Análisis estadístico sobre una base de datos de beísbol.

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

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

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

8 de abril de 2022

RESUMEN: Este es un ejemplo de plantilla generado en RStudio para mostrar como se vería el proyecto final. Se realizarán una serie de análisis cuantitativos y cualitativos sobre una base de datos de beísbol. Se exponen los resultados y algunas conclusiones que se pueden extraer de los mismos. Cualquier error puede ser notificado para su corrección final. ESTO NO ES UN RESUMEN, se los estoy diciendo literal XD.

Palabras clave: Proyecto, Estadistica, Rstudio, Beisbol

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. METODODOLOGÍA

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. Análisis 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, medido en hits por veces al bate.
- `X2`: tasa de carreras anotadas, medido en carreras anotadas por veces al bate.
- `X3`: tasa de dobles, medido en dobles por veces al bate.
- `X4`: tasa de triples, o los tripes por veces al bate.
- `X5`: tasa de jonrones, que son los jonrones por veces al bate.
- `X6`: tasa de ponches, medido como ponches por 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` para 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` que es para la desviación estándar, `IQR` es para el rango intercuartil y finalmente, el coeficiente de variación obtenido como `stad/media`.

Los resultados pueden ser apreciados en la tabla 1. De estos resultados hay varios puntos que podemos detacar. La varianza de los datos es muy baja indicativo que entre los datos hay pocos valores atípicos o muy dispersos, lo que se refleja en valores mas cercanos a la media.

Tabla 1: Resumen Estadístico de las variables

	Mínimo	25 %	Media	Mediana (50 %)	75 %	Máximo	RIC	Varianza	Desv. Estándar	Coef. Variación
X1	0.188	0.248	0.2805	0.290	0.308	0.367	0.060	0.0019	0.0440	0.1569
X2	0.064	0.119	0.1509	0.150	0.189	0.259	0.070	0.0018	0.0420	0.2784
X3	0.025	0.039	0.0464	0.045	0.053	0.068	0.014	0.0001	0.0105	0.2255
X4	0.001	0.007	0.0113	0.009	0.016	0.030	0.009	0.0000	0.0070	0.6165
X5	0.000	0.009	0.0243	0.013	0.039	0.085	0.030	0.0005	0.0223	0.9173
X6	0.000	0.062	0.1043	0.095	0.138	0.264	0.076	0.0040	0.0631	0.6044

Misma interpretación se puede extender a la desviación estándar pues es la raíz cuadrada de la varianza.

Una consecuencia de la baja varianza es que la media y la mediana son valores muy cercanos. Esto es particularmente útil al analizar el valor del `RIC`, que toma como medida central la mediana de los datos. Es decir, nos indica donde se encuentra el $50\,\%$ de los datos, cuánto mas bajo es el valor del RIc menos dispersos estan los datos.

Para un análisis mas detallado de como se relacionan las variables entre sí, podemos analizar los histogramas junto a los gráficos de cajas y bigotes.

2.1.4 Diagramas e histograma de los datos por cada variable

De la figura **??**, podemos establecer: para la variable `X1`, que 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, además, la ausencia de datos atípicos. Para la variable `X2`, se puede comprobar que ver simetría de los datos que se infería de la tabla **??**, particularmente 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.

Por su parte, para la variable `X3` y `X4`, Vemos que en general, ambos diagramas de caja son bastante parecidos, con la única diferencia siendo que el de triples está $0{,}03$ puntos corrido hacia arriba y los datos desde el primer cuartil hasta la mediana están muchos más dispersos. Otra diferencia es que el diagrama de cajas para los triples no cuenta con datos atípicos, en cambio los dobles si, que corresponde a $0{,}3$. Todo esto hace que el diagrama de los triples sea casi simétrico, y el de los dobles sea más chato entre el valor mínimo y la mediana, en comparación con lo que tenemos entre la mediana y el máximo valor.

