# Análisis estadístico sobre una base de datos de beísbol.

**Eduardo Gavazut**
Universidad Simón Bolívar
Caracas, Venezuela
13-10524@usb.ve

**Luis Riera**
Universidad Simón Bolívar
Caracas, Venezuela
16-10976@usb.ve

**Miguel Cordero**
Universidad Simón Bolívar
Caracas, Venezuela
15-10326@usb.ve

8 de abril de 2022

**Resumen:** En este documento se trata la descripción de los datos, la clase de la base de datos y los estadísticos para las variables

**Palabras clave:**

## 1. Planteamiento del problema

Se desea realizar un análisis estadístico completo sobre una base de datos con información sobre el rendimiento de algunos jugadores de beisbol y en particular estudiar la relación (si la hay) de los hits al bate respecto a las carreras, dobles, triples, jonrones y ponches de los jugadores.

### 1.1. Metodología

Para la realización de esta investigación se hará uso del sofware estadístico `R` en el entorno de desarrollo intergrado (IDE) `RStudio`. En este se iniciará por una descripción de los datos y variables almacenadas en el archivo fuente *Baseball.xlsx*, tales como: mínimo, media, cuantiles y desviación estándar. Para la media de las variables se obtendrá un intervalo de confianza del $95\,\%$. Como se desea estudiar la relación de la tasa de bateo respecto al resto de las variables, se buscará determinar la mejor distribución de probabilidad que se ajuste a esta variable. Finalmente, se estudiará la eficiencia del mejor modelo lineal de predicción que se ajuste a los datos y permita establecer si en efecto existe tal relación entre las variables y las implicaciones que tendría en las estrategias para futuros juegos de beisbol.

## 2. Descripción de los datos

Para la realización de este proyecto se contó con una archivo de excel con la información de algunos jugadores de la Major League Beisbol o MLB, el cual se almacenó en una variable

llamada `Baseball`:

## 2.1. Realizar un análisis descriptivo de los datos

### 2.1.1 ¿Qué clase es la base de datos?

Con el comando `class`, se pudo determinar el tipo de base de datos utilizada o lo que es equivalente, la clase de la variable `Baseball`.

El resultado que se obtuvo indica que es del tipo `tbl_df`, que es una subclase de la clase `data.frame`. `tbl_df` cumple con tener propiedades diferentes por defecto y se suele referir a ellas como `tibble`. Es una clase eficiente para trabajar con bases de datos grandes y su visualización.

### 2.1.2 Variables en la base de datos

Si se desea saber que tipo de variables están almacenadas en la base de datos, se puede utilizar el comando `str`. Esta función nos indica que se cuentan con $6$ variables denominadas `X1,X2,X3,X4,X5,X6`, y distribuidas de tal manera que representan la columnas de la base de datos. Cada una de estas variables tienen $45$ valores de tipo `double` o número decimal, que representan las $45$ observaciones aleatorias (una por fila) realizadas a jugadores de la (MLB) .

Cada variable representa la siguiente información:

- `X1`: tasa de bateo, (hit/veces al bate).
- `X2`: tasa de carreras anotadas, (carreras anotadas/veces al bate).
- `X3`: tasa de dobles, (dobles/ veces al bate).
- `X4`: tasa de triples, (tripes/ veces al bate).
- `X5`: tasa de jonrones, (jonrones/ veces al bate).
- `X6`: tasa de ponches, (ponches/ veces al bate).

### 2.1.3 Estadísticos

Para obtener los estadísticos de las seis (6) variables de esta base de datos, se inicia por guardar las $45$ observaciones en un vector que represente a cada variable.

Con los datos vectorizados se pueden aplicar las siguientes funciones: `mean` que permite obtener la media de los datos, `median` para obtener la mediana, `quantile` retornar los cuantiles al $0{,}25\,\%, 0{,}50\,\%$ y $0{,}75\,\%$ de cada variable, `min` para el valor mínimo, `max` para el valor máximo, `var` para la varianza, `sd` desviación estándar, `IQR` es para el rango intercuartil y finalmente, el coeficiente de variación obtenido como `stad/media`.

