Examen 2: Mariana Ávalos y Heráclito

Ejercicio 6

- Resumen de la tabla original

```
#install.packages("faraway")

library(faraway)

data = worldcup[c("Time", "Shots")]
summary(data)
cbind(sd(data$Time), sd(data$Shots))
```

```
> summary(data)
      Time            Shots
 Min.   :  1.0   Min.   : 0.000
 1st Qu.: 88.0   1st Qu.: 0.000
 Median :191.0   Median : 1.000
 Mean   :208.9   Mean   : 2.304
 3rd Qu.:270.0   3rd Qu.: 3.000
 Max.   :570.0   Max.   :27.000

> cbind(sd(data$Time), sd(data$Shots))
         [,1]    [,2]
[1,] 145.4336 3.34743
```

- Cree tres muestras aleatorias de 50 jugadores, con remplazo. Calcule la media, desviación estándar, rango y cuartiles, del tiempo dentro de la cancha y tiros de los jugadores seleccionados (*Time, Shots*). Explique el comportamiento de los resultados.

```
#1
n = NROW(data)
sampleIndex1 = sample(1:n, 50, replace = TRUE)
sample1 = data[sampleIndex1, ]

sampleIndex2 = sample(1:n, 50, replace = TRUE)
sample2 = data[sampleIndex2, ]

sampleIndex3 = sample(1:n, 50, replace = TRUE)
sample3 = data[sampleIndex3, ]

cbind(summary(sample1), summary(sample2), summary(sample3))
cbind(sd(sample1$Time), sd(sample1$Shots), sd(sample2$Time),
      sd(sample2$Shots), sd(sample3$Time), sd(sample3$Shots))
```

```
> cbind(summary(sample1), summary(sample2), summary(sample3))
     Time              Shots             Time              Shots             Time              Shots
 "Min.   :  7.00  " "Min.   : 0.00  " "Min.   :  3.00  " "Min.   :0.00  " "Min.   :  1.00  " "Min.   : 0.00  "
 "1st Qu.: 84.25  " "1st Qu.: 0.00  " "1st Qu.: 81.25  " "1st Qu.:0.00  " "1st Qu.: 71.25  " "1st Qu.: 0.00  "
 "Median :180.00  " "Median : 1.00  " "Median :243.00  " "Median :1.00  " "Median :269.00  " "Median : 1.00  "
 "Mean   :212.44  " "Mean   : 2.48  " "Mean   :207.26  " "Mean   :1.98  " "Mean   :229.48  " "Mean   : 2.46  "
 "3rd Qu.:270.00  " "3rd Qu.: 4.00  " "3rd Qu.:282.75  " "3rd Qu.:3.00  " "3rd Qu.:351.25  " "3rd Qu.: 3.75  "
 "Max.   :564.00  " "Max.   :18.00  " "Max.   :480.00  " "Max.   :8.00  " "Max.   :540.00  " "Max.   :14.00  "

> cbind(sd(sample1$Time), sd(sample1$Shots), sd(sample2$Time), sd(sample2$Shots), sd(sample3$Time), sd(sample3$Shots))
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
[1,] 151.2185 3.834537 145.2668 2.403144 147.1289 3.277692
```

Para el caso de las muestras aleatorias, observamos que, en el caso de los datos de tiempo, los valores de las tres muestras no varían demasiado respecto a los de la tabla original (proporcionalmente hablando en 2 de los 3 casos) puesto que, de acuerdo con los datos de los cuartiles, la distribución parece tender a estar centrada. Por el contrario, es claramente observable que los datos de tiros están sumamente sesgados (sesgo positivo), pues tenemos un rango de 0-27, con un 3er cuartil en 3, por lo tanto se denota una mayor variación en los valores tanto de la media como de la desviación estándar en proporción.

