A2 - Analítica Descriptiva e Inferencial

Solución

Semestre 2020.2

Índice

1. Lectura del fichero **3**

2. Rating de los jugadores **6**
2.1. Análisis visual . . . 6
2.2. Intervalo de confianza . . . 8

3. Diferencias entre jugadores **9**
3.1. Pregunta de investigación . . . 10
3.2. Representación visual . . . 10
3.3. Hipótesis nula y alternativa . . . 11
3.4. Método . . . 11
3.5. Cálculos . . . 12
3.6. Tabla de resultados . . . 13
3.7. Interpretación . . . 14

4. Comparación por pares **14**
4.1. Jugador más similar . . . 15
4.2. Muestras . . . 16
4.3. Hipótesis nula y alternativa . . . 17
4.4. Método . . . 17
4.5. Cálculos . . . 17
4.6. Interpretación . . . 18
4.7. Reflexión . . . 18

5. Comparación entre clubes **19**
5.1. Hipótesis nula y alternativa . . . 19
5.2. Método . . . 19
5.3. Cálculos . . . 19
5.4. Resultados e interpretación . . . 20

6. Resumen ejecutivo **20**

Introducción

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.

```
## [1] 17588    54
```

```
##        ID                      Name          Nationality
##  Min.   :    1   Danilo          :    5   England  : 1618
##  1st Qu.: 4398   Gabriel         :    5   Argentina: 1097
##  Median : 8794   Carlos Rodríguez:    4   Spain    : 1008
##  Mean   : 8794   Felipe          :    4   France   :  974
##  3rd Qu.:13191   Roberto         :    4   Brazil   :  921
##  Max.   :17588   Álvaro          :    3   Italy    :  751
##                  (Other)         :17563   (Other)  :11219
##  National_Position  National_Kit                Club       Club_Position
##         :16513     Min.   : 1.00   Free Agents    :  232   Sub    :7492
##  Sub    :  556     1st Qu.: 6.00   Angers SCO     :   33   Res    :3146
##  LCB    :   48     Median :12.00   Arsenal        :   33   RCB    : 633
##  GK     :   47     Mean   :12.22   AS Monaco      :   33   GK     : 632
##  RCB    :   46     3rd Qu.:18.00   Bor. M'gladbach:   33   LCB    : 631
##  LB     :   39     Max.   :36.00   Bournemouth    :   33   LB     : 549
##  (Other):  339     NA's   :16513   (Other)        :17191   (Other):4505
##     Club_Kit         Club_Joining   Contract_Expiry     Rating
##  Min.   : 1.00   07/01/2016: 1193   Min.   :2017    Min.   :45.00
##  1st Qu.: 9.00   07/01/2015:  907   1st Qu.:2017    1st Qu.:62.00
##  Median :18.00   07/01/2014:  558   Median :2019    Median :66.00
##  Mean   :21.29   01/01/2016:  412   Mean   :2019    Mean   :66.17
##  3rd Qu.:27.00   07/01/2013:  404   3rd Qu.:2020    3rd Qu.:71.00
##  Max.   :99.00   01/01/2015:  391   Max.   :2023    Max.   :94.00
##  NA's   :1       (Other)   :13723   NA's   :1
##      Height          Weight       Preffered_Foot      Birth_Date
##  Min.   :155.0   Min.   : 48.00   Left : 4094    02/29/1988:  160
##  1st Qu.:176.0   1st Qu.: 70.00   Right:13494    02/29/1984:  157
##  Median :181.0   Median : 75.00                  02/29/1992:  155
##  Mean   :181.1   Mean   : 75.25                  01/01/1996:   13
##  3rd Qu.:186.0   3rd Qu.: 80.00                  11/11/1996:   13
##  Max.   :207.0   Max.   :110.00                  01/08/1991:   12
##                                                  (Other)   :17078
##       Age        Preffered_Position           Work_Rate      Weak_foot
##  Min.   :16.00   CB     :2181       Medium / Medium:9897   Min.   :1.000
##  1st Qu.:21.00   GK     :2003       High / Medium  :2918   1st Qu.:3.000
##  Median :25.00   ST     :1825       Medium / High  :1534   Median :3.000
##  Mean   :25.14   CM     : 831       Medium / Low   : 845   Mean   :2.934
##  3rd Qu.:28.00   LB     : 808       High / High    : 747   3rd Qu.:3.000
##  Max.   :47.00   RB     : 689       High / Low     : 730   Max.   :5.000
##                  (Other):9251       (Other)        : 917
```

```
##   Skill_Moves     Ball_Control     Dribbling       Marking
##  Min.   :1.000   Min.   : 5.00   Min.   : 4.0   Min.   : 3.00
##  1st Qu.:2.000   1st Qu.:53.00   1st Qu.:47.0   1st Qu.:22.00
##  Median :2.000   Median :63.00   Median :60.0   Median :48.00
##  Mean   :2.303   Mean   :57.97   Mean   :54.8   Mean   :44.23
##  3rd Qu.:3.000   3rd Qu.:69.00   3rd Qu.:68.0   3rd Qu.:64.00
##  Max.   :5.000   Max.   :95.00   Max.   :97.0   Max.   :92.00
##
##  Sliding_Tackle  Standing_Tackle   Aggression      Reactions
##  Min.   : 5.00   Min.   : 3.00   Min.   : 2.00   Min.   :29.00
##  1st Qu.:23.00   1st Qu.:26.00   1st Qu.:44.00   1st Qu.:55.00
##  Median :51.00   Median :54.00   Median :59.00   Median :62.00
##  Mean   :45.57   Mean   :47.44   Mean   :55.92   Mean   :61.77
##  3rd Qu.:64.00   3rd Qu.:66.00   3rd Qu.:70.00   3rd Qu.:68.00
##  Max.   :95.00   Max.   :92.00   Max.   :96.00   Max.   :96.00
##
##  Attacking_Position Interceptions       Vision        Composure
##  Min.   : 2.00      Min.   : 3.00   Min.   :10.00   Min.   : 5.00
##  1st Qu.:37.00      1st Qu.:26.00   1st Qu.:43.00   1st Qu.:47.00
##  Median :54.00      Median :52.00   Median :54.00   Median :57.00
##  Mean   :49.59      Mean   :46.79   Mean   :52.71   Mean   :55.85
##  3rd Qu.:64.00      3rd Qu.:64.00   3rd Qu.:64.00   3rd Qu.:66.00
##  Max.   :94.00      Max.   :93.00   Max.   :94.00   Max.   :94.00
##
##     Crossing       Short_Pass      Long_Pass     Acceleration
##  Min.   : 6.00   Min.   :10.00   Min.   : 7.0   Min.   :11.00
##  1st Qu.:38.00   1st Qu.:52.00   1st Qu.:42.0   1st Qu.:57.00
##  Median :54.00   Median :62.00   Median :56.0   Median :68.00
##  Mean   :49.74   Mean   :58.12   Mean   :52.4   Mean   :65.29
##  3rd Qu.:64.00   3rd Qu.:68.00   3rd Qu.:64.0   3rd Qu.:75.00
##  Max.   :91.00   Max.   :92.00   Max.   :93.0   Max.   :96.00
##
##      Speed          Stamina         Strength        Balance
##  Min.   :11.00   Min.   :10.00   Min.   :20.00   Min.   :10.00
##  1st Qu.:58.00   1st Qu.:57.00   1st Qu.:57.00   1st Qu.:56.00
##  Median :68.00   Median :66.00   Median :66.00   Median :65.00
##  Mean   :65.48   Mean   :63.48   Mean   :65.09   Mean   :64.01
##  3rd Qu.:75.00   3rd Qu.:74.00   3rd Qu.:74.00   3rd Qu.:74.00
##  Max.   :96.00   Max.   :95.00   Max.   :98.00   Max.   :97.00
##
##     Agility         Jumping         Heading        Shot_Power
##  Min.   :11.00   Min.   :15.00   Min.   : 4.00   Min.   : 3.00
##  1st Qu.:55.00   1st Qu.:58.00   1st Qu.:45.00   1st Qu.:45.00
##  Median :65.00   Median :65.00   Median :56.00   Median :59.00
##  Mean   :63.21   Mean   :64.92   Mean   :52.39   Mean   :55.58
##  3rd Qu.:74.00   3rd Qu.:73.00   3rd Qu.:65.00   3rd Qu.:69.00
##  Max.   :96.00   Max.   :95.00   Max.   :94.00   Max.   :93.00
##
##    Finishing       Long_Shots       Curve       Freekick_Accuracy
##  Min.   : 2.00   Min.   : 4.0   Min.   : 6.00   Min.   : 4.00
##  1st Qu.:29.00   1st Qu.:32.0   1st Qu.:34.00   1st Qu.:31.00
##  Median :48.00   Median :52.0   Median :48.00   Median :42.00
##  Mean   :45.16   Mean   :47.4   Mean   :47.18   Mean   :43.38
##  3rd Qu.:61.00   3rd Qu.:63.0   3rd Qu.:62.00   3rd Qu.:57.00
```

```
##  Max.   :95.00   Max.   :91.0   Max.   :92.00   Max.   :93.00
##
##    Penalties        Volleys      GK_Positioning    GK_Diving
##  Min.   : 7.00   Min.   : 3.00   Min.   : 1.00   Min.   : 1.00
##  1st Qu.:39.00   1st Qu.:30.00   1st Qu.: 8.00   1st Qu.: 8.00
##  Median :50.00   Median :44.00   Median :11.00   Median :11.00
##  Mean   :49.17   Mean   :43.28   Mean   :16.61   Mean   :16.82
##  3rd Qu.:61.00   3rd Qu.:57.00   3rd Qu.:14.00   3rd Qu.:14.00
##  Max.   :96.00   Max.   :93.00   Max.   :91.00   Max.   :89.00
##
##    GK_Kicking     GK_Handling     GK_Reflexes
##  Min.   : 1.00   Min.   : 1.00   Min.   : 1.0
##  1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 8.0
##  Median :11.00   Median :11.00   Median :11.0
##  Mean   :16.46   Mean   :16.56   Mean   :16.9
##  3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.0
##  Max.   :95.00   Max.   :91.00   Max.   :90.0
##
## 'data.frame':    17588 obs. of  54 variables:
##  $ ID                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name              : Factor w/ 17347 levels "A.j. Delagarza",..: 3272 9929 12464 10273 10559 3902 1
##  $ Nationality       : Factor w/ 160 levels "Afghanistan",..: 122 6 20 155 59 139 121 158 143 14 ...
##  $ National_Position : Factor w/ 28 levels "","CAM","CB",..: 14 25 15 14 6 6 14 24 1 6 ...
##  $ National_Kit      : int  7 10 10 9 1 1 9 11 NA 1 ...
##  $ Club              : Factor w/ 634 levels "1. FC Heidenheim",..: 460 204 204 204 206 361 206 460 30
##  $ Club_Position     : Factor w/ 30 levels "","CAM","CB",..: 16 27 16 29 7 7 29 27 29 7 ...
##  $ Club_Kit          : int  7 10 11 9 1 1 9 11 9 13 ...
##  $ Club_Joining      : Factor w/ 1678 levels "","01/01/1993",..: 848 843 852 927 850 850 853 1247 855
##  $ 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    : Factor w/ 2 levels "Left","Right": 2 1 2 2 2 2 2 1 2 1 ...
##  $ Birth_Date        : Factor w/ 6063 levels "01/01/1982","01/01/1983",..: 623 2991 630 412 1490 5211
##  $ Age               : int  31 29 24 29 30 26 28 27 35 24 ...
##  $ Preffered_Position: Factor w/ 292 levels "CAM","CAM/CDM",..: 172 237 157 266 113 113 266 237 266 1
##  $ Work_Rate         : Factor w/ 9 levels "High / High",..: 2 9 3 3 9 9 3 3 8 9 ...
##  $ 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 ...
```

