

# Master en Big Data. Fundamentos matemáticos del análisis de datos.

## Sesión 2: Tipos de Variables y Análisis Exploratorio

Fernando San Segundo

Curso 2021-22. Última actualización: 2021-07-19



- 1 Trabajando con ficheros de datos.
- 2 Tipos de Variables.
- 3 Variables cuantitativas discretas.
- 4 Variables cuantitativas continuas,
- 5 Distribuciones.
- 6 Valores centrales, de posición y dispersión.
- 7 Factores.

RECUERDA: ANTES DE SEGUIR  
EJECUTA GIT PULL EN EL REPOSITORIO  
FMAD2122

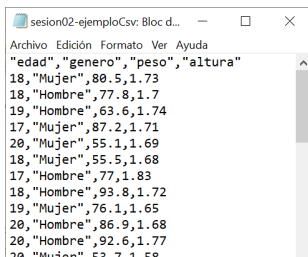
## Sección 1

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 normalmente importar datos procedentes de fuentes externas. Hoy aprenderemos a usar datos almacenados en:
  - ▶ ficheros de texto (csv, txt).
  - ▶ ficheros Excel Vamos a ver como leer estos ficheros para usar los datos en R y también veremos como guardar datos desde R en esos formatos.
- En otros recursos del curso trataremos formas alternativas de acceder a ficheros de otros programas estadísticos ( SAS, SPSS, etc.) o a datos no almacenados en ficheros (URLs, APIs, bases de datos tipo SQL, Web Scrapping, etc.)

## Ficheros de tipo csv o txt.

- El nombre csv proviene de *comma separated values*, valores separados por comas (pero 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 *línea* del fichero contiene una *fila* de la tabla y, dentro de esa fila, los elementos de cada columna de la tabla se separan mediante comas, espacios, 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",77,1.83
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

y la correspondiente tabla

También se usan ficheros txt para almacenar tablas de datos de esta manera.

# 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 las instrucciones de descarga de la primera sesión!:
  - ▶ Debes indicarle a RStudio el *Directorio de Trabajo*
  - ▶ El fichero descargado debe estar almacenado en la subcarpeta `data` del directorio de trabajo.
- Empieza abriendo ese fichero en un editor de texto (tipo *Bloc de Notas*) para hacer una exploración preliminar. Fíjate en si existe una *línea especial* en la cabecera y en el *tipo de separador* (coma, punto y coma, espacio).
- Para abrir ese fichero con R vamos a empezar usando una función del sistema básico de R:

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

Enseguida repetiremos esta operación usando el tidyverse.

- El resultado del comando básico de R `read.table` es un `data.frame`. 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.
  - ▶ *sep*: el carácter usado para separar las columnas de la tabla.

## Repaso de operaciones con data.frames.

- Recuerda que puedes seleccionar por filas con instrucciones como:

```
movies[7, ]
```

- Y por columnas de forma similar o también por nombre de variable usando \$:

```
tail(movies$Year, 20) # se muestran las 20 últimas
```

- Recuerda asimismo que puedes seleccionar por condiciones. Por ejemplo para ver el género de las películas de 2010 con:

```
movies$Genre[movies$Year == 2010]
```

- **Ejercicio:** Repite estas operaciones con dplyr. Además úsalo para localizar 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 es muy fácil de usar y más rápida que read.table para ficheros grandes. Explora esta tabla como hemos hecho con la primera versión.

```
library(tidyverse)
movies2 <- read_csv("./data/movies.csv")
```

- Ejercicio:** Ejecuta str(movies2). ¿Qué tipo de objeto usa el tidyverse para almacenar tablas en lugar del data.frame del R básico?
- 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 resultante en un editor de texto para ver el resultado.

```
set.seed(2019)
datos <-
  data.frame(A = sample(1:100, 10), B = sample(LETTERS, 10), C = rnorm(10))
head(datos, 2)
write_csv(datos, file = "./data/sesion02-guardarCsv.csv")
```

Las funciones write.table y write.csv de R básico funcionan de manera parecida.

- Es frecuente 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` (del tidyverse) de esta forma

```
library(readxl)
accidentes <- read_excel("./data/train_acc_2010.xls")
```

Puedes leer más sobre `readxl` [aquí](#).

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

## Sección 2

### Tipos de Variables.

# Tipos de Variables.

- Las tablas de datos que hemos leído en los ficheros de la sección previa pueden contener 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. Como ejemplos iniciales vamos a usar la tabla `mpg` contenida en la librería `tidyverse` y también una tabla con datos relativos a un estudio sobre enfermedades coronarias llevado a cabo en Framingham (UK). Puedes descargar el fichero csv desde [este enlace](#) y leer más detalles sobre el estudio [aquí](#)..

- **Ejercicio:** lee el fichero a una tabla de R llamada fhs (de Framingham Heart Study). Explora esa tabla con las funciones `str` y `glimpse`. Piensa en qué tipo de información contiene cada variable de la tabla. Lee también la documentación sobre `mpg` en [este enlace](#).

