

Sesión 2: Tipos de Variables y Análisis Exploratorio

MBD. Fundamentos matemáticos del análisis de datos.

Curso 2019-20. Última actualización: 2019-08-28

- 1 Trabajando con ficheros de datos.
- 2 Tipos de Variables.
- 3 Variables cuantitativas discretas: enteros.
- 4 Variables cuantitativas continuas,
- 5 Valores centrales, de posición y dispersión.
- 6 Factores.
- 7 Cadenas de caracteres (texto).

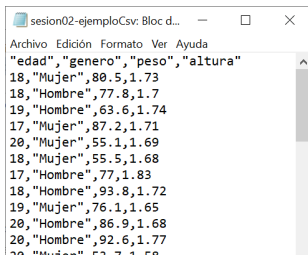
Sección 1

Trabajando con ficheros de datos.

- En la primera sesión hemos usado tablas de datos incorporadas en R (o en librerías). Pero para nuestro trabajo necesitaremos a menudo importar datos procedentes de fuentes externas. Hoy aprenderemos a usar datos almacenados en:
 - ▶ ficheros de texto
 - ▶ ficheros Excel
 - ▶ ficheros de otros programas estadísticos (SAS, SPSS, etc.)
 - ▶ ficheros RData propios de R Vamos a ver como leer estos ficheros para usar los datos en R y también veremos como guardar datos desde R en algunos de esos formatos.
- En otro momento del curso hablaremos de formas alternativas de acceder a datos no almacenados en ficheros (APIs, bases de datos tipo SQL, Web Scrapping, etc.)

Ficheros de tipo csv

- El nombre csv proviene de *comma separated values*, valores separados por comas, aunque vamos a ver enseguida que no hay que tomarse el nombre al pie de la letra.
- Un fichero csv es un fichero de *texto plano* que contiene una tabla de datos. Cada fila del fichero contiene una fila de la tabla y, dentro de esa fila, los elementos correspondientes a cada columna de la tabla se separan mediante comas o espacios o tabuladores, etc. La siguiente figura muestra uno de esos ficheros abierto en el *Bloc de Notas* de Windows y la tabla correspondiente (se muestran las primeras filas).



```
sesion02-ejemploCsv: Bloc d...
Archivo Edición Formato Ver Ayuda
"edad", "genero", "peso", "altura"
18, "Mujer", 80.5, 1.73
18, "Hombre", 77.8, 1.7
19, "Hombre", 63.6, 1.74
17, "Mujer", 87.2, 1.71
20, "Mujer", 55.1, 1.69
18, "Mujer", 55.5, 1.68
17, "Hombre", 77.1, 1.83
18, "Hombre", 93.8, 1.72
19, "Mujer", 76.1, 1.65
20, "Hombre", 86.9, 1.68
20, "Hombre", 92.6, 1.77
20, "Mujer", 53.7, 1.59
```

Fichero csv



edad	genero	peso	altura
18	Mujer	80.5	1.73
18	Hombre	77.8	1.70
19	Hombre	63.6	1.74
17	Mujer	87.2	1.71
20	Mujer	55.1	1.69
18	Mujer	55.5	1.68
17	Hombre	77.0	1.83
18	Hombre	93.8	1.72
19	Mujer	76.1	1.65
20	Hombre	86.9	1.68
20	Hombre	92.6	1.77

la correspondiente tabla

Ficheros csv con R.

- Vamos a empezar descargando uno de estos ficheros, llamado `movies.csv` que contiene datos sobre las [películas más taquilleras entre 2007 y 2011](#).
- Recuerda que debes indicarle a RStudio el *Directorio de Trabajo* y que el fichero descargado debe estar almacenado en la subcarpeta *datos* de ese directorio de trabajo.
- Empieza abriendo ese fichero en un editor de texto (tipo *Bloc de Notas*) para hacer una exploración preliminar.
- Para abrir ese fichero con R vamos a empezar usando:

```
movies = read.csv(file = "../datos/movies.csv", header = TRUE)
```

- El resultado de este comando es un `data.frame` de R. Las opciones de la función son:
 - ▶ *file*: el nombre y directorio del fichero relativo (a la carpeta de trabajo).
 - ▶ *header*: que puede ser `TRUE` o `FALSE`, para indicar si la primera fila del csv contiene los nombres de las variables.

Veremos más adelante otras opciones importantes de esta función y funciones similares.

- **Ejercicio:** Usa `head` y `str` para explorar la tabla. ¿Cuáles son sus dimensiones? ¿Cuál es la película más taquillera? ¿Cuál es el género de esa película?