De la gráfica para la variable `X5` podemos ver como a medida que nos vamos acercando a 1, la frecuencia de jonrones decae rapidamente, mientras que al incio es muy alta. De la gráfica para la variable `X6` podemos ver que la mayoria de los jugadores se ponchan menos de un $15\,\%$ de las veces que estan al bate.

2.2. Intervalo de confianza para la media de las variables

Con el uso de la función `t.test()` se puede encontrar el intervalo de confianza con una significancia de $0{,}03$ o $97\,\%$ de confianza para las variables estudiadas. Los resultados de aplicar esta función, se pueden visualizar en la tabla 2.

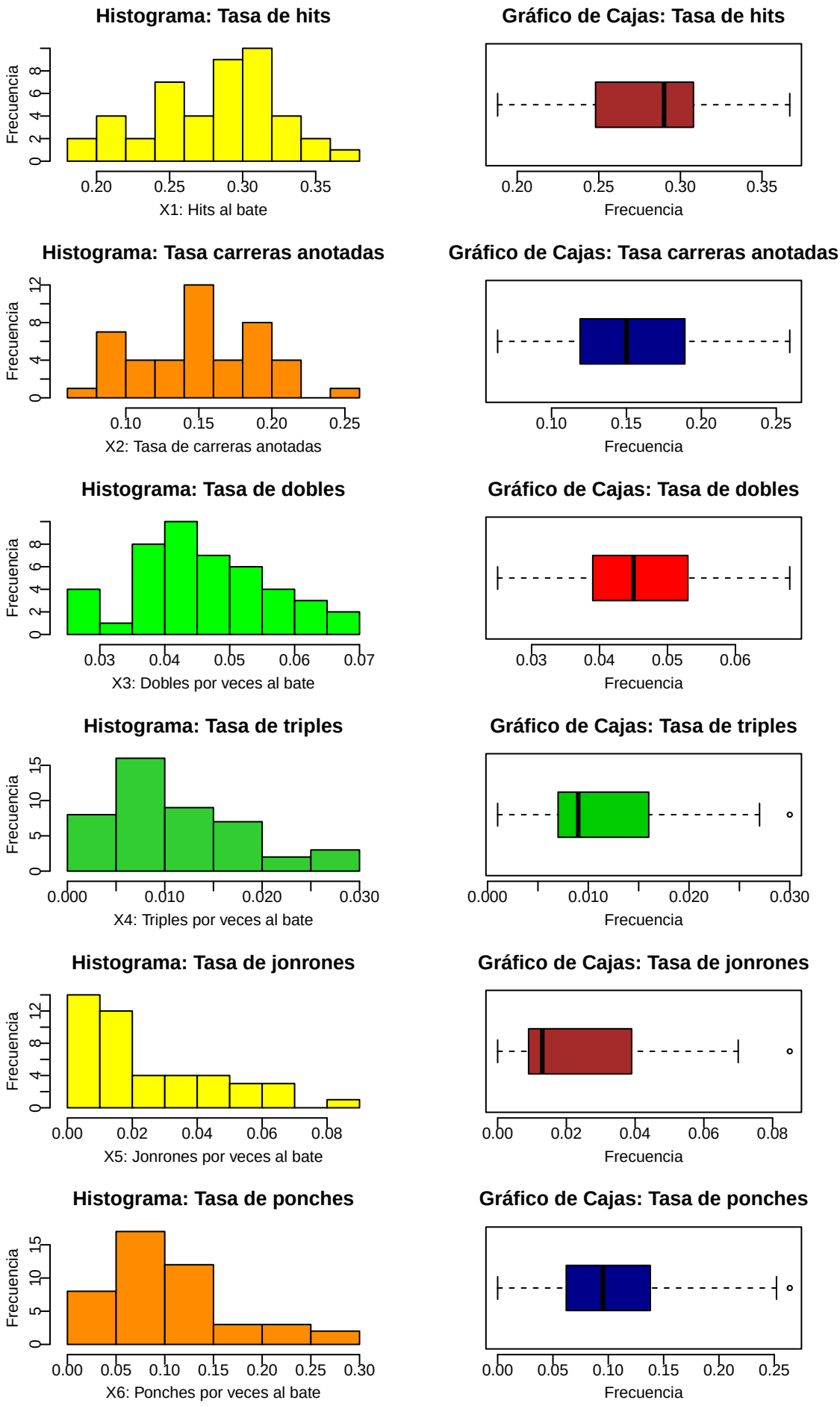

Figura 1: Histograma y gráfico de cajas para las variables

Tabla 2: Intervalos de confianza para las medias de las variables

	Limite inferior	Promedio	Limite Superior
Tasa de bateo	0.2658	0.2805	0.2952
Carreras anotadas	0.1368	0.1509	0.1649
Dobles	0.0429	0.0464	0.0498
Triples	0.0090	0.0113	0.0136
Jonrones	0.0168	0.0243	0.0317
Ponches	0.0833	0.1043	0.1254

Vemos que para cada variable, los intervalos de confianza son

1. Tasa de bateo (hits/veces al bate): $(0{,}2657556, 0{,}2951778)$.
2. Carreras anotadas (por veces al bate): $(0{,}1368441, 0{,}1649337)$.
3. Dobles (por veces al bate): $(0{,}04286130, 0{,}04984981)$.
4. Triples (por veces al bate): $(0{,}008962058, 0{,}013615720)$.
5. Jonrones (por veces al bate): $(0{,}01682441, 0{,}03170892)$.
6. Ponches (por veces al bate): $(0{,}08325165, 0{,}12541501)$.