Podemos ver la información por cada variable en las siguientes tablas:

```
##    Minimo   25%  Media Mediana / 50  75% Máximo   RIC Varianza Desv. Estándar
## X1  0.188 0.248 0.2805         0.29 0.06  0.308 0.367   0.0019          0.044
## X2  0.064 0.119 0.1509         0.15 0.07  0.189 0.259   0.0018          0.042
##    Coef. Variación
## X1          0.1569
## X2          0.2784
```

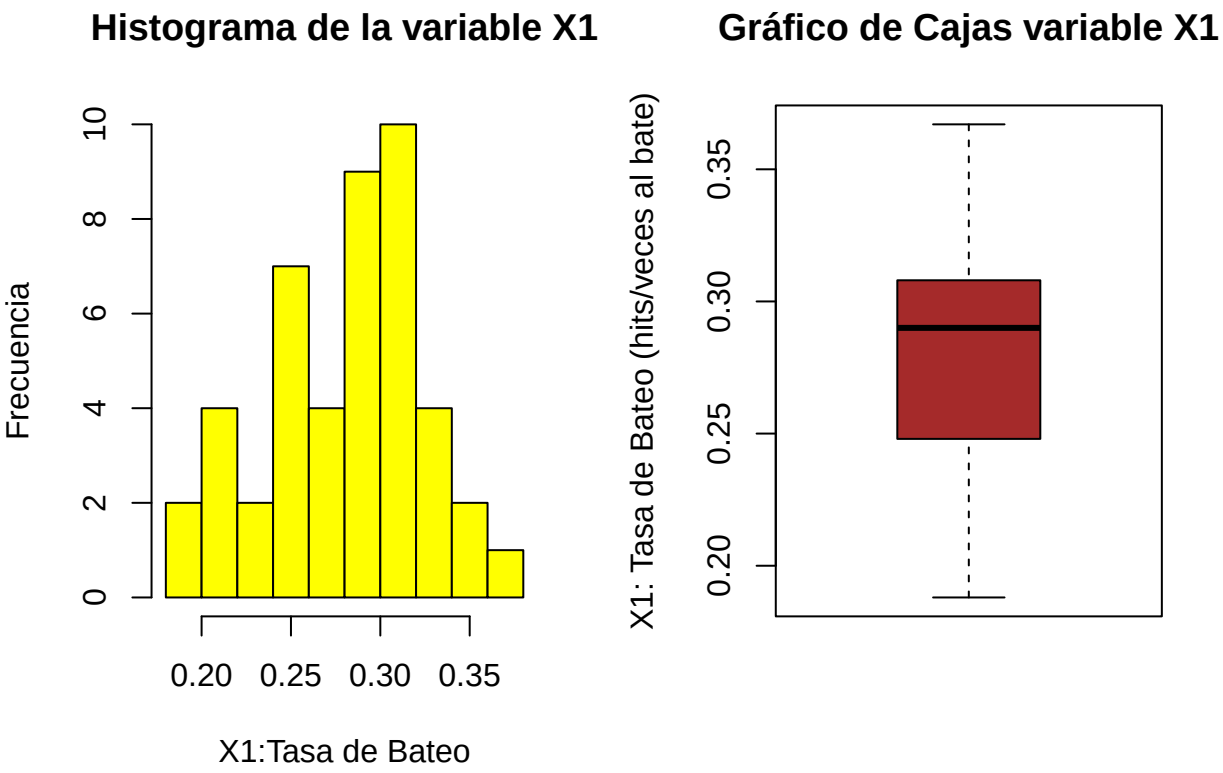


**Figura 1:** Histograma y gráfico de cajas para las variables X1

De estos resultados hay varios puntos que podemos detacar. La varianza de los datos es muy baja indicativo que entre los datos no hay valores atípicos; la media y la mediana son casi idénticos, de hecho en la tasa de carreras anotadas se puede decir que coinciden, esto indica que hay cierta simetría en los datos recolectados; sin embargo, la diferencia entre el máximo y el mínimo para la tasa de bateo es casi el doble en comparación con la tasa de carreras anotadas.

