

# Bioestadística en tiempo de pandemia

Antonio Falcó

Seminario 1

## Objetivo de los seminarios de Bioestadística

Vamos a trabajar para comprender, analizar e intentar predecir los efectos de la pandemia provocada por el COVID-19 en el ámbito de la Comunidad Valenciana.

## Evaluación de los seminarios

Cada alumno entregará un cuaderno de campo con las respuestas, notas y conclusiones a las cuestiones propuestas durante cada uno de los seminarios. El cuaderno de campo se entregará de forma individual y se puede descargar un modelo de la página

[https://github.com/afalco/Seminarios\\_Bioestadistica](https://github.com/afalco/Seminarios_Bioestadistica) El fichero se llama `cuaderno_de_campo.docx`. Cada alumno re-nombrará el fichero como `Nombre_Apellido1_Apellido2.docx`.

# El estudio de una epidemia

- ▶ La autoridades necesitan conocer la evolución de una epidemia para poder tomar decisiones que salvaguarden la salud de los individuos de una población.
- ▶ ¿qué necesitamos observar o medir para comprender y predecir los efectos de la enfermedad sobre los individuos que componen la población?
- ▶ ¿podemos predecir la evolución de una epidemia? o de forma más precisa:
- ▶ ¿podemos construir un modelo predictivo eficaz?
- ▶ ¿qué entendemos por modelo epidemiológico?

## Cuestiones científicas para el caso de estudio

- ▶ ¿Podemos confirmar empíricamente el aumento de la mortalidad en la Comunidad Valenciana debido al efecto del COVID-19?
- ▶ ¿Qué datos nos hacen falta? y ¿donde los podemos conseguir?
- ▶ ¿Los podemos procesar de manera sencilla?

# Las fuentes de datos

- ▶ <https://data.europa.eu/euodp/es/home>
- ▶ <https://dadesobertes.gva.es/>
- ▶ [https://afalco.github.io/DailyReport\\_COVID.html](https://afalco.github.io/DailyReport_COVID.html)
- ▶ [https://momo.isciii.es/public/momo/dashboard/momo\\_dashboard.html](https://momo.isciii.es/public/momo/dashboard/momo_dashboard.html)
- ▶ <https://www.ecdc.europa.eu/en/about-us/contact-ecdc>

# El formato de los datos

datos.gob.es

reutiliza la información pública

Datos tabulares

CSV

## Conceptos básicos

- Una sola tabla de datos por archivo.
- La primera fila puede contener exclusivamente el nombre de los campos (cabecera). Es opcional.
- Cada observación/registro es una línea.
- Cada variable/campo es una columna.
- Todas las filas deben contener el mismo número de campos.
- Cada campo está separado del siguiente por un carácter singular: ",", ";", ":", "!", "!" o "#". Tener en cuenta el separador decimal para elegir el carácter separador.
- Codificación de caracteres UTF-8.

## Diccionario de datos

- Cualquier información que no sea dato (esquemas, tablas, columnas, descripciones, comentarios o unidades de medida) debe incluirse en un diccionario de datos.
- Debe expresarse en un formato procesable, por ejemplo, BOM o SON-ED, mediante un vocabulario estandarizado que permita definir cada una de las características de los elementos del archivo de datos. W3C recomienda un vocabulario para la descripción de datos tabulares.
- De las propiedades que existen para describir, tablas, filas, columnas o celdas, es recomendable usar, al menos:

Para las tablas:  
Título de tabla ("table")  
Descripción ("description")  
Publicación ("publication")  
Ubicación del archivo que se describe ("url")

Para las columnas:  
Número de columna ("name")  
Título de columna ("title")  
Descripción ("description")  
Tipo de datos ("datatype")

## 3 Los nombres de las columnas

- No deben repetirse
- Nombres cortos
- Evitar caracteres especiales

## 4 Estructura vertical

- Siempre que sea posible, el fichero debe actualizarse incorporando nuevas filas y manteniendo fijas las columnas.

## 5 Valores desconocidos

- A las valores "desconocidos" o "no obtenidos" se les asignar un código común. Por ejemplo: "NA".
- El código que se utilice para indicar los valores desconocidos debe especificarse en el diccionario de datos.

## 6 No totales o subtotales

- No deben incluirse columnas o filas que representen agrupamientos de otras.

## 7 Tipo de datos por columna

- Usar un tipo de dato por columna.
- Las unidades deben describirse en el diccionario de datos.

## 8 Valores estandarizados

- Usar vocabularios y normas estandarizadas.
- Utilizar la misma codificación y normalización para el mismo tipo de dato publicado en diferentes datasets.

## 9 Campos codificados

- Categorizar los datos para poder entrar patrones o aplicar filtros, evitando esquemas de conceptos y códigos con valores preestablecidos.

## 10 Campos de texto

- Valores que incluyen comas o saltos de línea deben ir entre comillas.
- El contenido entre comillas dentro de un valor debe ir doblemente entrecomillado.
- Los campos con valores numéricos que incluyen ceros a la izquierda significativos deben ser de tipo texto.

## 11 Campos numéricos

- No usar separadores de millares.
- El separador decimal depende de la configuración regional (En España usar ",").
- Valores negativos irán precedidos del signo "-". No usar paréntesis.
- El diccionario de datos debe expresar las unidades de medida o de moneda asociadas a valores numéricos.

## 12 Campos fecha

- Usar el formato ISO 8601: YYYY-MM-DD (forma abreviada) o YYYY-MM-DDTHH:MM:SS (forma extendida).
- Usar formato de 24 horas.
- Para indicar un mes es aconsejable ajustar al último día del mes (2019-03-31).

## 13 Campos teléfono

- Mantener siempre el mismo formato.
- Se recomienda incluir código de país.

## 14 Campos dirección postal

- Usar campos separados para codificar los elementos necesarios para localizar una dirección postal: tipo de vía; nombre de la vía; localidad; código postal.

## 15 Coordenadas geográficas

- Especificar latitud y longitud en columnas separadas y en grados decimales.

id_registro	modelo	año	consumo	aceleracion	cambio	descripcion	precio_venta	moneda	fecha_venta	promedio_venta	codigo_vendedor	actividad_vendedor	tipo_via	nombre_via	localidad	c_postal	telefono	latitud	longitud
000034070234	Chevrolet Vectra	1970	18	NA	5	Manual	22640,23	EUR	1970-06-30T15:30:00	2,51	45.11	1	Calle	Alfonso, 23	Alicante	03050	+34 6760000	38,365171	-0,481492
000034070243	Buck Skylark 320	1970	15	NA	A	Automático	22189,00	EUR	1970-07-22T10:30:00	2,63	45.11	1	Avenida	Industria, 33	Oviedo	33011	NA	43,345679	-5,844701
000034070256	Plymouth Satellite	1970	18	11	A	Automático	28862,05	USD	1973-06-07T11:33:30	3,42	45.19	6	Calle	Carroas, 2	Madrid	28760	+34 6760003	40,416022	-3,702561
000034070258	Aene Rold ut	1970	16	12	M	Manual	29000,00	USD	1972-12-12T10:00:00	NA	45.11	1	Paseo	del Vagón, 23	Huelva	11170	+34 6960004	43,345679	-6,940041
000034070321	Ford Torino	1970	17	10,5	M	Manual	32090,15	EUR	1969-09-17T13:43:30	2,54	45.19	1	Plaza	Berlín, 1	Lugo	27294	+34 6760123	43,009923	-7,756602

codigo\_registro,modelo,año,consumo,aceleracion,cambio,descripcion,precio\_venta,moneda,fecha\_venta,promedio\_venta,codigo\_vendedor,actividad\_vendedor,tipo\_via,nombre\_via,localidad,c\_postal,telefono,latitud,longitud  
000034070234,Chevrolet Vectra,1970,18,NA,M,Manual,22640,23,EUR,1970-06-30T15:30:00,2,51,45.11,Venta de automóviles y vehículos de motor ligeros,Calle,Alfonso, 23, Alicante, 03050, +34 6760000,38,365171,-0,481492  
000034070243,Buck Skylark 320,1970,15,NA,A,Automático,22189,00,EUR,1970-07-22T10:30:00,2,63,45.11,Venta de automóviles y vehículos de motor ligeros,Avenida,Industria, 33, Oviedo, 33011, NA, 43,345679,-5,844701  
000034070256,Plymouth Satellite,1970,18,11,A,Automático,28862,05,USD,1973-06-07T11:33:30,3,42,45.19,Venta de otros vehículos de motor,Calle,Carroas, 2, Madrid, 28760, +34 6760003,40,416022,-3,702561  
000034070258,Aene Rold ut,1970,16,12,M,Manual,29000,00,USD,1972-12-12T10:00:00,NA,45.11,Venta de automóviles y vehículos de motor ligeros,Paseo del Vagón, 23, Huelva, 11170, +34 6960004,43,345679,-6,940041  
000034070321,Ford Torino,1970,17,10,5,M,Manual,32090,15,EUR,1969-09-17T13:43:30,2,54,45.19,Venta de otros vehículos de motor,Plaza,Berlín, 1, Lugo, 27294, +34 6760123,43,009923,-7,756602

Introducción

Captura de pantalla

[https://datos.gob.es/sites/default/files/doc/file/guia\\_csv\\_vf.pdf](https://datos.gob.es/sites/default/files/doc/file/guia_csv_vf.pdf)

Los ficheros en formato CSV se pueden abrir y manipular con

- ▶ Excel,
- ▶ LibreOffice,
- ▶ OpenOffice,
- ▶ Editor de texto plano.

en general se emplean para guardar datos que se cargan directamente.

Veamos dos ejemplos

- ▶ Los datos sobre la distribución geográfica en el mundo del COVID-19.
- ▶ Los datos de mortalidad diaria en España.

Para ello emplearemos el programa de estadística R que permite cargar directamente los datos.

```

#these libraries need to be loaded
library(utils)

#read the Dataset sheet into R. The dataset will be called data.
data <-
read.csv("https://opendata.ecdc.europa.eu/covid19/casedistribution/csv",
na.strings = "", fileEncoding = "UTF-8-BOM")

## Warning in scan(file = file, what = what, sep = sep, quote =
quote, dec = dec, :  entrada inv'alida encontrada en la conexi'on
de entrada
'https://opendata.ecdc.europa.eu/covid19/casedistribution/csv'

colnames(data)

## [1] "dateRep" "day"
## [3] "month" "year"
## [5] "cases" "deaths"
## [7] "countriesAndTerritories" "geoId"
## [9] "countryterritoryCode" "popData2018"
## [11] "continentExp"

```

```

library(utils)
library(httr)
#####
#download the dataset from the MOMO website to a local temporary file
GET("https://momo.isciii.es/public/momo/data",
    authenticate(":", ":", type="ntlm"),
    write_disk(tf <- tempfile(fileext = ".csv")))

## Response [https://momo.isciii.es/public/momo/data]
##   Date: 2020-04-24 06:12
##   Status: 200
##   Content-Type: text/csv; charset=utf-8
##   Size: 17.1 MB
## <ON DISK> /var/folders/c3/3tnzsl9d65d8djmr4m363m0000gt/T//RtmpP2mC08/file17c5

#read the Dataset sheet into R.
#The dataset will be called "data".
data <- read.csv(tf)
#####

```

```
colnames(data)
```

```
## [1] "ambito" "cod_ambito"
## [3] "cod_ine_ambito" "nombre_ambito"
## [5] "cod_sexo" "nombre_sexo"
## [7] "cod_gedad" "nombre_gedad"
## [9] "fecha_defuncion" "defunciones_observadas"
## [11] "defunciones_observadas_lim_inf" "defunciones_observadas_lim_sup"
## [13] "defunciones_esperadas" "defunciones_esperadas_q01"
## [15] "defunciones_esperadas_q99"
```

## ¿Cómo podemos saber que contiene cada columna? La función head()

```
head(data$ambito)
```

```
## [1] nacional nacional nacional nacional nacional nacional
## Levels: ccaa nacional
```

```
head(data$nombre_ambito)
```

```
## [1] <NA> <NA> <NA> <NA> <NA> <NA>
## 19 Levels: Andaluc\303\255a Arag\303\263n ... Rioja, La
```

```
head(data$fecha_defuncion)
```

```
## [1] 2018-04-06 2018-04-07 2018-04-08 2018-04-09 2018-04-10 2018-04-11
## 749 Levels: 2018-04-06 2018-04-07 2018-04-08 2018-04-09 ... 2020-04-23
```

```
head(data$defunciones_observadas)
```

```
## [1] 1163 1154 1068 1098 1155 1139
```

## Fijamos las fechas de defunción en formato de fecha (Año-Mes-Día)

```
data$fecha_defuncion <- as.Date(data$fecha_defuncion)
```

Creamos una nueva base de datos empleando la función de R `data.frame()` que seguimos llamando `data` y que contiene a las variables `data$nombre_ambito`, `data$nombre_sexo`, `data$nombre_gedad` y las columnas de la 9 a la 15 que contienen todas las variables numéricas: `data[9:15]`

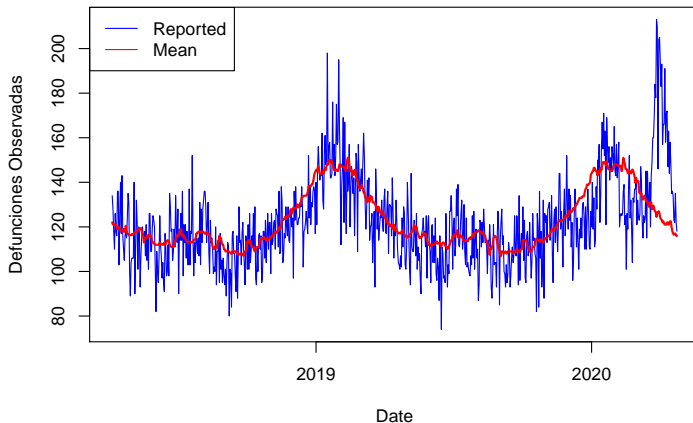
```
data <- data.frame(data$nombre_ambito,  
data$nombre_sexo,data$nombre_gedad,data[9:15])
```

Creamos una base de datos solo de la Comunidad Valenciana, para ello consideramos aquellas columnas de `data` cuyo valor para la variable

1. `data$nombre_ambito = Comunidad Valenciana`,
2. `data$nombre_sexo = todos`
3. `data$nombre_gedad = todos`

```
dataCV <- data[data$nombre_ambito=="Comunitat Valenciana"  
& data$nombre_sexo=="todos"  
& data$nombre_gedad=="todos",]  
dataCV <- na.exclude(dataCV) #Excluimos los datos en blanco NA
```

## Comunidad Valenciana



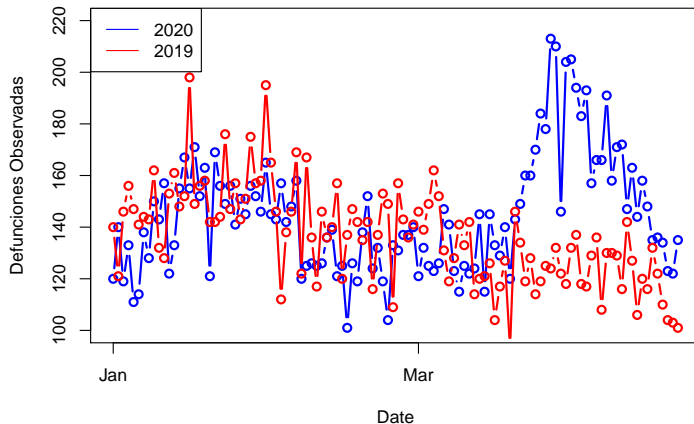
¿Qué podemos deducir de esta gráfica?

El gráfico anterior se generó con las funciones siguientes:

```
plot(dataCV$defunciones_observadas~as.Date(dataCV$fecha_defuncion,"%y/%m/%d"),
     type='l',ylab="Defunciones Observadas",
     xlab="Date",col="blue",main="Comunidad Valenciana")
lines(dataCV$defunciones_esperadas~as.Date(dataCV$fecha_defuncion,"%y/%m/%d")
      ,type='l',col="red",lwd=2)
legend("topleft", legend=c("Reported", "Mean"),
      lty=c(1,1), col=c("blue", "red"))
```

- Para poder comparar si realmente la pandemia provocada por el COVID-19 ha provocado un aumento de los fallecidos en la Comunidad Valenciana, vamos a comparar los mismos periodos de tiempo en 2019 y 2020

```
fecha_inicio_1 <- as.Date("2020-01-01")  
fecha_final_1 <- as.Date("2020-04-21")  
fecha_inicio_0 <- as.Date("2019-01-01")  
fecha_final_0 <- as.Date("2019-04-22")  
# 2019 es año bisiesto y hay 29 de Febrero
```



¿Qué podemos deducir de esta gráfica?

Construimos dos series diferentes para cada uno de los dos periodos de tiempo

```
dataCV1 <- dataCV[dataCV$fecha_defuncion>=fecha_inicio_1
& dataCV$fecha_defuncion <= fecha_final_1,]
dataCV0 <- dataCV[dataCV$fecha_defuncion>=fecha_inicio_0
& dataCV$fecha_defuncion <= fecha_final_0,]
```

El gráfico anterior se generó con las funciones siguientes:

```
plot(dataCV1$defunciones_observadas~as.Date(dataCV1$fecha_defuncion,"%y/%m/%d"),
,type='b',ylab="Defunciones Observadas",
,xlab="Date",col="blue",lwd=2,ylim=c(100,220))
lines(dataCV0$defunciones_observadas~as.Date(dataCV1$fecha_defuncion,"%y/%m/%d"),
,type='b',col="red",lwd=2)
legend("topleft", legend=c("2020", "2019"),
lty=c(1,1), col=c("blue", "red"))
```

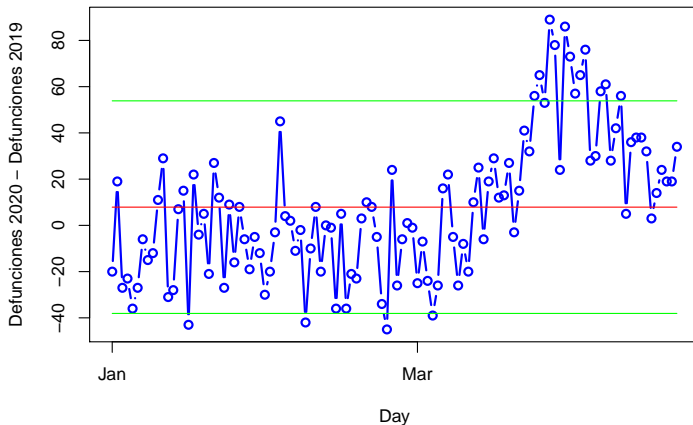
Calculamos la diferencia en los fallecimientos entre los dos periodos de tiempo de interés:

```
diferences <- (dataCV1$defunciones_observadas - dataCV0$defunciones_observadas)
```

Y construimos la serie temporal constante de la diferencia media y dos series temporales constantes que contienen  $\pm 1.5\sigma$  :

```
mu <- rep(mean(diferences),length(diferences))  
two_sigma <- rep(1.5*sd(diferences),length(diferences))
```

### Diferencias entre fallecidos 2020 – 2019 en la Comunidad Valenciana



¿Qué podemos deducir de esta gráfica?

El gráfico anterior se generó con las funciones siguientes:

```
plot(differences~as.Date(dataCV1$fecha_defuncion),type="b",col="blue",  
xlab = "Day",ylab="Defunciones 2020 - Defunciones 2019",  
main="Diferencias entre fallecidos \n 2020 - 2019 en la Comunidad Valenciana",lwd=2,  
lines(mu~as.Date(dataCV1$fecha_defuncion),type='l',col="red",lwd=1)  
lines(mu+two_sigma~as.Date(dataCV1$fecha_defuncion),type='l',col="green",lwd=1)  
lines(mu-two_sigma~as.Date(dataCV1$fecha_defuncion),type='l',col="green",lwd=1)
```