

FAEDE 1

2023-10-26

```
#Identificación del directorio de trabajo al iniciar R  
getwd()  
  
## [1] "D:/FAEDE"  
  
#Personalización del directorio de trabajo  
setwd("D:/FAEDE")  
library(e1071) #Cálculo de asimetría (skewness) y curtosis (kurtosis)
```

```
#Calculando con R...  
1+2 # Escribir 1+2 en la R-Consola y pulsar Enter  
  
## [1] 3  
  
3*7  
  
## [1] 21  
  
7/4  
  
## [1] 1.75  
  
7%%4  
  
## [1] 3  
  
abs(-4)  
  
## [1] 4  
  
log(0)  
  
## [1] -Inf  
  
2/0  
  
## [1] Inf  
  
sqrt(-4)  
  
## Warning in sqrt(-4): Se han producido NaNs  
  
## [1] NaN
```

```

#Estructuras de datos en R
#Vectores
c(1,2,3,4,5) # vector numérico

## [1] 1 2 3 4 5

c(F,T,T,F,F) # vector lógico

## [1] FALSE TRUE TRUE FALSE FALSE

c("Juan","Pepe","Pedro","Antonio") # vector de caracteres

## [1] "Juan" "Pepe" "Pedro" "Antonio"

1:5

## [1] 1 2 3 4 5

5:1

## [1] 5 4 3 2 1

v=c(1:5,10:5,12)
v=c(1:5,10:5,12);v

## [1] 1 2 3 4 5 10 9 8 7 6 5 12

v[4]

## [1] 4

x<-c(1,2,3)
y<-c(T,F,T)
c(x,y)

## [1] 1 2 3 1 0 1

seq(from=0, to=10)

## [1] 0 1 2 3 4 5 6 7 8 9 10

seq(0, 10, 0.5)

## [1] 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0
## [16] 7.5 8.0 8.5 9.0 9.5 10.0

seq(from=5, by=-0.5, length.out=7)

## [1] 5.0 4.5 4.0 3.5 3.0 2.5 2.0

```

```

#Factores
factor(letters[1:20], labels="letter")

## [1] letter1 letter2 letter3 letter4 letter5 letter6 letter7 letter8
## [9] letter9 letter10 letter11 letter12 letter13 letter14 letter15 letter16
## [17] letter17 letter18 letter19 letter20
## 20 Levels: letter1 letter2 letter3 letter4 letter5 letter6 letter7 ... letter20

```

#Listas

```
milista=list(hombre="Pedro",mujer="María",casados=T,n.hijos=3, edad.hijos=c(1,2,4))
```

```
milista
```

```
$hombre  
[1] "Pedro"
```

```
$mujer  
[1] "María"
```

```
$casados  
[1] TRUE
```

```
$n.hijos  
[1] 3
```

```
$edad.hijos  
[1] 1 2 4
```

#Matrices

```
matrix(1:12)
```

```
##      [,1]  
## [1,] 1  
## [2,] 2  
## [3,] 3  
## [4,] 4  
## [5,] 5  
## [6,] 6  
## [7,] 7  
## [8,] 8  
## [9,] 9  
## [10,] 10  
## [11,] 11  
## [12,] 12
```

```
matrix(1:12,nrow=3)
```

```
##      [,1] [,2] [,3] [,4]  
## [1,] 1 4 7 10  
## [2,] 2 5 8 11  
## [3,] 3 6 9 12
```

```
m <- matrix(1:12, nrow=3, byrow=T)
```

```
colnames(m) <- c("Dato 1", "Dato 2", "Dato 3", "Dato 4")
```

```
rownames(m) <- c("Primero", "Segundo", "Tercero")
```

#Identifica la utilidad de las funciones sobre matrices, ejecutando las líneas:

```
dim(m)
## [1] 3 4

length(m)
## [1] 12

m[1,]
## Dato 1 Dato 2 Dato 3 Dato 4
##      1      2      3      4

m[,1:3]
##      Dato 1 Dato 2 Dato 3
## Primero    1      2      3
## Segundo    5      6      7
## Tercero     9     10     11

m[-1,]
##      Dato 1 Dato 2 Dato 3 Dato 4
## Segundo    5      6      7      8
## Tercero     9     10     11     12

rbind(m,c(1,1,1,1))
##      Dato 1 Dato 2 Dato 3 Dato 4
## Primero    1      2      3      4
## Segundo    5      6      7      8
## Tercero     9     10     11     12
##           1      1      1      1

cbind(m,c(1,1,1))
##      Dato 1 Dato 2 Dato 3 Dato 4
## Primero    1      2      3      4 1
## Segundo    5      6      7      8 1
## Tercero     9     10     11     12 1
```

