ESTADÍSTICA DESCRIPTIVA

Instituto Tecnológico

ÍNDICE

TU RETO EN ESTA UNIDAD

En muchas situaciones cotidianas nos encontramos con información en forma de medidas estadísticas. Veamos algún ejemplo e intenta responder a las preguntas:

- Leemos en la prensa: la tasa de paro se sitúa en el 18.5 % según la última encuesta de población activa. ¿Cómo se calcula exactamente esa tasa?
- El bebé come muy bien, está en el percentil 90 de peso. ¿Qué significa eso?
- La desigualdad en la distribución de la riqueza crece año tras año. ¿Cómo se mide cuantitativamente la desigualdad?
- El salario medio en España es de 24000 Euros anuales mientras que el salario mediano se sitúa en 16000 euros anuales. ¿Qué diferencia hay entre el salario medio y el salario mediano? ¿Por qué son tan distintos?

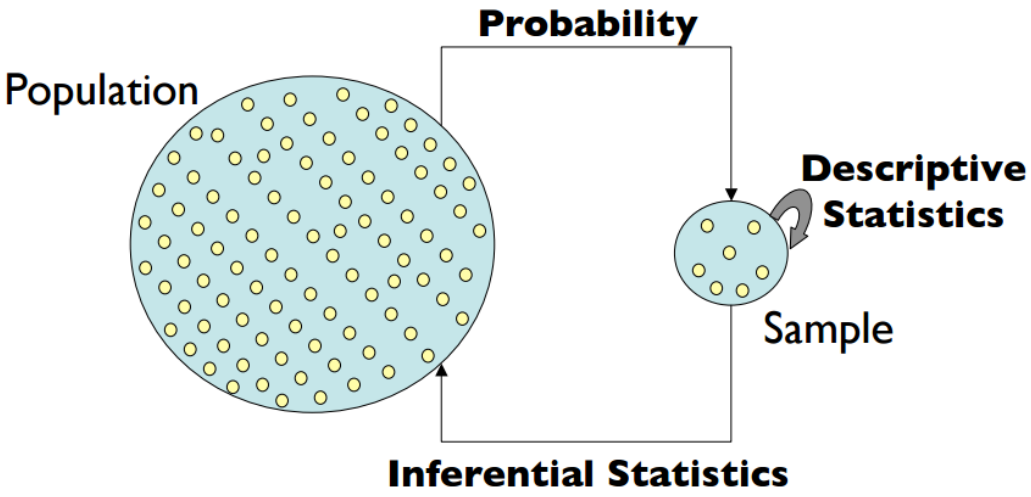
No te preocupes si no has sabido responder correctamente a las preguntas. Al final de la unidad serás capaz de responderlas y explicárselo a cualquiera que tenga dudas.

1. INTRODUCCIÓN

La definición de estadística según el diccionario de la Real Academia es:

1. Estudio de los datos cuantitativos de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de las sociedades humanas.
2. Rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades.

Las dos definiciones se corresponden con las dos ramas principales de la estadística: la estadística descriptiva y la estadística inferencial. En esta unidad nos vamos a centrar en la primera.



Áreas de trabajo de la estadística

1.1. ESTADÍSTICA DESCRIPTIVA Y ANÁLISIS EXPLORATORIO DE DATOS

La *estadística descriptiva* tiene por misión recolectar, presentar y caracterizar un conjunto de datos con el fin de describir apropiadamente las diversas características de ese conjunto.

Aprenderemos un conjunto de técnicas que permiten:

- Obtener, tabular, presentar, resumir y deducir propiedades del conjunto de datos en estudio.
- Representarlos gráficamente de forma adecuada para descubrir sus propiedades y relaciones.

El conjunto de técnicas para explorar los datos de forma sistemática es lo que los estadísticos llaman **Análisis Exploratorio de datos** (EDA por sus siglas en inglés). El análisis exploratorio no sigue unas reglas fijas, en realidad consiste en hacerse preguntas y buscar respuestas mediante visualizaciones, transformaciones, métricas y pequeños modelos. Normalmente estas respuestas nos hacen refinar las preguntas o hacernos nuevas preguntas.

El EDA es una parte importante del análisis de datos, incluso cuando las preguntas están claras desde el principio, ya que siempre va a ser necesario para explorar la calidad de los datos y proceder a su limpieza.

En resumen, el análisis exploratorio de datos es el primer paso, y el más crucial, a la hora de analizar unos datos. Antes de hacer inferencias/predicciones es esencial conocer y examinar nuestras variables para entre otras cosas:

- Encontrar errores.
- Encontrar valores anómalos.
- Ver patrones en los datos.
- Generar hipótesis.
- Encontrar violaciones a las suposiciones estadísticas.
- Y porque si no luego tendremos problemas.

1.2. ALGUNAS DEFINICIONES BÁSICAS

- **Población** es el conjunto de elementos, individuos o entes del que se pretende estudiar una serie de características o comportamientos.
- **Individuo** es cada uno de los elementos de la población y en Estadística el término puede referirse a cosas tan diversas como personas, provincias, empresas, edificios, etc. También se denomina observación.
- **Censo** es la información recogida para el estudio de una característica en todos los individuos de una población.
- **Muestra** es un subconjunto de elementos de la población, seleccionado para llevar a cabo un estudio estadístico sobre una o varias variables. Motivos tales como la economía, rapidez, calidad, imposibilidad o la observación destructiva hacen habitual el estudio de las características de una población a través de muestras.

1.3. TIPOS DE DATO

Una **variable** es una característica de los individuos de una población que nos interesa estudiar. Por ejemplo si estamos estudiando una población de personas: la altura, la edad, el color de piel, el peso, la presión arterial, o el grupo sanguíneo. Si nuestra población es el conjunto de hipotecas concedidas por un banco, las variables serán el capital del préstamo, el tipo de interés, el tipo de inmueble hipotecado, etc.

Dato es el conjunto de valores en bruto que toma variable.

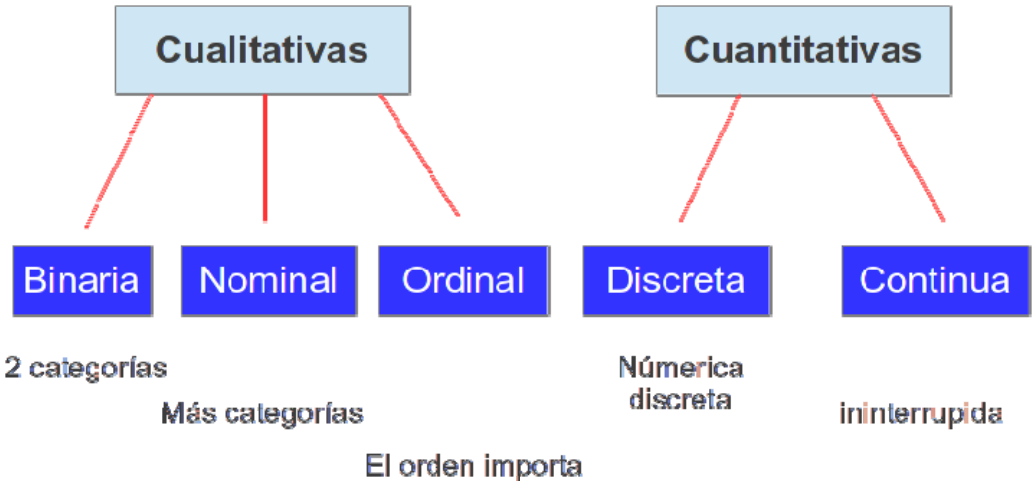
Podemos clasificar las variables según el tipo de datos que recogen como:

- **Cualitativas:** son aquellas en la que los valores posibles no expresan una cantidad sino la clasificación del dato en una categoría. Por ejemplo: color de ojos, el lugar de nacimiento, el género, el código postal (aunque sea un número) o una respuesta binaria a una pregunta tipo Si/No o Verdadero/Falso.

Las variables cualitativas pueden diferenciarse en:

- **Variable cualitativa ordinal:** la variable puede tomar distintos valores ordenados siguiendo una escala establecida, aunque no es necesario que el intervalo entre mediciones sea uniforme, por ejemplo: bajo, medio, alto.
- **Variable cualitativa nominal:** en esta variable los valores no pueden ser sometidos a un criterio de orden, como por ejemplo el color de pelo.

- **Cuantitativas:** aquellas cuyo resultado es un número. A su vez, las hay de dos tipos:
  - **Cuantitativas discretas:** cuando se toman valores aislados, normalmente resultados de conteos. Por ejemplo número de hermanos, número de empleados de una empresa o los puntos anotados por un jugador de baloncesto.
  - **Cuantitativas continuas:** cuando, entre dos valores cualesquiera, puede haber valores intermedios. Es decir, se toman todos los valores de un determinado intervalo. Por ejemplo: la altura de una persona, la temperatura en un instante temporal o el precio de un producto.



Las características a estudiar en un conjunto de datos pueden ser muy distintas y es importante conocer qué tipo de variables estamos manejando porque las técnicas estadísticas a utilizar dependerán de ello y por tanto la validez de las conclusiones que se extraigan de su análisis.

1.4. DATOS EJEMPLOS

Como siempre vamos a ilustrar todos los conceptos a partir de ejemplos basados en datos reales. En esta unidad vamos a utilizar datos de:

- **Encuesta Estructura Salarial** (Instituto Nacional de Estadística).
- **EPA:** Encuesta de Población Activa (Instituto Nacional de Estadística).
- **Datos antropométricos:** altura, peso, perímetro torácico, perímetro de muñeca, etc.

1.4.1. ENCUESTA DE POBLACIÓN ACTIVA (EPA)

Fuente: Instituto Nacional de Estadística.

http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=resultados&idp=1254735976595

En R podemos cargar los microdatos de la encuesta, es decir las respuestas detalladas de cada participante, usando el paquete MicroDatosEs.

```
# Si no tienes instalado el paquete instalalo
# install.packages('MicroDatosEs')

epa <- MicroDatosEs::epa2005("../Datasets/ine/EPAWEBT0416")
epa <- as.data.frame(epa)

#ccaa, prov, edad, sexo, nac, nforma, aoi, factorel
# Para simplificar, recodificamos la ocupacion a o="ocupado",
p="parado" e i="inactivo"
epa$AOI= factor(epa$AOI)
epa$ocupacion <- epa$AOI
lev=levels(epa$AOI)
levels(epa$ocupacion) = list(o =  lev[1:2], p = lev[3:4], i =
lev[5:7])
```

1.4.2. ENCUESTA DE ESTRUCTURA SALARIAL

Fuente: Instituto Nacional de Estadística.

http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177025&menu=resultados&idp=1254735976596

Cargamos en R los microdatos:

```
ees= MicroDatosEs::ees2010("../Datasets/ine/EES14_WEB")
ees=as.data.frame(ees)
```

1.4.3. ANTROPOMÉTRICOS

Proporcionados por la revista Journal of Statistics Education, Volume 11, Number 2 (July 2003).

http://ww2.amstat.org/publications/jse/v11n2/datasets.heinz.html

```
url="http://ww2.amstat.org/publications/jse/datasets/body.dat.txt
"
body <- read.table(url)
BodyMeasurements <-
c("Biacromial_diameter","Biiliac_diameter","Bitrochanteric_diamet
er","Chest_depth","Chest_diameter","Elbow_diameter","Wrist_diamet
er","Knee_diameter","Ankle_diameter","Shoulder_girth","Chest_girt
h","Waist_girth","Navel_girth","Hip_girth","Thigh_girth","Bicep_g
irth","Forearm_girth","Knee_girth","Calf_max_girth","Ankle_min_gi
rth","Wrist_min_girth","Age","Weight","Height","Gender")
names(body) <- BodyMeasurements
```

2. DISTRIBUCIÓN DE FRECUENCIAS

La forma más sencilla de visualizar y simplificar los datos sin apenas perder la información que contienen es mediante la distribución de frecuencias. Consiste en contar el número de veces que ocurre cada valor en unos datos. Debemos distinguir entre variables categóricas y numéricas.

