 3         14        4519
## 4         20       10272
## 5         28        5354
## 6         35        3951
```

```
tail(jugadores_similares)
```

```
##      index_left index_right
## 4017      17519       13982
## 4018      17523       12709
## 4019      17527       16729
## 4020      17529       13451
## 4021      17537       11709
## 4022      17550       14015
```

```
nrow(jugadores_similares)
```

```
## [1] 4022
```

4.2. Muestras

Llegados a este punto, tenemos dos muestras: **Left.sample**, con los jugadores de campo zurdos. Y **Right.paired**, que contiene los jugadores diestros más similares a los jugadores de la muestra **Left.sample**.

Mostrad las primeras filas de las dos muestras.

A partir del resultado de la aplicacion la funcion my.nn creada en el paso anterior generamos dos muestras (Left.sample y Right.paired) tal como es solicitado:

```
Left.sample = zurdos[jugadores_similares$index_left,]
Right.paired = diestros[jugadores_similares$index_right,]

cols = c('Name','Nationality','Club','Rating','Age','Height','Weight',
         'Dribbling','Ball_Control')
head(Left.sample[,cols])
```

```
##                 Name Nationality            Club Rating Age Height Weight
## 2       Lionel Messi   Argentina    FC Barcelona     93  29    179     72
## 8        Gareth Bale       Wales     Real Madrid     90  27    183     74
## 14        Mesut Özil     Germany         Arsenal     89  28    180     76
## 20 Antoine Griezmann      France Atlético Madrid     88  25    176     67
## 28 Giorgio Chiellini       Italy        Juventus     88  32    187     84
## 35   James Rodríguez    Colombia     Real Madrid     87  25    180     75
##    Dribbling Ball_Control
```

```
## 2         97           95
## 8         89           88
## 14        86           90
## 20        87           86
## 28        56           55
## 35        84           85
```

```
nrow(Left.sample)
```

```
## [1] 4022
```

```
head(Right.paired[,cols])
```

```
##                      Name Nationality            Club Rating Age Height Weight
## 4179  Juan Carlos Paredes     Ecuador  Olympiakos CFP     71  29    179     72
## 7929        Juan Arboleda    Colombia Deportes Tolima     67  27    183     74
## 4519        Efraín Juárez      Mexico       Monterrey     71  28    180     76
## 10272   Mayoro Ndoye-Baye     Senegal   RC Strasbourg     65  25    176     67
## 5354       Alex Wilkinson   Australia       Sydney FC     70  32    187     85
## 3951        Zeki Yildirim      Turkey     Antalyaspor     71  25    180     75
##       Dribbling Ball_Control
## 4179         69           67
## 7929         60           66
## 4519         64           70
## 10272        59           65
## 5354         42           60
## 3951         66           67
```

```
nrow(Right.paired)
```

```
## [1] 4022
```

4.3. Hipótesis nula y alternativa

A partir de las dos muestras, ¿podemos afirmar que los zurdos son mejores en Rating que los correspondientes jugadores diestros?

Procedemos de manera similar al punto 3, por lo que plantearemos las siguientes hipotesis:

El rating de los jugadores zurdos es mejor que el de los diestros?

Osea aqui queremos validar que:

rating_zurdos > rating_diestros

Por lo que las hipotesis nula (H0) y alternativa (H1) serian:

H0: **rating_zurdos <= rating_diestros** o **rating_zurdos-rating_diestros <= 0**

H1: **rating_zurdos > rating_diestros** o **rating_zurdos-rating_diestros > 0**

4.4. Método

El metodo que aplicaremos es *Contraste de hipotesis de dos muestras apareadas o emparejadas sobre la media.* Se los vincula a partir de las caracteristicas fisicas, ya que la muestra de diestros que se comparara contra la de zurdos fue obtenida en el punto 4.1 a partir de determinar las caracteristicas (edad, altura y peso) mas similares entre ambos grupos.

Por lo tanto sabiendo que tenemos dos muestras una de zurdos y otra de diestros y aplicaremos el metodo de dos muestras emparejadas, debemos calcular la diferencia entre los datos relacionados de estas dos muestras.

Por ejemplo teniendo ambas muestras:

```
x = Left.sample$Rating
```

```
y = Right.paired$Rating
```

Calcularemos el vector de diferencias de esta forma:

```
*di = xi - yi*
```

Por lo que ahora es como si fuera que tengamos una sola muestra: *D = (di, d2,… dn)*, y es entonces que reformulamos un poco las H0 y H1 de esta manera para trabajar con el metodo:

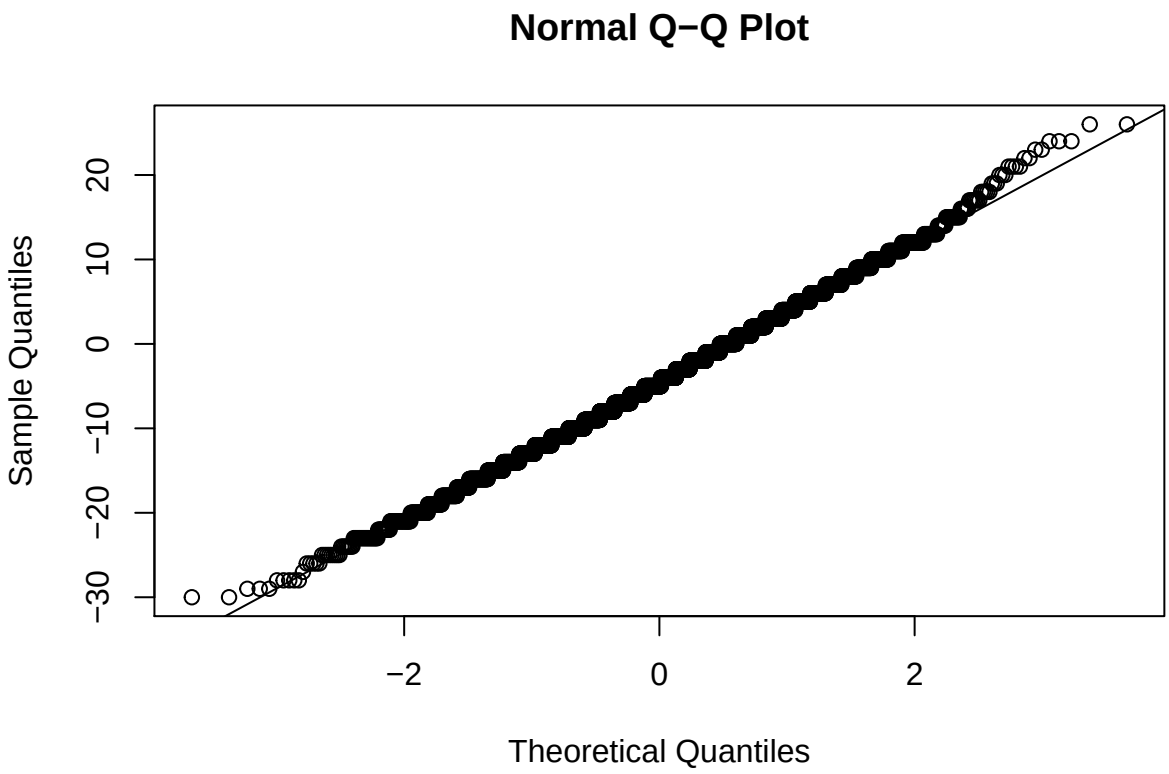
H0: $d = 0$

H1: $d > 0$

En cuanto a la normalidad de esta muestra, ya lo hemos visto antes, pero no esta de mas volver a mencionar que podemos basarnos en el Teorema del Limite Central, por lo que asumimos normalidad por el TLC, dado que el tamaño de la muestra es suficientemente grande (mayor a 30). De hecho la muestra de zurdos tiene un tamaño de 4022 y la de diestros que esta apareada tiene la misma cantidad. Entonces por el TLC, podemos asumir que las dos medias de cada muestra siguen distribuciones normales.

En caso de que igualmente quisieramos realizar un test rapido es posible aplicar el test de normalidad Shapiro-Wilk y dibujar el gráfico Q-Q

```
x = Left.sample$Rating
y = Right.paired$Rating
dif_rating=x-y

qqnorm(dif_rating)
qqline(dif_rating)
```




```
shapiro.test(dif_rating)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dif_rating
## W = 0.99788, p-value = 2.574e-05
```

Como vemos en la grafica los puntos estan casi todos pegados a la linea diagonal por lo que se puede asumir normalidad, mas alla de que con Shapiro-Wilk obtengamos un valor que nos permitiria rechazar la hipotesis nula de normalidad. Pero en este caso vamos a basarnos en el TLC y en el grafico Q-Q (que es un tipo de visualización que se utiliza para diagnosticar la desviación de los datos de la muestra en relación con una población normal, esta utima represenda en una linea diagonal mediante qqline) para asumir la normalidad.

Dicho todo esto continuemos en el siguiente punto la implementacion de los calculos.

4.5. Cálculos

Calcularemos aqui los siguientes puntos usando un nivel de confianza del 95%

- *estadístico de contraste(t)*
- *el valor crítico*
- *el valor p*

