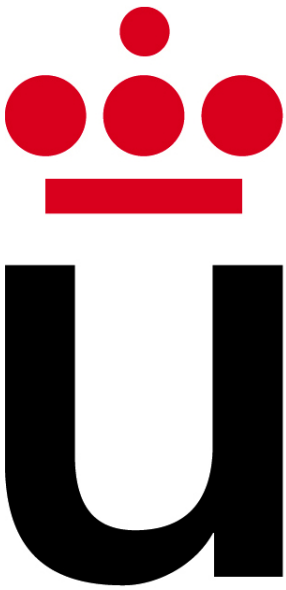
Caso Práctico I

Modelización y tratamiento de la incertidumbre

Estadística descriptiva

José Ignacio Escribano



Móstoles, 6 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 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 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: *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.

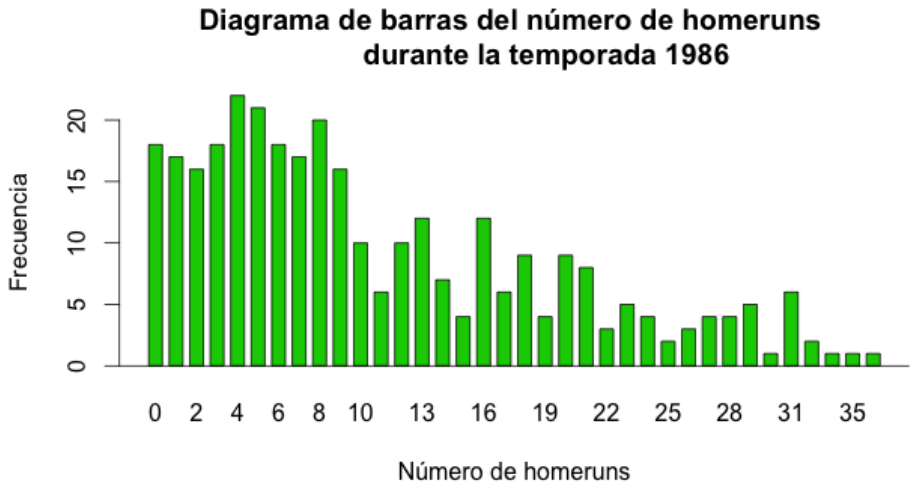


Figura 1: Diagrama de barras para el número de homeruns

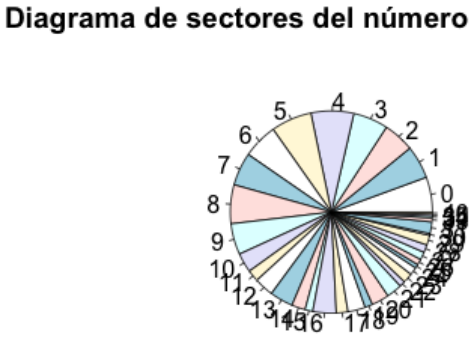


Figura 2: Diagrama de sectores para el número de homeruns

Para asegurarnos de que lo visto en los gráficos anteriores se corresponde con la realidad de los datos, calcualremos 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	10

Como estas medidas son más o menos similares, podemos decir que estas medidas son bastante representativas de estos datos.

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 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 **??**). 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.

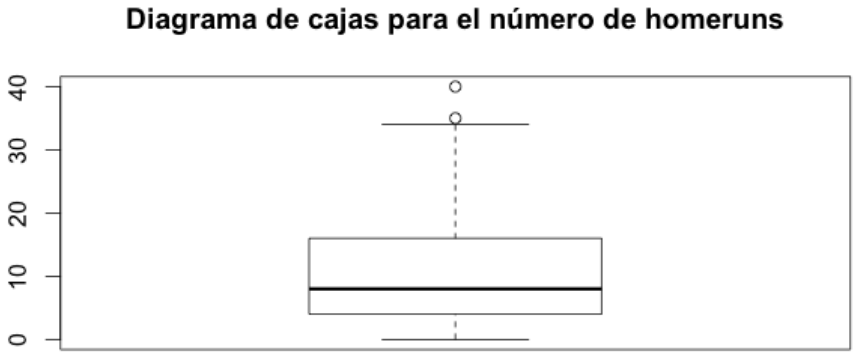


Figura 3: Diagrama de cajas para el número de homeruns

3. Análisis descriptivo de un conjunto de datos cuantitativos continuos