#Data frames

```
L3 <- LETTERS[1:3]
fac <- sample(L3, 10, replace = TRUE)
d <- data.frame(x = 1, y = 1:10, fac = fac)
is.data.frame(d)
```

```
## [1] TRUE
```

```
d
  x y fac
1 1 1  A
2 1 2  B
3 1 3  C
4 1 4  B
5 1 5  B
6 1 6  C
7 1 7  C
8 1 8  B
9 1 9  A
10 1 10 A
```

Repaso de Estadística Descriptiva Clasificación de Estadísticos Descriptivos

POSICIÓN	<ul style="list-style-type: none"> CENTRAL: MEDIA ARITMÉTICA, MEDIANA Y MODA NO CENTRAL: CUANTILES
DISPERSIÓN	<ul style="list-style-type: none"> ABSOLUTOS: RANGO, VARIANZA Y DESVIACIÓN TÍPICA RELATIVOS: COEFICIENTE DE VARIACIÓN DE PEARSON
FORMA DAN UNA IDEA DE LA FORMA DE LA DISTRIBUCIÓN	<ul style="list-style-type: none"> ASIMETRÍA: COEFICIENTE DE ASIMETRÍA APUNTAMIENTO O CURTOSIS: COEFICIENTE DE APUNTAMIENTO

Estadísticos descriptivos de posición (DAN UNA IDEA DE LA UBICACIÓN DE LOS DATOS)

Min.(mínimo), valor mínimo observado.

1st Qu. (primer cuartil), valor de la secuencia ordenada de observaciones que deja el 25% de las observaciones por debajo.

Median (mediana), valor que divide la secuencia ordenada en dos partes iguales.

Mean(media), valor obtenido al repartir equitativamente el total observado entre los individuos de la muestra.

3rd Qu. (tercer cuartil), valor de la secuencia ordenada de observaciones que deja el 75% de las observaciones por debajo.

Max (máximo), valor máximo observado.

moda, valor(es) que más se repite(n)

#Familia apply

`apply(d[,1:2],2,summary)` *#summary reporta medidas descriptivas de posición*

```
##      x      y
## Min.   1  1.00
## 1st Qu. 1  3.25 (el 25% de las observaciones están por debajo de 3.25)*
## Median 1  5.50 (el 50% de las observaciones están por debajo de 5.50)*
## Mean   1  5.50
## 3rd Qu. 1  7.75 (el 75% de las observaciones están por debajo de 7.75)*
## Max.   1 10.00
```

**Los cuantiles se han interpretado para la variable y.*

`tapply(dy,dfac,summary)`

```
## $A
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5      6      7      7      8      9
## $B
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.00   1.75   2.50   4.00   4.75  10.00
## $C
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.00   5.50   6.50   6.25   7.25   8.00
```

Estadísticos descriptivos de dispersión

(MIDEN LA VARIABILIDAD DE LOS DATOS)

- **ESTADÍSTICOS DE DISPERSIÓN ABSOLUTOS:** RANGO, RANGO INTERCUARTÍLICO, VARIANZA Y DESVIACIÓN TÍPICA (EXPRESADAS EN LAS MISMAS UNIDADES DE MEDIDA QUE LA VARIABLE)
- **ESTADÍSTICOS DE DISPERSIÓN RELATIVOS:** COEFICIENTE DE VARIACIÓN DE PEARSON (ADIMENSIONALES)

Rango, distancia entre el valor máximo y mínimo.

Rango intercuartílico, distancia entre tercer y primer cuartil.

var (varianza), media de los cuadrados de las desviaciones de los valores con respecto a la media de las observaciones, esto es, $s^2 = \sum (x_i - \bar{x})^2 / N$.

sd (desviación típica), raíz cuadrada positiva de la varianza.

```
apply(d[,c(1,2)], 2, sd)
```

```
##           x           y
## 0.00000 3.02765
```

*La variable x toma solo el valor 1, es decir, es constante $x_i = 1$ para todo i. Así pues la media $\bar{x} = 1$ y la desviación típica $s = 0$ ya que las desviaciones de los valores con respecto a su media $x_i - \bar{x} = 0$ para todo i

Coeficiente de variación de Pearson, dado por $CV = s/|\bar{x}|$.

Si el coeficiente de variación de una variable es mayor que 0.5, entonces su media no es representativa. Por el contrario, si es menor que 0.5, si es representativa.

Implementación en R del coeficiente de variación como objeto con la función CVar

```
CVar<-function(x)
{
  resultado<-sqrt(var(x))/abs(mean(x))
  return(resultado)
}
```

```
tapply(d$y,d$fac,CVar)
```

```
##           A           B           C
## 0.404061 1.020621 0.273252
```

*Las medias obtenidas en A y C son representativas (la media en C es más representativa que la media obtenida en A), pero la media en B no es representativa (coeficiente de variación = 1.020621 > 0.5).