### 2.1.4 Diagramas e histograma de los datos por cada variable

De la gráfica para la variable X1 podemos ver como los valores máximos de los datos se obtienen luego de la media, pero el mayor volumen de ellos se encuentra antes como bien se observa en el diagrama de caja que permite confirmar la ausencia de datos atípicos.

De la gráfica para la variable X2 podemos ver la simetría que se infería de la tabla anterior, con respecto al valor $0{,}15$ que coincide a su vez con la media de los datos. El diagrama de caja permite confirmar la ausencia de los valores atípicos.

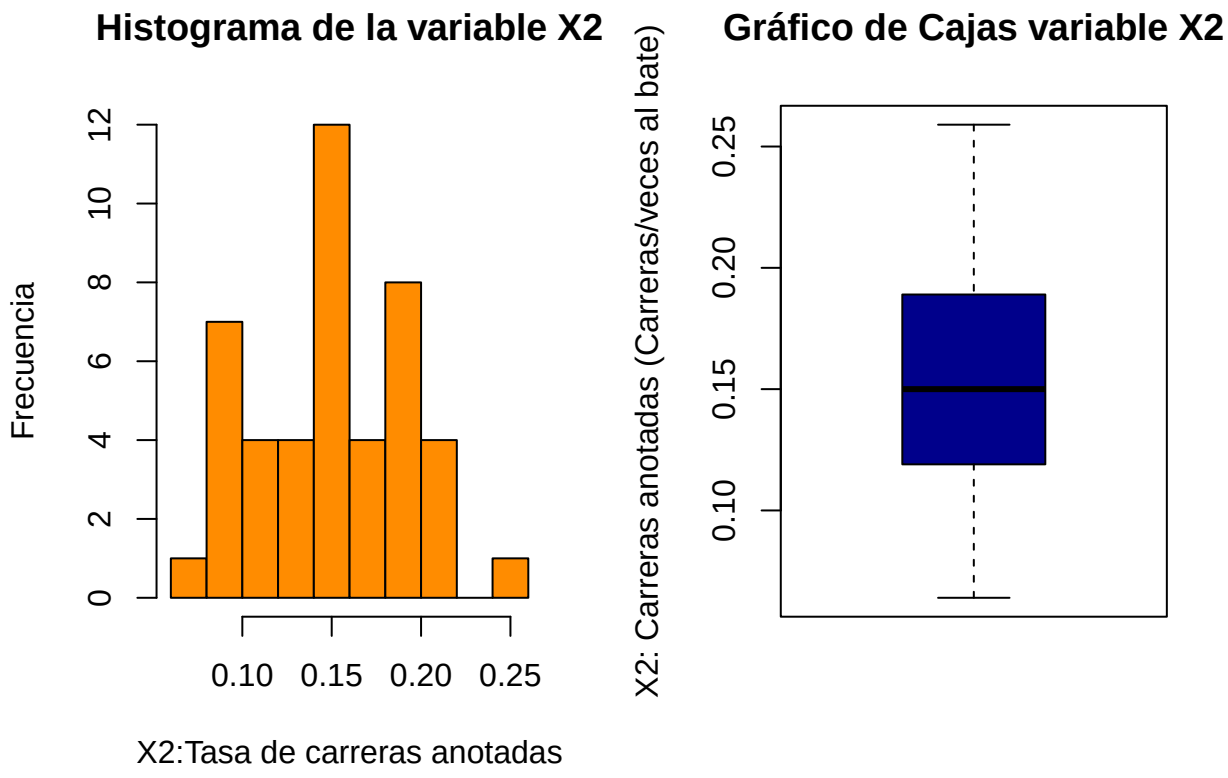


**Figura 2:** Histograma y gráfico de cajas para las variables X2