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHC
1	1	39	4	0	0	0	0	0	0	195	106.0	70.0	26.97	80	77	0
2	0	46	2	0	0	0	0	0	0	250	121.0	81.0	28.73	95	76	0
3	1	48	1	1	20	0	0	0	0	245	127.5	80.0	25.34	75	70	0
4	0	61	3	1	30	0	0	1	0	225	150.0	95.0	28.58	65	103	1
5	0	46	3	1	23	0	0	0	0	285	130.0	84.0	23.10	85	85	0
6	0	43	2	0	0	0	0	1	0	228	180.0	110.0	30.30	77	99	0
7	0	63	1	0	0	0	0	0	0	205	138.0	71.0	33.11	60	85	1
8	0	45	2	1	20	0	0	0	0	313	100.0	71.0	21.68	79	78	0
9	1	52	1	0	0	0	0	1	0	260	141.5	89.0	26.36	76	79	0
10	1	43	1	1	30	0	0	1	0	225	162.0	107.0	23.61	93	88	0
11	0	50	1	0	0	0	0	0	0	254	133.0	76.0	22.91	75	76	0
12	0	43	2	0	0	0	0	0	0	247	131.0	88.0	27.64	72	61	0
13	1	46	1	1	15	0	0	1	0	294	142.0	94.0	26.31	98	64	0

- 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.

# Variable cuantitativa

- Una *variable cuantitativa* (discreta o continua) es una variable que toma valores numéricos que *además* se han medido en alguna escala que permite interpretarlos y hacer operaciones aritméticas (sumas, productos, etc) con ellos.
- Una variable cuantitativa es *discreta* si se mide en una escala de unidades enteras (paso a paso, los valores se miden *contando*). Y la variable *continua* si la escala de medida se puede dividir arbitrariamente (se usan valores decimales). Pero como veremos en ejemplos, la división discreto/continuo también es sutil y se refiere en realidad a la forma en la que *usamos* la variable.
- Podría pensarse entonces que las variables cuantitativas son las variables numéricas y las cualitativas las no numéricas. La diferencia es, en realidad, un poco más sutil. Una variable es *cualitativa (nominal)* cuando *solo* se utiliza para establecer categorías, para *clasificar*. Podemos *representar* los valores de una de estas variables con números, pero el valor numérico concreto es arbitrario, es una *etiqueta*. Las variables cualitativas nominales también se denominan *factores*
- **Ejercicio:** Examina las variables `cty`, `disp`, `class` y `cyl` de la tabla `mpg`. ¿De qué tipo crees que es cada variable?

## Sección 3

Variables cuantitativas discretas.



## Tablas de frecuencia absolutas y relativas.

- La variable `cty` de `mpg` el número de millas por galón que el coche recorre en ciclo urbano (atención: valores más altos indican menor consumo de combustible). Fíjate en que los valores son un número entero de millas. En principio no hay nada que impida dar esos valores con decimales. Pero no es eso lo que se *ha decidido* hacer aquí, sino que se trata como una variable discreta.
- El primer paso con una variable discreta como esta es obtener una tabla de frecuencias (absolutas), que nos dirá qué valores toma la variable y cuántas veces toma cada valor. Usando `table`

```
table(mpg$cty)
```

```
 9 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 28 29 33 35  
5 20  8 21 19 24 19 16 26 20 11 23  4  3  5  2  3  2  1  1  1
```

- También se puede usar la función `count` de `dplyr` así (se omite el resultado):

```
mpg %>%  
  count(cty)
```

## Tabla de frecuencias relativas.

- A menudo, y especialmente cuando se usan para comparaciones, nos interesa más saber la fracción del total que corresponde a cada uno de los valores de una variable discreta. Cuando esas fracciones se expresan como *tanto por uno* obtenemos las frecuencias relativas, que es fácil convertir en porcentajes.
- Vamos a obtener una tabla de frecuencias relativas con R básico (para mejorar la presentación elegimos el número de cifras significativas con `signif`):

```
signif(prop.table(table(mpg$cty)), 2)
```

9	11	12	13	14	15	16	17	18	19
0.0210	0.0850	0.0340	0.0900	0.0810	0.1000	0.0810	0.0680	0.1100	0.0850
20	21	22	23	24	25	26	28	29	33
0.0470	0.0980	0.0170	0.0130	0.0210	0.0085	0.0130	0.0085	0.0043	0.0043
35									
0.0043									

- Se puede usar `dplyr` pero en este caso es más complicado que con el R básico.

```
mpg %>%  
  count(cty) %>%  
  mutate(cty, relFreq = prop.table(n), n=NULL)
```

Ejecuta el código por fases para entenderlo y analizar el papel de cada función y argumento.

## Propiedades de las frecuencias relativas.

- Las frecuencias relativas suman siempre 1. Compruébalo:

```
sum(prop.table(table(mpg$cty)))
```

**Ejercicio:** Piensa cómo hacer esto en `dplyr`.

- Además las frecuencias relativas están relacionadas con la idea de *probabilidad empírica*. Es decir, si elegimos aleatoriamente un valor de la variable `cty` y repetimos esa elección muchas veces, la probabilidad de cada uno de los distintos valores es la frecuencia relativa que hemos calculado.

## Frecuencias acumuladas.

- Las *frecuencias acumuladas* se usan con variables discretas para responder a la pregunta “¿cuántos valores hay que sean menores o iguales que ...?” En R se obtienen con:

```
cumsum(table(mpg$cty))
```

```
 9  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  28  29  33
 5  25  33  54  73  97 116 132 158 178 189 212 216 219 224 226 229 231 232 233
35
234
```

Por ejemplo, en la tabla hay 116 valores menores o iguales que 16. **Ejercicio:** (1) con `mutate` y una función adecuada haz esto en `dplyr`. (2) ¿para qué sirven las frecuencias *acumuladas relativas*? ¿son distintas de las *relativas acumuladas*?

## Sección 4

Variables cuantitativas continuas,

## Discreto vs continuo.

- Las tablas de frecuencias por valores no son útiles cuando hay muchos valores distintos. La tabla de frecuencias de `cty` ya era un poco excesiva. Pero si tratamos de calcular una tabla de frecuencia para la variable `age` de la tabla `fhs`

```
table(fhs$totChol)
```

puedes comprobar que la tabla que se obtiene no es una representación útil de la información.

