Avance del Postwork

Equipo 18

Ana Cristina Castillo Escobar

Edgar Balderas Loranca

José Alberto Cortes Ayala

Leandro Marcelo Pantoja Acosta

Marcos Yáñez Espíndola

Introduccion

Este documento recopila los resultados de los Postworks 1 - 4, así como las explicaciones paso por paso de cómo se alcanzaron los objetivos requeridos y el porqué de las decisiones que se tomaron para lograrlo. Así mismo, se incluye una breve interpretación de las salidas obtenidas, todo esto con el propósito de practicar los conceptos y comandos vistos en las sesiones en vivo y así reafirmar el conocimiento.

```
library(tidyverse)
library(knitr)
library(Matrix)
```

Desarrollo

Postwork 1

```
data <- read.csv("https://www.football-data.co.uk/mmz4281/1920/SP1.csv")
```

Importa los datos de soccer de la temporada 2019/2020 de la primera división de la liga española a R, los datos los puedes encontrar en el siguiente enlace: https://www.football-data.co.uk/spainm.php.

```
goles<-data %>% select(FTHG,FTAG)
```

Del data frame que resulta de importar los datos a R, extrae las columnas que contienen los números de goles anotados por los equipos que jugaron en casa (FTHG) y los goles anotados por los equipos que jugaron como visitante (FTAG)

Consulta cómo funciona la función table en R al ejecutar en la consola ?table

Posteriormente elabora tablas de frecuencias relativas para estimar las siguientes probabilidades:

- La probabilidad (marginal) de que el equipo que juega en casa anote x goles ($x = 0, 1, 2, \ldots$)

```
FTHG.P <- prop.table(table(goles$FTHG))
(FTHG.P <- round(FTHG.P,4))
```

```
     0      1      2      3      4      5      6 
0.2316 0.3474 0.2605 0.1000 0.0368 0.0211 0.0026 
```

- La probabilidad (marginal) de que el equipo que juega como visitante anote y goles $(y = 0, 1, 2, \ldots)$

```
FTAG.P <- prop.table(table(goles$FTAG))
(FTAG.P <- round(FTAG.P,4))
```

```
     0      1      2      3      4      5
0.3579 0.3526 0.2132 0.0474 0.0237 0.0053
```

- La probabilidad (conjunta) de que el equipo que juega en casa anote x goles y el equipo que juega como visitante anote y goles $(x = 0, 1, 2, \ldots, y = 0, 1, 2, \ldots)$

```
TPC <- with(goles,table(goles$FTHG,goles$FTAG))
PT.TPC <- prop.table(TPC)
(PT.TPC <- round(PT.TPC,4))
```

```
        0      1      2      3      4      5
  0 0.0868 0.0737 0.0395 0.0211 0.0053 0.0053
  1 0.1132 0.1289 0.0842 0.0132 0.0079 0.0000
  2 0.1026 0.0921 0.0526 0.0079 0.0053 0.0000
  3 0.0368 0.0368 0.0184 0.0053 0.0026 0.0000
  4 0.0105 0.0132 0.0105 0.0000 0.0026 0.0000
  5 0.0053 0.0079 0.0079 0.0000 0.0000 0.0000
  6 0.0026 0.0000 0.0000 0.0000 0.0000 0.0000
```

Postwork 2

Importa los datos de soccer de las temporadas 2017/2018, 2018/2019 y 2019/2020 de la primera división de la liga española a R, los datos los puedes encontrar en el siguiente enlace: https://www.football-data.co.uk/spainm.php

```
library(dplyr)

datos_1718<-"https://www.football-data.co.uk/mmz4281/1718/SP1.csv"
datos_1819<-"https://www.football-data.co.uk/mmz4281/1819/SP1.csv"
datos_1920<-"https://www.football-data.co.uk/mmz4281/1920/SP1.csv"

download.file(url = datos_1718, destfile = "Data/SP1-1718.csv", mode = "wb")
download.file(url = datos_1819, destfile = "Data/SP1-1819.csv", mode = "wb")
download.file(url = datos_1920, destfile = "Data/SP1-1920.csv", mode = "wb")

dir("Data")

lista <- lapply(paste("Data/",dir("Data"), sep = ""), read.csv) #Guardamos los archivos en lista
lista
```

Revisa la estructura de de los data frames al usar las funciones: str, head, View y summary.

