

NBA 2015

Maribel Crespí valero y Carlos Bosh Pino

7/11/2020

Contents

1	Miembros del grupo:	1
2	Introducción:	2
3	Estadísticos Descriptivos básicos globales: Medias, Medianas, Desviaciones típicas y Varianzas	2
4	Estadísticas descriptivas básicas por agrupaciones:	2
4.1	Comparaciones	2
4.2	Gráficos	2
5	Aplicaciones de estadística de inferencia a los datos	4
5.1	Carga de datos	4
5.2	Anova 1	4
5.3	Test KS	5
5.4	Segunda ANOVA	7
5.5	Comparación de medias dos a dos con ajuste de Bonferroni	7
6	Conclusiones:	8

1 Miembros del grupo:

Los autores de este proyecto somos Carlos Bosh Pino y María Isabel Crespí Valero; impulsores del grupo Ñery team. Se nos ocurrió tratar el tema del baloncesto viendo vídeos en youtube de Drafteados, donde hablan de la NBA y sacaban estadísticas constantemente. Además, como una de las integrantes juega al baloncesto, pues se acabó decidiendo que este era una buena cuestión para estudiar. Hubo consenso desde el principio y así se quedó.

2 Introducción:

En baloncesto abundan estadísticas de todo tipo, así que hemos pensado que era un tema bastante idóneo a tratar en este trabajo. La elección de la NBA es debido a que es la liga más atractiva a la hora de sacar datos, ya que en ésta es donde juegan los mejores jugadores de todo el mundo y por lo tanto es un foco de talento inmenso. Un ejemplo de la inmensa cantidad de estadísticas son actuaciones históricas donde jugadores han llegado a marcar 100 puntos, dar 30 asistencias o coger 42 rebotes. De jugadores que promedian triples-dobles por temporada o incluso aquellos que han conseguido llegar al cuadruple-doble en algún partido.

NOTA: Un triple-doble es conseguir doble dígito en al menos tres de los datos estadísticos cuantificables tradicionales. Estos son los puntos, los rebotes, las asistencias, los robos y los tapones. Un cuadruple-doble es lo mismo pero en cuatro datos estadísticos.

Nosotros nos vamos a encargar de estudiar todas estas estadísticas, ver que cómo se relacionan unas con otras y sacar las conclusiones pertinentes. Investigaremos qué es lo que hace a un equipo el dueño de su conferencia, comparar a los mejores jugadores para ver quién es realmente eficiente para su equipo, ver la evolución de la liga y de los equipos, entre otras muchas cosas.

3 Estadísticos Descriptivos básicos globales: Medias, Medianas, Desviaciones típicas y Varianzas

En esta sección, decidimos calcular las medias, medianas, desviaciones típicas y varianzas globales de todos los datos que teníamos en cada estadística en particular. Así, podemos apreciar información acerca de la liga de forma mucho más general. Se podría comparar con otro tipo de ligas y se confirmaría que la NBA es mucho más vistosa ya que cuenta con mucho mas volumen en cada apartado estadístico.

4 Estadísticas descriptivas básicas por agrupaciones:

Este es el sector donde nos hemos explayado más ya que como hemos dicho anteriormente, hay una cantidad inmensa de datos a tratar que no hemos podido por limitaciones de tiempo y de espacio, pues no queríamos alargar excesivamente el trabajo. Así y todo, hemos realizado gran cantidad de datos, comparaciones y gráficos que vamos a explicar a continuación.

4.1 Comparaciones

Para una primera toma de contacto, decidimos probar haciendo la media, mediana, máximos, mínimos, varianza y desviación típica conjunta de los cuatro mejores equipos de la liga. En concreto, tomamos como referencia las estadísticas Puntos, Asistencias y Rebotes.

También hemos comparado las estadísticas entre dos de los mejores bases de la liga, Stephen Curry y Damian Lillard. Podemos ver como Curry es superior en todas las estadísticas menos en tapones. Estos no es de extrañar puesto que el jugador de los Warriors fue MVP esa misma temporada. Aún así Lillard es considerado de los mejoresjugadores de la liga.

También hemos analizado las medias, desviaciones típicas, máximos y mínimos de todos los equipos de la liga de forma individual en este caso.

4.2 Gráficos

R studios nos ha dado la posibilidad de realizar gran cantidad de gráficos para representar de forma mucho más visual y óptima los datos que hemos ido opteniendo.

Así que hemos intentado aprovechar las facilidades que nos ha permitido R y hemos ido probando gráficos con las estadísticas que nos parecían más interesantes.