- En muchos ejemplos como este las diferencias entre valores consecutivos no son relevantes. Las preguntas relevantes pasan a ser las que se refieren a intervalos de valores. Para agrupar los valores en intervalos en R podemos usar la función `cut`.

```
cholLevels = cut(fhs$totChol, breaks = 10)  
head(cholLevels)
```

```
[1] (166,225] (225,284] (225,284] (225,284] (284,343] (225,284]  
10 Levels: (106,166] (166,225] (225,284] (284,343] (343,402] ... (637,697]
```

- La respuesta de R nos indica que ha dividido el *recorrido* de la variable (de mínimo a máximo) en 10 intervalos semiabierto de igual longitud. El primero incluye los valores entre 106 y 166, hasta el último que incluye los valores de 637 a 697. La variable `cholLevels` que hemos fabricado es un *factor ordenado*. Veremos más ejemplos cuando aprendamos más sobre factores.

- Con las variables puramente continuas no suele haber demasiado dudas a la hora de reconocerlas. Pero con las variables discretas el problema puede ser más complicado, porque depende esencialmente del número de valores distintos que tome la variable. Al final, en muchos casos, tratar a una variable como discreta o continua es decisión de quien realiza el análisis.
- Por ejemplo, la tabla de frecuencias de la variable agrupada `cholLevels` es mucho más informativa que la de la variable original `totChol`. Eso nos indica que seguramente es mejor tratar a `cholLevels` como una variable continua aunque sus valores sean enteros,

```
table(cholLevels)
```

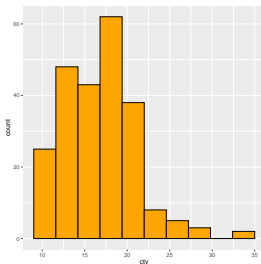
```
cholLevels
(106,166] (166,225] (225,284] (284,343] (343,402] (402,460]
      164      1555      1898       503        60         7
(460,519] (519,578] (578,637] (637,697]
      1         0         1         1
```

- En cualquier caso si una variable cuantitativa discreta solo toma cinco o menos valores en general es beneficioso pensar en ella como un *factor ordenado*, que discutiremos más adelante.

# Histogramas.

- Una forma común de representar gráficamente la tabla de frecuencias de una variable discreta que tome más de cinco valores distintos es mediante un *histograma*, que es un diagrama de barras.
- Lo dibujamos con ggplot, eligiendo primero los puntos de corte con cut:

```
# Histograma con ggplot2
cortes = seq(min(mpg$cty), max(mpg$cty), length.out = 11)
ggplot(data = mpg) +
  geom_histogram(mapping = aes(cty), breaks = cortes,
                 fill = "orange", color="black")
```



Fíjate en que el eje horizontal contiene los valores de la variable mientras que el eje vertical muestra las frecuencias.

- **Ejercicios:**

- ▶ ¿Cómo se usa la función `seq` que hemos empleado para fabricar los cortes del histograma? Averigua lo que hace la opción `by` de esa función `seq`.
- ▶ Fíjate en las opciones de color de `geom_histogram` y busca información sobre nombres de colores en R.
- ▶ Dibuja un histograma de `mpg$cy1`. ¿Por qué ocurre esto?
- ▶ Consulta [este enlace](#) para ver cómo se dibujan histogramas en R básico.

### ¿Cuántos intervalos se deben usar en la construcción de un histograma?

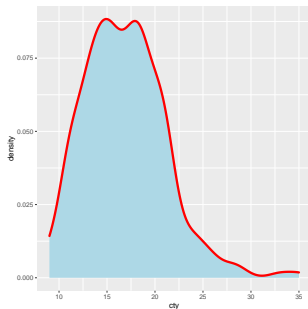
- No hay una regla fija. Aunque R y el resto de programas utilizan diversos algoritmos para determinar ese número, lo cierto es que la respuesta depende mucho de los datos concretos con los que trabajamos. Por eso normalmente es necesario experimentar un poco con diversos valores. En cualquier caso es *desaconsejable* utilizar menos de cinco intervalos (o más que  $\sqrt{n}$ , siendo  $n$  el número de datos).



## Curvas de densidad.

- La *curva de densidad* es un tipo de diagrama alternativo al histograma. Por ejemplo, para los datos de `cty` que venimos usando se obtiene con:

```
ggplot(mpg) +  
  geom_density(mapping = aes(cty), color="red", fill="lightblue", size=1.5)
```



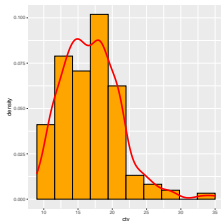
El eje horizontal contiene los valores de la variable. La altura de la curva indica la frecuencia de los valores.

- Ejercicio:** Usando los datos de `auto2` dibuja la curva de densidad de cada una de las variables `length`, `price`, `displacement` y `rep78`.
- ¿Curvas de densidad o histogramas?** ¡Ambos! En general, cuando analicemos un conjunto de datos es mejor empezar en la fase de exploración dibujando muchos gráficos. Al presentar

## Relación entre curvas de densidad e histogramas.

- En muestras de tamaño grande y usando una partición fina en subintervalos la curva de densidad se ajusta bastante a la forma o perfil del histograma como ilustra este ejemplo. Esa *forma* es lo que llamaremos **distribución** de la variable.

```
ggplot(mpg, aes(x = cty)) +  
  geom_histogram(aes(y=stat(density)),  
                 breaks = cortes, fill = "orange", color="black") +  
  geom_density(color="red", size=1.5)
```



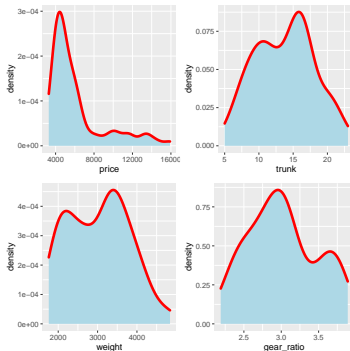
Este fenómeno es una manifestación más de esa separación borrosa que existe entre las variables discretas con muchos valores (el histograma es una representación discreta) y las variables continuas (la curva de densidad es una representación continua).