- Str

```
str(lista[[1]]) # Tiene 380 filas y 64 columnas
str(lista[[2]]) # Tiene 380 filas y 61 columnas
str(lista[[3]]) # Tiene 380 filas y 105 columnas
```

Observamos que la muestra contiene datos de tipo chr, int y numeric (double).

- Head

```
head(lista[[1]])
head(lista[[2]])
head(lista[[3]])
```

De acuerdo a estos registros, parece que BbAHh sólo guarda valores negativos y que las columnas FTR y HTR son datos de tipo categórico.

Además, notamos que a partir de B365H la mayoría de columnas son double. Si leemos el archivo que explica los datos contenidos en el csv, podemos observar que estas columnas pertenecen a porcentajes de apuestas, y por eso son valores de tipo double.

En el caso de los datos de tipo entero, se trata de cantidad de goles, y los de tipo char son nombres de equipos y fechas.

- View

```
View(lista[[1]])
View(lista[[2]])
View(lista[[3]])
```

- Summary

```
summary(lista[[1]]) # Casi no hay NAs
summary(lista[[2]]) # No hay NAs
summary(lista[[3]]) # Algunas columnas contienen varios NAs
```

Con la función select del paquete dplyr selecciona únicamente las columnas Date, HomeTeam, AwayTeam, FTHG, FTAG y FTR; esto para cada uno de los data frames. (Hint: también puedes usar lapply).

```
lista <- lapply(lista, select, Date, HomeTeam, AwayTeam, FTHG, FTAG, FTR)
```

Asegúrate de que los elementos de las columnas correspondientes de los nuevos data frames sean del mismo tipo (Hint 1: usa as.Date y mutate para arreglar las fechas). Con ayuda de la función rbind forma un único data frame que contenga las seis columnas mencionadas en el punto 3 (Hint 2: la función do.call podría ser utilizada).

```
lista[[1]]<-lista[[1]] %>% mutate(Date=as.Date(lista[[1]]$Date,format="%d/%m/%y"))
lista[[2]]<-lista[[2]] %>% mutate(Date=as.Date(lista[[2]]$Date,format="%d/%m/%Y"))
lista[[3]]<-lista[[3]] %>% mutate(Date=as.Date(lista[[3]]$Date,format="%d/%m/%Y"))


data <- do.call(rbind, lista)
View(data)
dim(data)
kable(head(data))
```

Las fechas en nuestros registros estaban identificadas como tipo char, por lo que hubo que cambiarlas a tipo date, pero tuvimos que realizar dicho cambio a cada elemento de la lista por separado, debido a que las fechas estaban en formatos distintos: unas tenían el año con cuatro cifras y otros con dos cifras.

Postwork 3

Con el último data frame obtenido en el postwork de la sesión 2, elabora tablas de frecuencias relativas para estimar las siguientes probabilidades:

- La probabilidad (marginal) de que el equipo que juega en casa anote x goles $(x = 0, 1, 2,)$
- La probabilidad (marginal) de que el equipo que juega como visitante anote y goles $(y = 0, 1, 2,)$
- La probabilidad (conjunta) de que el equipo que juega en casa anote x goles y el equipo que juega como visitante anote y goles $(x = 0, 1, 2, , y = 0, 1, 2,)$