---

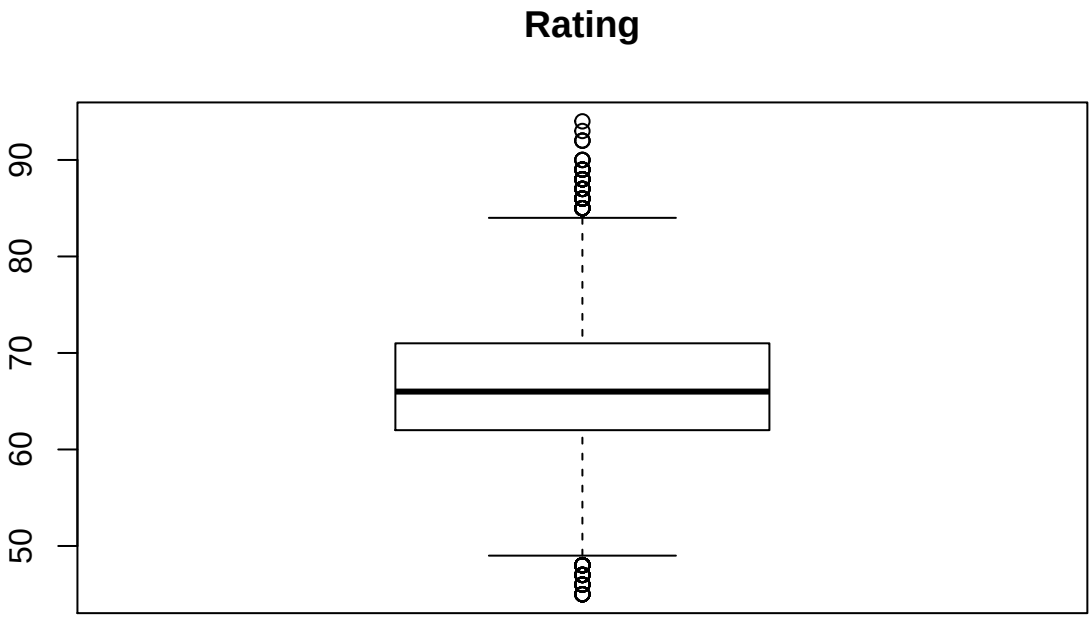
2. Rating de los jugadores

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.