- Ejercicio:** Para entender `stat(density)` lee esta sección de [R4DS](#)

## Sección 5

### Distribuciones.

- En un ejercicio previo proponíamos dibujar las curvas de densidad de varias variables de la tabla `auto2`. Si lo haces obtendrás curvas como estas:

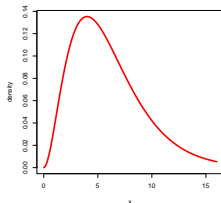


**Mira el código de esta sesión** para ver como dibujar esta “*tabla de figuras*”.

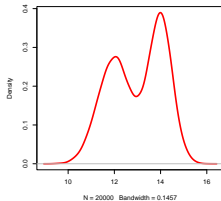
- Cada una de esas curvas muestra una *distribución de los valores*, que es una representación de la forma en la que se reparten los datos entre los valores posibles. Las distribuciones van a jugar un papel central en las próximas sesiones, así que vamos a desarrollar un poco de lenguaje para referirnos a ellas.

# Distribuciones unimodales y multimodales.

- Una distribución que presenta un único máximo se denomina *unimodal*



- Mientras que una distribución con dos máximos locales claramente definidos como la de la figura se denomina *bimodal* (o *multimodal* cuando son más de dos).

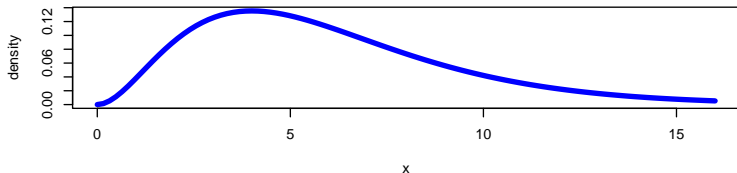


- La mayoría de las distribuciones con las que vamos a trabajar serán unimodales.

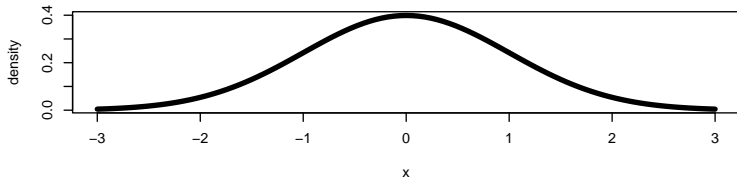
## Distribuciones asimétricas.

- Otra característica de las distribuciones en la que nos vamos a fijar a menudo es su simetría. Fíjate en que es la *cola* más larga de la distribución la que da nombre a la asimetría. En inglés esta característica se denomina **skewness**

### Asimétrica a derecha



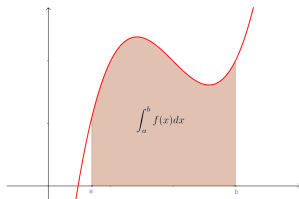
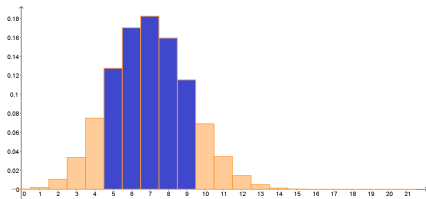
### Simétrica



### Asimétrica a izquierda

## Distribuciones y cálculo de probabilidad. Discusión informal.

- La distribución de una variable discreta con muchos valores distintos se representa adecuadamente mediante un histograma. Si las alturas de las barras del histograma representan frecuencias relativas entonces la probabilidad de que un valor elegido al azar caiga en el intervalo  $[a, b]$  se obtiene sumando el área de todas las barras desde  $a$  hasta  $b$ , como se ilustra en la parte izquierda de la figura.
- Por su parte, La distribución de una variable continua se representa mediante una curva de densidad. Y entonces la probabilidad de que un valor elegido al azar caiga en el intervalo  $[a, b]$  se obtiene calculando (con una integral) el área bajo la curva desde  $a$  hasta  $b$



- Don't panic! No hace falta que sepas calcular integrales. Pero es importante entender que el cálculo de probabilidades está estrechamente relacionado con el cálculo de áreas.

## Sección 6

Valores centrales, de posición y dispersión.



# La media aritmética

- Al trabajar con variables cuantitativas muchas veces trataremos de elegir un *valor central*. Es decir, un valor que sea *representativo* de los valores que toma la variable. Un buen valor central debería ser la respuesta a esta pregunta: “*si elijo un valor de la variable al azar ¿lo más probable es que se parezca a ... ?*”
- El valor central más conocido es la *media aritmética*. Dados  $n$  números  $x_1, x_2, \dots, x_n$  su media aritmética es:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

- En R se utiliza la función `mean` para calcular la media de los valores de un vector numérico. Con este código vamos a elegir 20 números al azar entre 0 y 100 (con remplazamiento). Recuerda que los paréntesis sirven para que R muestre el resultado.

```
(muestra = sample(0:100, size = 20, replace = TRUE))
```

```
[1] 24 41 68 16 60 75 18 25 8 89 45 71 22 93 71 66 83 46  
[19] 68 66
```

```
(media = mean(muestra))
```

```
[1] 52.75
```

- Ejercicio:** Repítelo varias veces y mira las medias que obtienes. ¡Al pensar en esas medias estás empezando a hacer Estadística!

