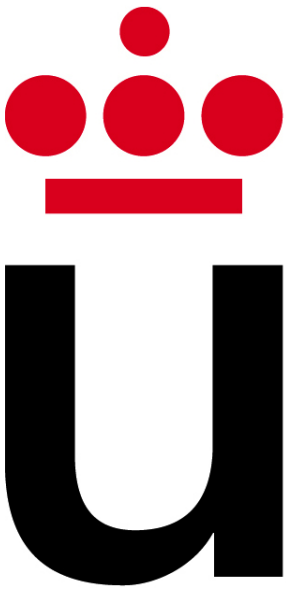
Caso Práctico I

Modelización y tratamiento de la incertidumbre

Estadística descriptiva

José Ignacio Escribano



Móstoles, 10 de octubre de 2015

Índice de figuras

Índice

1. Introducción

Para la realización de este caso práctico requerimos dos conjuntos de datos: un conjunto cuantitativo discreto y otro cuantitativo continuo. Cuando dispongamos de estos conjuntos de datos procederemos a realizar un análisis descriptivo de cada uno de éstos.

Para ello, necesitamos encontrar una base de datos que disponga de los conjuntos de datos requeridos anteriormente. La base de datos elegida se centra en el estudio de la liga norteamericana de béisbol de la temporada 1986. Los datos se encuentran disponibles en el siguiente enlace: `http://lib.stat.cmu.edu/datasets/baseball.data`[1]. En el archivo se encuentran distintos datos relativos a equipos, pitchers y hitters. Nos centraremos en estos últimos. De entre las 24 variables disponibles en el fichero utilizaremos como variable cuantitativa discreta el número de homeruns durante la temporada 1986, y como variable cuantitativa continua el salario (en miles de dólares) al comienzo de la temporada 1987.

2. Análisis descriptivo de un conjunto de datos cuantitativos discretos

El conjunto de datos cuantitativo discreto elegido es **el número de homeruns de los "hitters" de la liga norteamericana de béisbol durante la temporada 1986**.

Para realizar el análisis descriptivo de este conjunto de datos, realizaremos distintos tipos de gráficos como el gráfico de barras y sectores, entre otros. Posteriormente, daremos una serie de medidas, tanto de centralización, posición como dispersión.

Comencemos realizando el diagrama de barras para ver como se distribuye la frecuencia de cada uno de los posibles valores de la variable número de homeruns. En la Figura 1 se muestra este diagrama. Se puede observar que la frecuencia se mantiene más o menos constante entre los 0 y 9 homeruns, mientras que a partir de 10, la frecuencia desciende de manera considerable. Y la realización de 31 homeruns o más es bastante infrecuente.

Si realizamos el diagrama de sectores (Figura 2), veremos que el número de posibles valores es 41: desde 0 hasta 40 homeruns, que es el máximo. Debido a este gran número de valores, el gráfico no se aprecia bien.

[1] El archivo se encuentra en forma de "shell archive", que es una forma de fichero autoextraíble, en el que se agrupan varios archivos, tanto de datos como descripciones de las variables. Para poder extraer estos ficheros, ejecutaremos en una terminal el comando **sh** seguido del nombre del fichero. Esto creará cuatro ficheros nuevos: *pitcher.final*, *team.final*, *data.des.form*, *hitter.final*. De estos ficheros sólo nos interesan los dos últimos: en el primero se encuentran las descripciones de las variables; y en el segundo, los datos que utilizaremos para nuestro análisis descriptivo.