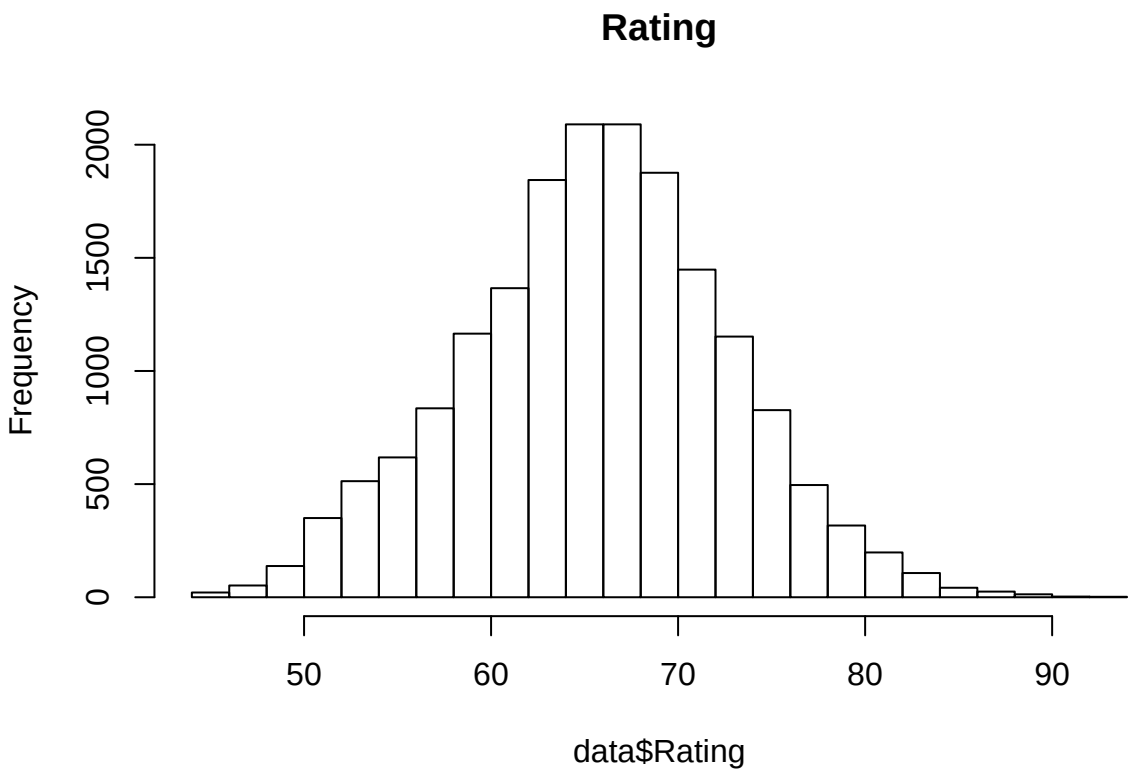
2.1. Análisis visual

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.

```r
boxplot(data$Rating, main="Rating")
```




```
hist( data$Rating, main="Rating", breaks=20)
```



Interpretación: Los valores de Rating se distribuyen entre los 40 y 100 puntos. A partir del histograma, se observa que sigue una distribución normal, centrada en el valor 66 aproximadamente.

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.

Solución:

```
n<-nrow( data )
alpha<-1-0.95
#Error típico
errorTipico <- sd(data$Rating) / sqrt( n )
errorTipico
```

```
## [1] 0.05340842
```

```
#Valor z
t<-qt( 1-alpha/2, df=n-1)
t
```

```
## [1] 1.960099
```

```r
#Margen de error
error<- t * errorTipico
error
```

```
## [1] 0.1046858
```

```r
#Intervalo
c( mean(data$Rating) - error, mean(data$Rating) + error )
```

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

```r
#Comprobación con el test t de Student.
t.test( data$Rating, conf.level=0.95 )
```

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

Interpretación:

Se ha calculado el intervalo de confianza del 95 % de Rating, asumiendo distribución normal y varianza poblacional desconocida y aplicando por tanto, distribución t Student con n-1 grados de libertad. Este intervalo es:

66.0615071, 66.2708786

La interpretación del intervalo de confianza es: El 95 % de los intervalos que construimos a partir de infinitas muestras de la población contendrían el valor medio de la variable Rating.

---

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 zurdos (no porteros) (**Preferred_Foot** igual a Left). La segunda muestra contiene todos los jugadores diestros (no porteros) (**Preferred_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 comparar. No deberías escribir el mismo código tres veces.

Seguid los pasos que se especifican a continuación.

3.1. Pregunta de investigación

Formulad la/s pregunta/s de investigación que se plantea/n en este apartado.

Pregunta de investigación:

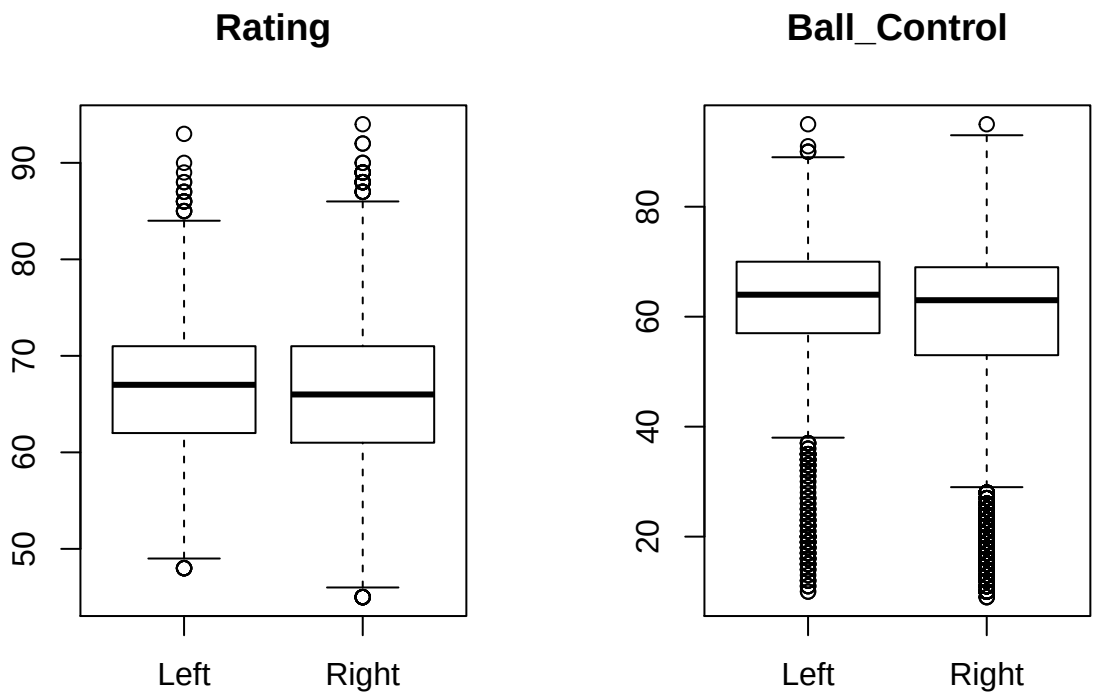
¿Los jugadores de campo zurdos tienen mejor valor en Rating, Ball_Control y Dribbling que los jugadores diestros?

3.2. Representación visual

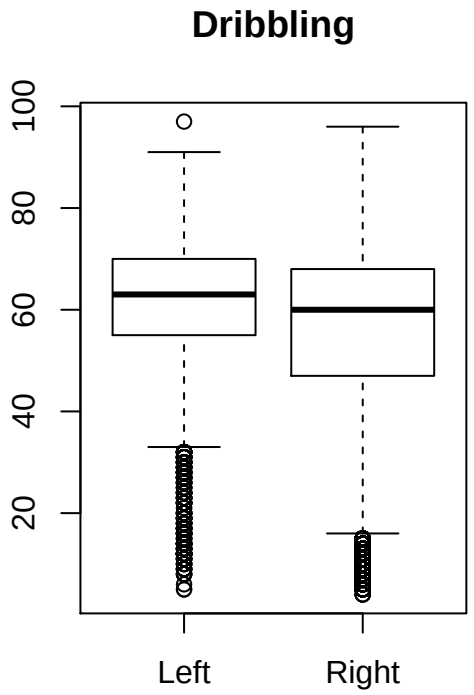
Representad visualmente, mediante el gráfico que sea más apropiado el valor de estas variables en jugadores diestros y zurdos. Se deben mostrar los valores de forma comparativa entre zurdos y diestros. Interpretad los gráficos.

```
#jugadores de campo: Left y Right
player <- data[ data$Club_Position!="GK", ]
Left <- player[ player$Preffered_Foot=="Left",]
Right <- player[ player$Preffered_Foot=="Right",]

par( mfrow=c(1,2))
boxplot( Left$Rating, Right$Rating, names=c("Left","Right"), main="Rating" )
boxplot( Left$Ball_Control, Right$Ball_Control, names=c("Left","Right"), main="Ball_Control" )
```



```
boxplot( Left$Dribbling, Right$Dribbling, names=c("Left","Right"), main="Dribbling" )
par( mfrow=c(1,1))
```



Interpretación: Se observan ligeras diferencias en las variables Rating y Ball_Control entre los jugadores diestros y zurdos, siendo los zurdos un poco mejores que los diestros. En Dribbling, se observan diferencias mayores, también a favor de los zurdos. Debemos aplicar un contraste de hipótesis para poder analizar si estas diferencias observadas son significativas.

3.3. Hipótesis nula y alternativa

Escribid la/s hipótesis nula/s y la/s hipótesis alternativa/s.

Respuesta: En las tres variables, las hipótesis nula y alternativa son:

$H0 : \mu_{Left} = \mu_{Right}$

$H1 : \mu_{Left} > \mu_{Right}$

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), b) si podéis asumir normalidad y por qué, c) si el test es paramétrico o no paramétrico, d) si el test es bilateral o unilateral, e) si se puede asumir homocedasticidad o heterocedasticidad.

Respuesta: a) Se trata de contrastes de dos muestras independientes sobre la media. b) Asumimos normalidad por el teorema del límite central. c) Como podemos asumir normalidad, podemos aplicar un test paramétrico sobre la diferencia de medias de dos muestras independientes. d) El test es unilateral, ya que nos preguntamos si los zurdos son mejores que los diestros. Es unilateral por la derecha. e) Comprobamos si podemos asumir homocedasticidad.

```
var.test( Left$Rating, Right$Rating )
```

```
##
##  F test to compare two variances
##
## data:  Left$Rating and Right$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( Left$Dribbling, Right$Dribbling )
```

```
##
##  F test to compare two variances
##
## data:  Left$Dribbling and Right$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
```

```
var.test( Left$Ball_Control, Right$Ball_Control )
```

```
##
##  F test to compare two variances
##
## data:  Left$Ball_Control and Right$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
```

Interpretación del test de homoscedasticidad: En las tres variables, el resultado nos da un valor $p<0.001$. Por tanto, debemos rechazar la hipótesis nula de igualdad de varianzas. Debemos considerar que las varianzas son distintas.