## La media y los valores atípicos.

- La media aritmética se usa mucho porque su definición matemática es muy sencilla. Pero tiene un problema: su valor se ve muy afectado por la presencia de valores anormalmente grandes o pequeños de la variable, los llamados *valores atípicos* (que muy pronto vamos a definir con más rigor).
- Por ejemplo, si elegimos 99 números al azar de 0 a 100, su media es parecida a 50 como cabría esperar.

```
set.seed(2019)
muestra = sample(0:100, size = 99, replace = TRUE)
(media = mean(muestra))
```

```
[1] 51.64646
```

Pero si ahora añadimos un único valor igual a 1000 y volvemos a calcular la media:

```
muestra2 = c(muestra, 1000)
(media2 = mean(muestra2))
```

```
[1] 61.13
```

Las dos muestras son esencialmente idénticas, así que al elegir un valor representativo (en el sentido probabilístico que hemos discutido) nos gustaría obtener respuestas mucho más parecidas. Pero como ilustra este ejemplo, la media se ve muy afectada por ese único valor atípico.

## La mediana.

- Para paliar este problema podemos usar la *mediana* como sustituto o alternativa a la media. La mediana de un conjunto de valores se define con esta receta: se ordenan los números de menor a mayor y se toma el valor que queda en el centro (si hay una cantidad impar de valores; si son pares se promedian los dos valores centrales).
- Por ejemplo, para calcular la mediana de estos 17 valores:

```
set.seed(2019)
(valores = sample(1:100, 17, replace = TRUE))
```

```
[1] 25 42 69 17 61 76 19 26 9 90 46 72 23 94 72 67 84
```

los ordenamos y tomamos el valor central

```
(ordenados = sort(valores))
(mediana = ordenados[9])
```

```
[1] 9 17 19 23 25 26 42 46 61 67 69 72 72 76 84 90 94
[1] 61
```

Aunque lo mejor, claro, es usar la función de R:

```
median(valores)
```

```
[1] 61
```

## Más detalles sobre la mediana.

- Por su propia construcción (en términos de posiciones/rangos y no de tamaños) debería estar claro que la mediana no se ve muy afectada por la presencia de valores atípicos. Si volvemos a la muestra de 99 valores que construimos antes y miramos su mediana antes y después de añadir el valor atípico 1000 se obtiene:

```
median(muestra)
```

```
[1] 54
```

```
median(muestra2)
```

```
[1] 54
```

En este ejemplo concreto la mediana no se ve afectada en absoluto por ese valor.

- Y entonces, ¿por qué no se usa siempre la mediana en lugar de la media aritmética? Pues porque la definición de la mediana utiliza unas matemáticas bastante más complicadas que la media aritmética. Gracias al ordenador la importancia de los métodos basados en la mediana ha ido aumentando. Pero los métodos clásicos que usan la media aritmética siguen siendo los más comunes.
- En cualquier caso, recomendamos explorar bien los datos, identificar posibles valores atípicos y comprobar si la media y la mediana son parecidas.

- **Cuartiles y percentiles.** La mediana se puede pensar como el valor que es mayor o igual que el 50 % de los valores de nuestro conjunto de datos. De la misma forma se definen el *primer y tercer cuartil* como los valores que son mayores o iguales respectivamente que el 25 % y el 75 % de los valores. Esto se generaliza a la noción de percentil. El percentil 43, por ejemplo, es un valor que deja por debajo al 43 % de los valores. En R ese percentil se calcula para una variable como `cty` de `mpg` mediante:

```
quantile(mpg$cty, probs = 0.43)
```

```
43%  
16
```

- El *recorrido* (del inglés *range*) de un conjunto de valores es el intervalo que va del valor mínimo al máximo. Cuidado: en español se usa a menudo erróneamente *rango*. Pero los rangos en Estadística son otra cosa (que corresponde a *rank* en inglés).
- Una manera sencilla de obtener varias de estas medidas de posición es usando `summary`

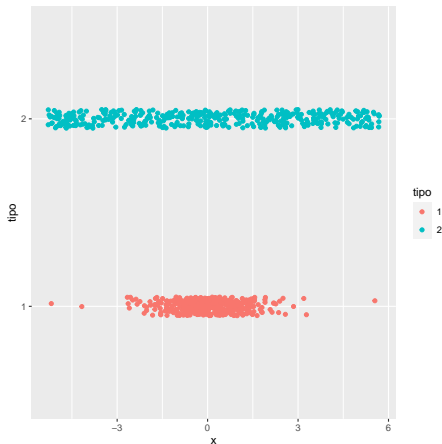
```
summary(mpg$cty)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9.00	14.00	17.00	16.86	19.00	35.00

- **Ejercicio.** Prueba a aplicar la función a todo el `data.frame` con `summary(mpg)`. Ejecuta también `str(mpg)` y `glimpse(mpg)`.

# Dispersión

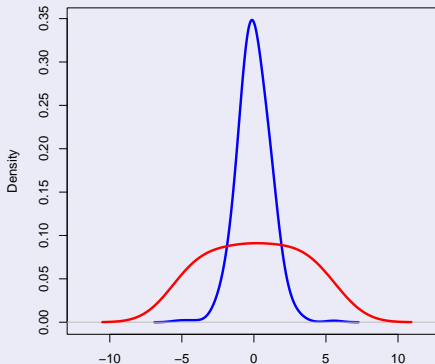
- La siguiente figura muestra dos conjuntos de valores, ambas con media 0 y el mismo número de puntos. Las coordenadas verticales de los puntos se han modificado con un poco de *ruido* para hacer mejor la visualización de ambos conjuntos. Puedes mirar el código de esta sesión para ver como se ha construido esta figura.



- ¿Qué diferencia a estos dos conjuntos de datos?

## Dispersión en los diagramas de densidad.

- En esta otra figura puedes ver juntas las curvas de densidad de esos dos conjuntos de datos.



Este diagrama ilustra lo que la mostraban las nubes de puntos de la figura previa: los puntos de un conjunto están más agrupados en torno al centro que los del otro conjunto. La diferencia entre ambos conjuntos está en la **dispersión**.

- La idea intuitiva de la dispersión está clara: se trata de medir como de agrupados en torno al centro están los valores. ¿Como se define rigurosamente? Al igual que sucedía con los valores centrales (media vs mediana) hay varias posibilidades.

## Recorrido intercuartílico. Valores atípicos.

- La primera forma de medir la dispersión que vamos a ver se basa en observar la diferencia entre los cuartiles tercero (75 %) y primero (25 %). Ese número es el *recorrido intercuartílico*; un intervalo de longitud IQR desde el primer hasta el tercer cuartil contiene siempre al 50 % central de los valores. En R se calcula con:

```
IQR(mpg$cty)
```

```
[1] 5
```

Fíjate en que es precisamente la diferencia entre esos dos cuartiles.

```
summary(mpg$cty)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9.00	14.00	17.00	16.86	19.00	35.00

- El IQR nos permite dar una definición formal de los **valores atípicos**: son aquellos valores que superan  $(\text{tercer cuartil}) + 1.5 \cdot \text{IQR}$  o quedan por debajo de  $(\text{primer cuartil}) - 1.5 \cdot \text{IQR}$ . Por ejemplo, en `mpg$cty` será atípico cualquier valor fuera de este intervalo

```
unnamed(quantile(mpg$cty, probs = c(1/4, 3/4)) + c(-1, 1) * 1.5 * IQR(mpg$cty))
```

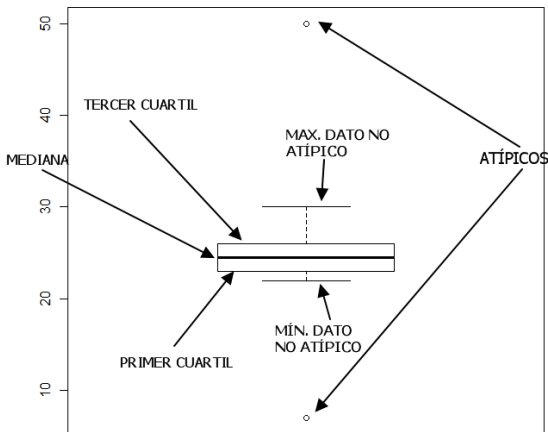
```
[1] 6.5 26.5
```

- Ejercicio:** Mira el código de esta sesión y calcula el IQR de los dos conjuntos de datos de las páginas previas.



# Boxplots

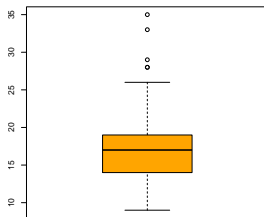
- El *boxplot* o diagrama de caja y bigotes (box and whiskers) es una forma de representar gráficamente estas medidas de posición. La estructura de un boxplot es la que se describe en esta figura:



## Boxplots con R básico.

- Dibujar un boxplot con R básico es muy sencillo. Por ejemplo, para la variable `cty`:

```
bxp_cty = boxplot(mpg$cty, col="orange")
```



- Además al darle un nombre al boxplot podemos usar ese nombre para acceder a varios componentes del boxplot. Por ejemplo los valores atípicos en el vector de datos:

```
bxp_cty$out
```

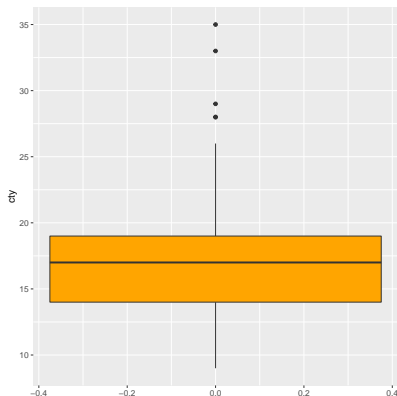
```
[1] 28 28 33 35 29
```

- **Ejercicio:** ¿Cuáles son las *posiciones que ocupan* los valores atípicos de esta variable? Además haz un boxplot de `speed` en accidentes y busca en la ayuda del boxplot como dibujar el gráfico sin los valores atípicos.

# Boxplot con ggplot

- En este caso usamos:

```
ggplot(mpg) +  
  geom_boxplot(mapping = aes(y = cty), fill="orange")
```

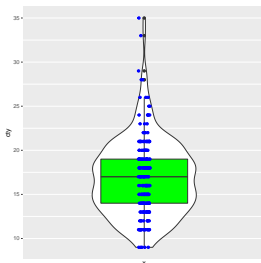


A diferencia de lo que sucede con el boxplot de R básico, ggplot no proporciona una forma directa de localizar los posibles valores atípicos.