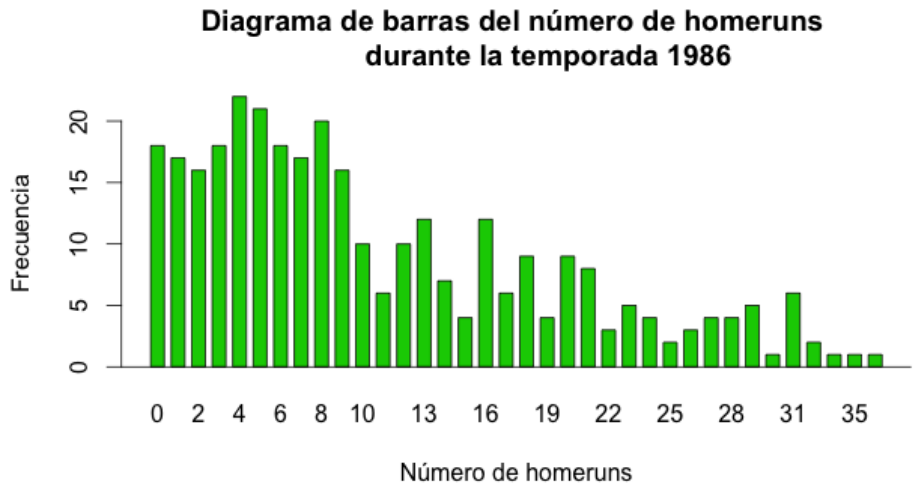


Figura 1: Diagrama de barras para el número de homeruns

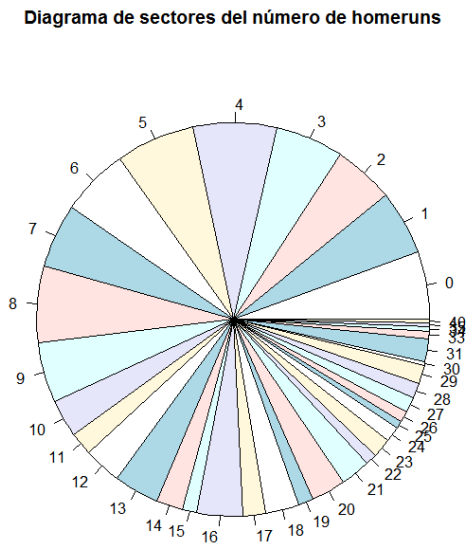


Figura 2: Diagrama de sectores para el número de homeruns

Para asegurarnos de que lo visto en los gráficos anteriores se corresponde con la realidad de los datos, calcularemos una serie de medidas de distintos tipos: de centralización, de posición y de dispersión.

Las medidas de centralización que consideraremos son la media aritmética, la mediana y la moda. Estas medidas son:

Media	10.77019
Mediana	8
Moda	4

La media y la mediana son bastante parecidas, por lo que se puede suponer que son medidas representativas de los datos. Sin embargo, la moda está muy alejada de las otras dos medias, pero este valor es el que más se repite.

Las medidas de posición que consideraremos serán los cuartiles y los deciles. Entre los primeros, optaremos por el primer (Q_1) y el tercer cuartil (Q_3); y de los segundos, optaremos por el segundo (D_2) y el noveno decil (D_9).

Q_1	4
Q_3	16
D_2	3
D_9	24

De los cuartiles Q_1 y Q_3 concluimos que el $25\,\%$ de los jugadores consiguen 4 o menos homeruns y que el $75\,\%$ es menor o igual a 16. De la misma forma, el $20\,\%$ de los jugadores consiguen menos de 3 homeruns; y el $90\,\%$ hace 24 o menos homeruns.

Las medidas de dispersión son la varianza y la cuasivarianza, la desviación estándar, el rango y el rango intercuartílico.

Varianza	75.61178
Cuasivarianza	75.84733
Desviación estándar	8.709037
Rango	40
Rango intercuartílico	12

De los resultados anteriores vemos que la varianza es muy elevada, lo que hace que haya gran variabilidad de los datos. El rango también es muy grande debido a todos los posibles valores que toman los datos: entre 0 y 40.

Con el rango intercuartílico podemos obtener el diagrama de cajas (Figura 3). Podemos observar que el rango intercuartílico se sitúa entre 4 y 16 homeruns y la mediana es de 8 homeruns. Hay que destacar que hay dos datos atípicos, por lo que conseguir 35 ó 40 homeruns es poco común en los jugadores de la temporada 1986.