Ahora, para visualizar un poco mejor estos intervalos, pasemos a graficarlos con ayuda de la librería ggplot2:

Vemos que en general, los intervalos de confianza más estrechos son los de dobles y triples, lo que nos indica que en general, con una probabilidad del 97 %, podemos asegurar que los jugadores de la MLB tendrán un promedio de triples y dobles que puede ser estimado con bastante certeza, pero vemos que las carreras anotadas, los ponches y la tasa de bateo tienen un intervalo de confianza mucho más grande, por lo que no podemos asegurar que el promedio será estimado de forma tan certera.

2.3. Promedio de bateo

Se desea probar con un nivel de significancia de $\alpha = 0{,}05$, que el promedio de bateo es inferior a $0{,}300$.

Como hipótesis nula H_0, supongamos que la media de bateo, $\overline{X1}$, es igual a $0{,}3$. Y como hipótesis alternativa, H_a, que el promedio de bateo es superior a $0{,}3$, $\overline{X1} > 0{,}3$.

Suponiendo que los datos presentan una distribución normal, podemos aplicar el comando `t.test`.

Con este función, se obtuvo que el valor para el estadístico t es $-23{,}811$, con 44 grados libertad. Como el $p - valor$ es bastante alto, de hecho es igual $0, 9976$ (que representa un $99{,}76\,\%$), se cumple que $\alpha = 0{,}05 < 99{,}76$ y por lo tanto la hipótesis alternativa se rechaza, mas aún, se rechaza para todo nivel de significancia porque se necesita un valor para α más alto que el $p - valor$ para rechazar la hipótesis nula.

Se afirma entonces, con total seguridad, que la tasa de bateo es inferior a $0{,}300$.