2.1. VARIABLES CATEGÓRICAS

Frecuencias absolutas (conteo) y **frecuencias relativas** (porcentaje de observaciones respecto al total).

```
# Tabla de frecuencias de la variable aoi (tipo de ocupación) de
la EPA
#absolutas
table(epa$ocupacion)
```

```
##
##     o     p     i
##  1859 66145 66408
```

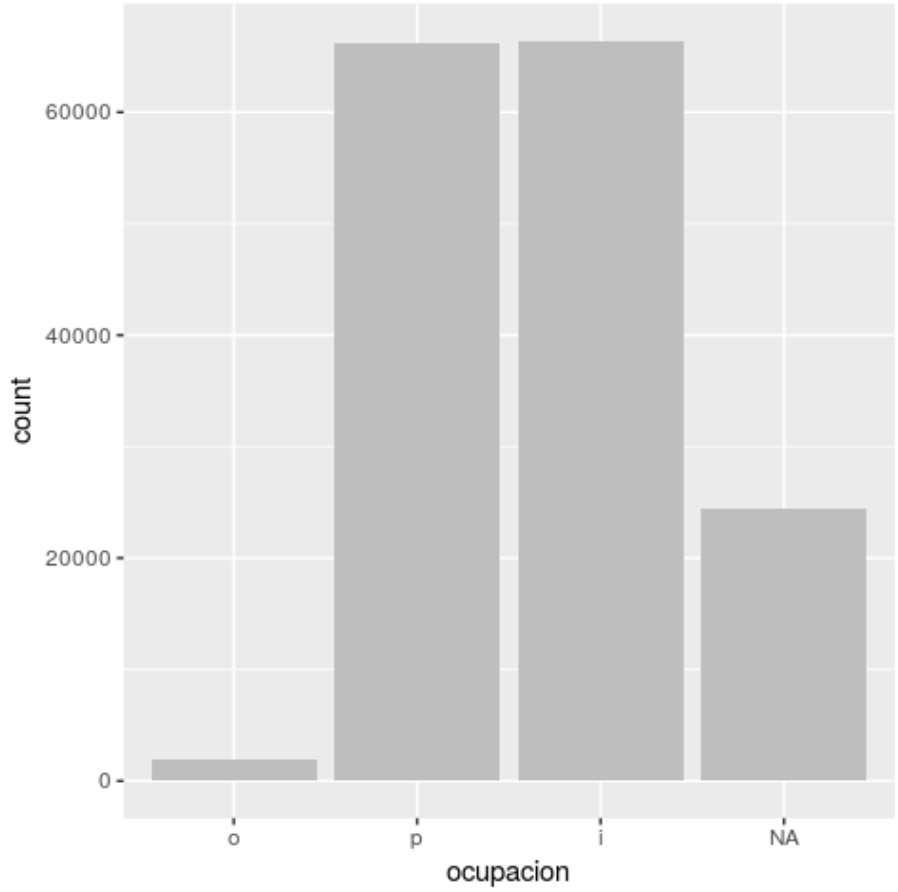
```
#relativas
prop.table(table(epa$ocupacion))
```

```
##
##           o           p           i
## 0.01383061 0.49210636 0.49406303
```

Se define la **Moda** como el valor de la variable con mayor frecuencia. En el ejemplo la categoría moda es: i.

De forma gráfica mediante ggplot, construimos un gráfico de frecuencias mediante geom_bar()

```
ggplot(epa) + geom_bar(aes(ocupacion),fill="grey")
```



Los NA que aparecen en la gráfica corresponden a encuestados menores de 16 años. El comando table no cuenta los NA a no ser que se lo digamos explícitamente.

```
table(epa$ocupacion,useNA='ifany')

##
##     o     p     i  <NA>
##  1859 66145 66408 24503
```

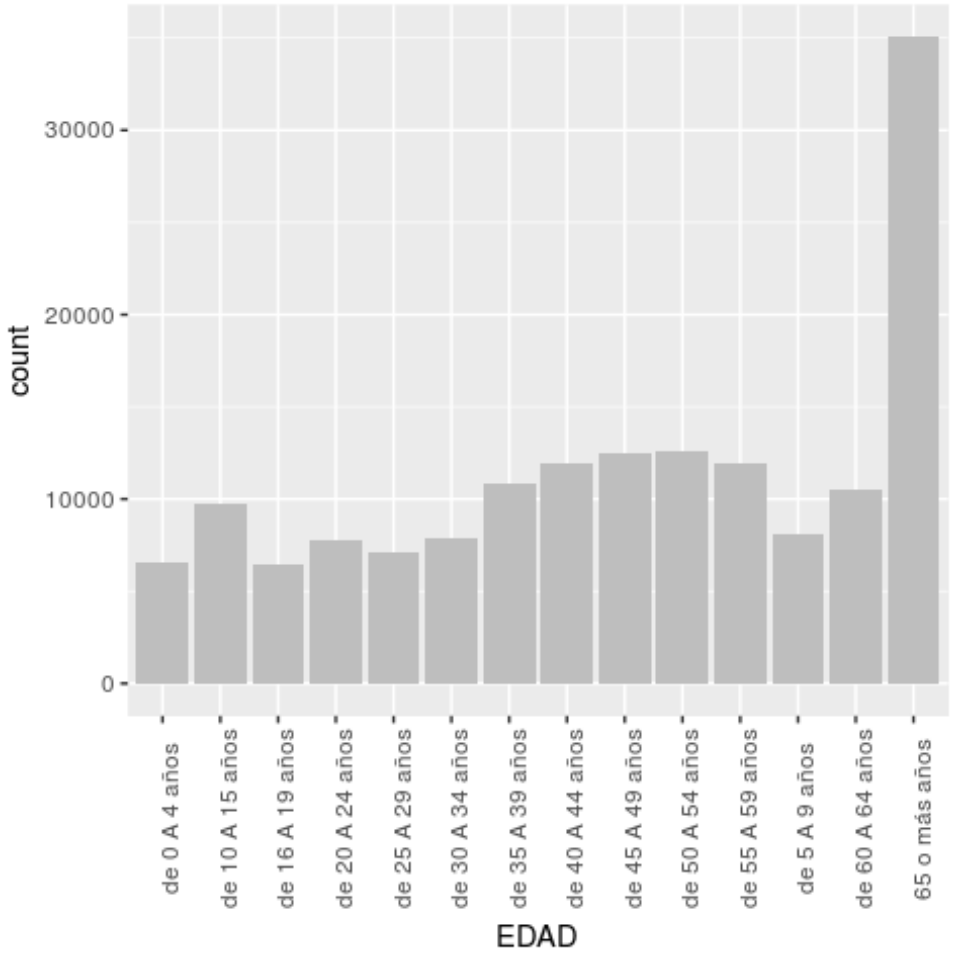
Un tipo especial de variable categórica son aquellas que tienen un orden, por ejemplo los grupos de edad.

```
table(epa$EDAD)
```

```
##
##   65 o más años   de 0 A 4 años de 10 A 15 años de 16 A 19
años de 20 A 24 años de 25 A 29 años
##           35054            6615            9779
6441            7733            7099
## de 30 A 34 años de 35 A 39 años de 40 A 44 años de 45 A 49
años de 50 A 54 años de 55 A 59 años
##            7855           10828           11920
12499           12567           11938
##   de 5 A 9 años de 60 A 64 años
##            8109           10478
```

En las representaciones gráficas debemos tener cuidado que las categorías queden ordenadas correctamente

```
# Hay que convertir la EDAD a factor y reordenar niveles
epa$EDAD=factor(epa$EDAD)
epa$EDAD=factor(epa$EDAD,levels=levels(epa$EDAD)[c(2:14,1)])
ggplot(epa) + geom_bar(aes(EDAD),fill="grey") +
  theme(axis.text.x=element_text(angle = 90))
```



2.2. VARIABLES CATEGÓRICAS MÚLTIPLES

Podemos estudiar las distribuciones de frecuencias de múltiples variables mediante tablas de contingencia en varias dimensiones.

```
# simplemente pasamos varios argumentos a la función table
t<- table(epa$ocupacion,epa$SEXO); t

##
##     Mujer Varón
##   o  1252   607
##   p 38553 27592
##   i 30798 35610
```

Las frecuencias marginales, es decir las frecuencias de cada variable considerada de forma individual se obtienen mediante.

```
# Por filas (variable 1)
margin.table(t,1)

##
##     o     p     i
##  1859 66145 66408

# Por columnas (variable 2)
margin.table(t,2)

##
## Mujer Varón
## 70603 63809
```

Para obtener las frecuencias relativas, tanto absolutas como marginales.

```
# Frecuencias relativas
prop.table(t)

##
##           Mujer       Varón
##   o 0.009314645 0.004515966
##   p 0.286827069 0.205279291
##   i 0.229131328 0.264931703

# Frecuencias relativas marginales por filas
prop.table(t,1)

##
##         Mujer     Varón
##   o 0.6734804 0.3265196
##   p 0.5828558 0.4171442
##   i 0.4637694 0.5362306

# Frecuencias relativas marginales por columnas
prop.table(t,2)

##
##           Mujer       Varón
##   o 0.017732958 0.009512765
##   p 0.546053284 0.432415490
##   i 0.436213759 0.558071745
```

Veamos ahora un ejemplo con 3 variables.

```
t<- table(epa$ocupacion,epa$SEXO,epa$NFORMA);
t

## , ,  = Analfabetos
##
##
##     Mujer Varón
##   o     8     7
##   p  1583   628
##   i    86    86
##
## , ,  = Educación primaria
##
##
##     Mujer Varón
##   o   176   100
##   p  9969  6697
##   i  1552  2478
##
## , ,  = Educación primaria incompleta
##
##
##     Mujer Varón
##   o    54    30
##   p  5027  3216
##   i   345   448
##
## , ,  = Educación superior
##
##
##     Mujer Varón
##   o   222    97
##   p  5208  4313
##   i 13794 12149
##
## , ,  = Primera etapa de educación secundaria
##
##
##     Mujer Varón
##   o   545   264
##   p 10449  7838
```

```
##   i  7932 12352
##
## , ,  = Segunda etapa de educación secundaria, orientación ge-
neral
##
##
##     Mujer Varón
##   o   132    76
##   p  4417  3691
##   i  3894  4672
##
## , ,  = Segunda etapa de educación secundaria, orientación pro-
fesional
##
##
##     Mujer Varón
##   o   115    33
##   p  1900  1209
##   i  3195  3425
```

2.3. GRÁFICOS DE TABLAS MULTIDIMENSIONALES

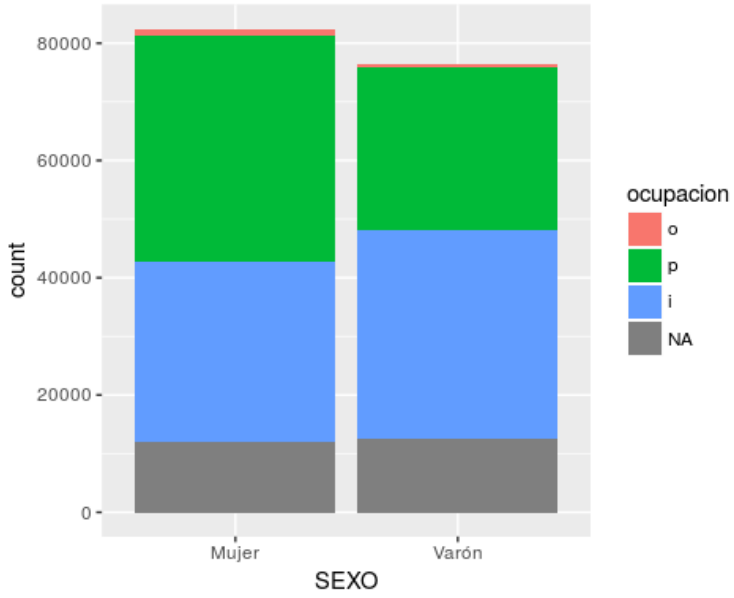
Veamos ahora algún ejemplo de representación gráfica de las tablas de frecuencias multidimensionales.

```
t=table(epa$ocupacion,epa$SEXO);
```

Suelen representarse con diagramas de barras, de diferentes tipos:

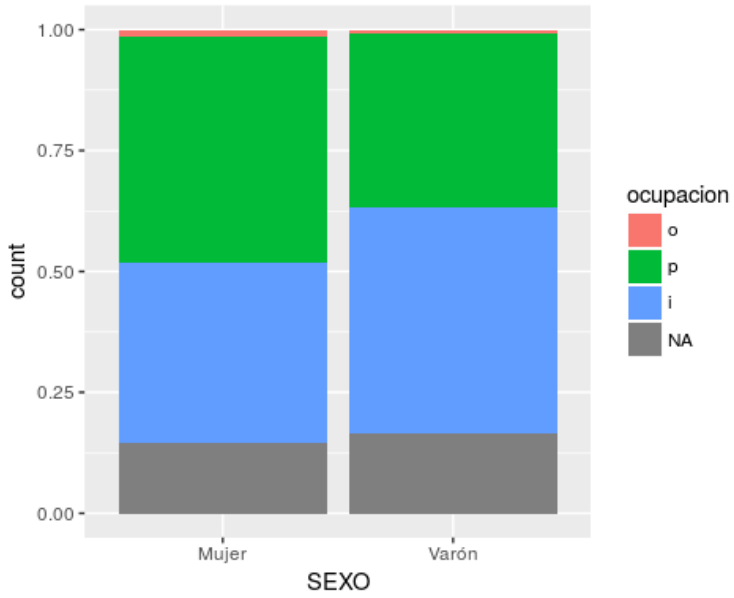
- **Barras apiladas** (position="stack): la altura de cada barra principal es proporcional a cada una de las frecuencias de la variable 1. Dentro de cada barra los segmentos de cada color son proporcionales a las frecuencias conjuntas de la variable 1 y 2.

```
ggplot(epa,aes(SEXO,fill=ocupacion)) + geom_bar(position="stack")
```



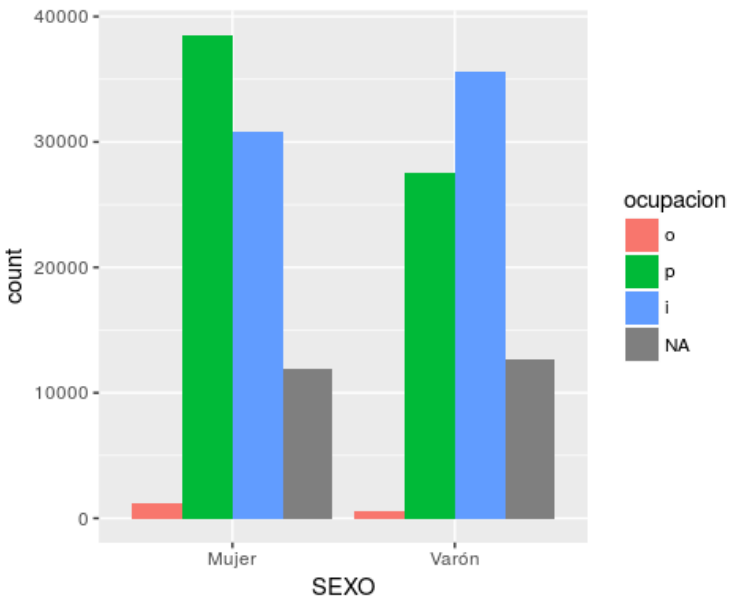
- **Barras apiladas normalizadas** (position="fill"): puede observarse la distribución de la variable 2 relativa a la variable 1, pero no se ve la distribución global de la variable 1.

```
ggplot(epa,aes(SEXO,fill=ocupacion)) + geom_bar(position="fill")
```



- **Barras paralelas** (position="dodge"): la atura de cada barra indica la frecuencia conjunta. Esta es la mejor representación para observar las frecuencias conjuntas, pero es difícil visualizar las marginales.

```
ggplot(epa,aes(SEXO,fill=ocupacion)) + geom_bar(position="dodge")
```



2.4. VARIABLES NUMÉRICAS CONTINUAS-HISTOGRAMAS

Para estudiar la distribución de frecuencias en variables continuas dividimos en intervalos y contamos las frecuencias por intervalos. La división en intervalos no tiene porqué ser regular, es decir podemos dividir en intervalos de distinta longitud La representación gráfica asociada de denomina histograma. En el caso de intervalos no regulares debe representarse la densidad de frecuencia, que se calcula como la frecuencia relativa dividida por la longitud del intervalo.