Hemos observado el número de jugadores por equipo, donde hemos visto que están bastante nivelados y de hecho es un resultado bastante esperable porque los equipos no buscan plantillas excesivamente largas a causa de un posible exceso en salarios, ni tampoco demasiada reducida ya que a largo terminio las lesiones machacarían el equipo.

También hicimos un gráfico de la frecuencia de asistencias por tiempo jugado en cancha. Evidentemente, a más minutos, más asistencias.

Probamos de hacer un gráfico multidimensional que comparase las asistencia/pérdida con los robos/pérdida de los mejores jugadores de la liga. Así podremos ver cuál de ellos es más eficiente en general, o en alguno de esos aspectos en concreto. Destacan Harden en las asistencias y Curry en los robos.

4.2.1 Gráficos relacionando la altura

Puesto que la altura es una de las características más típicas con las que se relaciona el baloncesto, hicimos bastantes gráficos acerca de dicho dato.

El primero, simplemente nos enseña la frecuencia de rebotes por jugador sin tener en cuenta nada más. Vemos como oscila bastante. Bastante lógico pensar que las posiciones del juego más cerca del aro cogerán menos en detrimento de los que juegan más lejos. En el mismo chunk hicimos también otro gráfico que está vez sí relacionaba la altura con los rebotes. Entonces se confirma lo dicho anteriormente.

Seguidamente hemos relacionado la altura con lanzar tiros libres. Verificamos que los jugadores más altos suelen ser los que más lanzan. Esto es debido a que juegan más cerca del aro y al haber mucho contacto es más fácil que te piten la falta.

Relacionar la altura con los triples anotados también era interesante de examinar. Está claro que es mucho más probable que las posiciones donde se juegan más alejadas del aro, sean las que más volumen de triples lancen. Pero es que además, los jugadores más bajos tienen ventaja ya que al tener el centro de gravedad más bajo, su mecánica es más rápida y efectiva. Así lo refleja el gráfico.

Finalmente hemos desmentido el tópico de que los jugadores más altos son siempre los que más juegan en el baloncesto. Claramente si eres alto es una gran ventaja para este deporte, sin embargo también conlleva sus desventajas. El baloncesto es también un deporte muy explosivo, y los jugadores bajitos destacan mejor en esa faceta que los altos.

4.2.2 Más datos

Como ya hicimos bastante gráficos con los datos más habituales (puntos, asistencias, rebotes, etc), decidimos ver desde qué universidades llegaban los jugadores a la nba. En general está bastante repartido, y por lo visto es bastante común la llegada de jugadores que no han pasado por la universidad, sino que han llegado a través de otras ligas o del instituto(aunque esto último no es tan común).

Consideramos que era también importante mirar la cantidad de puntos que metían cada uno de los equipos de la liga. Además, aunque no salía directamente en el data set, lo relacionamos con sus posiciones al finalizar la liga. Pudimos comprobar que los equipos que mejor atacaban no eran los mejores de sus conferencias, pero sí que estaban en posiciones ventajosas. Por otro lado, los equipos que menos anotaban eran claramente los peores de sus respectivas conferencias. Como conclusión, para que un equipo sea un **contender** deberá de destacar sobre los dos lados de la cancha.

Hicimos un plot que relacionaba puntos, canastas intentadas y canastas anotadas.

También un boxplot que compara la edad con los puntos anotados. Vemos como la edad de máximo rendimiento, por lo general es la que comprende entre los 25 y 30 años, que coincide con la etapa de máximo rendimiento físico de los seres humanos. También hay que añadir, que los jugadores jóvenes tienen

inexperiencia y por lo tanto es normal que no anoten tanto, por lo general. Los jugadores veteranos tienen la desventaja del deterioro del físico.

Por último, hicimos también un gráfico del porcentaje de canastas anotadas y los minutos totales jugados. Vemos como es un gráfico bastante regular aunque con alguna desviación puntual. Esto no se adecua a lo que nos esperabamos exactamente, ya que uno podría pensar que los jugadores que más juegan es porque son más eficientes y por tanto tienen un porcentaje mayor. Vemos sin embargo, que esto no tiene por qué ser así. Una posible explicación es que las estrellas que más juegan en cancha son las que asumen más volumen de tiro y por lo tanto es normal que no tengan tan buenos porcentajes. Los jugadores que juegan menos, son menos talentosos pero irán con mucho más cuidado para seleccionar un buen tiro y así poder ganarse más minutos de su entrenador.