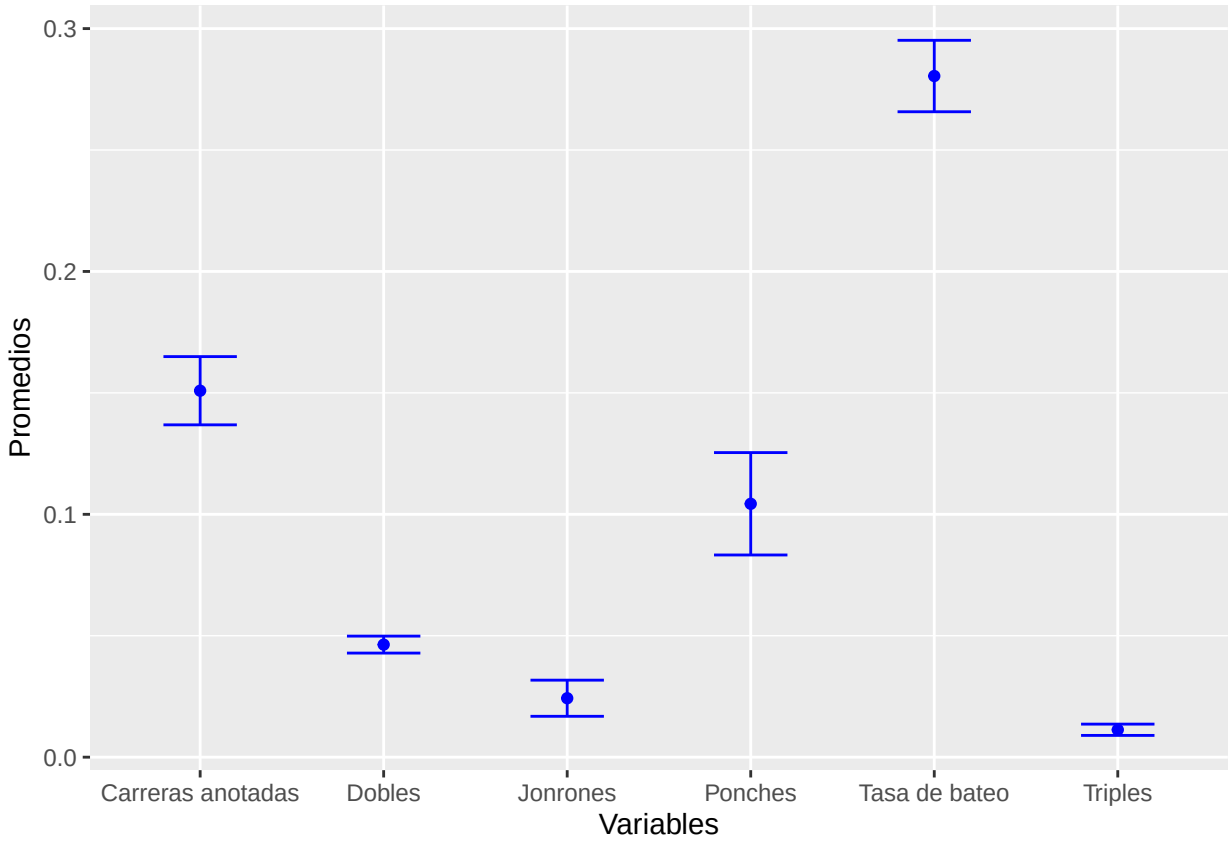

Figura 2: Representación gráfica de los intervalos

2.4. Comparación entre las tasas de ponches y las de jonrones

Queremos extraer la tasa de jonrones y de ponches al bate, estas variables corresponden a X5 y X6, respectivamente, entonces extraigámoslas de la base de datos

Ahora, como no tenemos conocimiento acerca de las varianzas poblacionales, usaremos el test de Welch tal y como es explicado en Heumann, Schomaker (2017) para comparar las medias. En este caso, haremos una prueba de hipótesis, donde tomaremos como hipótesis

$$H_0 : \mu_{\text{jonrones}} - \mu_{\text{ponches}} = 0 \text{ vs. } H_a : \mu_{\text{jonrones}} - \mu_{\text{ponches}} \neq 0$$

es decir, queremos determinar si las tasas de jonrones y ponches son distintas. Ahora, usemos el `t.test()` para determinar cuál de estas hipótesis es aceptada:

```
##
##  Welch Two Sample t-test
##
## data:  jonrones and ponches
## t = -8.032, df = 54.799, p-value = 7.929e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.10004554 -0.06008779
## sample estimates:
##  mean of x  mean of y
```

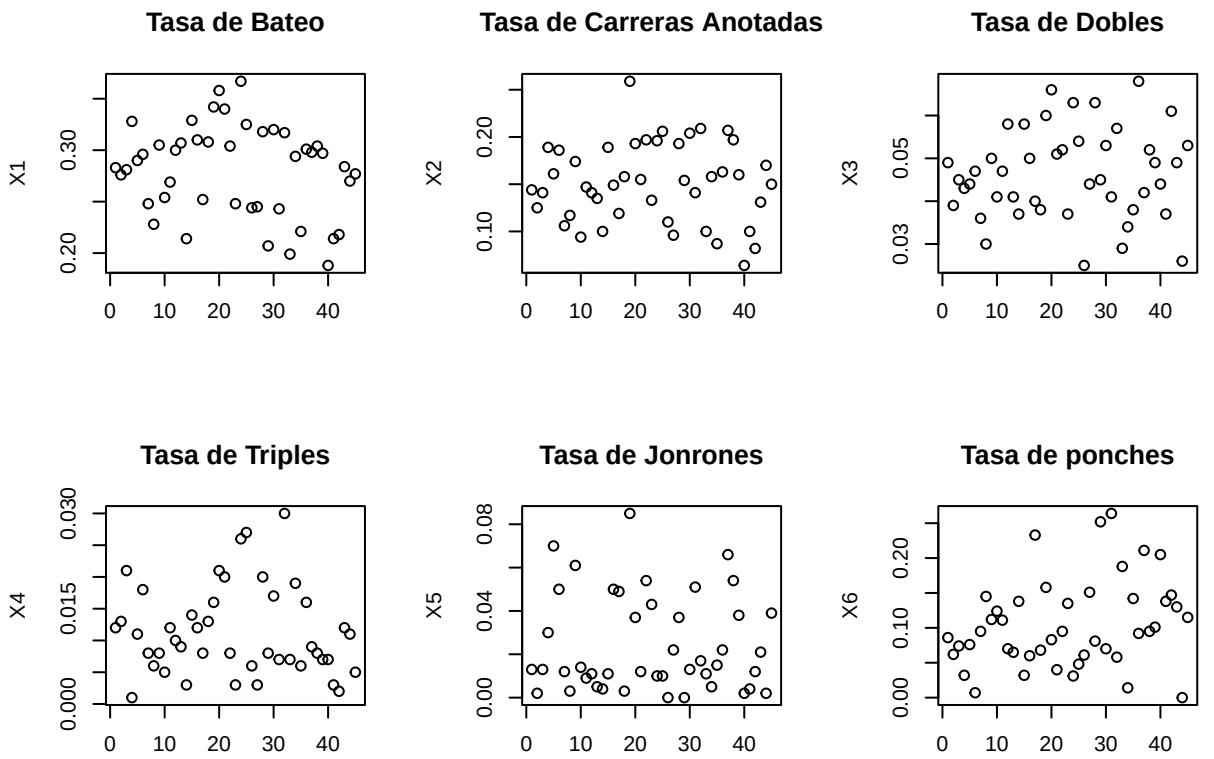

Figura 3: Gráfico de Dispersión

```
## 0.02426667 0.10433333
```

Vemos que el p-valor es extremadamente pequeño, mucho más que el $\alpha = 0{,}01$ que es razonable utilizar para nuestra prueba de hipótesis. Adicionalmente, vemos que en el intervalo de confianza no se incluye el cero. Otra cosa que podemos hacer es evaluar el estadístico de prueba con el comando `qt()` (vemos por lo anterior que $dt = 55$ y $\alpha = 0{,}05$):

```
qt(0.975, 55)
```

```
## [1] 2.004045
```

Como $t = -8$, vemos que el estadístico cae en la región de rechazo (porque es de cola doble).