3.5. Cálculos

Calcular el estadístico de contraste, el valor crítico y el valor p.

Respuesta:

```
my.ttest.onetail <-function( x1, x2, cl=0.95, equalvar=TRUE ){
  mean1<-mean(x1)
  mean2<-mean(x2)
  n1<-length(x1)
  n2<-length(x2)
  sd1 <- sd(x1)
  sd2 <- sd(x2)
  if (equalvar==TRUE){
    s <-sqrt( ((n1-1)*sd1^2 + (n2-1)*sd2^2 )/(n1+n2-2) )
    Sb <- s*sqrt(1/n1 + 1/n2)
    df<-n1+n2-2
  }
  else{  #equalvar==FALSE
    Sb <- sqrt( sd1^2/n1 + sd2^2/n2 )
    denom <- ( (sd1^2/n1)^2/(n1-1) + (sd2^2/n2)^2/(n2-2))
    df <- ((sd1^2/n1 + sd2^2/n2)^2) / denom
  }
  #Estadístico de contraste
  t <- abs( ( mean(x1) - mean(x2) ) /  Sb )
  #Valor crítico
  alfa=1-cl
  tcritical <- qt( alfa, df, lower.tail=FALSE )
  #Valor p
  pvalue <- pt( t, df, lower.tail=FALSE)
  #Valores
  info <- data.frame( mean1, mean2, n1, n2, t, tcritical, pvalue, df )
  return (info)
}



i1 <- my.ttest.onetail( Left$Rating, Right$Rating, equalvar=FALSE )
i2 <- my.ttest.onetail( Left$Ball_Control, Right$Ball_Control, equalvar = FALSE )
i3 <- my.ttest.onetail( Left$Dribbling, Right$Dribbling, equalvar = FALSE )
iall<-rbind(i1,i2,i3)
colnames( iall ) <- c("mean_Left", "mean_Right", "n_Left", "n_Right", "obs_value", "critical", "pvalue")
output <- cbind( var=c("Rating","BallControl", "Dribbling"), iall)
```

3.6. Tabla de resultados

Incorporad una tabla de resultados con este formato. 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).

```
library(kableExtra)

out <- data.frame( var=c("Rating", "BallControl", "Dribbling"), mean_Left=c(0,0,0),
                   mean_Right=c(0,0,0), n_Left=c(0,0,0), n_Right=c(0,0,0), obs_value=c(0,0,0),
                   critical=c(0,0,0), pvalue=c(0,0,0))
out %>% kable() %>% kable_styling()
```

var	mean_Left	mean_Right	n_Left	n_Right	obs_value	critical	pvalue
Rating	0	0	0	0	0	0	0
BallControl	0	0	0	0	0	0	0
Dribbling	0	0	0	0	0	0	0

Respuesta:

var	mean_Left	mean_Right	n_Left	n_Right	obs_value	critical	pvalue	NA
Rating	66.58155	65.85820	4022	12934	5.933765	1.645065	0	7218.284
BallControl	62.16335	58.47727	4022	12934	15.181923	1.645030	0	8623.723
Dribbling	60.15266	55.09688	4022	12934	18.137562	1.645036	0	8359.341

3.7. Interpretación

A partir de los resultados obtenidos, realizad la interpretación de los mismos, dando respuesta a las preguntas formuladas.

Interpretación: En todos los casos el valor p<0.001, lo cual significa que podemos rechazar la hipótesis nula a favor de la hipótesis alternativa. Podemos concluir con un 95 % de nivel de confianza que los jugadores de campo zurdos son significativamente mejores que los jugadores de campo diestros, en relación a las medidas de Rating, Ball_Control y Dribbling.

---

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. Seguid el esquema que se especifica a continuación.

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 devuelve el índice (posición) del jugador de la muestra **sample** que se parece más a **x**.

Solución:

```
set.seed(10)
#This funcion gets: a player with set of features, and a sample with the same features
#Given player x, finds in sample, the most similar player. Returns the index of the
#most similar player in the sample.
my.knn <- function( x, sample ){
#x<-player
#sample<-Right.data
   distances <- rep(0, nrow(sample))
   for ( i in 1:nrow(sample)){
       distances[i] <- euclidean( x, sample[i,])
   }
#   print(distances)
   return (which.min(distances))
#   idx <- which( distances==min(distances) )
#   ii <- sample.int( length(idx),1)    #just in case there are several players with the same distance
#   return(  idx[ii] )    #getting one of the minimum values
}
```

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.