- Cree tres muestras sistemáticas de 100 jugadores con intervalo de 10 entre muestras. Calcule la media, desviación estándar, rango y cuartiles del tiempo dentro de la cancha y tiros de los jugadores seleccionados (time, shots). Explique el comportamiento de los resultados.

```
#2
first1 = sample(1:n, 1)
selection1 = seq(from = first1, by = 10, length = 100)
selection1 = selection1%%n

sample4 = data[selection1, ]

first2 = sample(1:n, 1)
selection2 = seq(from = first2, by = 10, length = 100)
selection2 = selection2%%n

sample5 = data[selection2, ]

first3 = sample(1:n, 1)
selection3 = seq(from = first3, by = 10, length = 100)
selection3 = selection3%%n

sample6 = data[selection3, ]


cbind(summary(sample4), summary(sample5), summary(sample6))
cbind(sd(sample4$Time), sd(sample4$Shots), sd(sample5$Time),
      sd(sample5$Shots), sd(sample6$Time), sd(sample6$Shots))
```

```
> cbind(summary(sample4), summary(sample5), summary(sample6))
     Time              Shots              Time              Shots              Time              Shots        
 "Min.   :  5.0  " "Min.   : 0.00  " "Min.   :  1.0  " "Min.   : 0.0  " "Min.   :  1.00  " "Min.   : 0.0  "
 "1st Qu.: 90.0  " "1st Qu.: 0.00  " "1st Qu.: 90.0  " "1st Qu.: 0.0  " "1st Qu.: 88.75  " "1st Qu.: 0.0  "
 "Median :180.0  " "Median : 1.00  " "Median :188.0  " "Median : 1.0  " "Median :189.50  " "Median : 1.0  "
 "Mean   :207.4  " "Mean   : 2.19  " "Mean   :201.1  " "Mean   : 2.2  " "Mean   :207.87  " "Mean   : 2.2  "
 "3rd Qu.:270.0  " "3rd Qu.: 3.00  " "3rd Qu.:270.0  " "3rd Qu.: 3.0  " "3rd Qu.:270.00  " "3rd Qu.: 3.0  "
 "Max.   :567.0  " "Max.   :11.00  " "Max.   :540.0  " "Max.   :11.0  " "Max.   :570.00  " "Max.   :27.0  "

> cbind(sd(sample4$Time), sd(sample4$Shots), sd(sample5$Time), sd(sample5$Shots), sd(sample6$Time), sd(sample6$Shots))
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
[1,] 141.0641 2.740254 134.4674 2.792486 144.4425 3.524804
```

Con respecto al muestreo sistemático, se observa menor variación entre los valores de media y desviación estándar obtenidos entre las muestras, en ambas columnas y una menor diferencia contrastando con los datos de la tabla original, ya que incluso los valores de los cuartiles, resultan bastante similares a en comparación con el muestreo anterior.

- Cree tres muestras de 50 jugadores, utilizando el muestreo estratificado, y considerando la información de la columna Position como estratos. Calcule la media, desviación estándar, rango y cuartiles del tiempo dentro de la cancha y tiros de los jugadores seleccionados (time, shots). Explique el comportamiento de los resultados y compárelos con los resultados de las muestras aleatorias de los incisos anteriores

```r
#3
sampleSize = 50
tmp = worldcup[c("Time", "Shots", "Position")]
tmp = tmp[order(tmp$Position),]
levels(tmp$Position) #[1] "Defender"   "Forward"    "Goalkeeper" "Midfielder"
num_def = NROW(tmp[tmp$Position=="Defender", ])
p_def = num_def/n
qty_def = as.integer(p_def*sampleSize)

num_gk = NROW(tmp[tmp$Position=="Goalkeeper", ])
p_gk = num_gk/n
qty_gk = as.integer(round(p_gk*sampleSize, digits = 0))

num_fwd = NROW(tmp[tmp$Position=="Forward", ])
p_fwd = num_fwd/n
qty_fwd = as.integer(round(p_fwd*sampleSize, digits = 0))

num_mf = NROW(tmp[tmp$Position=="Midfielder", ])
p_mf = num_mf/n
qty_mf = as.integer(round(p_mf*sampleSize, digits = 0))


defIndex1 = sample(1:num_def, qty_def, replace = TRUE)
gkIndex1 = sample((num_def+1):(num_def+num_gk) , qty_gk, replace = TRUE)
fwdIndex1 = sample((num_def+num_gk+1):(num_def+num_gk+num_fwd) , qty_fwd, replace = TRUE)
mfIndex1 = sample((num_def+num_gk+ num_fwd+1):n , qty_mf, replace = TRUE)

sampleIndex1 = c(defIndex1, gkIndex1, fwdIndex1, mfIndex1)


defIndex2 = sample(1:num_def, qty_def, replace = TRUE)
gkIndex2 = sample((num_def+1):(num_def+num_gk) , qty_gk, replace = TRUE)
fwdIndex2 = sample((num_def+num_gk+1):(num_def+num_gk+num_fwd) , qty_fwd, replace = TRUE)
mfIndex2 = sample((num_def+num_gk+ num_fwd+1):n , qty_mf, replace = TRUE)

sampleIndex2 = c(defIndex2, gkIndex2, fwdIndex2, mfIndex2)

```

```r
defIndex3 = sample(1:num_def, qty_def, replace = TRUE)
gkIndex3 = sample((num_def+1):(num_def+num_gk) , qty_gk, replace = TRUE)
fwdIndex3 = sample((num_def+num_gk+1):(num_def+num_gk+num_fwd) , qty_fwd, replace = TRUE)
mfIndex3 = sample((num_def+num_gk+ num_fwd+1):n , qty_mf, replace = TRUE)

sampleIndex3 = c(defIndex3, gkIndex3, fwdIndex3, mfIndex3)

tmp1 = tmp[c("Time", "Shots")]

sample7 = tmp1[sampleIndex1, ]
sample8 = tmp1[sampleIndex2, ]
sample9 = tmp1[sampleIndex3, ]

cbind(summary(sample7), summary(sample8), summary(sample9))
cbind(sd(sample7$Time), sd(sample7$Shots), sd(sample8$Time),
      sd(sample8$Shots), sd(sample9$Time), sd(sample9$Shots))
```

```
> cbind(summary(sample7), summary(sample8), summary(sample9))
      Time            Shots             Time            Shots             Time            Shots
 "Min.   :  1.0  " "Min.   : 0.000  " "Min.   :  7.0  " "Min.   : 0.000  " "Min.   :  1.0  " "Min.   : 0.000  "
 "1st Qu.: 60.0  " "1st Qu.: 0.000  " "1st Qu.:104.0  " "1st Qu.: 0.000  " "1st Qu.: 76.0  " "1st Qu.: 0.000  "
 "Median :152.0  " "Median : 1.000  " "Median :222.0  " "Median : 2.000  " "Median :225.0  " "Median : 1.000  "
 "Mean   :177.7  " "Mean   : 1.837  " "Mean   :229.3  " "Mean   : 2.898  " "Mean   :219.5  " "Mean   : 2.653  "
 "3rd Qu.:270.0  " "3rd Qu.: 3.000  " "3rd Qu.:292.0  " "3rd Qu.: 4.000  " "3rd Qu.:358.0  " "3rd Qu.: 4.000  "
 "Max.   :450.0  " "Max.   :11.000  " "Max.   :540.0  " "Max.   :22.000  " "Max.   :516.0  " "Max.   :15.000  "

> cbind(sd(sample7$Time), sd(sample7$Shots), sd(sample8$Time), sd(sample8$Shots), sd(sample9$Time), sd(sample9$Shots))
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
[1,] 127.3668 2.519482 143.8275 4.297504 153.9214 3.626011
```

En comparación con los muestreos aleatorios, estas muestras presentaron mayor variabilidad entre sí y para con los datos originales, siendo así la que más difiere en básicamente todos los conceptos obtenidos.

- Con base a los resultados anteriores, recomiende la mejor forma de realizar un muestreo para analizar el tiempo pasado por los jugadores dentro de la cancha y la cantidad de tiros realizados.

De acuerdo con lo observado y previamente presentado, concluimos que el mejor tipo de muestreo para este caso de análisis sería el muestreo aleatorio sistemático, puesto que fue el que presentó una menor variación entre cada una de las muestras, y con los valores de la tabla inicial. Esto probablemente se deba a que con este método se usó un tamaño de muestra más grande, por lo que de igual forma recomendamos un análisis más amplio de los resultados de cada tipo de muestreo, a través de una gráfica de distribución muestral y junto con la variación de los tamaños muestrales.