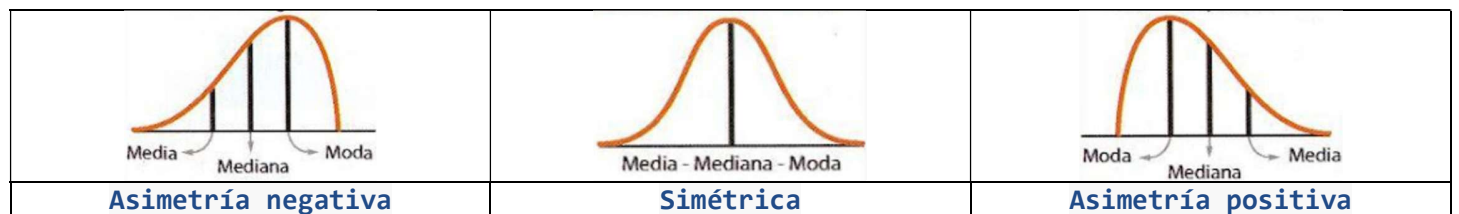
Estadísticos descriptivos de forma

(DAN UNA IDEA DE LA FORMA DE LA DISTRIBUCIÓN)

- **ESTADÍSTICOS DE ASIMETRÍA:** COEFICIENTE DE ASIMETRÍA
- **ESTADÍSTICOS DE APUNTAMIENTO O CURTOSIS:** COEFICIENTE DE APUNTAMIENTO

skewness (coeficiente de asimetría), mide el grado de asimetría de la distribución con respecto a la media, y viene dado por $g_1 = \sum \frac{(x_i - \bar{x})^3}{N} / s^3$.

Si $g_1 > 0$, la distribución es asimétrica positiva (asimetría a la derecha).
 Si $g_1 < 0$, la distribución es asimétrica negativa (asimetría a la izquierda).
 Si la distribución es simétrica, $g_1 = 0$.



Simétrica: Moda = Mediana = Media.
 asimétrica positiva: Moda \leq Mediana < Media.
 asimétrica negativa: Moda \geq Mediana > Media.

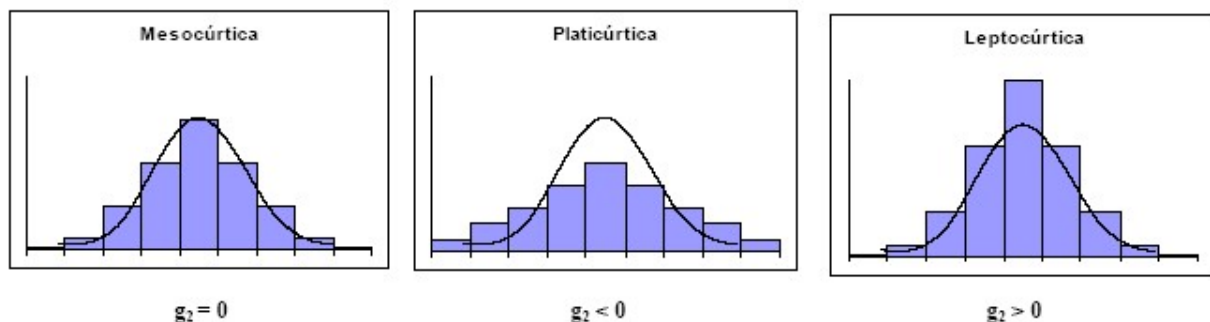
```
lapply(d[, -3], skewness)
```

```
## $x
## [1] NaN
##
## $y
## [1] 0
```

*La variable x toma solo el valor 1, es decir, es constante $x_i = 1$ para todo i . Así que la media $\bar{x} = 1$ y la desviación típica $s = 0$ ya que las desviaciones de los valores con respecto a su media $x_i - \bar{x} = 0$ para todo i , por lo que numerador y denominador $g_1 = \frac{\sum \frac{(x_i - \bar{x})^3}{N}}{s^3}$ es cero.

kurtosis (coeficiente de apuntamiento o curtosis), en distribuciones unimodales simétricas o ligeramente asimétricas, mide el grado de apuntamiento de la distribución tomando como referencia la distribución normal, y viene dado por $g_2 = \left(\frac{\sum \frac{(x_i - \bar{x})^4}{N}}{s^4} \right) - 3$.