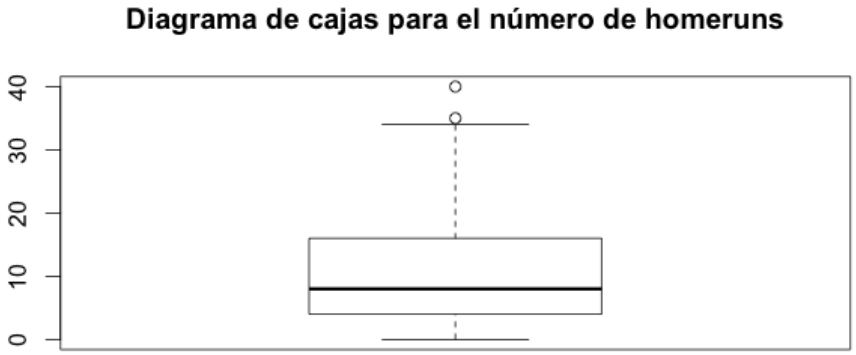


Figura 3: Diagrama de cajas para el número de homeruns

Por último, calculamos las medidas de asimetría y el curtosis. Los datos se muestran a continuación:

Asimetría	0.8963904
Curtosis	0.001943684

Puesto que el coeficiente de asimetría es mayor que cero, tenemos asimetría a la derecha. Con respecto al curtosis, tenemos una distribución de los datos casi normal ya es práctica cero.

3. Análisis descriptivo de un conjunto de datos cuantitativos continuos

El conjunto de datos continuo elegido es **el salario (en miles de dólares) de los "hitters" al comienzo de la temporada 1987**. De igual modo que en el caso discreto, daremos distintos gráficos y distintas medidas para realizar el análisis descriptivo.

Si echamos un vistazo a este conjunto de datos, veremos que hay datos que no están disponibles (se muestran como *NA*), por lo que debemos eliminarlos a la hora de realizar el análisis descriptivo.

Comencemos, de nuevo, representando los datos mediante una serie de gráficos. En este caso, utilizaremos un histograma y un gráfico de tallos y hojas.

El histograma (Figura 4) muestra que el salario más frecuente se encuentra entre 0 y 200 miles de dólares. Según va aumentando el salario, la frecuencia va disminuyendo, hasta el máximo que se encuentra en torno a 2500 miles de dólares, que es muy poco frecuente.

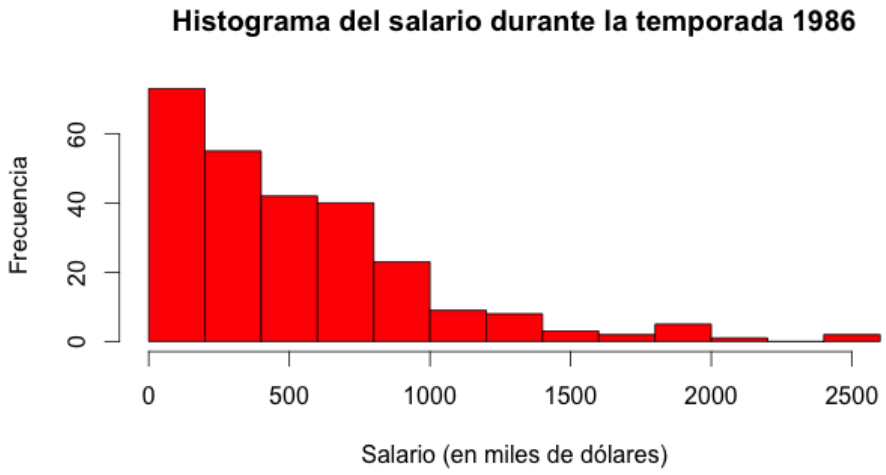


Figura 4: Histograma para el salario

El diagrama de tallos y hojas se puede ver a continuación:

```
0 | 7777777888888999999999
1 | 000000001111112223344444
1 | 555566666777888889999
2 | 00000011222333444
2 | 5555555567888889
3 | 00000012333444
```

 3 | 555677899
 4 | 000022233333
 4 | 55555888899
 5 | 00001233344
 5 | 5556889
 6 | 0000013334
 6 | 56678
 7 | 0000013344444
 7 | 555555557788889
 8 | 0023
 8 | 55558888
 9 | 0002334
 9 | 556
10 | 000144
10 | 5
11 | 0
11 | 588
12 | 024
12 | 6
13 | 0001
13 | 5
14 |
14 | 5
15 | 0
15 |
16 | 0
16 | 7
17 |
17 |
18 | 0
18 | 6
19 | 034
19 | 8
20 |
20 |
21 | 3
21 |
22 |
22 |
23 |
23 |
24 | 1
24 | 6

Como en el caso discreto, calcularemos una serie de medidas, que serán de centralización, de posición y de dispersión.

En el caso de las medidas de centralización utilizaremos la media y la mediana. Los resultados se pueden ver a continuación:

Media	535.9259
Mediana	425

Vemos que tanto la mediana como la media están cerca el uno del otro por lo que podemos considerar que ambas medidas son representativas de este conjunto de datos.

Las medidas de posición que consideraremos serán los cuartiles y los deciles: de los primeros elegiremos el primero (Q_1) y el tercero (Q_3); y de los segundos, el segundo (D_2) y el noveno decil (D_9). Los datos se muestran a continuación:

Q_1	190
Q_3	750
D_2	155
D_9	1048.667

Podemos decir que el $25\,\%$ de los jugadores cobró 190 o menos, y el $75\,\%$ cobró 750 o menos. De igual forma, con los deciles, podemos decir que el $20\,\%$ cobraba menos de 155 y el $90\,\%$ cobró menos de 1048.667.

Las medidas de dispersión que utilizaremos son la varianza y la cuasivarianza, la desviación estándar y el rango y el rango intercuartílico. Los datos se muestran a continuación:

Varianza	202734.3
Cuasivarianza	203508.1
Desviación estándar	451.1187
Rango	2392.5
Rango intercuartílico	560

Vemos que existe una gran varianza, por lo que existe una gran dispersión de los datos, que comprobamos mirando el rango que es muy amplio. Con todos estos datos representamos el diagrama de cajas (Figura 5). En él se muestra que el rango intercuartílico se sitúa entre 190 y 750. También observamos que existen una gran cantidad de datos atípicos, por lo que encontrar jugadores que cobraran entre 1500 y 2500 al comienzo de

Figura 5: Diagrama de cajas para el salario

la temporada 1987 era bastante infrecuente.

Por último, calculamos las medidas de asimetría y el curtosis. Los datos se muestran a continuación:

Asimetría	1.570888
Curtosis	2.933014

Puesto que el coeficiente de asimetría es mayor que cero, tenemos asimetría a la derecha. Con respecto al curtosis, tenemos una distribución de los datos leptocúrtica.

4. Conclusiones

En este caso práctico hemos visto cómo realizar tanto un análisis descriptivo para una variable discreta como para una variable continua. Tanto los gráficos como las medidas nos han dado una serie de interpretaciones acerca de los datos. Por último, hemos aprendido a utilizar de forma básica el lenguaje de programación R, un lenguaje muy potente utilizado en multitud de situaciones.

5. Código R

```
# Seleccionamos el archivo "hitter.final"
filename <- file.choose()

# Cargamos los datos del fichero de texto como CSV
data <- read.csv(filename,sep="\t", header=FALSE,
    col.names=c("name", "bat_1986", "hits_1986", "homeruns_1986",
        "runs_1986", "runs_batted_1986", "walks_1986", "years_major_leagues",
        "times_at_bat_career", "hits_career", "homeruns_career", "runs_career",
        "runs_batted_career", "walks_career", "league_end_1986",
        "division_end_1986", "team_end_1986", "position_1986",
        "put_outs_1986", "assits_1986", "errors_1986", "annual_salary_1987",
        "league_beginning_1987", "team_beginning_1987"))

# Seleccionamos las variables que nos interesan:
# el número de homeruns como variable discreta,
# y el salario como la variable continua
homeruns <- data$homeruns_1986
salary <- data$annual_salary_1987

#-------------------------------------------------------------------------------
# Diagramas de la variable discreta
#-------------------------------------------------------------------------------

# Diagrama de barras
aux <- as.data.frame(table(homeruns))
ybar <- aux$Freq
xbar <- as.character(aux$homeruns)
barplot(ybar, names.arg=xbar, space=0.5, col=3, main="Diagrama de barras
    del número de homeruns durante la temporada 1986",
    xlab="Número de homeruns", ylab="Frecuencia" )
abline(h=0)

# Diagrama de sectores
pie(ybar, labels=xbar, main="Diagrama de sectores del número de homeruns")

#-------------------------------------------------------------------------------
# Diagramas de la variable continua
#-------------------------------------------------------------------------------

# Histograma
hist(salary, breaks=,main="Histograma del salario durante la temporada 1986",
```

```
    xlab="Salario (en miles de dólares)", ylab="Frecuencia", col=2)

# Diagrama de tallos y hojas
stem(salary, scale=4)

#-------------------------------------------------------------------------------
# Medidas
#-------------------------------------------------------------------------------

#-------------------------------------------------------------------------------
# Medidas de centralización
#-------------------------------------------------------------------------------

# Media

# Eliminamos los NA, si lo hubiera
homeruns_mean <- mean(homeruns, na.rm = TRUE)
salary_mean <- mean(salary, na.rm = TRUE)

# Mediana
homeruns_median <- median(homeruns, na.rm = TRUE)
salary_median <- median(salary, na.rm = TRUE)

# Moda

# Definimos una función para la moda
mode <- function(x) {
  temp <- table(as.vector(x))
  return(names(temp)[temp == max(temp)])
}

homeruns_mode <- mode(homeruns);

#-------------------------------------------------------------------------------
# Medidas de posición
#-------------------------------------------------------------------------------

homeruns_quantiles <- quantile(homeruns, c(0.25, 0.75, 0.2, 0.9), na.rm = TRUE)
salary_quantiles <- quantile(salary, c(0.25, 0.75, 0.2, 0.9), na.rm = TRUE)

#-------------------------------------------------------------------------------
# Medidas de dispersión
#-------------------------------------------------------------------------------
```

```
# Cuasivarianza
homeruns_qvar <- var(homeruns, na.rm = TRUE)
salary_qvar <- var(salary, na.rm = TRUE)

# Varianza

# Definimos una función para la varianza
variance <- function(x) {
  n <- sum(!is.na(x))
  return ((n-1)*var(x, na.rm = TRUE)/n)
}

homeruns_var <- variance(homeruns)
salary_var <- variance(salary)

# Desviación estándar

homeruns_std = sqrt(var(homeruns, na.rm = TRUE))
salary_std = sqrt(var(salary, na.rm = TRUE))

# Rango
homeruns_range <- max(homeruns, na.rm = TRUE) - min(homeruns, na.rm = TRUE)
salary_range <- max(salary, na.rm = TRUE) - min(salary, na.rm = TRUE)

# Rango intercuartílico

homeruns_iqr <- IQR(homeruns, na.rm = TRUE)
salary_iqr <- IQR(salary, na.rm = TRUE)

#-------------------------------------------------------------------------
# Diagrama de cajas
#-------------------------------------------------------------------------


boxplot(homeruns, range=1.5, main="Diagrama de cajas para el número de homeruns")
boxplot(salary, range=1.5, main="Diagrama de cajas para el salario")

#-------------------------------------------------------------------------
# Coeficientes de asimetría y curtosis
#-------------------------------------------------------------------------
```

```
# Cargamos la libería fBasics
library(fBasics)

# Asimetría
homeruns_skeness <- skewness(homeruns, na.rm = TRUE)
salary_skeness <- skewness(salary, na.rm = TRUE)

# Curtosis
homeruns_kurtosis <- kurtosis(homeruns, na.rm = TRUE)
salary_kurtosis <- kurtosis(salary, na.rm = TRUE)
```