Consideremos una variable x que toma 50 valores.

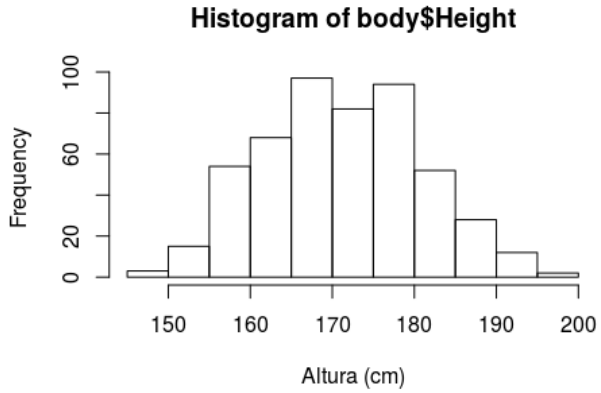
Sea x= 32.8, 39.5, 17.3, 9.4, 10.5, 38.9, 8.5, 39.4, 31.5, 25.2, 19.1, 10.1, 36.2, 30.3, 15.2, 30.3, 36.2, 32.8, 36.4, 33.3, 16.8, 36.4, 21.9, 33.2, 36.4, 38.9, 39.4, 10.5, 33.3, 38.9, 38, 26.2, 38.8, 36.2, 14.9, 6.3, 33.2, 36.4, 33.3, 8.7, 30.3, 38, 38.2, 38, 9.4, 6.3, 21.9, 29.5, 31.5, 38.

Para calcular su histograma manualmente deberíamos rellenar la siguiente tabla:

Intervalo	Frecuencia absoluta	Frecuencia relativa	Longitud intervalo	Densidad absoluta	Densidad relativa
0-10	5	5/50 = 0.1	10	5/10=0.5	0.1/10=0.01
10-20	10	10/50 = 0.2	10	10/10=1	0.2/10=0.02
...	n_i	$f_i = n_i/N$	l_i	$D_i = n_i/l_i$	$d_i = f_i/l_i$
Total	N=50	1	-	-	-

Veamos algún ejemplo con datos reales, de cómo calcular y representar un histograma con R.

```
h<- hist(body$Height,xlab="Altura (cm)")
```



La función ***hist*** calcula más que un gráfico.

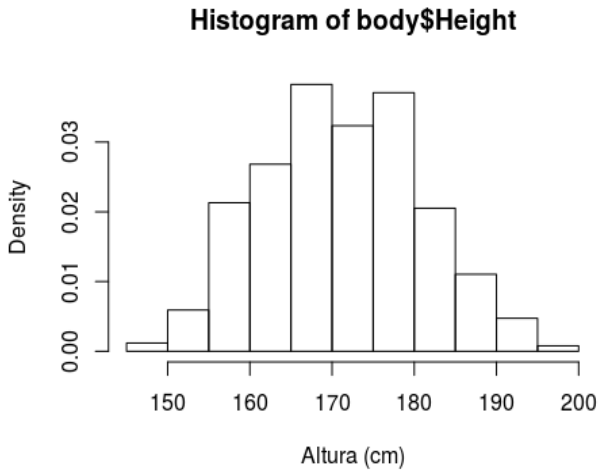
```
h

## $breaks
##  [1] 145 150 155 160 165 170 175 180 185 190 195 200
##
## $counts
##  [1]  3 15 54 68 97 82 94 52 28 12  2
##
## $density
```

```
##  [1] 0.0011834320 0.0059171598 0.0213017751 0.0268244576
0.0382642998 0.0323471400 0.0370808679
##  [8] 0.0205128205 0.0110453649 0.0047337278 0.0007889546
##
## $mids
##  [1] 147.5 152.5 157.5 162.5 167.5 172.5 177.5 182.5 187.5
192.5 197.5
##
## $xname
## [1] "body$Height"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

2.5. DENSIDAD DE FRECUENCIA

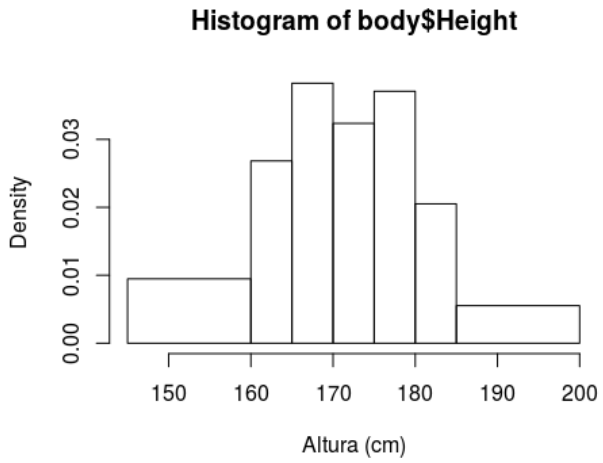
En el histograma se puede mostrar la densidad de frecuencia en lugar de la frecuencia absoluta. La densidad de frecuencia en un histograma se define como la frecuencia relativa dividida por la longitud del intervalo.

```
h <- hist(body$Height,probability = TRUE, xlab="Altura (cm)")
```



Si los intervalos no son regulares, hist mostrará por defecto la densidad de frecuencia.

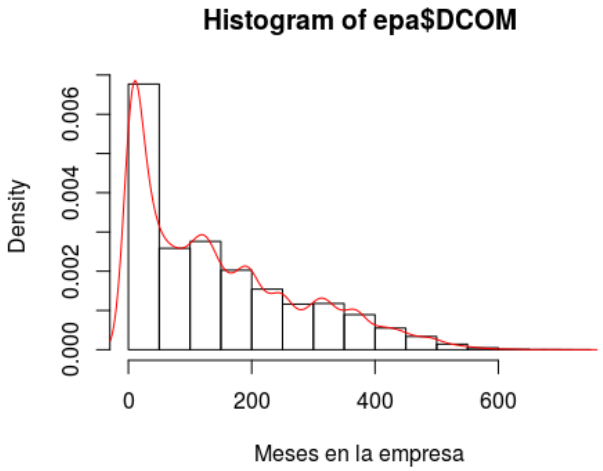
```
h <- hist(body$Height,breaks=c(145,160,165,170,175,180,185,200),
xlab="Altura (cm)")
```



2.5.1. AJUSTE DE DENSIDAD

Los histogramas a veces son ruidosos y su forma depende de la elección de los intervalos. Muchas veces útil hacer un ajuste suave de la densidad de la distribución de una variable. En R lo hacemos con la función *density*.

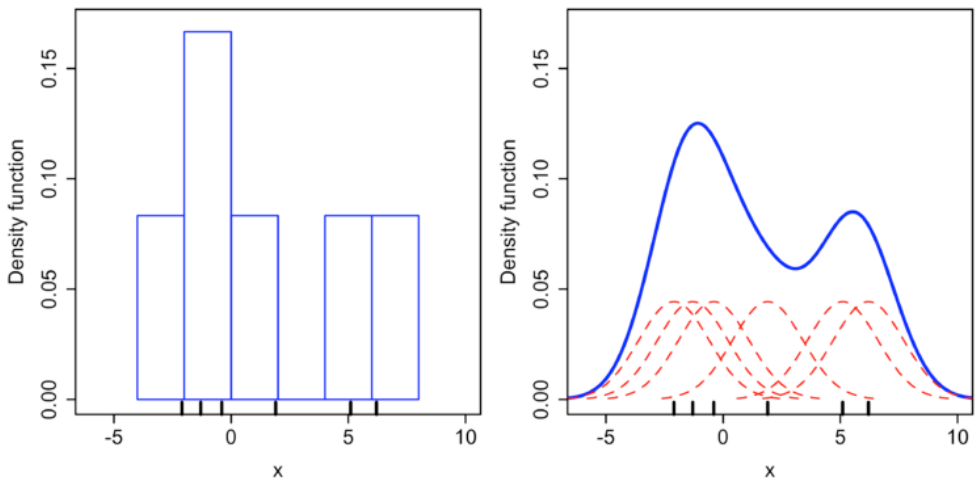
```
h <- hist(epa$DCOM,probability = TRUE, xlab="Meses en la empre-
sa")
d <- density(epa$DCOM,na.rm = TRUE)
lines(d,col="red")
```



Realiza el ajuste mediante el método de estimación de densidad Kernel Consiste en un ajuste de la función de densidad mediante la superposición de unas funciones de un tipo dado (suelen ser gaussianas, rectangulares, cosenos, etc.), que se denominan *funciones kernel*.

$$\hat{f}(x) = \frac{1}{n b_w} \sum_{i=1}^{n} K\left(\frac{x - x_i}{b_w}\right)$$

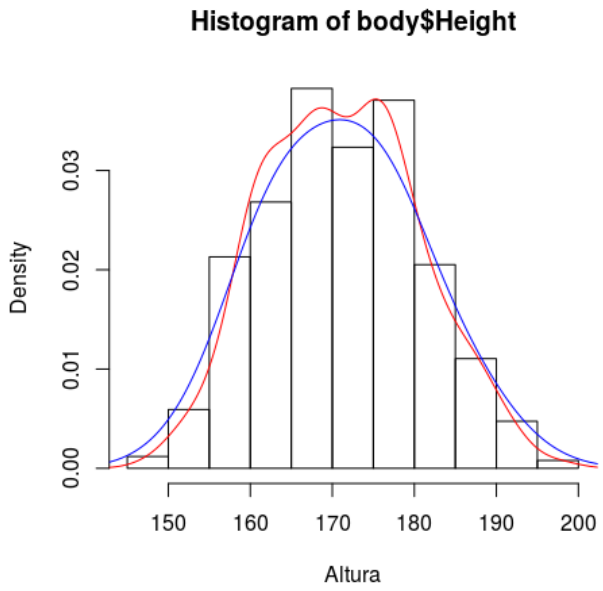
donde el parámetro $\mathbf{b_w}$ se denomina ancho de banda y representa la suavidad del ajuste. Muchas veces es necesario elegir el parámetro $\mathbf{b_w}$ manualmente.



Estimación de densidad por funciones *kernel*

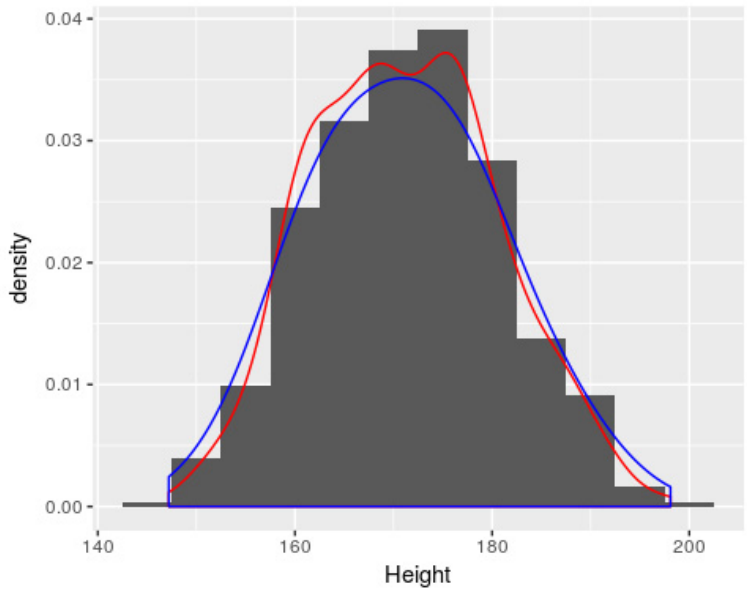
En nuestro caso de ejemplo, veamos los ajustes de densidad obtenidos usando diferentes anchos de banda.