Para cualquiera de estos casos, podemos concluir que la hipótesis nula se rechaza, es decir que hay suficiente evidencia para creer que $\mu_{\text{jonrones}} - \mu_{\text{ponches}} \neq 0$, además, como el intervalo de confianza es negativo, concluimos que $\mu_{\text{ponches}} > \mu_{\text{jonrones}}$ con un nivel de confianza del $95\,\%$.

2.5. Gráfico de dispersión y matrix de correlación

2.6. Matriz de Correlación:

2.7. ANOVA para la Tasa de Bateo

Para realizar el analiza de varianza sobre la variable `X1` O la tasa de bateo. Primero dividimos los datos en 3 grupos:

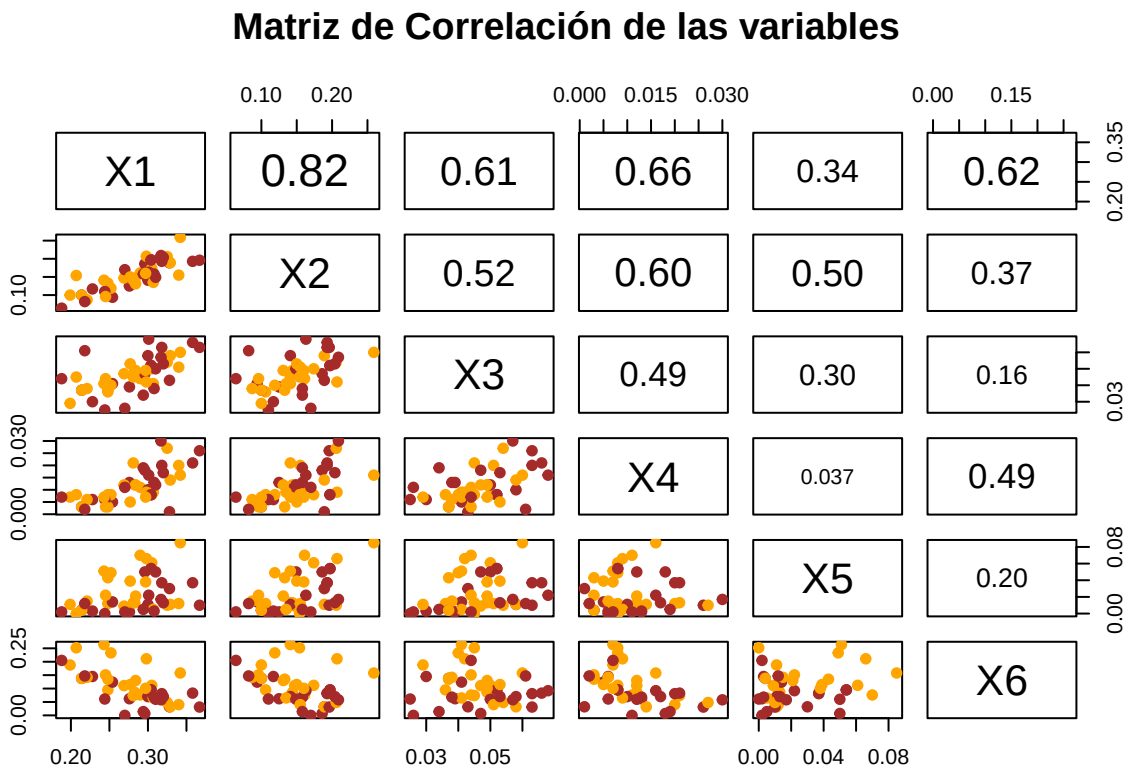

Figura 4: Matrix de correlación

Tabla 3: Tabla a dos factores para las medias de las variables

	X1	X2	X3	X4	X5	X6
Grupo1	0.1935000	0.0820000	0.0365000	0.0070000	0.00650	0.1965000
Grupo2	0.2580400	0.1327600	0.0414400	0.0087200	0.02228	0.1204800
Grupo3	0.3212778	0.1837222	0.0542778	0.0153333	0.02900	0.0716667

- Grupo 1: los bateadores con una tasa de bateo igual a $(X1 < 0{,}200)$.
- Grupo 2: los bateadores con una tasa de bateo igual a $(0{,}200 \leq X1 < 0{,}300)$.
- Grupo 3: los bateadores con una tasa de bateo igual a $(0{,}300 \leq X1)$

Con esta agrupación se decidió por considerar un análisis de varianza con bloques aleatorizados donde los bloques serán los grupos y los tratamientos o métodos las distintas variables de la base de datos.

Con la tabla 3 podemos apreciar las medias de los valores agrupados.Con esto valores, se puede aplicar el comando `anova` de `R` para obtener la tabla ANDEVA detallada en la tabla 4, donde se obtiene que el p-valor para lo grupos es de $0{,}6198$ que es alto, indicando que la hipótesis nula para los grupos no se puede rechazar por lo que las medias por grupos son iguales. Sin embargo, para las medias por variable o método se obtuvo un p-valor de $0{,}0004$ que es bastante bajo, incluso significativo indicando que las medias son distintas tal como se esperaba por los datos analizados.

Con esto podemos afirmar que los promedios de las tasas de las otras variables son iguales por cada grupo.

Tabla 4: Tabla a dos factores para las medias de las variables

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
grupos	2	0.0019826	0.0009913	0.5020692	0.6197546
variables	5	0.1334024	0.0266805	13.5132135	0.0003524
Residuals	10	0.0197440	0.0019744	NA	NA

3. Referencias

Codigos utilizados en este informe:

```
# Función para obtener un resumen estadístico completo de cada variable
estadisticos<- function(variables){
  # Inicializamos las variables
  k<- length(variables)
  # Minimo
  minimo <- rep(0,k)
  # Media
  media <- rep(0,k)
  # Mediana
  mediana<- rep(0,k)
  # Cuartile 1: 25%
  q1 <-rep(0,k)
  # Cuartile 3: 75%
  q3 <- rep(0,k)
  # Maximo
  maximo <- rep(0,k)
  # Rango Intercuartile
  ric <- rep(0,k)
  # Varianza
  varianza <- rep(0,k)
  # Desviación estándar
  stad <-rep(0,k)
  # Coeficiente de variación
  coef_var <-  rep(0,k)

  for(i in 1:k){
    # Minimo
    minimo[i] <- min(variables[,i])
    # Media
    media[i] <- mean(variables[,i])
    # Mediana
    mediana[i]<- median(variables[,i])
```

```
    # Cuartile 1: 25%
    q1[i] <- quantile(variables[,i],0.25)
    # Cuartile 3: 75%
    q3[i] <- quantile(variables[,i],0.75)
    # Maximo
    maximo[i] <- max(variables[,i])
    # Rango Intercuartile
    ric[i] <- IQR(variables[,i])
    # Varianza
    varianza[i] <- var(variables[,i])
    # Desviación estándar
    stad[i] <- sd(variables[,i])
    # Coeficiente de variación
    coef_var <-  stad/media
  }

  # Unimos los valores obtenidos
  estadisticos <- cbind(round(minimo, digits=4),round(q1, digits = 4),
                        round(media, digits=4), round(mediana,digits=4),
                        round(q3, digits=4), round(maximo, digits=4),
                        round(ric, digits=4),round(varianza, digits=4),
                        round(stad, digits=4), round(coef_var, digits=4))
  # Definimos los nombres de las columnas y filas
  rownames(estadisticos) <- c("X1", "X2", "X3", "X4", "X5", "X6")
  colnames(estadisticos) <- c("Minimo", "25%", "Media", "Mediana / 50" ,
                              "75%", "Máximo", "RIC","Varianza",
                              "Desv. Estándar","Coef. Variación")
  # Mostramos el arreglo
  return(estadisticos)

}
variables <-as.data.frame(Baseball)
```