```
x = Left.sample$Rating
y = Right.paired$Rating

#diferencia de muestras:
dif_rating = x-y
head(dif_rating)
```

```
## [1] 22 23 18 23 18 16
```

Si no se conoce la varianza de la población, como es nuestro caso, entonces hay que estimarla a partir de la desviación de la muestra. En este caso el estadístico de contraste sigue una distribución t de Student con n – *1* grados de libertad. Teniendo en cuenta, claro, que se aplicara este test a la muestra de las diferencias, la formula para su cualculo es:

tobs = mean(D) / (sD / sqrt(nD))

donde D es la media de las diferencias, sD es la desviación estándar y nD es el número de elementos de la muestra de las diferencias.

Veamos el codigo R:

```
nD <- length(dif_rating)
alfa <- 0.05 # unilateral a la derecha
meanD = mean(dif_rating);
sD = sd( dif_rating )
tobs = (meanD)/(sD/sqrt(nD))

#Región de aceptación unilateral por la derecha
tcrit.L = '-INF'
tcrit.U = qt(1-alfa, df=nD-1)

#Cálculo del valor p unilateral
pvalue = pt(tobs, lower.tail=FALSE, df=nD-1)

# Visualizar resultado
data.frame(tcrit.L, tcrit.U, tobs, pvalue)
```

```
##   tcrit.L  tcrit.U      tobs pvalue
## 1    -INF 1.645233 -33.84937      1
```

Con los valores obtenidos vemos que el valor observado cae dentro de la zona de la aceptacion de la hipotesis nula, por ende, podemos inferir que los zurdos no son necesariamente mejores que los diestros en igualdad de condiciones (edad, peso y altura). De hecho se reconfirma con el pvalue cuyo valor es mayor al error maximo permitido (alpha), lo cual indica que no se puede refutar la hipotesis nula.

Validemos lo calculado arriba mediante el uso t.test:

```
t.test( Left.sample$Rating, Right.paired$Rating, paired=TRUE, alternative="greater")
```

```
##
##  Paired t-test
##
## data:  Left.sample$Rating and Right.paired$Rating
```