Decidimos calcular estos tres datos juntos en una sola función, ya que más adelante, en el Postwork 4, volveremos a realizar estas operaciones en cada una de las muestras generadas con el método Bootstrap.

```
calcular_probabilidades<-function(df){
  probabilidad_conjunta <- table(df)/dim(df)[1]
  probabilidad_conjunta

  marginal_FTHG <- apply(probabilidad_conjunta, 1, sum)
  marginal_FTHG

  marginal_FTAG <- apply(probabilidad_conjunta, 2, sum)
  marginal_FTAG
  probabilidad_conjunta<-as.data.frame(probabilidad_conjunta)
  marginal_FTHG<-as.data.frame(marginal_FTHG)
  marginal_FTHG$goles<-rownames(marginal_FTHG)
  rownames(marginal_FTHG) <- c()
  marginal_FTAG<-as.data.frame(marginal_FTAG)
  marginal_FTAG$goles<-rownames(marginal_FTAG)
  rownames(marginal_FTAG) <- c()
  return(list(marginal_FTHG,marginal_FTAG,probabilidad_conjunta))
}

goles<-data %>% select(FTHG,FTAG)
probabilidades<-calcular_probabilidades(goles)
```

- La probabilidad (marginal) de que el equipo que juega en casa anote x goles $(x = 0, 1, 2,)$.

```
marginal_FTHG<-probabilidades[[1]]
kable(marginal_FTHG[,2:1], col.names = c("Goles","Probabilidad"))
```

Goles	Probabilidad
0	0.2324561
1	0.3271930
2	0.2666667
3	0.1122807
4	0.0350877
5	0.0192982
6	0.0052632
7	0.0008772
8	0.0008772

- La probabilidad (marginal) de que el equipo que juega como visitante anote y goles $(y = 0, 1, 2,)$.

```
marginal_FTAG<-probabilidades[[2]]
kable(marginal_FTAG[,2:1], col.names = c("Goles","Probabilidad"))
```

Goles	Probabilidad
0	0.3517544
1	0.3403509
2	0.2122807
3	0.0543860
4	0.0289474
5	0.0096491
6	0.0026316

- La probabilidad (conjunta) de que el equipo que juega en casa anote x goles y el equipo que juega como visitante anote y goles $(x = 0, 1, 2, , y = 0, 1, 2,)$.