Si $g_2 = 0$, la distribución es mesocúrtica (normal).
 Si $g_2 > 0$, la distribución es leptocúrtica (más apuntada que la normal).
 Si $g_2 < 0$, la distribución es platicúrtica (menos apuntada que la normal).



```
sapply(d[, 1:2], kurtosis)
```

```
##      x      y
##    NaN -1.561636
```

#Lectura e introducción de datos en R

```
library(readr) #Importar fichero ElPulso.txt (desde Environment)
```

```
ElPulso <- read_table2("ElPulso.txt")
```

```
##
```

```
## — Column specification —————
```

```
## cols(
```

```
##   Pulse1 = col_double(),
```

```
##   Pulse2 = col_double(),
```

```
##   Ran = col_double(),
```

```
##   Smokes = col_double(),
```

```
##   Sex = col_double(),
```

```
##   Height = col_double(),
```

```
##   Weight = col_double(),
```

```
##   Activity = col_double()
```

```
## )
```

#estructura de un dataframe

```
names(ElPulso)
```

```
## [1] "Pulse1" "Pulse2" "Ran" "Smokes" "Sex" "Height" "Weight"
```

```
## [8] "Activity"
```

```
ElPulso[,1]
```

```
## # A tibble: 92 × 1
```

```
##   Pulse1
```

```
##   <dbl>
```

```
## 1     64
```

```
## 2     58
```

```
## 3     62
```

```
## 4     66
```

```
## 5     64
```

```
## 6     74
```

```
## 7     84
```

```
## 8     68
```

```
## 9     62
```

```
## 10    76
```

```
## # i 82 more rows
```

```
ElPulso$Activity
```

```
## [1] 2 2 3 1 2 1 3 2 2 2 1 2 3 2 3 3 2 3 2 3 2 2 2 2
```

```
## [26] 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 3 2 3 3 3 2 3 3
```

```
## [51] 2 2 2 NA 3 2 2 2 3 2 1 2 2 2 3 2 2 2 2 3 2 2 2 2
```

```
## [76] 2 2 1 3 2 2 2 2 2 2 1 3 3 1 1 2 3 2
```

```
ElPulso[, "Smokes"] # también ElPulso[["Smokes"]]
```

```
## # A tibble: 92 × 1
```

```
##   Smokes
```

```
##   <dbl>
```

```
## 1     2
```

```
## 2     2
```



```
## 3      1
## 4      1
## 5      2
## 6      2
## 7      2
## 8      2
## 9      2
## 10     2
## # i 82 more rows
```

```
summary(ElPulso)
```

```
##      Pulse1      Pulse2      Ran      Smokes      Sex
## Min.   : 48.00   Min.   : 50   Min.   :1.00   Min.   :1.000   Min.   :1.00
## 1st Qu.: 64.00   1st Qu.: 68   1st Qu.:1.00   1st Qu.:1.000   1st Qu.:1.00
## Median : 71.00   Median : 76   Median :2.00   Median :2.000   Median :1.00
## Mean   : 72.87   Mean    : 80   Mean    :1.62   Mean    :1.696   Mean    :1.38
## 3rd Qu.: 80.00   3rd Qu.: 85   3rd Qu.:2.00   3rd Qu.:2.000   3rd Qu.:2.00
## Max.   :100.00   Max.    :140   Max.    :2.00   Max.    :2.000   Max.    :2.00
##
##      Height      Weight      Activity
## Min.   :61.00   Min.   : 95.0   Min.   :1.000
## 1st Qu.:66.00   1st Qu.:125.0   1st Qu.:2.000
## Median :69.00   Median :145.0   Median :2.000
## Mean   :68.72   Mean    :145.2   Mean    :2.132
## 3rd Qu.:72.00   3rd Qu.:155.5   3rd Qu.:2.000
## Max.   :75.00   Max.    :215.0   Max.    :3.000
##
##                      NA's      :1
```

```
#Etiquetar los codigos de las variables cualitativas
```

```
ElPulso$Actividad=factor(ElPulso$Activity, levels=1:3, labels=c("Suave","Moderada","Alta"))
ElPulso$Fumar <- factor(ElPulso$Smokes, levels=1:2, labels=c("Fuma","No Fuma"))
ElPulso$Sexo <- factor(ElPulso$Sex, levels=1:2, labels=c("Hombre","Mujer"))
ElPulso$Correr <- factor(ElPulso$Ran,levels=c(1,2), labels=c("No","Si"))
```

```
#Calcular nuevas variables
```

```
ElPulso$Peso.kg <- with(ElPulso, round(ElPulso$Weight*0.454,1))
ElPulso$Altura.cm <- with(ElPulso, round(ElPulso$Height*2.54,1))
```

```
#Transformar una variable "cuanti" en "cuali", agrupando
```

```
library("car") #cargar libreria car para usar la funcion recode
```

```
## Loading required package: carData
```

```
summary(ElPulso$Peso.kg)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  43.10   56.80   65.80   65.90   70.62   97.60
```

```
ElPulso$Peso.int<-recode(ElPulso$Peso.kg, '40:60="N";60.1:80="S";80.1:100="M"', as.factor=T)
```

```
#Renombrar variables
```

```
names(ElPulso)[c(1,2)] <- c("Pulso1","Pulso2")
attach(ElPulso) #carga en la memoria las variables del dataframe
detach(ElPulso) #para desactivar attach
```