Usando readr para leer y escribir ficheros csv.

- La librería `readr`, que forma parte del `tidyverse`, incluye la función `read_csv`, que a menudo es muy fácil de usar y muy rápida para ficheros grandes. Explora esta tabla como hemos hecho con la primera versión.

```
library(tidyverse)
movies2 = read_csv("../datos/movies.csv")
```

- También puedes usar `readr` para crear ficheros csv a partir de una tabla (por ejemplo un `data.frame`) en R. El siguiente código genera primero una tabla con tres variables A, B y C y a continuación guarda esa tabla a un fichero csv. Asegúrate de abrir el fichero csv en un editor de texto para ver el resultado.

```
datos =
  data.frame(A = sample(1:100, 10), B = sample(LETTERS, 10), C = rnorm(10))
head(datos, 2)
write_csv(datos, path = "../datos/sesion02-guardarCsv.csv")
```

```
##      A B      C
## 1 25 N -0.1114757
## 2 42 Q -2.3553230
```

- Las funciones `write.table` y `write.csv` de R funcionan de manera parecida. Veremos algún ejemplo de uso más adelante.

- Las hojas de cálculo y en particular Excel son una herramienta muy utilizada. Por eso no es infrecuente encontrarse con ficheros de datos que se han almacenado en alguno de los formatos propios de diferentes versiones de Excel.
- Descarga para usar como ejemplo este fichero en formato xls, que contiene datos sobre accidentes ferroviarios ocurridos en 2010 en los Estados Unidos. Puedes encontrar más detalles sobre el fichero en [este documento auxiliar](#).
- Para leer esos datos vamos a usar la librería `readxl` de esta forma

```
library(readxl)
accidentes = read_excel("../datos/train_acc_2010.xls")
```

- **Ejercicio:** exporta esta tabla de R a un fichero en formato csv llamado `accidentes.csv`.

Ficheros de otros programas estadísticos.

- Aunque existen muchos otros programas estadísticos, aquí solo vamos a ver como se usa la librería `haven` del `tidyverse` para importar en R ficheros de datos de SPSS, Stata y SAS. Si necesitas importar datos almacenados en un formato propio de otro programa lo mejor es buscar en Internet algo como *import from ... to R*. Recuerda empezar cargando la librería.

```
library(haven)
```

Fichero SAV de SPSS

- Descarga el fichero `CH10_Planet_distances_and_y.SAV` a la carpeta `datos` desde [este enlace](#) y ábrelo con:

```
library(haven)
planetas = read_spss("../datos/CH10_Planet_distances_and_y.SAV")
head(planetas, 3) # Veamos las tres primeras filas.
```

```
## # A tibble: 3 x 4
##   Planet   PositionNumber Distancefromsunmillionmiles Lengthofyearearthyears
##   <chr>         <dbl>             <dbl>                 <dbl>
## 1 Mercury         1              36                 0.24
## 2 Venus           2              67                 0.61
## 3 Earth           3              93                  1
```

Ficheros sas7bdat de SAS y dta de Stata

- Usa [este enlace](#) para descargar el fichero `transport.sas7bdat` a la carpeta `datos` y ábrelo con:

```
transport = read_sas("../datos/transport.sas7bdat")
head(transport, 3)
```

```
## # A tibble: 3 x 4
##   AUTOTIME BUSTIME DTIME  AUTO
##   <dbl>    <dbl> <dbl> <dbl>
## 1    52.9      4.40 -48.5     0
## 2     4.10     28.5  24.4     0
## 3     4.10     86.9  82.8     1
```

- Procede de forma análoga con [este fichero](#) llamado `auto2.dta` en formato de Stata

```
auto2 = read_dta("../datos/auto2.dta")
head(auto2, 3)
```

```
## # A tibble: 3 x 13
##   make      price  mpg rep78 headroom trunk  weight  length  turn displacement gear_ratio  foreign weightsq
##   <chr>    <dbl> <dbl> <dbl>   <dbl> <dbl>   <dbl> <dbl> <dbl>   <dbl>    <dbl>   <dbl+lbl> <dbl>
## 1 AMC Concord  4099    22     3     2.5    11   2930   186    40      121      3.58 0 [Domestic] 8584900
## 2 AMC Pacer    4749    17     3     3      11  3350   173    40      258      2.53 0 [Domestic] 11222500
## 3 AMC Spirit   3799    22    NA     3      12  2640   168    35      121      3.08 0 [Domestic] 6969600
```

- Como ves todos los casos se gestionan de forma muy parecida. En ejemplos posteriores veremos otras situaciones; como tratar por ejemplo con ficheros comprimidos tipo zip.

Ficheros RData.

- R también posee su propio formato de almacenamiento de objetos, usando ficheros tipo RData. Estos ficheros pueden contener varias tablas de datos, variables y otros objetos de R. Por ejemplo podemos guardar la tabla de accidentes ferroviarios y la de planetas que hemos usado antes mediante:

```
save("accidentes", "planetas", file = "../datos/accidentes_planetas.RData")
```

- Fíjate en que hemos añadido la extensión RData manualmente, porque R no lo hace por defecto. Ahora vamos a eliminar por ejemplo la tabla planetas con:

```
rm(planetas)
```

Comprueba mirando el panel de entorno que en efecto la tabla ha desaparecido y si intentas usarla R lanzará un mensaje de error. Y ahora para recuperar esos datos usa:

```
load(file = "../datos/accidentes_planetas.RData")  
head(planetas, 3)
```

```
## # A tibble: 3 x 4  
##   Planet PositionNumber Distancefromsunmillionmiles Lengthofyearearthyears  
##   <chr>          <dbl>                <dbl>                <dbl>  
## 1 Mercury           1                     36                 0.24  
## 2 Venus             2                     67                 0.61  
## 3 Earth             3                     93                  1
```

- Para aprender más sobre manejo de ficheros con R recomendamos consultar (Boehmke 2016) y la [hoja de referencia](#) creada por RStudio.

Sección 2

Tipos de Variables.

- Los tablas de datos que hemos leído en los ficheros de la sección previa contienen variables de distintos tipos: números enteros, con decimales, fechas, variables binarias de tipo sí/no, hombre/mujer, ubicaciones, etc. Existen muchos tipos de datos distintos, que permiten distintas operaciones con ellos.
- En las próximas secciones vamos a conocer las categorías básicas de datos y las formas más adecuadas de describirlos.

- Los datos que han ido apareciendo en nuestros ejemplos se pueden clasificar en:
 - ▶ **Datos Cuantitativos (Numéricos):** que a su vez se dividen en **discretos** y **continuos**.
 - ▶ **Datos Cualitativos (Factores):** que pueden ser o no ordenados.
- Esta es la clasificación tradicional en muchos cursos de introducción a la Estadística y enseguida vamos a ver ejemplos para entender la diferencia entre estos tipos de datos, Pero queremos subrayar que existen muchos tipos de datos estructurados de uso frecuente que superan esta clasificación tradicional (fechas, imágenes, ficheros de audio o vídeo).
- Primero vamos a aprender a analizar variables individuales, por separado, antes de preguntarnos por las relaciones entre ellas.

Sección 3

Variables cuantitativas discretas: enteros.

Tablas de frecuencia absolutas y relativas

Sección 4

Variables cuantitativas continuas,

Discreto vs continuo.

Histogramas

Curvas de densidad

Sección 5

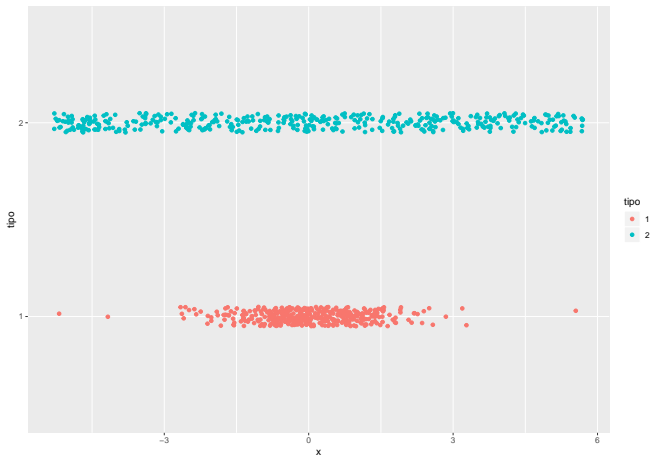
Valores centrales, de posición y dispersión.

Medidas de posición

Boxplots y violinplots.

Dispersión

- La siguiente figura contiene los boxplots de dos muestras, ambas con media 0 y el mismo número de puntos. ¿Qué diferencia a estas muestras?



Sección 6

Factores.

De nuevo, tablas de frecuencia.

Moda y representación gráfica de factores.

Factores ordenados.

Sección 7

Cadenas de caracteres (texto).

Tareas para casa



- Lee el capítulo 1 del libro.

Enlaces

[Resumen de importación de datos elaborado por RStudio.](#)

Bibliografía

Boehmke, B. C. (2016). *Data Wrangling with R* (p. 508). Springer.
doi:10.1007/978-3-319-45599-0