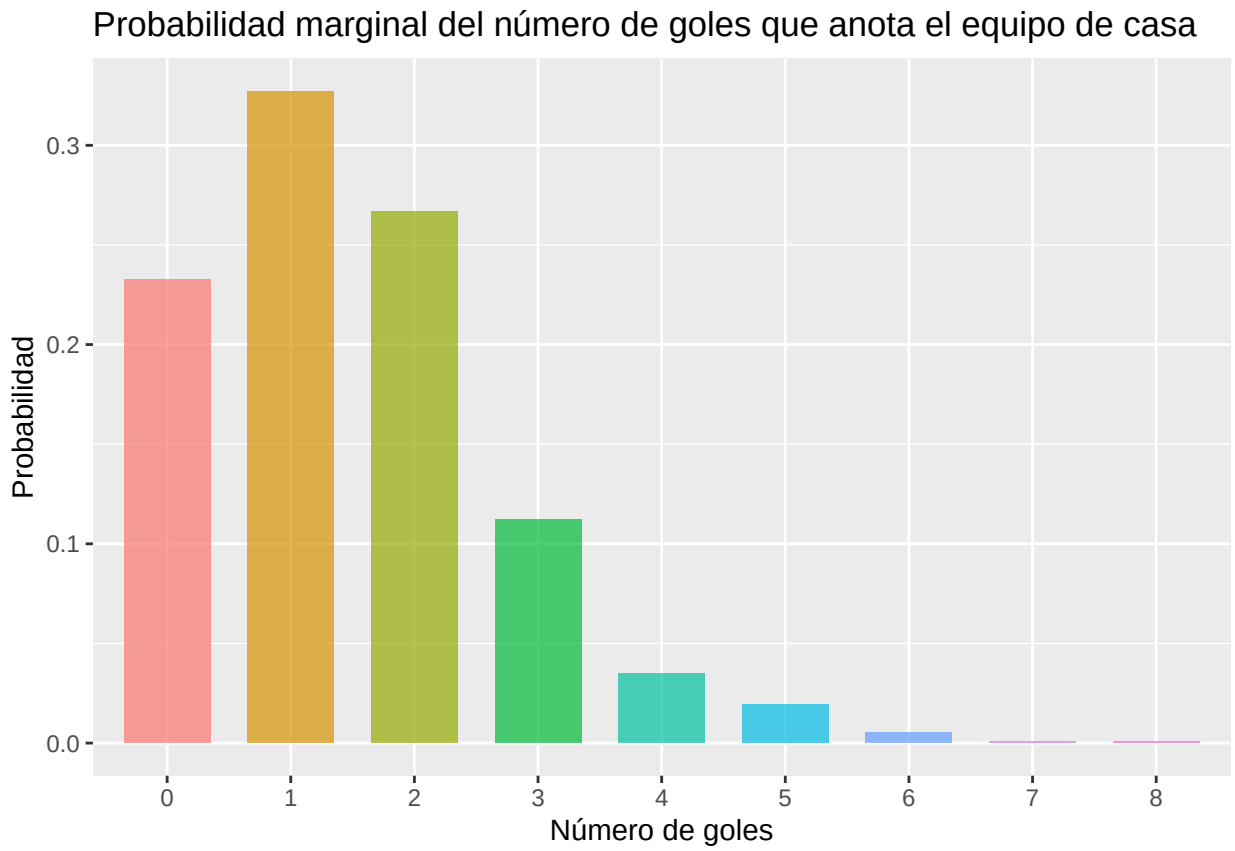
```
probabilidad_conjunta<-probabilidades[[3]]
probabilidad_conjunta_matrix<-as.matrix(sparseMatrix(i = as.integer(probabilidad_conjunta$FTHG),
                                                     j= as.integer(probabilidad_conjunta$FTAG),
                                                     x= probabilidad_conjunta$Freq))
rownames(probabilidad_conjunta_matrix)<-c(0,1,2,3,4,5,6,7,8)
colnames(probabilidad_conjunta_matrix)<-c(0,1,2,3,4,5,6)
round(probabilidad_conjunta_matrix,5) #La pasamos a matriz para tener una mejor visualizacion
```

```
        0       1       2       3       4       5       6
0 0.07807 0.08070 0.04561 0.01842 0.00526 0.00439 0.00000
1 0.11579 0.11491 0.06842 0.01754 0.00877 0.00175 0.00000
2 0.08772 0.09386 0.06140 0.01140 0.00877 0.00175 0.00175
3 0.04474 0.03246 0.02456 0.00614 0.00175 0.00175 0.00088
4 0.01404 0.01053 0.00702 0.00000 0.00351 0.00000 0.00000
5 0.00877 0.00526 0.00439 0.00000 0.00088 0.00000 0.00000
6 0.00263 0.00175 0.00000 0.00088 0.00000 0.00000 0.00000
7 0.00000 0.00088 0.00000 0.00000 0.00000 0.00000 0.00000
8 0.00000 0.00000 0.00088 0.00000 0.00000 0.00000 0.00000
```

Realiza lo siguiente:

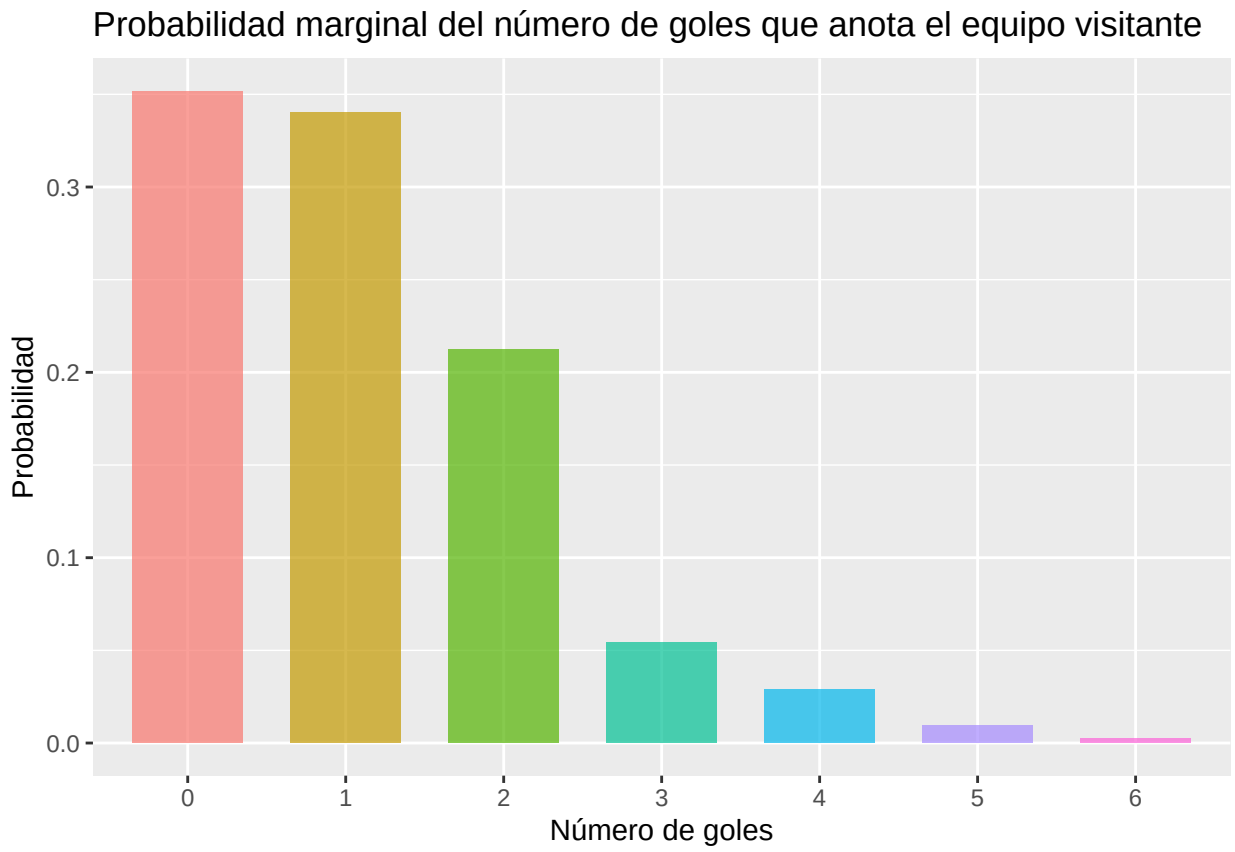
- Un gráfico de barras para las probabilidades marginales estimadas del número de goles que anota el equipo de casa.

```
marginal_FTHG %>%
  ggplot()+
  aes(x=goles,y=marginal_FTHG, fill=goles)+
  geom_bar(stat = "identity", width=0.7, alpha=0.7) +
  theme(legend.position="none")+
  xlab("Número de goles")+
  ylab("Probabilidad")+
  ggtitle("Probabilidad marginal del número de goles que anota el equipo de casa")
```