```
h <- hist(body$Height,probability = TRUE, xlab="Altura")
d1 <- density(body$Height,bw= 2.5,na.rm = TRUE)
d2 <- density(body$Height,bw= 5,na.rm = TRUE)
lines(d1,col="red")
lines(d2,col="blue")
```



Pueden realizarse los mismos gráficos usando ggplot.

```
ggplot(body,aes(x=Height)) +
  geom_histogram(aes(y=..density..),binwidth=5) +
  geom_density(color="red",bw=2.5) +
  geom_density(color="blue",bw=5)
```



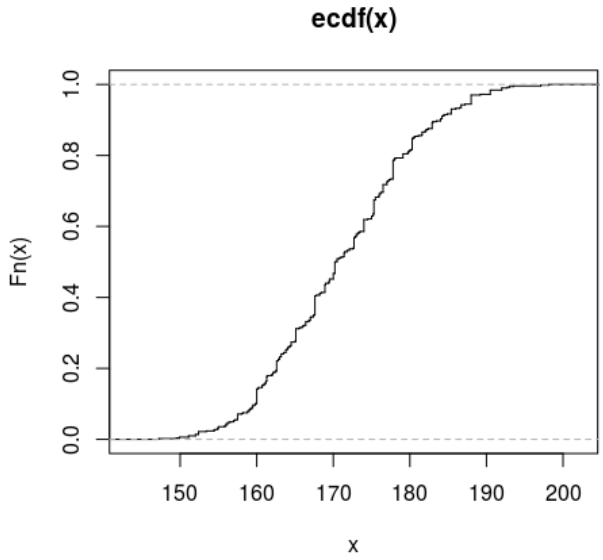
2.6. DISTRIBUCIÓN ACUMULADA

Cuando trabajamos con variables cuantitativas y con cualitativas ordinales es interesante conocer las distribuciones de frecuencias acumuladas.

La distribución de frecuencias acumulada, tiene sentido para variables numéricas y para categóricas ordenadas. Cuenta el número de eventos con valor igual o menor que cada posible valor de la variable.

En R, la función `ecdf` calcula una función de distribución acumulada.

```
x=body$Height
fac=ecdf(x)
# ojo, ecdf devuelve una función
plot(fac, verticals = TRUE, do.points = FALSE)
```


ecdf(x)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
150
160
170
180
190
200
x

3. MEDIDAS NUMÉRICAS

Aunque la distribución de frecuencias nos da mucha información sobre nuestros datos, es deseable poder caracterizar los aspectos principales de una distribución de frecuencias mediante unas medidas numéricas que nos permitan resumir las características principales.

De esta manera podemos comparar diferentes conjuntos de datos o distribuciones mediante el análisis de las medidas numéricas adecuadas. Podemos distinguir diferentes tipos de medidas que nos serán necesarias en función del fenómeno que queremos analizar.:

- Centralidad.
- Posición.
- Dispersión.
- Forma.
- Concentración.

Existen una serie de preguntas genéricas que nos debemos hacer para toda medida de síntesis

- ¿Intervienen todos los datos?
- ¿Con qué tipo de datos se puede calcular?
- ¿Es única?
- ¿Es robusta?
- ¿Qué representatividad tiene?
- ¿Cómo se interpreta?
- ¿Cómo se comporta al transformar los datos originales?

4. MEDIDAS DE CENTRALIDAD/POSICIÓN

Son las medidas que buscan situar dónde se encuentra situada la distribución de frecuencias, bien sean sus valores más representativos o centrales, bien sean sus zonas intermedias o sus extremas.

4.1. MEDIA

Dado un conjunto de datos $\{x_1, x_2, \ldots, x_n\}$, la media se define como:

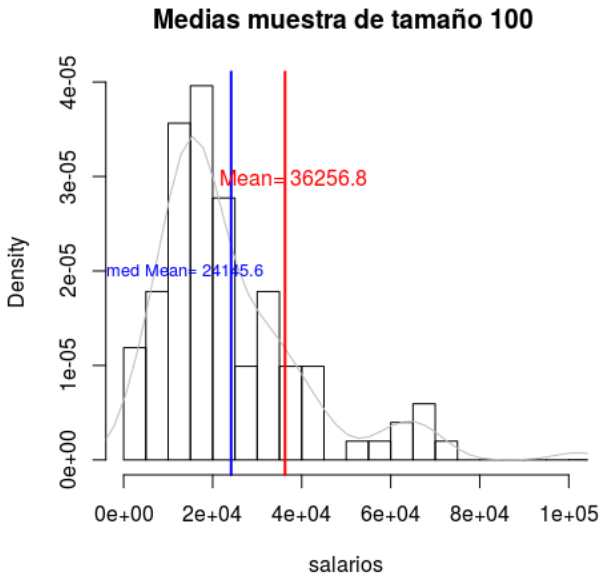
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

El mayor problema de la media es que es una medida muy sensible a "outliers" o valores anómalos. Eso quiere decir que la media no es una medida robusta.

Por ejemplo si miramos los salarios de la Encuesta de Estructura Salarial:

```
mean(ees$SALBRUTO)

## [1] 24782.68

summary(ees$SALBRUTO)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##      23.8   13150.0   20450.0   24780.0   31020.0 1247000.0
```

Se observa que el valor máximo es muy superior a la media. Si tomamos una muestra que contiene el máximo o no, obtenemos valores muy distintos para la media muestral.



4.2. OTROS TIPOS DE MEDIA

Media Truncada (trimmed mean)

Eliminamos una fracción α por arriba y abajo de los datos ordenados. Esto la hace poco sensible a los valores anómalos.

```
mean(salarios,trim = 0.01)

## [1] 24359.6

mean(salarios,trim = 0.1)

## [1] 21725.28
```

Media Ponderada

Los distintos elementos tienen distinta importancia.

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i \, x_i}{\sum_{i=1}^{n} w_i}$$

```
# La variable FACTOTAL se denomina factor de elevación y se usa
para corregir los errores de muestreo de la encuesta
weighted.mean(ees$SALBRUTO,ees$FACTOTAL)

## [1] 21309.06
```

Geométrica

Relevante cuando el conjunto de números es interpretable por su producto. Por ejemplo tasas de crecimiento ($x_{t+1} = t_1 t_2 \ldots t_n x_0$).

$$\overline{t} = \left(\prod_{i=1}^{n} t_i \right)^{\frac{1}{n}}$$

Mediana

La mediana de una serie de datos $\{x_1, x_2, \ldots, x_n\}$ es el valor tal que la mitad de las x's son mayores que él y la otra mitad son menores:

- **Si n es impar,** entonces la mediana es el valor central $x_{(n+1)/2}$ de la serie ordenada.
- **Si n es par,** la mediana es el promedio de los dos valores centrales.

La mediana es muy poco sensible los valores anómalos, en este sentido es una medida robusta. Para comparar variables que contienen muchos valores anómalos, la mediana es más útil que la media.

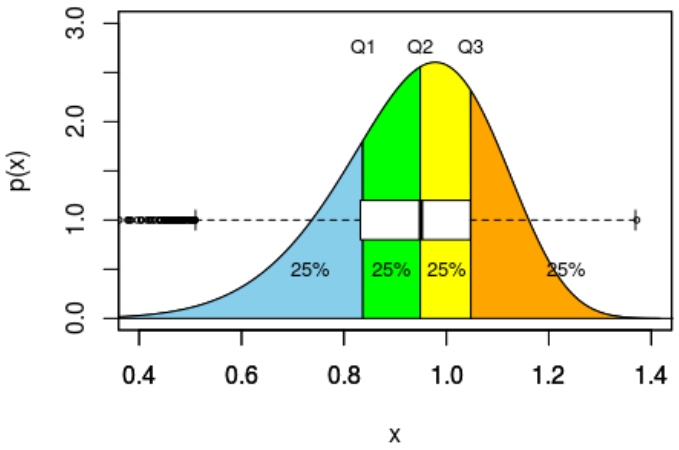
En R se calcula mediante la función *median*.

```
x=rnorm(100)
# Mediana según definición
x=sort(x)
mean(x[50:51])

## [1] -0.1176227

# Usando la función apropiada de R
median(x)

## [1] -0.1176227
```

4.3. CUANTILES

- El cuantil p, que denotamos q_p, de una variable, es el valor tal que una fracción p de observaciones se encuentran por debajo y la fracción (1 − p) se encuentran por encima de este valor.
- Los percentiles, p_k se definen de forma equivalente expresados en porcentaje en lugar de fracciones
- Los cuartiles Q_1, Q_2 y Q_3 corresponden a los cuantiles 0.25, 0.5 y 0.75

Veamos gráficamente para una distribución de frecuencias dada, la relación de los cuartiles con las áreas encerradas por la densidad de frecuencia.



Para calcular cuantiles en R lo hacemos mediante la función `quantile`

```
x=ees$SALBRUTO
quantile(x,probs = 0.6)
```

```
##      60%
## 23879.73
```

```
# Si no especificamos probs, por defecto muestra el
0,0.25,0.5,0.75 y 1
quantile(x)
```

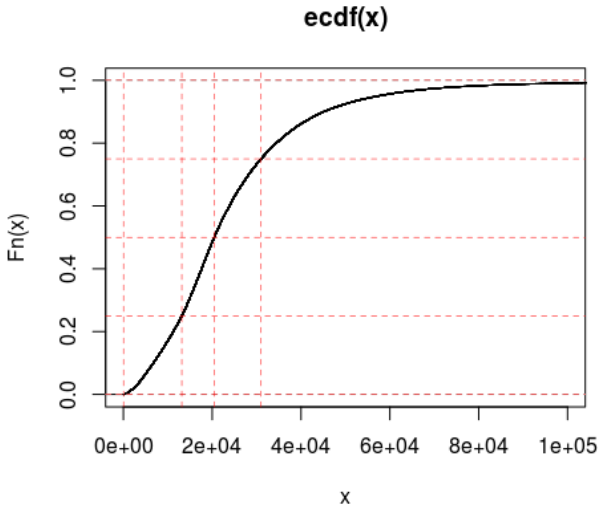
```
##          0%         25%         50%         75%        100%
##      23.80   13151.56   20446.32   31020.01 1247382.83
```

```
# Podemos especificar un conjunto de probs mediante seq
# Calculo los cuantiles 0,0.1,0.2, ..., 0.9,1
quantile(x,probs=seq(0,1,.1))
```

```
##          0%         10%         20%         30%         40%
50%         60%         70%         80%
##      23.80    6743.15   11244.72   14755.90   17630.97
20446.32   23879.73   28196.40   34413.18
##         90%        100%
##   45102.78 1247382.83
```

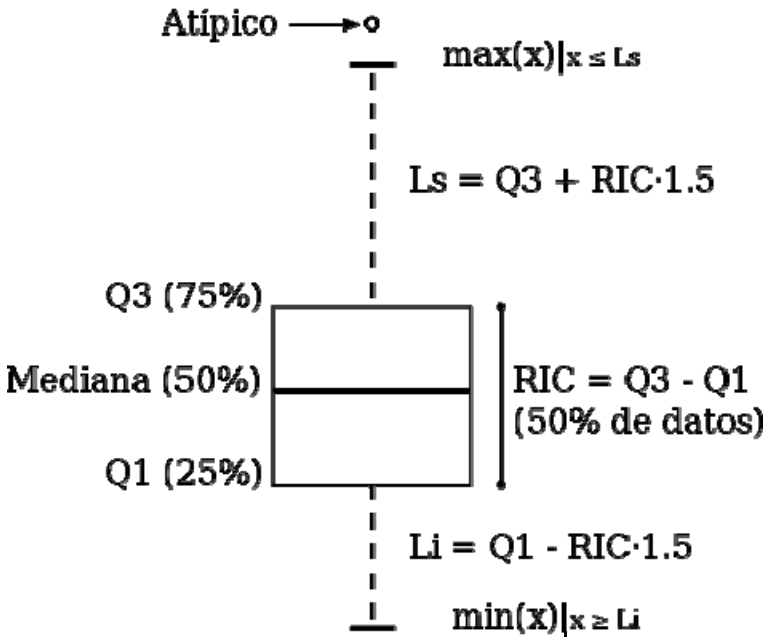
Los cuantiles son en realidad los inversos de la distribución acumulada de frecuencias.