5 Aplicaciones de estadística de inferencia a los datos

EL link con los datos es: NBA 2015

5.1 Carga de datos

```
data=read.csv("players_stats.csv.xls",header=TRUE)

colnames(data)<-c("Nombre","Partidos jugados","Minutos","Puntos",
                 "Canastas anotadas","Canastas intentadas",
                 "Porcentaje canastas","3P anotados",
                 "3P intentados","Porcentaje triples",
                 "Tiros libres anotados","Tiros libres intentados",
                 "Porcentaje tiros libres ","Rebotes ofensivos",
                 "Rebotes defensivos","Rebotes",
                 "Asistencias","Robos","Tapones","Perdidas","Faltas en ataque",
                 "Faltas en defensa","Asistencias/Perdida","Robo/Pérdida",
                 "Edad","Nacimiento","Fecha de nacimiento","Instituto",
                 "Experiencia","Peso","Posición","Equipo","Altura","BMI")
```

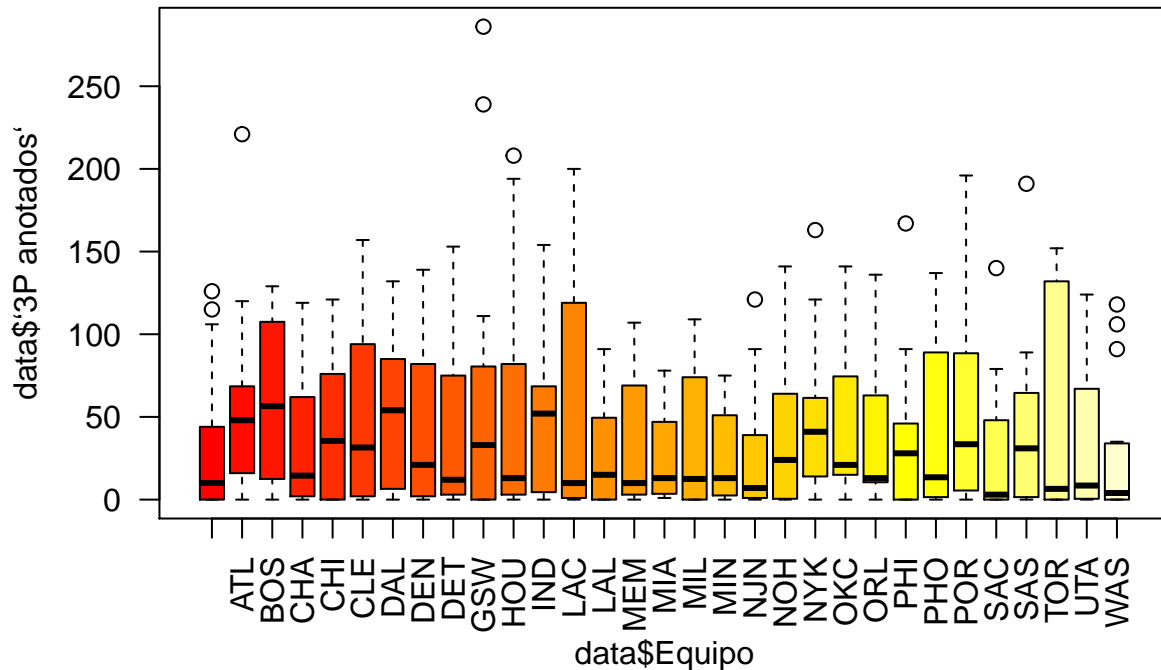
5.2 Anova 1

A continuación vamos a realizar un análisis de la varianza(*ANOVA*) para comparar entre las siguientes hipótesis:

$$\begin{cases} H_0 : \text{Media Equipo 1} = \text{Media Equipo 2} \\ H_1 : \text{Media Equipo 1} \neq \text{Media Equipo 2} \end{cases}$$

Primero, se puede ver un boxplot, donde se pueden apreciar qué variables hemos escogido para realizar el *ANOVA*.

```
boxplot(data$`3P anotados`~data$Equipo,col=heat.colors(32),las=2)
```



Este gráfico muestra los triples anotados por los equipos estudiados. Como vemos, el equipo con más cantidad de triples anotados son Los Ángeles Clippers. El equipo que más triples ha anotado en un partido son los Golden State Warriors, ya que en el gráfico aparecen dos valores atípicos asignados a ellos; valores bastante superiores a los de sus rivales. De hecho esto se puede justificar ya que Golden State cuentan en su plantilla con los “*Splash Brothers*” (Stephen Curry y Klay Thompson), considerados de los mejores triplistas de la historia.