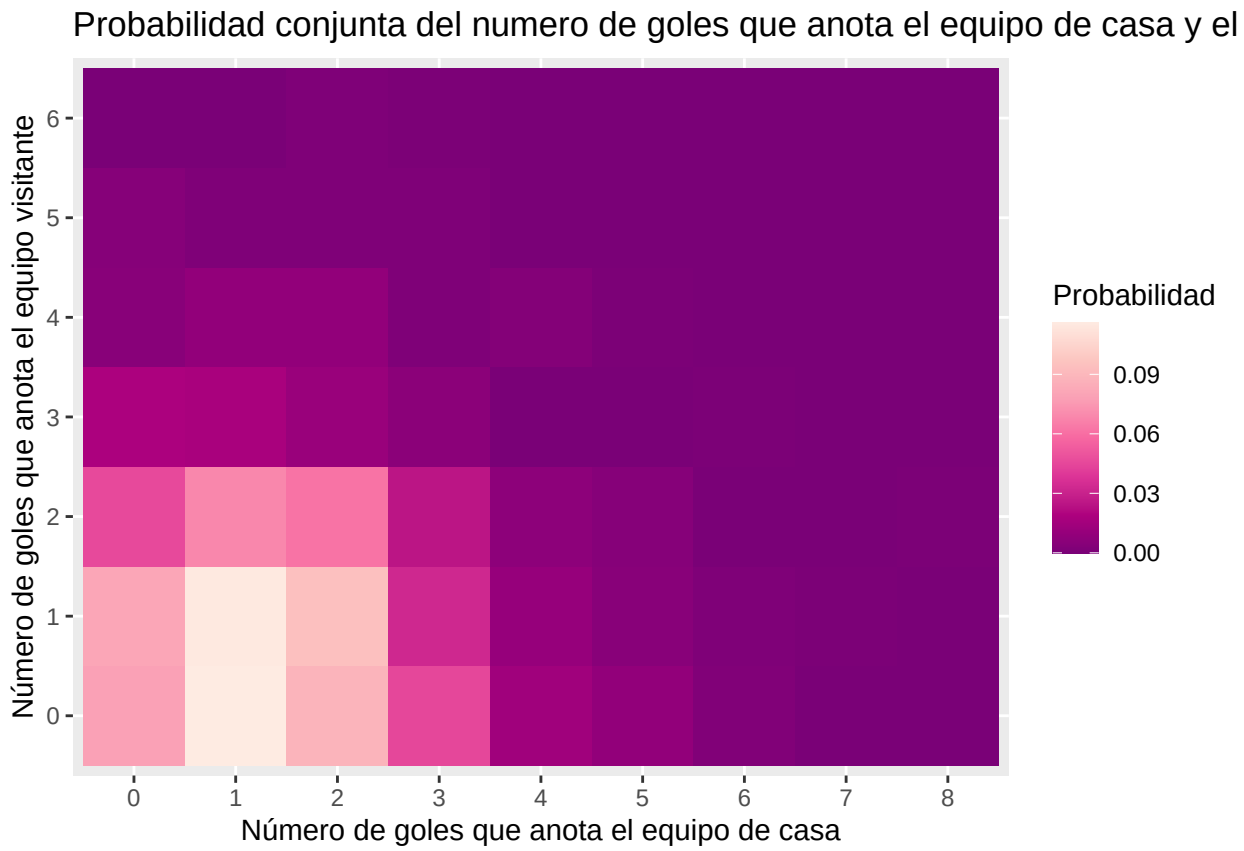
- Un gráfico de barras para las probabilidades marginales estimadas del número de goles que anota el equipo visitante.

```
marginal_FTAG %>%
  ggplot()+
  aes(x=goles,y=marginal_FTAG, fill=goles)+
  geom_bar(stat = "identity", width=0.7, alpha=0.7) +
  theme(legend.position="none")+
  xlab("Número de goles")+
  ylab("Probabilidad")+
  ggtitle("Probabilidad marginal del número de goles que anota el equipo visitante")
```



- Un HeatMap para las probabilidades conjuntas estimadas de los números de goles que anotan el equipo de casa y el equipo visitante en un partido.

```
probabilidad_conjunta %>% ggplot()+
  aes(FTHG , FTAG        , fill= Freq)+
  geom_tile()+
  scale_fill_distiller(palette = "RdPu")+
  xlab("Número de goles que anota el equipo de casa")+
  ylab("Número de goles que anota el equipo visitante")+
  labs(fill = "Probabilidad")+
  ggtitle("Probabilidad conjunta del numero de goles que anota el equipo de casa y el equipo visitante")
```



Postwork 4

Ya hemos estimado las probabilidades conjuntas de que el equipo de casa anote X=x goles ($x = 0, 1, \ldots, 8$), y el equipo visitante anote Y=y goles ($y = 0, 1, \ldots, 6$), en un partido. Obtén una tabla de cocientes al dividir estas probabilidades conjuntas por el producto de las probabilidades marginales correspondientes.

```
producto<-merge(x=marginal_FTHG,y=marginal_FTAG,by=NULL) %>%
  mutate(producto=marginal_FTHG*marginal_FTAG, FTHG=goles.x,FTAG=goles.y) %>%
  select(FTHG,FTAG,producto)

cocientes<-merge(x=producto,y=probabilidad_conjunta, by=c("FTHG","FTAG")) %>%
  mutate(cociente=Freq/producto) %>%
  select(FTHG,FTAG,cociente)

cocientes_matrix<-as.matrix(sparseMatrix(i = as.integer(cocientes$FTHG)+1,
                                         j= as.integer(cocientes$FTAG)+1,
                                         x= cocientes$cociente))
rownames(cocientes_matrix)<-c(0,1,2,3,4,5,6,7,8)
colnames(cocientes_matrix)<-c(0,1,2,3,4,5,6)
round(cocientes_matrix,5)
```

```
        0       1       2       3       4       5       6
0 0.95478 1.02004 0.92437 1.45709 0.78216 1.95540 0.00000
1 1.00606 1.03190 0.98509 0.98590 0.92615 0.55569 0.00000
2 0.93516 1.03415 1.08471 0.78629 1.13636 0.68182 2.50000
3 1.13272 0.84931 1.03048 1.00554 0.53977 1.61932 2.96875
4 1.13716 0.88144 0.94215 0.00000 3.45455 0.00000 0.00000
5 1.29222 0.80131 1.07062 0.00000 1.57025 0.00000 0.00000
6 1.42145 0.97938 0.00000 3.06452 0.00000 0.00000 0.00000
7 0.00000 2.93814 0.00000 0.00000 0.00000 0.00000 0.00000
8 0.00000 0.00000 4.71074 0.00000 0.00000 0.00000 0.00000
```

Recordando la condición de independencia, la cual dice que todos los cocientes deben ser igual a 1, podemos decir a partir de estos datos que estas variables no son independientes, sin embargo podríamos suponer que esto se debe solo al azar debido a que tenemos solo una muestra.

Mediante un procedimiento de boostrap, obtén más cocientes similares a los obtenidos en la tabla del punto anterior. Esto para tener una idea de las distribuciones de la cual vienen los cocientes en la tabla anterior. Menciona en cuáles casos le parece razonable suponer que los cocientes de la tabla en el punto 1, son iguales a 1 (en tal caso tendríamos independencia de las variables aleatorias X y Y).

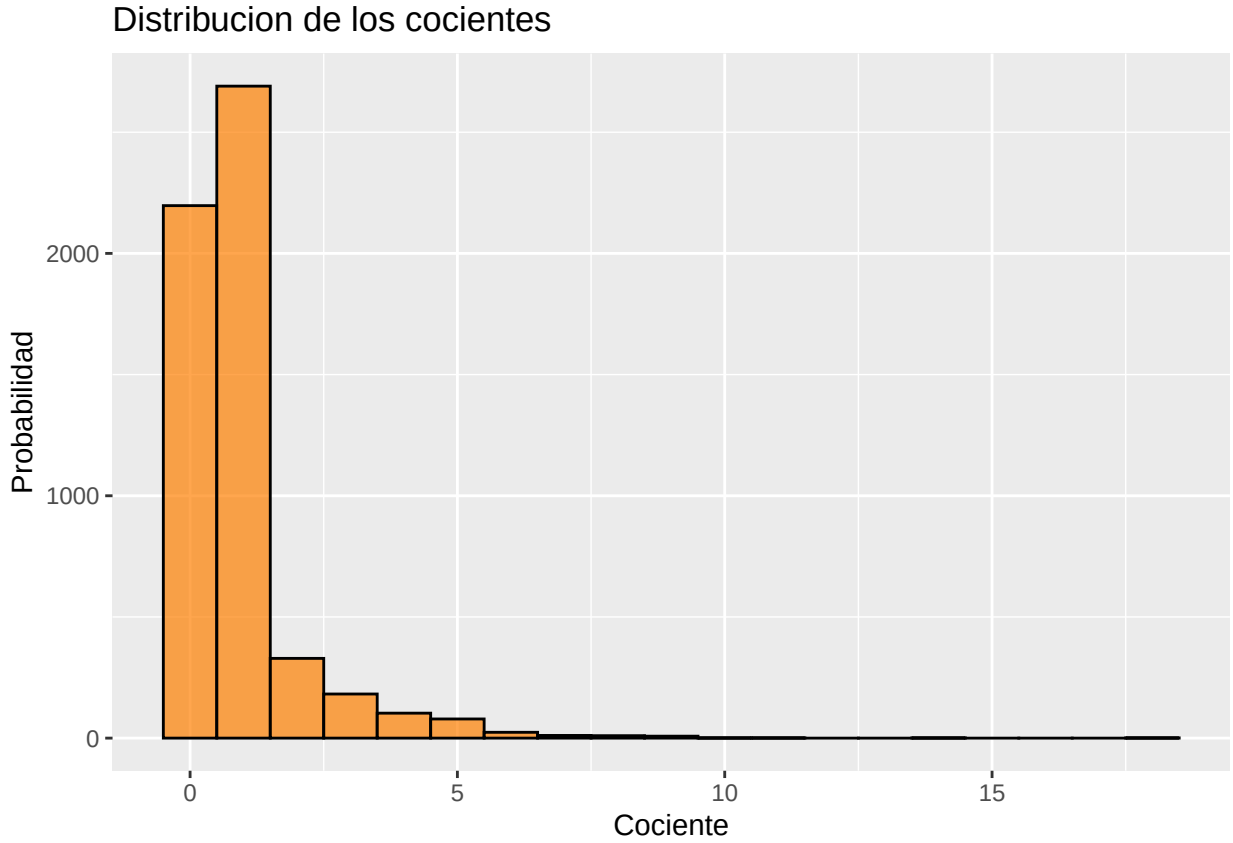
Usando el método de bootstraping, generamos 100 muestras de 1140 las cuales tendrán la misma distribución que nuestra muestra original, una vez hecho esto, sacamos los cocientes para cada muestra y los almacenamos en un Dataframe para así poder construir un histograma y ver la distribución de estos.

```r
library(rsample)

goles_boot <- bootstraps(goles, times = 100)
cocientes<-NULL

for(i in goles_boot$splits){
  probabilidades<-calcular_probabilidades(as.data.frame(i))
  marginal_FTHG<-probabilidades[[1]]
  marginal_FTAG<-probabilidades[[2]]
  probabilidad_conjunta<-probabilidades[[3]]
  producto<-merge(x=marginal_FTHG,y=marginal_FTAG,by=NULL) %>%
    mutate(producto=marginal_FTHG*marginal_FTAG, FTHG=goles.x,FTAG=goles.y) %>%
    select(FTHG,FTAG,producto)
  cocientes_muestra<-merge(x=producto,y=probabilidad_conjunta, by=c("FTHG","FTAG")) %>%
    mutate(cociente=Freq/producto) %>%
    select(cociente)
  cocientes<-rbind(cocientes,cocientes_muestra)
}

cocientes %>% ggplot()+
  aes(x=cociente)+
  geom_histogram(binwidth = 1, fill="darkorange1", col="Black", alpha=0.7)+
  xlab("Cociente")+
  ylab("Probabilidad")+
  ggtitle("Distribucion de los cocientes")
```



Observando esta distribución, podemos apreciar que existe una gran cantidad de cocientes muy lejanos a 1, de hecho, la mayoría son iguales a 0, por lo que de acuerdo con estos datos podemos decir que estas 2 variables no son independientes.

Conclusión

Durante el desarrollo de estos Postworks aprendimos a importar múltiples archivos csv a R, ya fuera que estuvieran localmente almacenados en la computadora o los descargáramos con comandos de R directamente desde su ubicación en internet, para observar algunas de sus características, corregir algunos tipos de datos que no fueran los adecuados y en general, manipular los dataframes.

Con datos ya limpios y organizados, visualizamos probabilidades con la ayuda de barplots y heatmaps, para finalmente, por medio del método Bootstrap, conocer más sobre la distribución de nuestros datos y determinar la dependencia de nuestras variables aleatorias.