```
## t = -33.849, df = 4021, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -4.658234       Inf
## sample estimates:
## mean of the differences
##               -4.442317
```

Como reconfirmamos aqui con un p-value > alpha, no podemos rechazar la hipotesis nula, por lo que se infiere o no hay evidencias suficientes que afirmen que los jugadores zurdos son mejores que los derechos que posean caracteristicas similares de edad, altura y peso, con un nivel de confianza del 95%.

4.6. Interpretación

Tal como fue explicado en el punto anterior, al aparear las muestras de zurdos y diestros a traves de su edad y caracteristicas fisicas notamos que bajo esa igualdad o similitud de variables los zurdos no son mejores que los diestros. Ya que la hipotesis nula no pudo ser rechazara al caer el valor observado en su area de aceptacion y el p-value es mayor al error maximo permitido.

4.7. Reflexión

Lo que vemos es que al realizar contrastes de muestras emparejadas de poblaciones podemos llegar a tener resultados totalmente opuestos a cuando contrastamos muestras de esa misma poblacion pero independientes entre ellas.

En las muestras independientes obtuvimos por ejemplo que hay evidencias que indican que en general los zurdos son mejores que los diestros, pero al emparejarlos por edad, peso y altura hubiesemos creido que las diferencias se harian mas notorias que el rating de un zurdo hubiese sido en la mayoria de los casos mayor que el de los diestros, de hecho esa es la "idea popular", pero no.

Con estos dos contrastes de hipotesis podemos concluir varias cosas:

1. que la inferencia estadistica, en particular el contraste de hipotesis, es una gran herramienta o metodologia para aceptar o refutar cualquier idea, hasta las mas instauradas en el publico general
2. que es evidente que el zurdo en general es mejor que el diestros para ciertos parametros o variables de su juego, sin importar la edad, el peso, la altura o cualquier atributo fisico de los mismos.
3. pero que esa mejor aptitud en general no es asi cuando ambos grupos tienen caracteristicas fisicas similares o igualadas.

Se podrian realizar planter mas hipotesis y realizar otros contrastes para entender porque se da el punto 2 y porque el 3. Por ejemplo compararlos apareados pero sin la edad, solo por peso y altura, y poder validar si los zurdos son mejores o no a cierta edad. O viceversa entender si la mejora o no del zurdo tiene que ver con el peso y altura de los mismos. En futbol muchas veces influye la estatura de los jugadores. Es de publico conocimiento que los jugadores de futbol (al contrario que en el basquet) de gran altura en general, o al menos se supone, que no son mas habiles. Sera que en los jugadores zurdos son mas bajos que los diestros?. Y si los apareamos sin la estatura obtenemos valores diferentes?

Como vemos la inferencia tambien es un proceso iterativo, el haber realizado los contrastes del ejercicio 3 y 4, puede llevarnos a plantear mas hipotesis para entender aun mas los datos, en este caso las muestras de cierta poblacion.

5. Comparación entre clubes

Es bien conocida la rivalidad entre los clubes de Barcelona y Madrid. Se desea calcular si el porcentaje de jugadores con un Rating superior a 90 es diferente en Barcelona y Madrid con un nivel de confianza del 97 %.

Para ello, seguid los pasos que se especifican a continuación.

5.1. Hipótesis nula y alternativa

Siendo que nuestra pregunta de investigacion claramente se trata de un contrates sobre la proporcion, donde queremos entender si la cantidad de jugadores de un club con cierta caracteristica (Rating > 90) es diferente a la del otro club, planteramos la hipotesis nula y alternativa para este caso.

Formulamos entonces la hipótesis nula e hipótesis alternativa

```
H0: p_bcn = p_rm
```

```
H1: p_bcn <> p_rm
```

Donde:

```
p_bcn = proporcion de jugadores del Barcelona con rating mayor a 90.
```

```
p_rm = proporcion de jugadores del Real Madrid con rating mayor a 90.
```

5.2. Método

Dado que queremos inferir si el rating entre los jugadores de ambos equipos es diferente realizaremos un Contraste de Hipotesis de dos muestras sobre la proporcion aplicando un test bilateral.

5.3. Cálculos

Dado que este contrate es sobre la proporcion de jugadores con rating mayor a 90 entre ambos clubes, incluiremos a todos los jugadores, no solo los de campo.

Aqui calculamos las proporciones para ambas muestras:

```
bcn_players = fifa2017['FC Barcelona' == fifa2017$Club,]
bcn_players_rating_greater_90 = bcn_players[bcn_players$Rating > 90,]
bcn_ratio_greater_90 = nrow(bcn_players_rating_greater_90) / nrow(bcn_players)


rm_players = fifa2017['Real Madrid' == fifa2017$Club,]
rm_players_rating_greater_90 = rm_players[rm_players$Rating > 90,]
rm_ratio_greater_90 = nrow(rm_players_rating_greater_90) / nrow(rm_players)
```

Realizando un calculo simple sobre la muestra, la proporcion de jugadores del Barcelona con rating mayor a 90 es:

```
bcn_ratio_greater_90
```

```
## [1] 0.09090909
```

Mientras que la del Madrid es:

```
rm_ratio_greater_90
```

```
## [1] 0.03030303
```

Por lo que la proporcion es de 3 a 1 para esta muestra particular. Pero que sucede con esto si lo inferimos mediante el contraste? Veamos:

Calculamos a partir de la proporción observada en la muestra: el estadístico de contraste (valor observado), los valores criticos y el valor p.