Solución:

```
#For each element in sample1, finds the most similar in sample2
my.knn.sample <- function( sample1, sample2 ){
  pairs <- data.frame(sample1=0, sample2=0)
  for (i in 1:nrow(sample1)){
    #i<-1
```

```
    player <- sample1[i,]
    #print(player)
    idx.similar<- my.knn( player, sample2 )
    pairs <- rbind( pairs, data.frame(sample1=i, sample2=idx.similar ))
  }
  pairs <- pairs[-1,]
  return (pairs)
}

#Prueba
Left     <- data[ data$Preffered_Foot=="Left", ][1:5,]
Left.data <- Left[ ,  c("Age", "Weight", "Height")]
Right     <- data[ data$Preffered_Foot=="Right", ][1:100,]
Right.data<- Right[ , c("Age", "Weight", "Height")]
my.knn.sample( Left.data, Right.data )
```

```
##   sample1 sample2
## 2       1      29
## 3       2      17
## 4       3      82
## 5       4      29
## 6       5       2
```

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.

Nota: Puede que el algoritmo del vecino más cercano necesite bastante tiempo en computarse, ya que las muestras contienen muchos datos. Si es así, os recomendamos que apliquéis el cálculo sobre una muestra de 100 jugadores zurdos. Podéis también tomar una muestra de 200 jugadores diestros. Así evitaréis el exceso de tiempo computacional requerido y la actividad se considera igualmente válida.

```
#Muestras originales Lef y Right
Left     <- data[ data$Preffered_Foot=="Left" & data$Club_Position!="GK", ][1:100,]
Left.data <- Left[ ,  c("Age", "Weight", "Height")]
Right     <- data[ data$Preffered_Foot=="Right"& data$Club_Position!="GK", ][1:200,]
Right.data<- Right[ , c("Age", "Weight", "Height")]
pairs <- my.knn.sample( Left.data, Right.data )
head(pairs)
```

```
##   sample1 sample2
## 2       1     100
## 3       2     102
## 4       3      25
## 5       4     165
## 6       5     198
## 7       6      46
```

```
Left.sample <- Left
Right.paired <- Right[ pairs$sample2, ]
head( Left.sample[,c("Name", "Age", "Weight", "Height", "Rating")])
```

```
##                 Name Age Weight Height Rating
## 2       Lionel Messi  29     72    179     93
## 8        Gareth Bale  27     74    183     90
## 14        Mesut Özil  28     76    180     89
## 20 Antoine Griezmann  25     67    176     88
## 28 Giorgio Chiellini  32     84    187     88
## 35   James Rodríguez  25     75    180     87
```

```
head( Right.paired[,c("Name", "Age", "Weight", "Height", "Rating")] )
```

```
##                  Name Age Weight Height Rating
## 142     Juan Cuadrado  28     71    179     83
## 146     Ander Herrera  27     73    182     83
## 38       Arturo Vidal  29     75    180     87
## 231   Nathaniel Clyne  25     67    175     82
## 279             Kaká  34     83    186     82
## 70  Christian Eriksen  24     76    180     85
```

```
head( paste( Left.sample$Rating, Right.paired$Rating ) )
```

```
## [1] "93 83" "90 83" "89 87" "88 82" "88 82" "87 85"
```

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? Escribid la hipótesis nula y alternativa.

Respuesta:

$H0 : \mu_{Left} = \mu_{Right}$

$H1 : \mu_{Left} > \mu_{Right}$

4.4. Método

Explicad el método que aplicaréis y realizad los cálculos oportunos.

Respuesta: Nos encontramos ahora en un caso de muestras apareadas. Por tanto, aplicamos un contraste de dos muestras apareadas, que se reducirá a un contraste de una muestra sobre las diferencias de Rating entre cada par de jugadores. Es un contraste unilateral, puesto que nos preguntamos si los zurdos son mejores que los diestros.

4.5. Cálculos

Realizad los cálculos de: estadístico de contraste, valor crítico y valor p.Usad un nivel de confianza del 95 %.

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

Respuesta:

```
  alfa <- 1-0.95
  dif <- Left.sample$Rating - Right.paired$Rating
  dif
```

```
##   [1]  10   7   2   6   6   2   4   5   2   1   4  -1   1   3  -3   1   2
##  [18]  -1   2  -3   1   2   1  -5  -4  -4   0   1   1   0   0   0  -1   1
##  [35]  -3  -2  -2  -7  -2  -8  -3   0   0  -7  -4  -2  -4  -3  -2  -3  -2
##  [52]  -2  -3  -2  -1  -1  -2  -5  -6  -4  -8  -5  -4  -4  -8  -5 -11  -3
##  [69]  -1  -1  -3  -2  -2  -6  -6  -8  -8  -8  -2  -3  -4  -4  -6  -4  -5
##  [86]  -4  -4  -3  -8  -6  -2  -3  -6  -7  -5  -4  -8  -7  -7  -6
```

```
n<-length(dif)

t <- mean(dif) / (sd(dif)/sqrt(n))
tcritical <- qt( alfa, lower.tail=FALSE, df=n-1)
pvalue <- pt( t, lower.tail=FALSE, df=n-1)
t; tcritical; pvalue
```

```
## [1] -6.311914
```

```
## [1] 1.660391
```

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

```
#Comprobación
t.test( dif, alternative="greater", conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  dif
## t = -6.3119, df = 99, p-value = 1
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -3.006075       Inf
## sample estimates:
## mean of x
##     -2.38
```

4.6. Interpretación

Interpretad el resultado obtenido.

Respuesta: A partir de la comparación entre las dos muestras apareadas, obtenemos un valor p>0.05. Por tanto, no podemos rechazar la hipótesis nula de igualdad de Rating en jugadores zurdos y diestros. Concluimos por tanto que el valor de Rating en los jugadores zurdos es equivalente al valor de Rating de jugadores diestros con las mismas características físicas.

4.7. Reflexión

Comparad el análisis realizado en el apartado 3, con el realizado en este apartado. Explicad también qué sentido puede tener realizar una comparación de este tipo en relación al contraste de la pregunta 3.

Respuesta: Este resultado nos sugiere que cuando se comparan jugadores similares en peso, altura y edad, no aparecen diferencias significativas entre jugadores diestros y zurdos. Por tanto, podríamos decir que las variables de las características físicas podrían estar sesgando el resultado obtenido anteriormente.

---

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

Escribid la hipótesis nula y alternativa.

Respuesta:

$H0 : p_{FCB} = p_{Madrid}$

$H1 : p_{FCB} \neq p_{Madrid}$

siendo p el porcentaje de jugadores con Rating superior a 90.

5.2. Método

Explicad qué método aplicaréis para dar respuesta a la pregunta formulada. Justificad vuestra elección.

Respuesta: Se trata de un contraste de dos muestras independientes sobre la proporción.

5.3. Cálculos

Preparad las muestras y realizad los cálculos oportunos. Al igual que anteriormente, no podéis usar funciones que ya realicen este contraste automáticamente. Sí podéis usar **pnorm**, **qnorm**, **pt**, **qt**.

Respuesta:

```
FCB <- data[ grep( "Barcelona", data$Club ), ]
Madrid <- data[ grep("Real Madrid", data$Club), ]
n1 <- nrow(FCB)
n2 <- nrow(Madrid)
p1 <- sum( FCB$Rating > 90 ) / n1
p2 <- sum( Madrid$Rating > 90 ) / n2

alfa = 1-0.97
p<- (n1*p1 + n2*p2) /(n1+n2)
z<- (p1-p2) / sqrt( p*(1-p)*(1/n1 + 1/n2))
zcrit <- qnorm(alfa/2)
pvalue <- pnorm(abs(z),lower.tail=FALSE)*2
z;abs(zcrit);pvalue
```

```
## [1] 1.031754
```

```
## [1] 2.17009
```

```
## [1] 0.3021874
```

```
#Comprobación
success <- c(p1*n1,p2*n2)
n <- c(n1,n2)
prop.test( success, n, alternative="two.sided", correct=FALSE)
```

```
## Warning in prop.test(success, n, alternative = "two.sided", correct =
## FALSE): Chi-squared approximation may be incorrect
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of n
## X-squared = 1.0645, df = 1, p-value = 0.3022
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.05359157  0.17480369
## sample estimates:
##     prop 1     prop 2
## 0.09090909 0.03030303
```

5.4. Resultados e interpretación

Escribid la interpretación de los resultados y dad respuesta a la pregunta formulada.

Respuesta: El valor p del contraste sobre la diferencia de dos proporciones da como resultado 0.3021874. Este valor no es inferior a 0.03, que es el nivel de significación fijado (para un nivel de confianza del 97 %). Por tanto, no podemos rechazar la hipótesis nula de igualdad de proporciones de jugadores con rating superior a 90 entre los clubes de Madrid y Barcelona.

6. Resumen ejecutivo

Enumerad brevemente las conclusiones obtenidas en este análisis inferencial. Explicad qué resultados se han obtenido y con qué nivel de confianza podéis extraer las conclusiones.

Respuesta:

1. La variable Rating se distribuye de forma normal. El intervalo de confianza de Rating al 95 % es de (66.02, 62.27)
2. Al comparar la muestra de jugadores de campo zurdos con diestros, podemos concluir que el Rating, Ball_Control y Dribbling es mejor en los jugadores zurdos que en los diestros con un nivel de confianza del 95 % ($p<0.001$).
3. Sin embargo, si realizamos un matching entre jugadores zurdos y diestros en función de similitud por edad, peso y altura, no se encuentran diferencias significativas en Rating al 95 %. Esto puede sugerir que las características personales pueden sesgar el resultado. Cuando se comparan jugadores físicamente similares, no parece que ser zurdo conduzca a mejor Rating.
4. No existen diferencias significativas al 97 % entre los clubes de Barcelona y Madrid en relación al porcentaje de jugadores con un Rating superior a 90. Ambos clubes tienen un porcentaje equivalente de buenos jugadores (Rating>90) en relación a la plantilla de jugadores.

Puntuación de la actividad

- Apartado 1 (10 %)
- Apartado 2 (10 %)

- Apartado 3 (20 %)
- Apartado 4 (20 %)
- Apartado 5 (20 %)
- Apartado 6 (10 %)
- Calidad del informe dinámico (10 %)