A continuación realizamos el anova comparando los triples anotados por equipo.

```
data=na.omit(data)
sol_anova=aov(data$`3P anotados`~data$Equipo)
summary(sol_anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## data$Equipo  29  42521    1466   0.582  0.961
## Residuals   392 987571     2519
```

Comprobamos que el p-valor es muy grande. Concluimos, por tanto, que no tenemos evidencias suficientes para rechazar la hipótesis nula, y por lo tanto la media de los equipos sí que es distinta.

5.3 Test KS

A continuación vamos a realizar el test de Kolmogorov-Smirnov que utilizaremos para verificar si nuestro conjunto de datos se adapta a una distribución normal o no. Por consiguiente nuestro contraste será:

$$\begin{cases} H_0 : \text{Los datos proceden de una distribución normal} \\ H_1 : \text{Los datos no proceden de una distribución normal} \end{cases}$$

Para la realización de este apartado hemos cogido 4 equipos y hemos realizado 4 test ks, uno para cada muestra. La elección del número de equipos es simplemente por comodidad y porque hemos pensado que con esta cantidad ya sería suficiente para llegar a conclusiones más generales. Los equipos seleccionados son los que se encuentran en la parte alta de la clasificación.

```
Mejores_equipos=c("GSW","CLE","HOU","ATL")
data4=data[as.character(data$Equipo)%in% Mejores_equipos,c("Equipo","Puntos",
"Asistencias","Rebotes")]

data4=na.omit(data4)
X=data4$Puntos
F=as.factor(as.character(data4$Equipo))
levels(F)
```

```
## [1] "ATL" "CLE" "GSW" "HOU"
```

```
muestra1=X[F=="GSW"]
ks.test(muestra1,"pnorm",mean(muestra1),sd(muestra1))
```

```
##
## One-sample Kolmogorov-Smirnov test
##
## data: muestra1
## D = 0.17068, p-value = 0.7134
## alternative hypothesis: two-sided
```

```
muestra2=X[F=="CLE"]
ks.test(muestra2,"pnorm",mean(muestra2),sd(muestra2))
```

```
##
## One-sample Kolmogorov-Smirnov test
##
## data: muestra2
## D = 0.30367, p-value = 0.177
## alternative hypothesis: two-sided
```

```
muestra3=X[F=="HOU"]
ks.test(muestra3,"pnorm",mean(muestra3),sd(muestra3))
```

```
##
## One-sample Kolmogorov-Smirnov test
##
## data: muestra3
## D = 0.20377, p-value = 0.6797
## alternative hypothesis: two-sided
```

```
muestra4=X[F=="ATL"]
ks.test(muestra4,"pnorm",mean(muestra4),sd(muestra4))
```

```
##
## One-sample Kolmogorov-Smirnov test
##
## data: muestra4
## D = 0.20777, p-value = 0.4743
## alternative hypothesis: two-sided
```

5.4 Segunda ANOVA

A continuación vamos a realizar un segundo anova entre los puntos anotados por cada equipo.

```
sola_aov=aov(X~F)
summary(sola_aov)
```

```
##           Df    Sum Sq Mean Sq F value Pr(>F)
## F           3     70653   23551   0.076  0.973
## Residuals   49  15206645   310340
```

De nuevo obtenemos un p-valor muy muy grande, y llegamos a la misma conclusión que en el primer anova; la media de puntos entre los cuatro equipos es distinta.