```
x=ees$SALBRUTO
plot(ecdf(x),xlim=c(0,1e5))
abline(v=quantile(x,seq(0,1,.25)),lty=2,lwd=0.5,col=2)
abline(h=seq(0,1,.25),lty=2,lwd=0.5,col=2)
```



4.3.1. BOXPLOTS

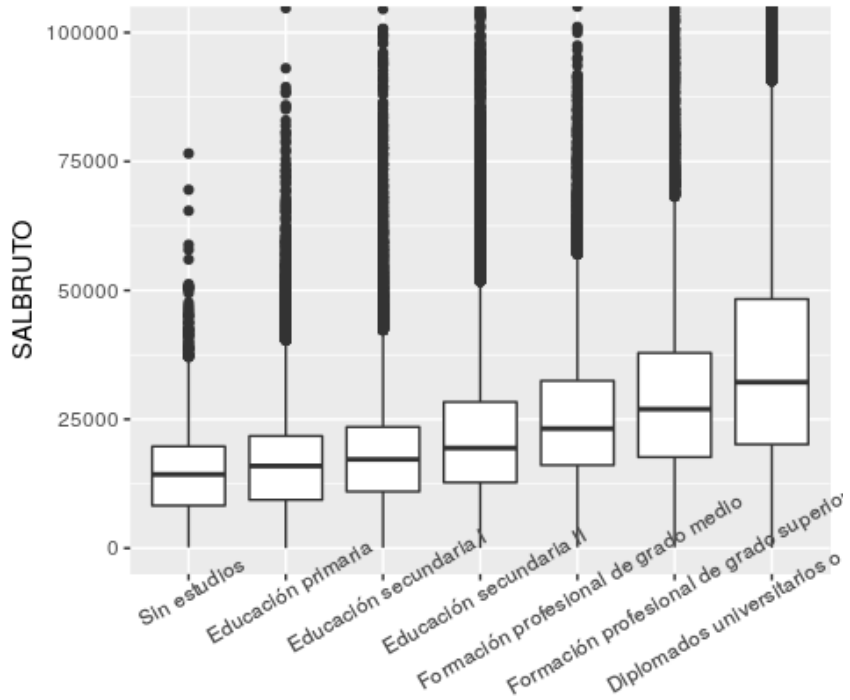
A partir de los cuantiles se elaboran los boxplots o diagramas de caja, que son muy útiles para comparar las distribuciones de datos pertenecientes a distintas muestras o categorías. Se trata de representar la distribución de datos mediante una caja y unos segmentos, cuyos límites se corresponden con medidas de posición tal y como se muestra en la siguiente figura.



Significado de un *boxplot*

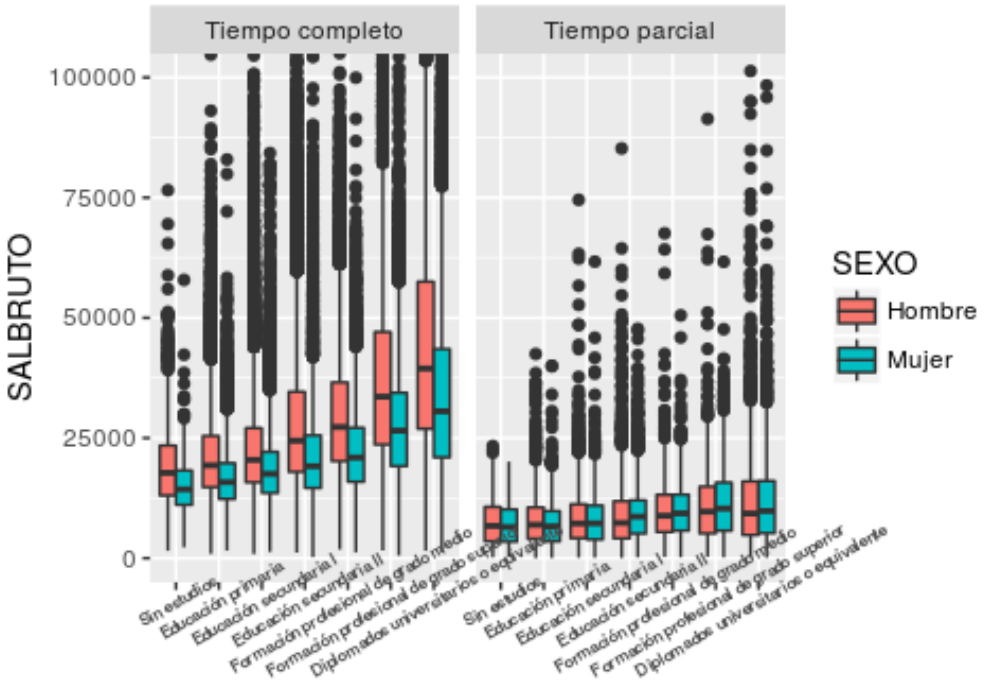
Veamos un ejemplo con los datos de la encuesta de estructura salarial, donde mostramos los salarios de los encuestados divididos según su nivel de estudios.

```
# Reordenamos los niveles por orden creciente de salario mediante
reorder(ESTU,SALBRUTO)
ggplot(ees) + geom_boxplot(aes(reorder(ESTU,SALBRUTO),SALBRUTO))
+ theme(axis.text.x = element_text(angle=30)) +
  coord_cartesian(ylim=c(0,1e5)) + xlab("")
```



Podemos hacer el mismo gráfico distinguiendo además por sexo y tipo de jornada, gracias al coloreado y los facets.

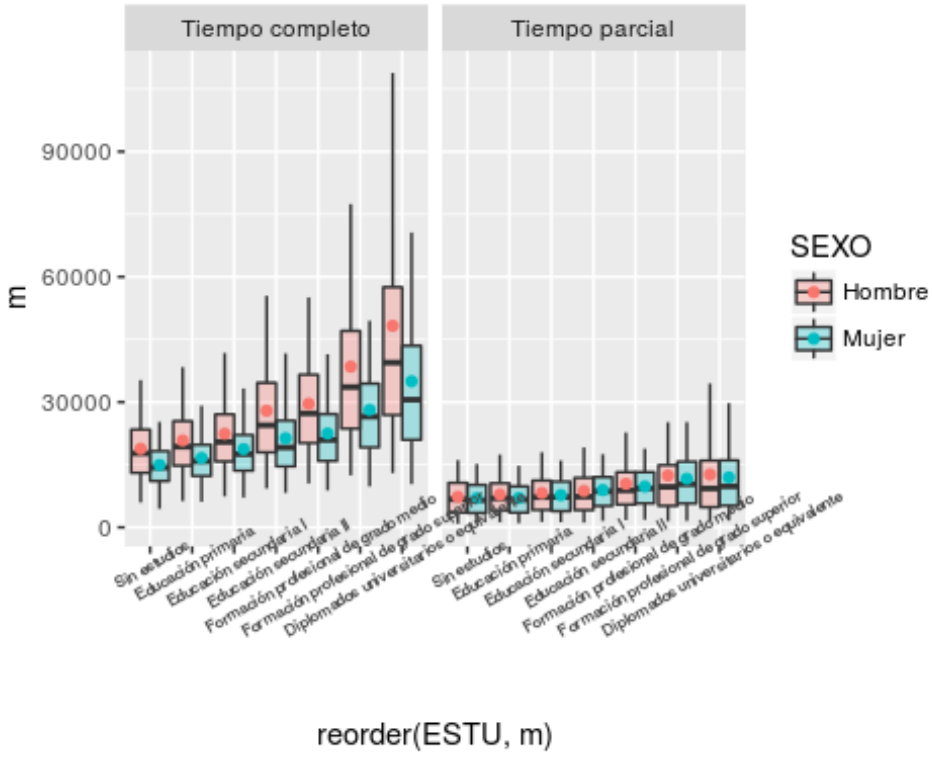
```
ggplot(ees) +
geom_boxplot(aes(reorder(ESTU,SALBRUTO),SALBRUTO,fill=SEXO)) +
facet_wrap(~TIPOJOR) +
  theme(axis.text.x = element_text(size=6,angle=30)) + xlab("") +
  coord_cartesian(ylim=c(0,1e5))
```



Podemos realizar también un boxplot a medida, es decir donde los límites de las cajas y de los segmentos signifiquen cosas distintas a lo habitual. Por ejemplo hagamos un boxplot con cuantiles 0.05, 0.25, 0.5, 0.75 y 0.95. Además pintamos las medias mediante puntos. Con el paquete ggplot es posible, haciendo stat="identity" dentro de geom_boxplot y especificando manualmente los límites de cajas y segmentos: ymin, lower, middle e ymax.

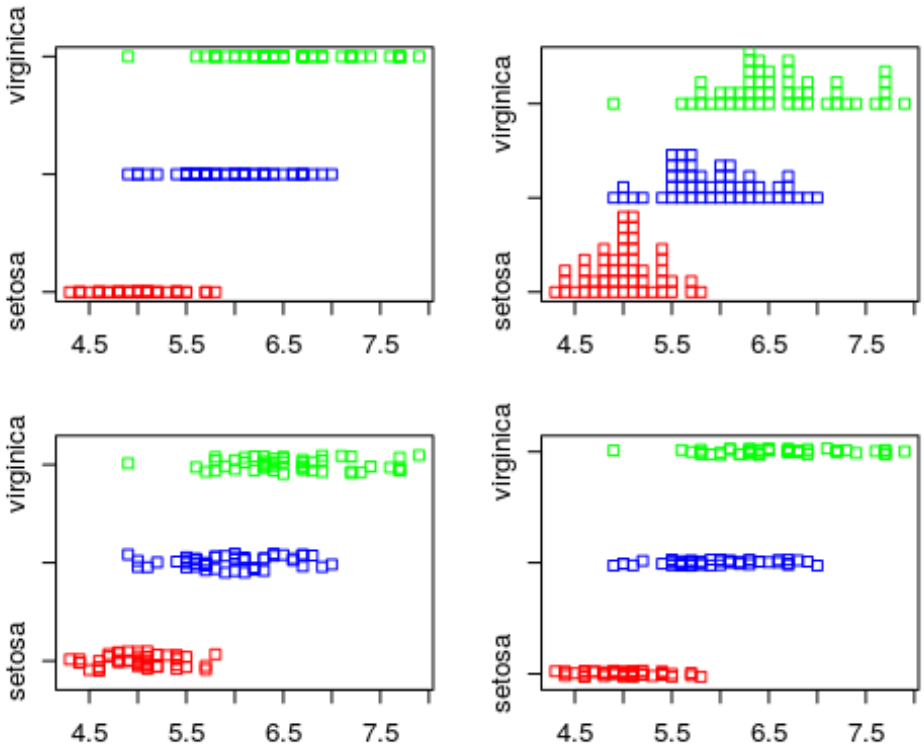
```
# En primer lugar escribimos una función que me calcula las mé-
tricas
seven_nums <-  function(x,p=c(0,0.05,0.25,0.5,0.75,0.95,1)){
  q=quantile(x,p,na.rm=TRUE)
  res
=data.frame(num=length(x),m=mean(x,na.rm=TRUE),sd=sd(x,na.rm=TRUE
),t(q))
  names(res)[-(1:3)] =paste0("q",round(p*100))
  res
}
tmp <- ees %>% group_by(ESTU,SEXO,TIPOJOR) %>% do({
  seven_nums(.$SALBRUTO)
})

ggplot(tmp) +
```

```
geom_boxplot(aes(x=reorder(ESTU,m),ymin=q5,lower=q25,middle=q50,
                 up-
per=q75,ymax=q95,fill=SEXO),alpha=0.3,stat="identity") +
  geom_point(aes(reorder(ESTU,m),m,color=SEXO), posi-
tion=position_dodge(width=.9)) +
  facet_wrap(~ TIPOJOR) +
  theme(axis.text.x = element_text(size=6,angle=30))
```



4.4. STRIPCHART

Cuando tenemos pocos datos, a veces es más útil el stripchart que muestra la distribución de todos los puntos.



5. MEDIDAS DE DISPERSIÓN

Es importante completar la información proporcionada por las medidas de posición con medidas de dispersión que midan el grado de variabilidad de las variables.

5.1. VARIANZA / DESVIACIÓN TÍPICA

La medida de dispersión más habitual es la varianza. Debemos de distinguir entre

Varianza poblacional:

$$\sigma^2 = \frac{1}{n}\sum (x_i - \mu)^2$$

Varianza muestral:

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

dependiendo de si se aplica a una población con la media conocida o a una muestra de la que conocemos una estimación de la media poblacional, que es la media muestral.

En las fórmulas $\bar{x}$ es la media de la muestra y μ la media de la población.

La **Desviación típica** es simplemente la raíz cuadrada de la varianza, y tiene las mismas unidades que los datos.

$$\sigma = \sqrt{\sigma^2}$$

La varianza, al igual que la media, está muy influenciada por los valores anómalos.

Autotexto: ¿Por qué n-1?

Una pregunta muy habitual es porque en la varianza muestral dividimos por n-1 y no por n. Respondamos de un modo práctico, a partir de los resultados de una simulación. Este es un enfoque que vamos a utilizar habitualmente a lo largo del curso. No vamos a hacer demostraciones matemáticas complicadas, pero vamos a hacer comprobaciones mediante experimentos con R. Este enfoque se denomina por muchos autores "Estadística Moderna".

Genero 10000 muestras de una población normal con media 0 y $\sigma = 1$.

Calculo $\sum_i (x_i - \overline{x})^2$.

```
varn <- function(x){sum((x-mean(x))^2)}
res=NULL
# Genero 10000 muestras de tamaño 10 y calculo la suma de cuadra-
dos
n=10
s<- replicate(10000,varn(rnorm(n)))
mean(s)
```

```
## [1] 9.052651
```

```
# Para una muestra de tamaño 100
n=100
s<- replicate(10000,varn(rnorm(n)))
mean(s)
```

```
## [1] 98.9566
```

Se observa que el valor medio de $\sum_i (x_i - \overline{x})^2$ es aproximadamente 9 o 99, en lugar de 10 o 100. Por eso en la definición de la varianza muestral se divide por n-1.

La diferencia radica en que en el caso de la población la media real es conocida y para la muestra tenemos una estimación de esta calculada a partir de los datos.

5.2. COEFICIENTE DE VARIACIÓN

Se define como

$$CV = \frac{\sigma}{|\overline{x}|}$$

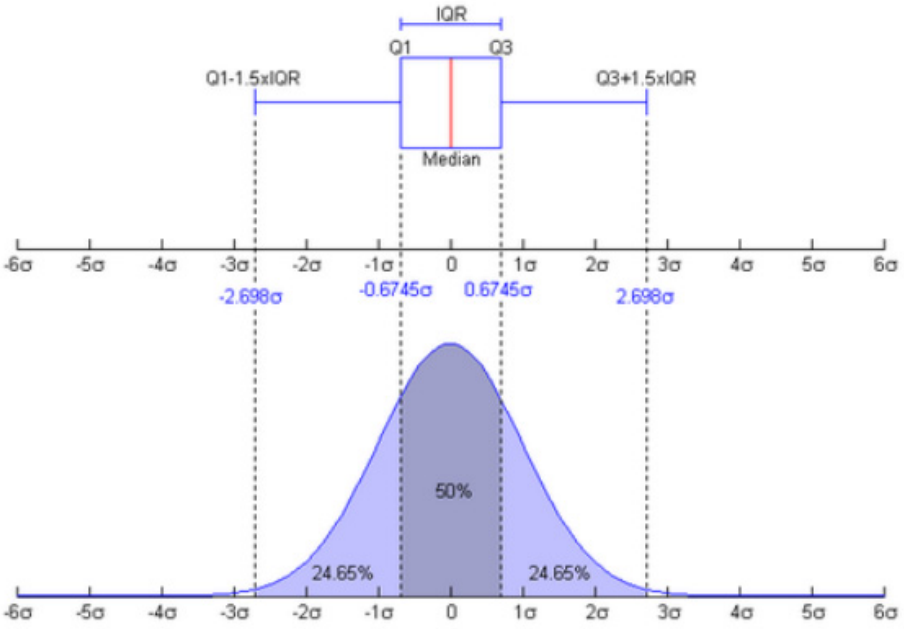
Simplemente escala la desviación típica a la magnitud de $\overline{x}$. Es útil para comparar la dispersión en variables con diferentes órdenes de magnitud, por ejemplo porque han sido expresadas en unidades distintas.

5.3. INTERVALO INTERCUARTÍLICO

$$IQR = Q_3 - Q_1 = q_{0.75} - q_{0.25} = P_{75} - P_{25}$$

Es una medida robusta, a la que afectan poco los valores anómalos.

En él se basa el criterio de detección de valores anómalos que usan los boxplots: x_i es anómalamente grande si $x_i \geq q_{0.75} + 1.5IQR$ y anómalamente pequeño si $x_i \leq q_{0.25} - 1.5IQR$. Este criterio se basa en las propiedades de la distribución normal y según él si los valores están normalmente distribuidos solo el 0.70 % serían clasificados como anómalos y se usa el IQR para que los propios valores anómalos no influyan en la medida de la dispersión.



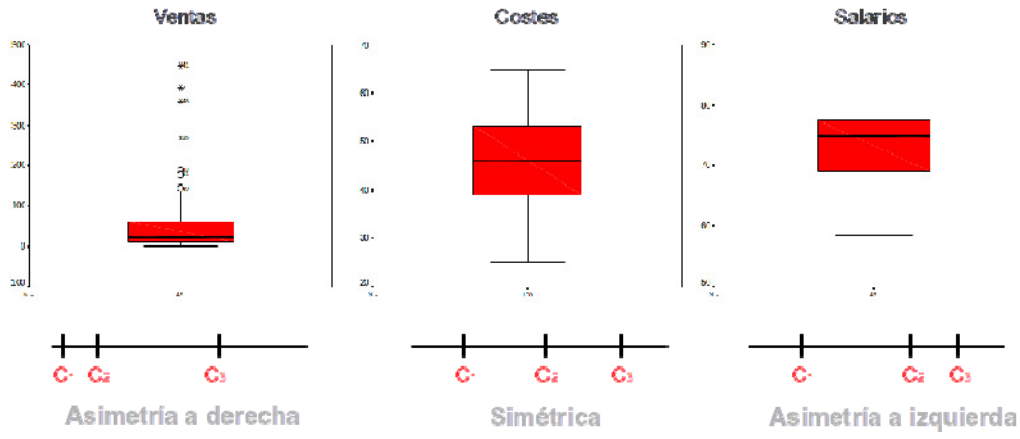
IQR y valores anómalos

6. MEDIDAS DE FORMA

Las medidas de posición y dispersión son las más habituales a la hora de sintetizar la información contenida en unos datos. Sin embargo en algunas ocasiones es conveniente obtener más información sobre la forma de las distribuciones.

6.1. ASIMETRÍA (SKEWNESS)

Las distribuciones de frecuencia pueden clasificarse según su simetría de la siguiente manera:



Tipos de simetría/asimetría en las distribuciones de frecuencia

El *skewness* o coeficiente de asimetría es un indicador numérico que clasifica la simetría de una distribución de frecuencias según su valor.

$$Sk = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^3$$

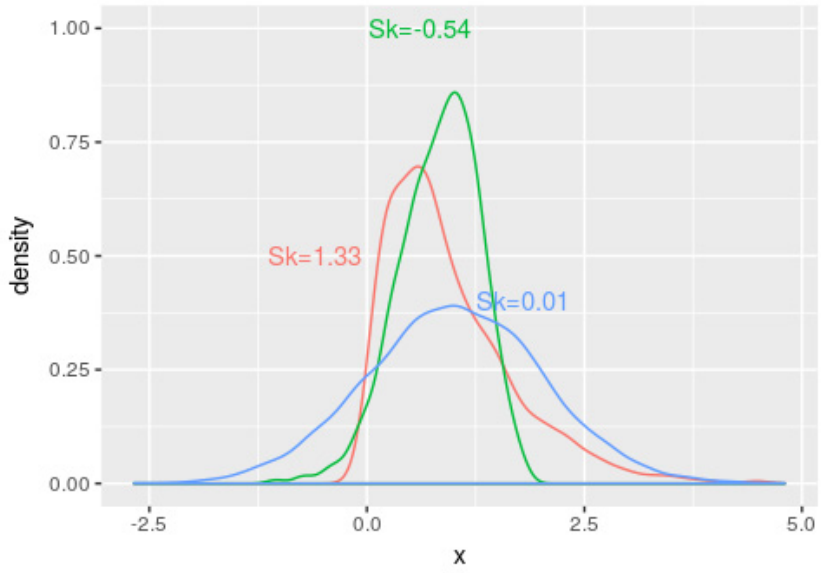
- Si Sk<0 --> Asimétrica a izquierda.
- Si Sk=0 --> Simétrica.
- Si Sk>0 --> Asimétrica a derecha.

En R podemos calcularlo usando la función `skewness` contenida en el paquete e1071.

En el ejemplo siguiente, generamos 3 distribuciones de números aleatorios con diferentes propiedades de simetría y calculamos su skewness.

La distribución de Weibull depende de dos parámetros: forma(shape) y escala(scale) y en función de sus valores genera números distribuidos con diferentes propiedades de simetría. En cambio la distribución normal genera números simétricamente distribuidos.

```
library(e1071)
df=rbind(data.frame(asim="right",x=rweibull(1000,shape =
1.3,scale= 1)),
          data.frame(asim="left",x=rweibull(1000,shape =
7,scale=3)-2),
            data.frame(asim="norm",x=rnorm(10000,mean=1,sd=1)))
skew= df %>% group_by(asim) %>% summarise(sk=skewness(x))
ggplot(df) + geom_density(aes(x,color=asim)) +
  geom_text(aes(1.2*(as.numeric(asim)-
1.5),c(0.5,1,0.4),label=paste0("Sk=",round(sk,2)), col-
or=asim),data=skew) +
  scale_color_hue(guide=FALSE)
```



6.2. CURTOSIS

Mide si una distribución es más "picuda" que una normal o más plana

$$K = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^4 - 3$$

- Si K>0 --> Leptocúrtica (más apuntada).
- Si K=0 --> Mesocúrtica.
- Si K<0 --> Platicúrtica (más plana).

En realidad los valores que más afectan a la curtosis son los valores extremos (las colas de la distribución):

- Platicúrtica(K<0)--> colas pequeñas.
- Leptocúrtica (K>0) --> colas largas.
- Una curtosis grande indica la presencia de valores anómalos, respecto a una distribución normal.

En R calculamos la curtosis mediante la función `kurtosis` del paquete e1071.

```
# Uniforme  - platicúrtica
# Es intervalo ha elegido para que tenga varianza unidad
xu=runif(100000,min = -1.73,max = 1.73)
kurtosis(xu)
```

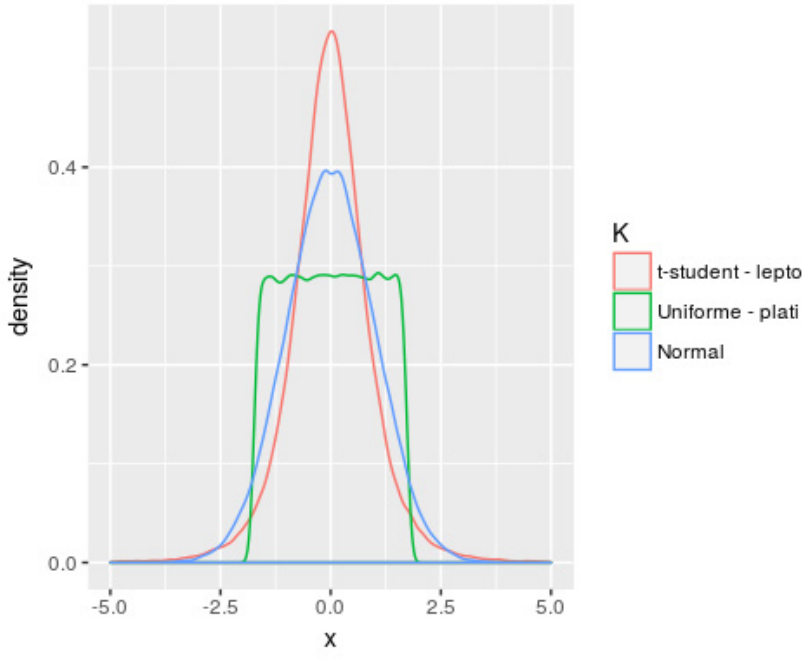
```
## [1] -1.19799
```

```
# t-student - leptocúrtica
xt=rt(100000,df=4)
# Escalo para que tenga varianza unidad
xt=xt/sd(xt)
kurtosis(xt)
```

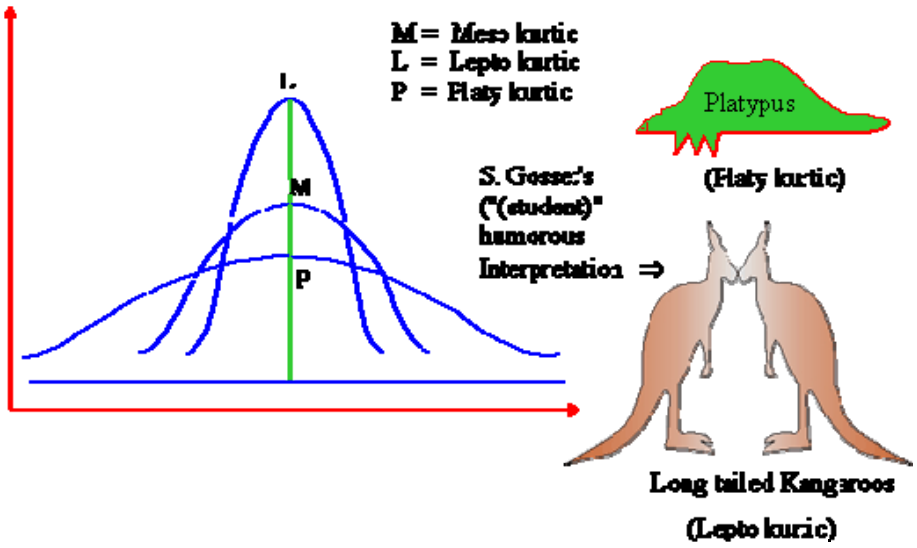
```
## [1] 12.76337
```

Gráficamente

```
df=rbind(data.frame(K="t-student - lepto",x=xt),
         data.frame(K="Uniforme - plati",x=xu),
           data.frame(K="Normal",x=rnorm(100000,mean=0,sd=1)))
ggplot(df) + geom_density(aes(x,color=K)) + xlim(-5,5)
```



Un truco mnemotécnico nos lo da la siguiente imagen, creada originalmente por el estadístico William Gosset, más conocido por su seudónimo Student, descubridor de la distribución t de Student.

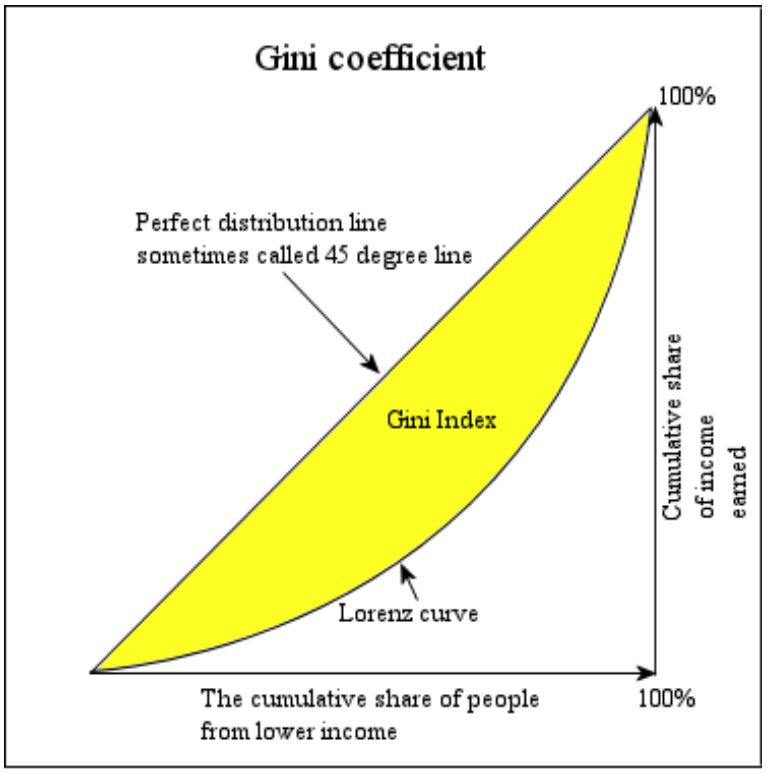


kurtosis

7. MEDIDAS DE CONCENTRACIÓN

7.1. CURVA DE LORENTZ E ÍNDICE DE GINI

Se pretende medir la distribución entre sus componentes de una determinada variable: por ejemplo riqueza entre personas



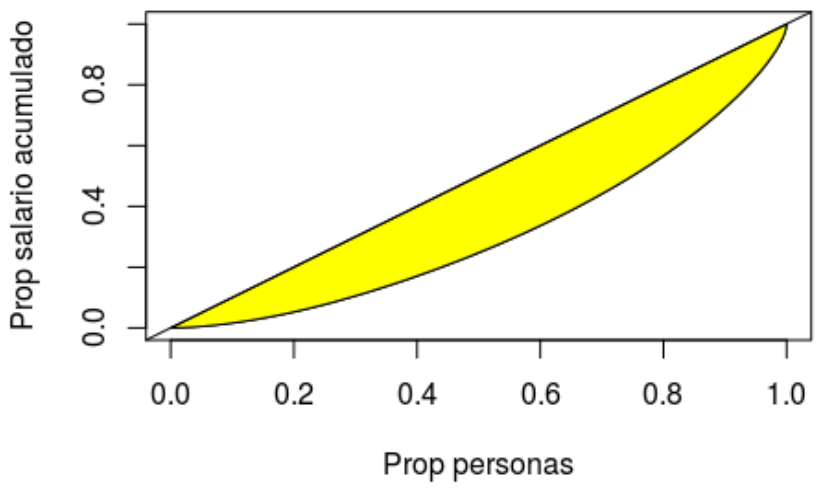
Curva de Lorent e índice de Gini

Veamos cómo se calcula la curva de Lorentz por ejemplo para la distribución de salarios de la encuesta de estructura salarial:

- Se ordena en sentido creciente el vector de salarios de las personas encuestadas. Y se calcula la suma acumulada de salarios.
- Se muestra en el eje x, la proporción de personas y en el eje y la proporción de salario acumulado.
- Si la curva de Lorentz pasa por ejemplo por el punto (0.5,0.3) quiere decir que el 50% de las personas más pobres poseen un 30% de la riqueza total.

El **índice de Gini** mide cuanto se aleja la distribución de salarios de la equipartición (recta y=x). Representa el área de la zona sombreada de la figura.

```
x=ees$SALBRUTO
n=length(x)
x=sort(x)
cumx=cumsum(x)
plot((1:n)/n,cumx/cumx[n],type="l",xlab="Prop perso-
nas",ylab="Prop salario acumulado")
polygon(c(0,(1:n)/n,0),c(0,cumx/cumx[n],0),col="yellow")
abline(0,1)
```

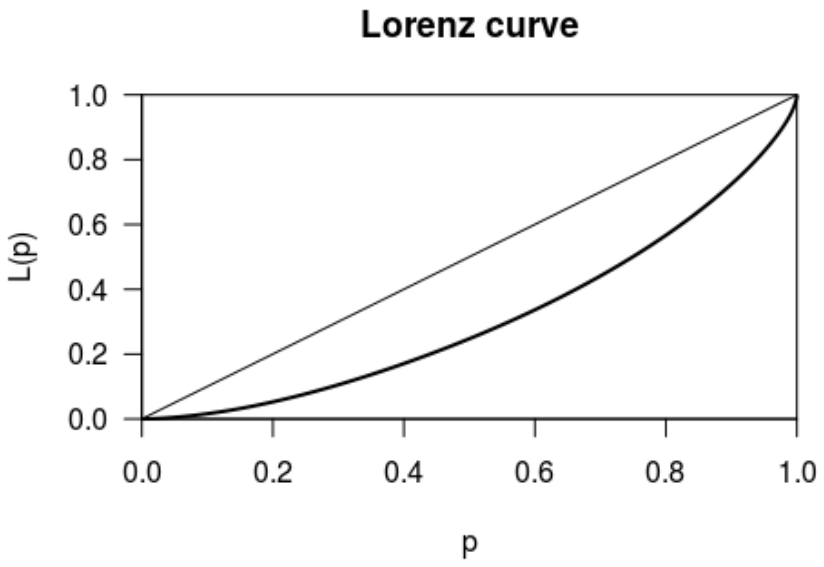


En R, podemos calcular el índice de Gini y la curva de Lorentz mediante le paquete `ineq`.

```
library('ineq')
# Indice de GIni
Gini(ees$SALBRUTO)
```

```
## [1] 0.375856
```

```
# Curva de Lorentz
plot(Lc(ees$SALBRUTO))
```



Podemos hacernos más preguntas sobre la desigualdad en los salarios. Respondamos a algunas.

¿Están más repartidos los salarios en hombres o mujeres?

```
ees %>% group_by(SEXO) %>% summarise(gini=Gini(SALBRUTO))
```

```
## # A tibble: 2 x 2
##     SEXO      gini
##    <chr>     <dbl>
## 1 Hombre 0.3630556
## 2  Mujer 0.3764620
```

¿Y en grupos por estudios terminados?

```
ees %>% group_by(ESTU) %>% summarise(gini=Gini(SALBRUTO))
```

```
## # A tibble: 7 x 2
##                                       ESTU      gini
##                                      <chr>     <dbl>
## 1 Diplomados universitarios o equivalente 0.3773374
## 2                      Educación primaria 0.3281608
## 3                   Educación secundaria I 0.3206992
## 4                  Educación secundaria II 0.3394566
## 5    Formación profesional de grado medio 0.2962608
## 6 Formación profesional de grado superior 0.3335782
## 7                             Sin estudios 0.3315690
```

8. DOS O MÁS VARIABLES NUMÉRICAS

8.1. COVARIANZA

Para detectar relaciones entre dos variables continuas, las medidas más básicas son la covarianza y la correlación lineal.

Covarianza:

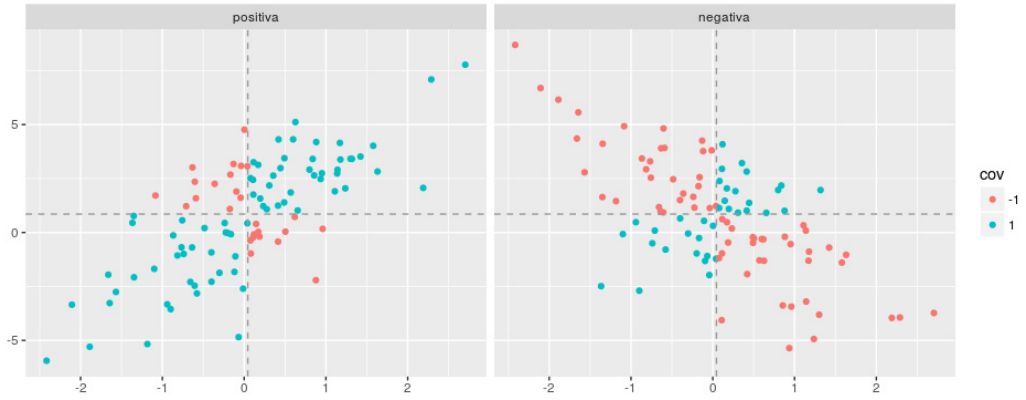
$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Mide cuanto se espera que una variable cambie cuando cambia la otra, suponiendo que ambas están relacionadas linealmente. Si dos variables son independientes su covarianza es 0.

La tabla siguiente nos indica el signo de las contribuciones de cada observación a la covarianza:

Contribución cov	$y < \overline{y}$	$y > \overline{y}$
$x < \overline{x}$	+	-
$x > \overline{x}$	-	+

En el gráfico a continuación, se muestran los puntos (x,y) de dos casos. El de la izquierda con covarianza positiva y de la derecha con covarianza negativa. Los puntos están coloreados en azul cuando hacen una contribución positiva a la covarianza y en rojo cuando su contribución es negativa. Se observa la diferente proporción de puntos rojos y azules en cada uno de los casos.



8.2. CORRELACIÓN

Para tener una medida de relación que no dependa de la escala de cada variable, usamos la correlación lineal.

Se define a la correlación lineal o coeficiente de correlación de Pearson como:

$$r(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

La correlación lineal varía entre -1 y 1

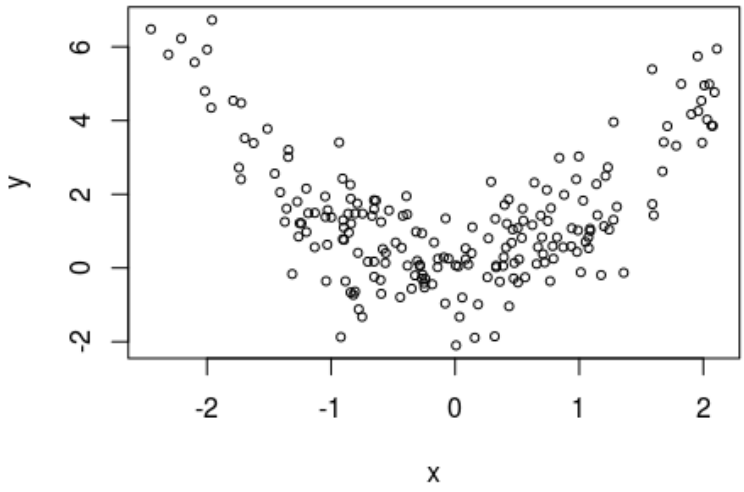
- Si r = 1 si x e y están perfectamente correlacionadas de forma positiva.
- Si r = 1 si x e y están perfectamente correlacionadas de forma negativa.
- Si r = 0 son independientes linealmente. Esto no quiere decir que no pueda existir una relación no lineal entre las variables.

Para calcular la covarianza y correlación en R se usan las funciones `cov` y `cor`

```
x=rnorm(100,sd=4); y=2*x + rnorm(100)
cov(x,y)

## [1] 34.20254

cor(x,y)

## [1] 0.993144
```

Sin embargo la covarianza y la correlación nula no son una condición suficiente para la independencia de dos variables. Es posible que dos variables estén relacionadas de forma no monótona y en ese caso la correlación lineal podría ser nula. Veamos un ejemplo de dos variables x e y que están relacionadas por una relación cuadrática y la correlación lineal es aproximadamente 0.

```
set.seed(2)
x=rnorm(200);
#
y=x^2 + rnorm(200)
cor(x,y)

## [1] 0.02552473

plot(x,y,cex=0.7)
```



8.3. MATRICES DE CORRELACIÓN

Cuando tenemos más de dos variables, entre las cuales quiero calcular la correlación, podemos hablar de la matriz de covarianza o la matriz de correlación.

En R, se puede pasar como argumento de `cov` o `cor` una matriz o un data frame numérico. En ese caso devuelve una matriz con las covarianzas o correlaciones lineales de las diferentes combinaciones de columnas.

$C_{ij} = cov(x_i, x_j)$

```
tmp <- body %>%  select(Age,Gender,Weight, Height,
Chest_diameter,Chest_depth,                                  Bitro-
chanteric_diameter,Wrist_min_girth, Ankle_min_girth)
tmp$Gender=factor(tmp$Gender)
levels(tmp$Gender)=c("W","M")

cor(tmp[c(1,3,4,5)])
```

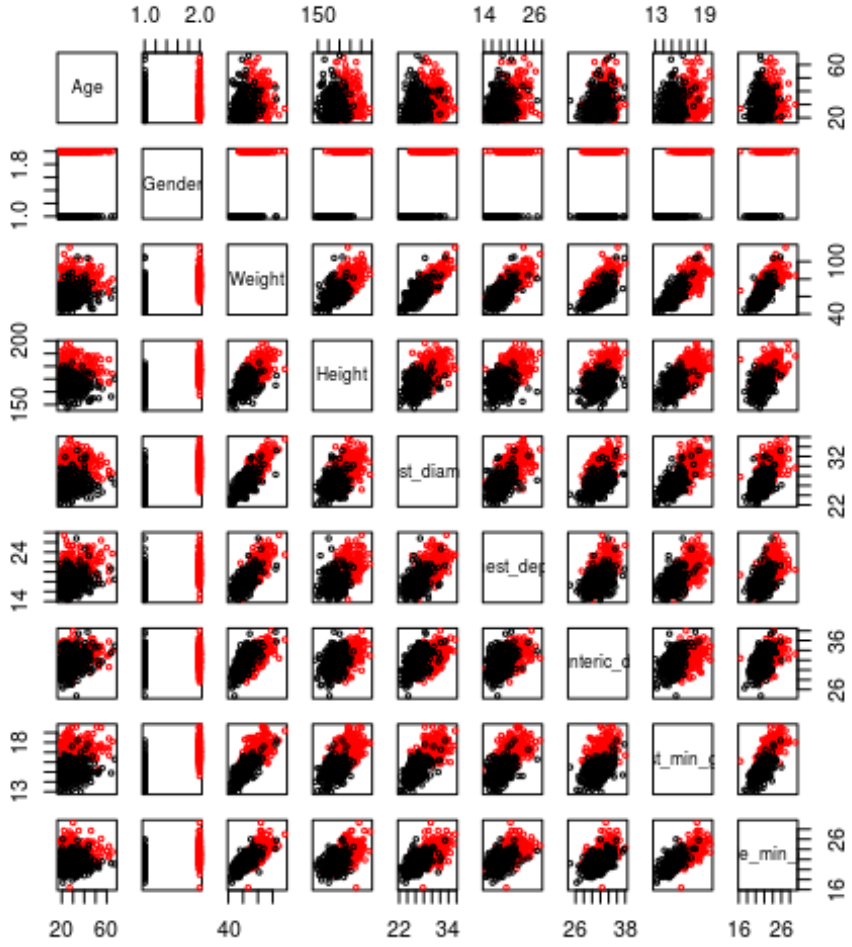
```
##                       Age    Weight     Height Chest_diameter
## Age            1.00000000 0.2072652 0.06788349      0.1928877
## Weight         0.20726524 1.0000000 0.71730108      0.8314645
## Height         0.06788349 0.7173011 1.00000000      0.6268931
## Chest_diameter 0.19288765 0.8314645 0.62689315      1.0000000
```

8.4. GRÁFICOS DE CORRELACIÓN

8.4.1. GRÁFICOS DE DISPERSIÓN

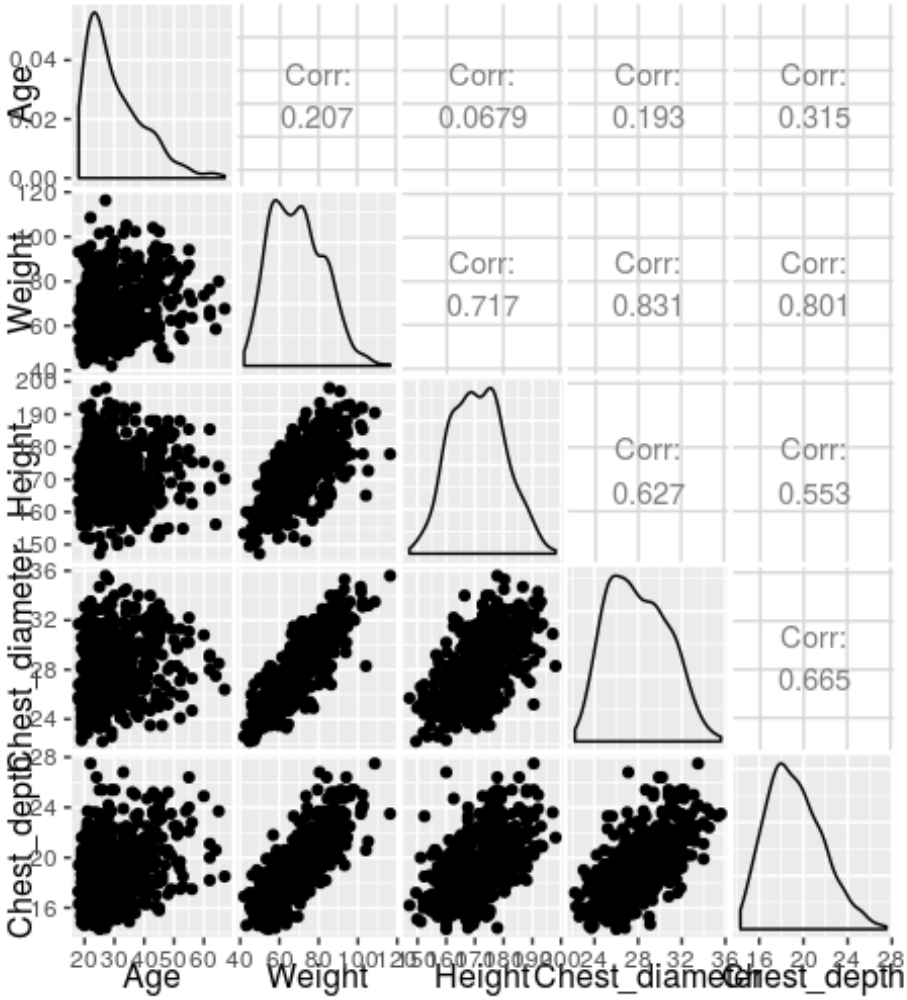
Con múltiples variables podemos hacer una matriz de gráficos de dispersión.

```
pairs(tmp,col=as.numeric(tmp$Gender),cex=0.5,alpha=0.3)
```



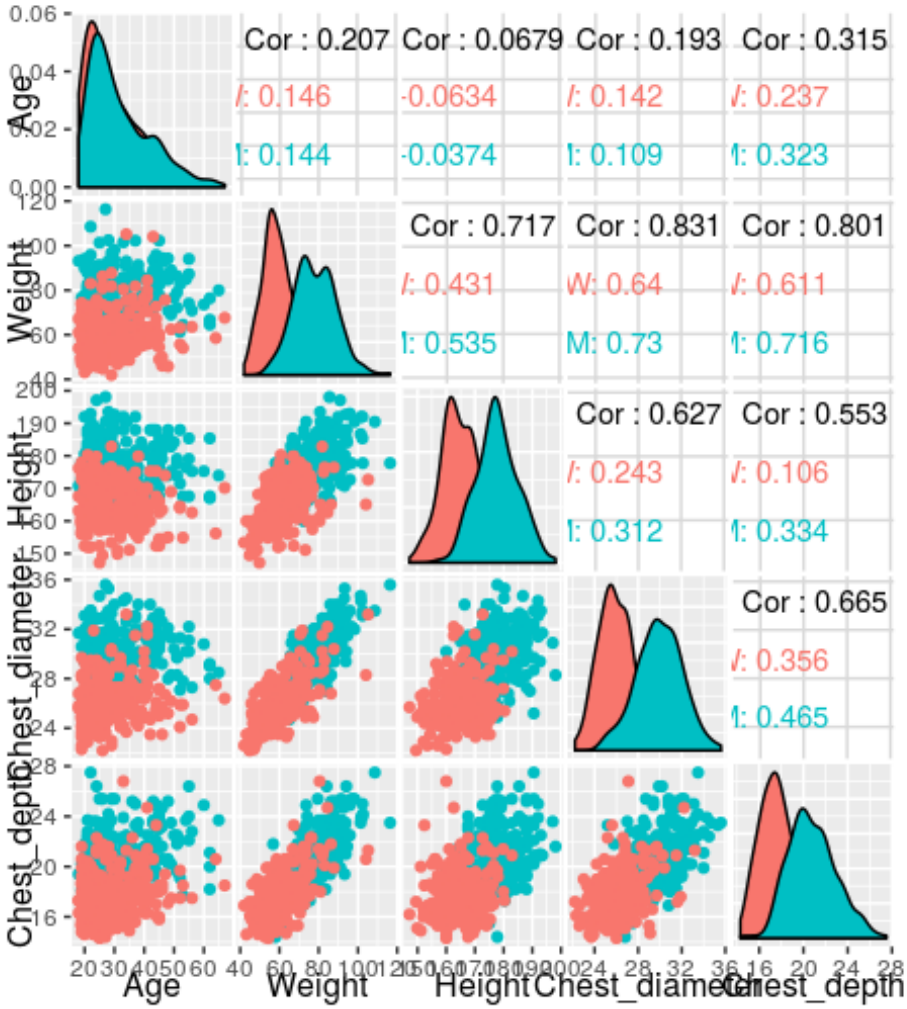
Con `ggplot`, se hace mediante la función `ggpairs` del paquete `GGally`. Muestra los gráficos de dispersión de las diferentes combinaciones de variables en el triángulo inferior, mientras que utiliza la diagonal para mostrar las distribuciones univariantes y el triángulo superior para mostrar los valores numéricos del coeficiente de correlación.

```
library(GGally)
ggpairs(tmp, columns=c(1,3:6))
```



Si se elige una variable para agrupar y colorear, calcula también las correlaciones por grupos.

```
ggpairs(tmp, mapping=ggplot2::aes(color = Gen-
der),columns=c(1,3:6))
```


Cor : 0.207
/: 0.146
1: 0.144
Cor : 0.0679
-0.0634
-0.0374
Cor : 0.193
/: 0.142
1: 0.109
Cor : 0.315
/: 0.237
1: 0.323
Cor : 0.717
/: 0.431
1: 0.535
Cor : 0.831
W: 0.64
M: 0.73
Cor : 0.801
/: 0.611
1: 0.716
Cor : 0.627
/: 0.243
1: 0.312
Cor : 0.553
/: 0.106
1: 0.334
Cor : 0.665
/: 0.356
1: 0.465
Age
Weight
Height
Chest_diameter
Chest_depth

¿QUÉ HAS APRENDIDO?

En esta unidad has aprendido a exprimir la información que está contenida en un conjunto de datos.

En concreto, has aprendido a:

- Calcular, representar y comparar distribuciones de frecuencias.
- Sintetizar los datos mediante medidas de:
  - Posición.
  - Dispersión.
  - Forma.
  - Desigualdad.
- Detectar correlaciones entre variables de forma tanto numérica como gráfica.

Ahora ya dispones de todo el conocimiento para realizar el análisis exploratorio de datos, que es un elemento fundamental de la ciencia de datos, ya que es a través del cual serás capaz de hacerte las preguntas adecuadas que posteriormente responderás usando las técnicas de inferencia y modelización que vas a aprender en las siguientes unidades.

Para terminar, vuelve a las preguntas del inicio de la unidad y verás como eres capaz de dar una respuesta para todas ellas.

AUTOCOMPROBACIÓN

1. **La variable ESTU de la encuesta de estructura salarial, que representa el nivel de estudios de los encuestados es**

   a) Una variable cuantitativa continua.

   b) Una variable cuantitativa discreta.

   c) Una variable cualitativa ordinal.

   d) Ninguna de las respuestas anteriores.

2. **La variable TIPOCON de la encuesta de estructura salarial representa el tipo de contrato. ¿Cuál es la categoría más frecuente?**

   a) Duración indefinida.

   b) Duración determinada.

   c) Obra y servicios.

   d) Indeterminada.

3. **Que comando de R has utilizado para responder a la pregunta anterior**

   a) hist.

   b) density.

   c) table.

   d) mean.

4. La media es una medida

a) Sensible a los valores anómalos.

b) Robusta frente a los valores anómalos.

c) Insensible a los valores anómalos.

d) Mal definida, si la variable puede tomar el valor 0.

5. Di cuál de las siguientes relaciones se representaría mejor con un boxplot

a) La altura de un conjunto de personas en función del peso.

b) La evolución temporal de la cotización de un valor en la bolsa.

c) La proporción de personas según el grupo de edad al que pertenecen.

d) La distribución de alturas de un conjunto de personas en función del sexo.

6. Si el cuantil 0.3 una cierta variable es 5, ¿qué podemos decir del cuantil 0.2 de esa misma variable?

a) Es menor que 5.

b) Es menor o igual que 5.

c) Es mayor que 5.

d) Es mayor o igual que 5.

7. El rango intercuartílico es una medida de:

a) Forma.

b) Centralidad.

c) Desigualdad.

d) Dispersión.

8. Si una distribución tiene skewness menor de 0:

a) Es simétrica.

b) Es asimétrica a izquierdas.

c) Es asimétrica a derechas.

d) No sé cómo es su simetría.

9. Si dos variables son independientes

a) Su covarianza es 0.

b) Su coeficiente de correlación es 0.

c) Su coeficiente de correlación es menor de 1.

d) Todas las respuestas anteriores son ciertas.

10. Si la correlación lineal entre dos variables x,y es 1

a) $y= a + bx$ con $b>0$.

b) $y= a + bx$ con $b=0$.

c) $y= a + bx$ con $b<0$.

d) $y= a + bx$ con $a>0$.

SOLUCIONARIO

1.	c	**2.**	a	**3.**	c	**4.**	a	**5.**	d
6.	b	**7.**	d	**8.**	b	**9.**	d	**10.**	a

BIBLIOGRAFÍA

- Introduction to Probability and Statistics Using R, G. Jay Kerns.
- R for Data Science, Garrett Grolemund, Hadley Wickham: http://r4ds.had.co.nz/
- Probabilidad y estadística para ingeniería y ciencias / Jay L. Devore.