# Boxplot con datos y violinplots.

- Aunque los boxplots son muy útiles y se usan mucho, tienen sus limitaciones. La principal de ellas es que no muestran la distribución subyacente de valores (¡la *forma*!). Hay dos formas de combatir ese problema:
  - ▶ incorporando los valores al gráfico
  - ▶ usando un *violinplot* que es básicamente un boxplot con curvas de densidad añadidas.
- Esos dos remedios se pueden aplicar por separado o combinados como hemos hecho en este gráfico de ggplot:

```
ggplot(mpg) +  
  geom_violin(mapping = aes(x=0, y = cty)) +  
  scale_x_discrete(breaks = c()) +  
  geom_boxplot(mapping = aes(y = cty), fill="green") +  
  geom_jitter(aes(x=0, y = cty),  
              position = position_jitter(w=0.05, h= 0), col="blue")
```



## Desviación absoluta mediana

- Además del rango intercuartílico se usan otras medidas de dispersión. Las más comunes están basadas en la idea de medir la *desviación absoluta* de cada dato individual con respecto a un valor central. Si los datos son  $x_1, x_2, \dots, x_n$  y el valor central (puede ser la media o la mediana) es  $c$  entonces las desviaciones absolutas son:

$$d_1 = |x_1 - c|, \quad d_2 = |x_2 - c|, \quad \dots \quad d_n = |x_n - c|$$

Usamos valores absolutos para evitar que las desviaciones por exceso compensen a las desviaciones por defecto. Aunque el valor absoluto es complicado. . .

- Para obtener una medida de dispersión buscamos un valor representativo (valor central) de las desviaciones absolutas. Por ejemplo, la mediana de las desviaciones absolutas respecto de la mediana (usando  $c$  igual a la mediana de los datos). Este valor es la **desviación absoluta mediana (MAD)**, que en R se calcula así:

```
library(readxl)
accidentes = read_excel("./data/train_acc_2010.xls")
mad(accidentes$Speed, constant = 1)
```

```
[1] 4
```

- Ejercicio:** calcula este valor a partir de los datos usando `median`. Mira lo que sucede si quitas la opción `constant = 1`. Lee la ayuda de `mad` y [esta página](#).
- Ejercicio:** calcula `mean(accidentes$Speed - mean(accidentes$Speed))`. Cambia de variable y de tabla y repite el cálculo. ¿Por qué pasa esto?

## Varianza y desviación típica (definición poblacional).

- Usar los valores IQR o MAD como medidas de dispersión tiene inconvenientes similares a los que planteamos al comparar media y mediana. Las matemáticas de medianas y valores absolutos son complicadas y los resultados teóricos asociadas son más difíciles. Por esa razón tradicionalmente la Estadística ha usado una medida de dispersión basada en la media y con matemáticas más simples.
- Dados los valores numéricos  $x_1, x_2, \dots, x_n$  su **varianza (en sentido poblacional)** es la *media de las desviaciones cuadráticas respecto de la media aritmética  $\bar{x}$* :

$$\sigma_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

y la **desviación típica (también poblacional)** es la raíz cuadrada de la varianza:

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

Elevamos al cuadrado para conseguir un efecto similar al valor absoluto, pero usando una función derivable. Pero al hacerlo cambiamos las unidades y tenemos que tomar raíz cuadrada para obtener la dispersión en las unidades originales del problema.

## Varianza y desviación típica muestrales.

- Al introducir la varianza y desviación típica les hemos puesto el apellido de *poblacionales*. Y tal vez te hayas dado cuenta de que no hemos dicho como calcular esos valores en R.
- La razón para hacer esto empezará a quedar más clara en las próximas sesiones y tiene que ver con la esencia misma de la Estadística. La misión fundamental de la Estadística se puede resumir en **obtener información fiable sobre una población a partir de muestras**. En particular, para obtener información sobre la media y la dispersión en la población usaremos la media y la dispersión en la muestra. Y ese es el problema: usar en la muestra la misma fórmula de la varianza que en la población *no funciona bien para averiguar la varianza de la población*.
- El remedio es sencillo, afortunadamente. Hay que usar en la muestra las fórmulas *muestrales* para la varianza y desviación típica:

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \qquad s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

como ves, la diferencia estriba en que dividimos por  $n - 1$  en lugar de  $n$ , siendo  $n$  el **tamaño de la muestra** (no el de la población).

## Las funciones var y sd.

- Para calcular la varianza muestral en R usamos `var`:

```
var(mpg$displ)
```

```
[1] 1.669158
```

puedes comprobar que esto es lo mismo que:

```
n = length(mpg$displ)
media = mean(mpg$displ)
sum((mpg$displ - media)^2) / (n - 1)
```

```
[1] 1.669158
```

- Y para calcular la desviación típica muestral usamos `sd` (de *standard deviation*):

```
sd(mpg$displ)
```

```
[1] 1.291959
```

puedes comprobar que esto es lo mismo que:

```
sqrt(var(mpg$displ))
```

```
[1] 1.291959
```



## Sección 7

### Factores.

## Tablas de frecuencia y terminología para factores.

- Hasta ahora hemos centrado la discusión en variables cuantitativas, con valores numéricos. Vamos a tratar aquí muy brevemente las variables cualitativas o factores. Habrá muchas ocasiones en las próximas sesiones de practicar su uso.
- Al trabajar con factores sigue teniendo sentido usar tablas de frecuencias absolutas y relativas (pero no acumuladas, en general):

```
table(accidentes$TrkType)
```

Industry	Main	Not rptd	Siding	Yard
247	975	3	56	1340

```
prop.table(table(accidentes$TrkType))
```

Industry	Main	Not rptd	Siding	Yard
0.094238840	0.371995422	0.001144601	0.021365891	0.511255246

- **¡Con factores genéricos los valores centrales o la dispersión no tienen sentido!**
- Los distintos valores de un factor se suelen llamar *niveles (levels)* del factor. Un factor *dicotómico* (o binario) tiene dos niveles; es decir, solo toma dos valores distintos. Si toma más de dos valores el factor es *politómico*. El nivel más frecuente es la *moda* del factor.