5.5 Comparación de medias dos a dos con ajuste de Bonferroni

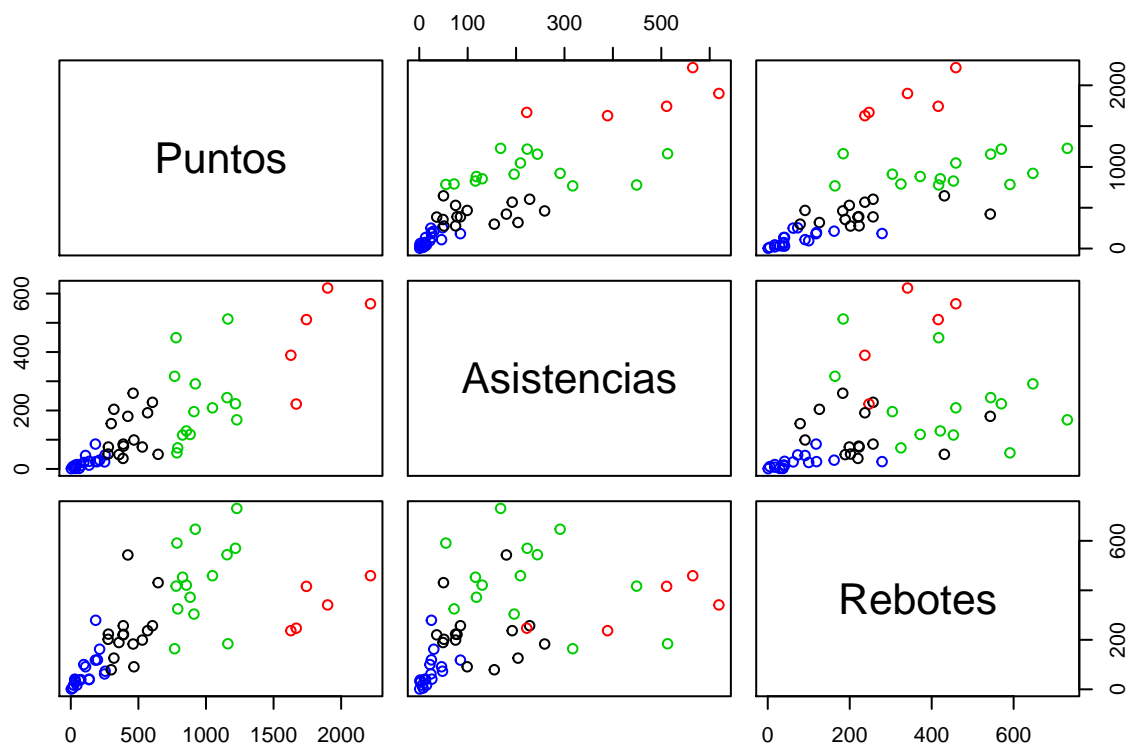
```
Mejores_equipos=c("GSW","CLE","HOU","ATL")
data2=data[as.character(data$Equipo)%in% Mejores_equipos,c("Equipo","Puntos",
                                                           "Asistencias","Rebotes")]
data2=na.omit(data2)
pairwise.t.test(X,F,p.adjust.method = "bonferroni")
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data:  X and F
##
##      ATL CLE GSW
## CLE 1    -    -
## GSW 1    1    -
## HOU 1    1    1
##
## P value adjustment method: bonferroni
```

```
cluster4=kmeans(data4[,2:4],centers = 4)
table(cluster4$cluster,F)
```

```
##      F
##      ATL CLE GSW HOU
## 1     5   2   5   3
## 2     0   2   2   1
## 3     6   2   4   2
## 4     4   6   4   5
```

```
plot(data2[,2:4],col=cluster4$cluster)
```



6 Conclusiones:

En el transcurso de la realización de este trabajo se han llevado a cabo varios estudios estadísticos muy interesantes, aprovechando que en el mundo NBA se pueden hacer gran cantidad de análisis y comparaciones. Aprovechando esto, hemos cotejado a los mejores jugadores de la liga, ver en qué aspecto destacaban cada uno de ellos, comprobando así que tanto los altos como los bajos pueden destacar, aunque de forma totalmente distinta. También hemos explorado los datos de los equipos de la liga, analizando cuáles de ellos podían ser serios candidatos al anillo, y por contra, cuáles eran los que necesitaban un cambio de dinámica. Hemos analizado estadísticas cruciales como la altura y las hemos relacionado con otras como el triple, los tiros libres anotados, los robos, etc. También hemos hecho una parte de estadísticas de inferencia a los datos. Así que con los últimos temas dados en el temario, decidimos realizar varios test ANOVA, un test de Kolmogorov-Smirnov, una comparación de medias dos a dos con ajuste de Bonferroni y un Clustering.

Podemos decir que gracias a este proyecto hemos aprendido mucho sobre las herramientas de R, pues hemos podido averiguar de primera mano cómo se hacen los análisis estadísticos, aplicando las técnicas y conceptos dados en clase. Este trabajo nos ha ayudado a comprender mejor la asignatura en general.

En nuestra opinión, R es una herramienta muy útil con un gran abanico de posibilidades; hemos podido personalizar a nuestro antojo la gran cantidad de funcionalidades que se nos ofrecían ya que cuentan con un número muy extenso de variables con las cuales ir jugando. También hemos podido realizar fácilmente operaciones que a mano hubiesen sido de lo más tediosas. Por último, cabe mencionar que a pesar de las circunstancias, hemos podido interactuar más con nuestro profesor, el cuál ha supuesto una gran ayuda a la hora de realizar el proyecto.