```
alpha = 0.03
n1_bcn = nrow(bcn_players)
n2_rm = nrow(rm_players)

# Calculo del parámetro de la proporción del fenómeno en la población.:
p<-(n1_bcn*bcn_ratio_greater_90 + n2_rm*rm_ratio_greater_90) / (n1_bcn+n2_rm)

# Calculo del valor observado
zobs <- (bcn_ratio_greater_90-rm_ratio_greater_90)  /
        sqrt(p*(1-p)*(1/n1_bcn+1/n2_rm))

# Calculo de los limites de aceptacion
zcritL <- qnorm(alpha/2, lower.tail=TRUE)
zcritU <- qnorm(1-alpha/2, lower.tail=TRUE)

# Determinamos la probabilidad de P(z   -1.04)
pvalue<- pnorm(zobs, lower.tail=FALSE)*2 # Multiplicamos x2 porque el test es bilateral

# Visualicemos via un dataframe los valores obtenidos
data.frame(zobs, zcritL, zcritU, pvalue)
```

```
##       zobs   zcritL  zcritU    pvalue
## 1 1.031754 -2.17009 2.17009 0.3021874
```

Se comprueba que el valor observado 1.03 está dentro de la región de aceptación de la hipótesis nula y, por lo tanto, no podemos rechazar H0.

El valor p es P(z>=2.17009)*2 = 0.3021874. Dado que no es inferior al nivel de significancia fijado de 0.03, no podemos rechazar la hipótesis nula

Por lo tanto, se concluye que no contamos con evidencia suficiente para confirmar la proporcion de jugadores con rating superior a 90 sea diferente entre ambos clubes.

Reconfirmemos el resultado con prop.test

```r
success<-c( bcn_ratio_greater_90*n1_bcn, rm_ratio_greater_90*n2_rm)
nn<-c(n1_bcn,n2_rm)
prop.test(success, nn,correct=FALSE, conf.level = 0.97)
```

```
## Warning in prop.test(success, nn, correct = FALSE, conf.level = 0.97): Chi-
## squared approximation may be incorrect
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 1.0645, df = 1, p-value = 0.3022
## alternative hypothesis: two.sided
## 97 percent confidence interval:
##  -0.06583462  0.18704674
## sample estimates:
##     prop 1     prop 2
## 0.09090909 0.03030303
```

Nuevamente con los resultados obtenidos, no podemos rechazar la hipótesis nula con un nivel de confianza del 97 %.

5.4. Resultados e interpretación

Como vimos el valor observado esta dentro del rango , zcrit.L - zcrit.U, se acepta la hipótesis nula y se rechaza la Hipotesis Alternativa en la que se dice que: El Porcentaje de jugadores con Rating > 90 es diferente entre los equipos del Madrid y el Barcelona.Tambien se corrobora esto con el p_value (0.03022) que es mayor que el error maximo permitido o nivel de significancia (alpha= 0.03).

Todo lo comentado en el parrafo anterior, es si nos basamos exclusivamente en los numeros, pero como vemos el P-value esta al limite, que hubiese pasado si subiamos el nivel de significancia? Seguramente el valor observado permaneceria dentro de la zona de aceptacion de la hipotesis nula, pero el el p-value seria menor al alpha. En ese caso para por ejemplo un nivel de confianza del 95 %, se podría rechazar la hipótesis nula, puesto que el valor p es 0.3022, inferior a alpha=0.05. Por lo tanto, se podría concluir que la proporción de jugadores con rating mayor a 90 es diferente entre ambos equipos con un nivel de confianza del 95 %.

Esto lo mencionamos ya que como vimos en los numeros reales, ya sabemos de ante mano que la proporcion (Rating>90) de jugadores del Barcelona es mayor a los del Madrid. Pero claro, esos son solo los hechos de una muestra particular, no una inferencia sobre las proporciones en la poblacion como hemos demostrado arriba.

6. Resumen ejecutivo

A modo de resumen sobre las preguntas de investigacion planteadas en cada punto tenemos.

1. En general los jugadores zurdos puntuan mejor que los diestros en variables como el Rating, Ball Control y Dribbling.

   Se pudo reunir evidencia suficiente con un nivel de confianza del 95% para afirmar esta hipotesis.

2. Los jugadores zurdos con edades, estatura y peso similares a los diestros son mejores en Rating, Ball Control y Dribbling?

   No se ha encontrado evidencia que nos permita asegurar con un nivel de confianza del 95% que los zurdos sean mejores que los diestros bajo estas condiciones.

3. La proporcion de jugadores con rating elevado en Barcelona es diferente al del Real Madrid?

   La evidencia reunida no corrobara con un nivel de confianza del 97% que la proporcion de ratings mayores a 90 sea diferente entre ambos clubes.

Referencias

https://www.rdocumentation.org/packages/proxy/versions/0.4-25/topics/dist

https://medium.com/@sorenlind/create-pdf-reports-using-r-r-markdown-latex-and-knitr-on-windows-10-952b0c48bfa9