## Operaciones básicas con factores.

- Podemos crear un factor a partir de un vector de strings con la función `factor`:

```
(ardeida = factor(c("martinete", "garzaReal", "avetorillo", "garzaReal",  
                  "cangrejera", "martinete", "martinete"), ))
```

```
[1] martinete  garzaReal  avetorillo  garzaReal  cangrejera  martinete  
[7] martinete  
Levels: avetorillo cangrejera garzaReal martinete
```

Dos detalles sobre la salida: fíjate en la ausencia de comillas y en que el orden de los niveles es alfabético. Si quieres otro orden (por ejemplo para las tablas de frecuencia) puedes hacerlo explícito:

```
(ardeida = factor(c("martinete", "garzaReal", "avetorillo", "garzaReal",  
                  "cangrejera", "martinete", "martinete"),  
levels = c("garzaReal", "martinete", "cangrejera", "avetorillo")))
```

```
[1] martinete  garzaReal  avetorillo  garzaReal  cangrejera  martinete  
[7] martinete  
Levels: garzaReal martinete cangrejera avetorillo
```

- El factor puede ser *ordenado* si además incorporamos la opción `ordered = TRUE`. *No se debe confundir con el uso de levels para fijar un orden “estético” de los niveles. En un factor ordenado el orden aporta información relevante sobre los niveles.*

# Más funciones que generan factores

- Ya hemos visto que el resultado de `cut` es un factor ordenado cuyos niveles son los intervalos en que se divide el recorrido de la variable.
- La función `gl` sirve para generar factores a medida y es un complemento para otras funciones como `rep`. Un ejemplo en el que fabricamos un factor con tres niveles y 4 repeticiones:

```
gl(n = 3, k = 4, labels = c("piedra", "papel", "tijera"))
```

```
[1] piedra piedra piedra piedra papel  papel  papel  papel  tijera  
[10] tijera tijera tijera  
Levels: piedra papel tijera
```

- A veces, por cuestiones de diseño del experimento o del conjunto de datos, queremos que los niveles del factor aparezcan intercalados.

```
gl(n = 3, k=1, length = 30, labels = c("piedra", "papel", "tijera"))
```

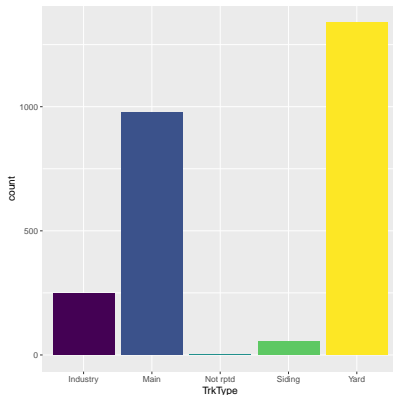
```
[1] piedra papel  tijera piedra papel  tijera piedra papel  tijera  
[10] piedra papel  tijera piedra papel  tijera piedra papel  tijera  
[19] piedra papel  tijera piedra papel  tijera piedra papel  tijera  
[28] piedra papel  tijera  
Levels: piedra papel tijera
```

- Puedes consultar el Capítulo 7 de (Boehmke 2016), el [Capítulo 12 \(15 en la versión online\)](#) de (Wickham and Grolemund 2016), o el Capítulo 6 de (Matloff 2011).

## Gráficos para factores: diagramas de barras.

- El diagrama más simple y a menudo el más efectivo para representar la tabla de frecuencias de los niveles de un factor es el diagrama de barras:

```
library(viridisLite)
ggplot(accidentes) +
  geom_bar(mapping = aes(x = TrkType, fill= viridis(5)))
```



- Se desaconseja el uso de diagramas de tarta (*piecharts* en inglés).

## Factores dicotómicos: variables binarias (de Bernouilli).

- Un factor dicotómico solo tiene dos niveles (valores). Esos valores se pueden representar como sí/no, cierto/falso, éxito/fallo, etc. Los factores dicotómicos también se llaman a menudo variables binarias o de Bernouilli. Los factores con más de dos niveles se llaman *politómicos*.
- *Media como proporción*. Si pensamos en los niveles de un factor binario como cierto/falso, entonces al igual que hace R podemos identificar cierto/falso con 1/0. Esta representación numérica tiene una ventaja importante. Aunque hemos dicho que los valores centrales no tienen sentido en factores genéricos, en este caso concreto la media aritmética de un vector de unos y ceros *mide la proporción* de valores iguales a 1.
- Veamos un ejemplo. La variable `male` de la tabla `fhs` es un factor binario que vale 1 si el paciente es un hombre y 0 si es una mujer. Para averiguar la proporción de hombres en ese estudio basta con hacer:

```
fhs = read_csv("./data/framingham.csv")
mean(fhs$male)
```

```
[1] 0.4292453
```

Aproximadamente el 43 % de los pacientes son hombres.

## Enlaces

- [Código de esta sesión](#)
- [Cookbook for R](#)
- [Página web de ggplot2](#), que contiene el [resumen \(chuleta\)](#) elaborado por RStudio.
- [Resumen sobre importación de datos a R \(chuleta\)](#) elaborado por RStudio.
- Web del libro [PostData](#) y los tutoriales asociados. Para esta sesión se recomienda el Capítulo 2.

## Bibliografía

Boehmke, B. C. (2016). *Data Wrangling with R* (p. 508). Springer.

<https://doi.org/10.1007/978-3-319-45599-0>

Matloff, N. S. (2011). *The art of R programming : tour of statistical software design* (p. 373). No Starch Press. <https://doi.org/10.1080/09332480.2012.685374>

Wickham, H., & Golemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc.