R-studio

- R-Studio entorno completo. Linea de comandos. Gráficos. Editor.
- Incluye R.
- Como una calculadora.
- 3+4, 5-3, etc..
- x<-3; x
- x <- c(1,2,3,4,5); x
- x2 <- seq(2,18,2); x2; length(x2)
- x5 <- c(1:4, 8:10, seq(-7,5,by=2), rep(x,length=8)); x5; length(x5)
- ## [1] 1 2 3 4 8 9 10-7-5-3-1 1 3 5 1 2 3 4 5 1 2 3
- ## [1] 22

x5[10] Décimo elemento del vector x5
[1]-3
x5[c(10,15,1)] Décimo, decimoquinto y primer elemento de x5
[1]-3 1 1

x6 <- x5[-10]; x6; length(x6) Eliminar la décima componente

[1] 1 2 3 4 8 9 10-7-5-1 1 3 5 1 2 3 4 5 1 2 3
[1] 21

Por otra parte, también se puede acceder a los elementos de un vector que cumplen una determinada condición. Por ejemplo, si se desea conocer el valor de los elementos de x5 que son menores que 0, se utiliza la función which() primero para saber la posición de los elementos que cumplen la condición:
ind <- which(x5<0); ind
[1] 8 9 10 11
x5[ind]
[1]-7-5-3-1

El conjunto de datos *stackloss* contiene los datos operativos de una planta de oxidación de amoniaco a ácido nítrico. Cárgalo e imprime en pantalla las 6 primeras filas, ¿Cuál es la temperatura del agua de la quinta muestra?

```
data(stackloss); head(stackloss)
```

```
stackloss$Air.Flow[5]
```

¿Cuántas observaciones tiene el conjunto de datos?

```
dim(stackloss)
```

¿Cuántas variables tiene el conjunto de datos?

Explora lo que significa cada variable: nombre, tipo de datos, unidades de medida, etc.

```
help(stackloss)
```

- **Frecuencia absoluta** (n_i)**:** Es el número de repeticiones que presenta una observación.
- **Frecuencia relativa** (f_i)**:** Es la frecuencia absoluta, dividida por el número total de datos.
- **Frecuencia absoluta acumulada** (N_i)**:** Es la suma de los distintos valores de la frecuencia absoluta tomando como referencia un individuo dado.
- **Frecuencia relativa acumulada** (F_i)**:** Es el resultado de dividir cada frecuencia absoluta acumulada por el número total de datos.

Ejemplo 1: El conjunto de datos para el control de calidad del agua de diferentes reactores es el siguiente, en que cada número representa el reactor que se eligió como el mejor:

1,5,3,1,2,3,4,5,1,4,2,4,4,5,1,4,2,4,2,2

Reactor	Frec. absoluta	Frec. relativa	Frec. abs. acumulada	Frec. rel. acumulada
1	4	0.20	4	0.20
2	5	0.25	9	0.45
3	2	0.10	11	0.55
4	6	0.30	17	0.85
5	3	0.15	20	1.00

```
ni <- table(stackloss$Air.Flow)
fi <- table(stackloss$Air.Flow)/length(stackloss$Air.Flow)
Ni <- cumsum(ni)
Fi <- cumsum(fi)
Tabla_Frec = cbind(ni,fi,Ni,Fi); Tabla_Frec
```

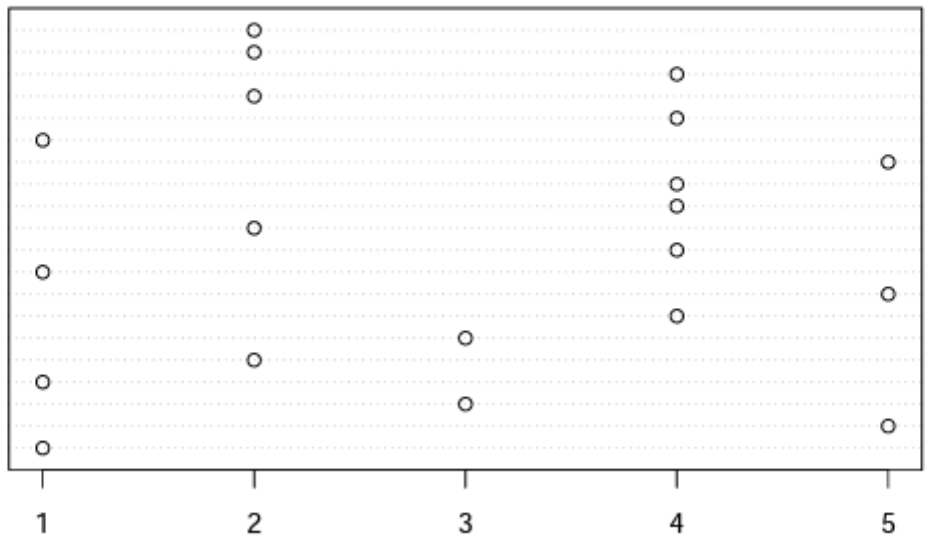
Realiza el gráfico de tallos y hojas del flujo de aire. ¿Cuáles son los números que representan las hojas del tallo con el número 7?

```
>stem(datos_2,scale=2)
The decimal point is 1 digit(s) to the right of the |
  7 | 6                16 | 0003357789
  8 | 7                17 | 0112445668
  9 | 7                18 | 0011346
 10 | 15               19 | 034699
 11 | 058              20 | 0178
 12 | 013              21 | 8
 13 | 133455           22 | 189
 14 | 12356899         23 | 7
 15 | 001344678888     24 | 5
```

```
stem(stackloss$Air.Flow)
```

Realiza el gráfico de puntos de la temperatura del agua cuando el flujo de aire es 58. ¿Cuál es la temperatura máxima? (23)

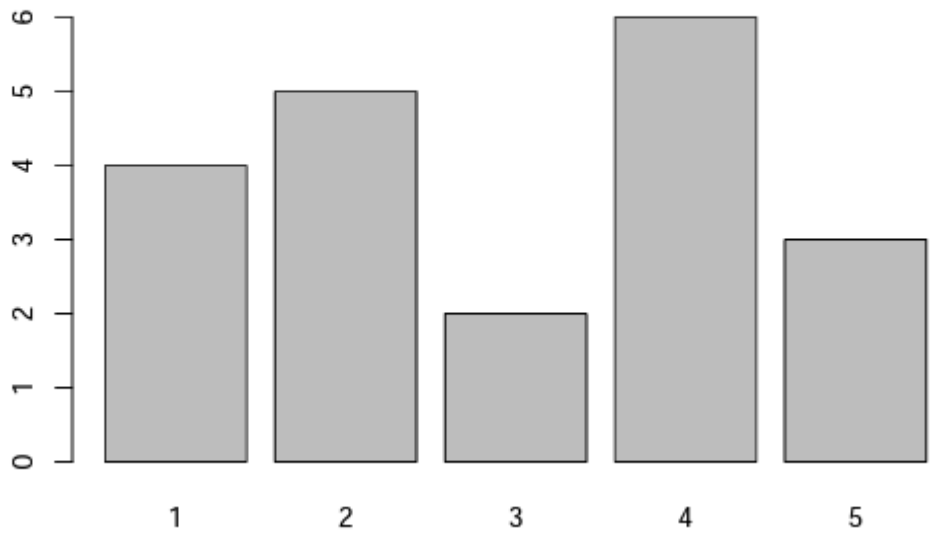
```
dotchart(datos_1)
```




```
Temp <- stackloss$Water.Temp[stackloss$Air.Flow==58]
dotchart(Temp)
```

Realiza el gráfico de barras de la concentración de ácido. ¿Cuántas veces se ha medido 87? (4)

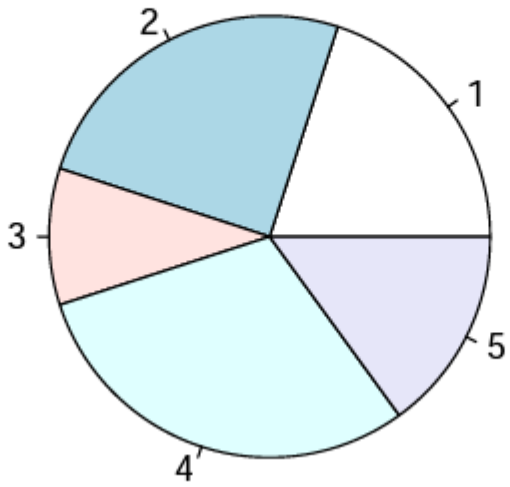
```
barplot(table(datos_1))
```



barplot(table(stackloss$Acid.Conc.))

Realiza el gráfico de sectores del flujo de aire. ¿Cuál medida de flujo es la que más se repite? (58)

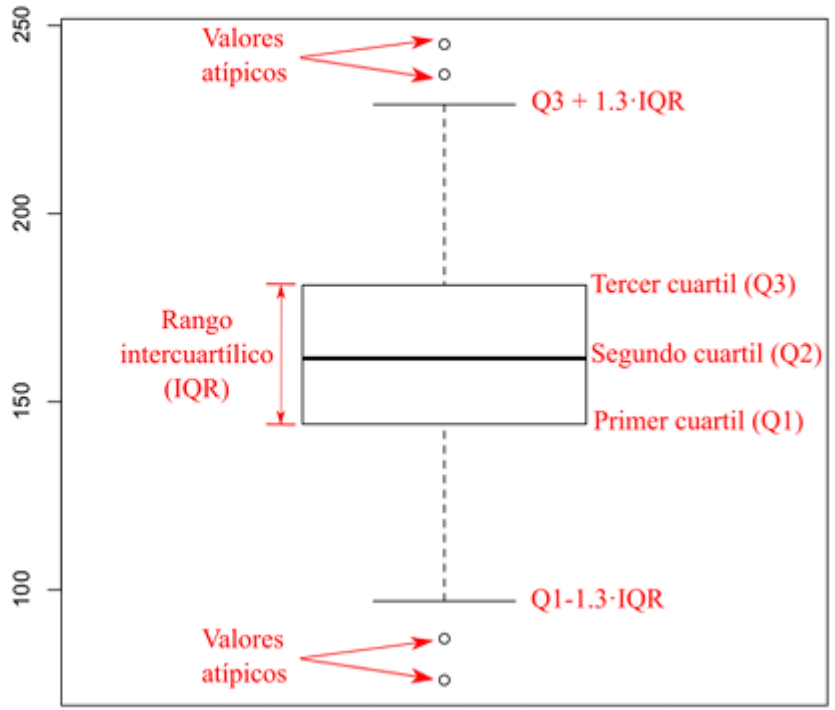
```
pie(table(datos_1))
```




```
pie(table(stackloss$Air.Flow))
```

Realiza el diagrama de cajas de todas las variables. ¿En cuáles variables hay valores extremos?(Flujo de aire y Pérdida de amoniaco)

```
bp = boxplot(datos_2); bp
```




```
bp = boxplot(stackloss)
```

Media

También conocida como el valor medio, se define como la suma de todos los valores de cada observación (x_i), dividido por el número total de observaciones del conjunto de datos (N).

$$\bar{X} = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

```
sum(datos_1)/length(datos_1)
```

```
## [1] 2.95
```

```
mean(datos_1)
```

```
## [1] 2.95
```

¿Cuál es la media del flujo de aire?

```
mean(stackloss$Air.Flow)
```

Mediana

La mediana es el dato que ocupa la posición central en la muestra ordenada de menor a mayor; es un punto que divide la muestra ordenada en dos grupos iguales (deja el 50 % de los valores por debajo y el otro 50 % por encima). Para calcularla, se ordenan los datos de menor a mayor, y el dato central es el que ocupa la posición $\frac{N+1}{2}$ donde N es el número total de datos. Si N es impar, la mediana es el mismo dato central; si N es par, existen dos datos centrales, por lo que la mediana es el promedio de ambos. Igualmente, existe una función que la calcula directamente: median().

¿Cuál es la mediana de la temperatura del agua?

```
median(stackloss$Water.Temp)
```

Cuantiles

Los cuantiles son valores de la lista de datos que la dividen en partes iguales, es decir, en intervalos, que comprenden el mismo número de valores. Los más usados son los percentiles, los deciles y los cuartiles. Los percentiles son 99 valores que dividen en cien partes iguales el conjunto de datos ordenados. Por ejemplo, el percentil de orden 15 deja por debajo al 15 % de las observaciones y por encima quedan el 85 %. Los deciles son los nueve valores que dividen el conjunto de datos ordenados en diez partes iguales; son un caso particular de los percentiles. Los cuartiles son los tres valores que dividen el conjunto de datos ordenados en cuatro partes iguales; son también un caso particular de los percentiles. En *R*, cualquiera de estos se calcula con la función quantile(), donde adicionalmente se ha de especificar el cuantil o los cuantiles deseados (como un valor entre 0 y 1) de la siguiente forma:

```
quantile(datos_2,0.95)  Percentil de orden 95
```

¿Cuál es el primer cuartil de la concentración de ácido?

```
quantile(stackloss$Acid.Conc.,0.25)
```

¿Cuál es el valor (percentil) en el cual el 18% de las observaciones de la pérdida de amoniaco son menores y el 82% son mayores?

```
quantile(stackloss$stack.loss,0.18)
```

Finalmente, el rango intercuartílico es la extensión cubierta por la mitad central de los datos ordenados, excluyendo la cuarta parte inicial (los que son inferiores al primer cuartil) y la cuarta parte final (los que son superiores al tercer cuartil). La función IQR() calcula directamente el rango intercuartílico.

```
quantile(datos_2,0.75) - quantile(datos_2,0.25)
```

```
##  75%
## 36.5
```

```
IQR(datos_2)
```

```
## [1] 36.5
```

¿Cuál es el rango intercuartílico de la temperatura del agua?

```
IQR(stackloss$Water.Temp)
```

Varianza y desviación típica

Estas medidas miden cuán lejos difieren los datos de la media. Específicamente, expresan "el promedio de la distancia de cada punto respecto de la media". La varianza se calcula según:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2$$

donde x_i es el valor de cada observación, $\overline{X}$ es la media y N es el número total de datos. Nótese que las unidades de la varianza están expresadas al cuadrado; por tanto, si se tienen datos de longitud (en mm), la varianza resulta con unidades de superficie (en mm^2), lo cual no tiene mucho sentido. Así pues, se dispone de la desviación estándar o típica, que no es más que la raíz cuadrada de la varianza; de esta forma, las unidades de la medida de dispersión son las mismas de los datos.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2}$$

¿Cuál es la desviación típica (corregida) de la concentración de ácido?

```
sd(stackloss$Acid.Conc.)
```

```r
sum((datos_1-mean(datos_1))2)/length(datos_1)  Varianza
```

```
## [1] 1.9475
```

```r
sqrt(sum((datos_1-mean(datos_1))2)/length(datos_1))  Desviación típica
```

```
## [1] 1.395529
```

En *R*, la varianza y la desviación estándar se pueden calcular mediante las funciones var () y sd(), respectivamente; sin embargo, estas funciones utilizan $N-1$ (o $\sqrt{N-1}$) en el denominador, en lugar de N (o $\sqrt{N}$), para poderlas usar como estimadores no sesgados en inferencia estadística. Estas medidas se conocen como; varianza y la desviación típica corregidas. Por tanto, para conocer la varianza y la desviación típica sin corregir, se tienen que multiplicar por los factores $\frac{N-1}{N}$ y $\sqrt{\frac{N-1}{N}}$, respectivamente.

En *R*, la varianza y la desviación estándar se pueden calcular mediante las funciones var () y sd(), respectivamente; sin embargo, estas funciones utilizan $N-1$ (o $\sqrt{N-1}$) en el denominador, en lugar de N (o $\sqrt{N}$), para poderlas usar como estimadores no sesgados en inferencia estadística. Estas medidas se conocen como; varianza y la desviación típica corregidas. Por tanto, para conocer la varianza y la desviación típica sin corregir, se tienen que multiplicar por los factores $\frac{N-1}{N}$ y $\sqrt{\frac{N-1}{N}}$, respectivamente.

```
N = length(datos_1) var(datos_1)   Varianza corregida
```

```
## [1] 2.05
```

```
((N-1)/N)*var(datos_1)   Varianza NO corregida
```

```
## [1] 1.9475
```

```
sd(datos_1)   Desviación típica corregida
```

```
## [1] 1.431782
```

```
sqrt((N-1)/N)*sd(datos_1)   Desviación típica NO corregida
```

```
## [1] 1.395529
```

¿Cuál es la desviación típica (corregida) de la concentración de ácido?

```
sd(stackloss$Acid.Conc.)
```

¿Cuál es la varianza (corregida) del flujo de aire?

```
var(stackloss$Air.Flow)
```

El conjunto de datos *DatosCoches.txt* contiene información sobre diferentes marcas de coches. Cárgalo e imprime en pantalla las 6 primeras filas, ¿Cuál es el modelo de la cuarta muestra? (amc)

```
Datos = read.table("DatosCoches.txt",header=TRUE, sep="\t",dec=".")
head(Datos)
```

```
dim(Datos)
```

```
dim(Datos)
```

```
ni <- table(Datos$Origin)
fi <- table(Datos$Origin)/length(Datos$Origin)
Tabla_Frec = cbind(ni,fi);
Tabla_Frec
```

```
breaks <- seq(8,27,by=3);
breaks
```

```
Acel.grp <- cut(Datos$Acceleration, breaks, right=FALSE)#
ni <- table(Acel.grp) fi <- table(Acel.grp)/length(Acel.grp)
 Ni <- cumsum(ni) Fi <- cumsum(fi)
Tabla_Frec_grp = cbind(ni,fi,Ni,Fi); Tabla_Frec_grp
```

Calcula la tabla de frecuencias para las millas por galón (MPG) con los datos agrupados cada 10 unidades, comenzando por 5 (no incluido).

```
mpg <- na.omit(Datos$MPG)
breaks <- seq(5,45,by=10); breaks
```

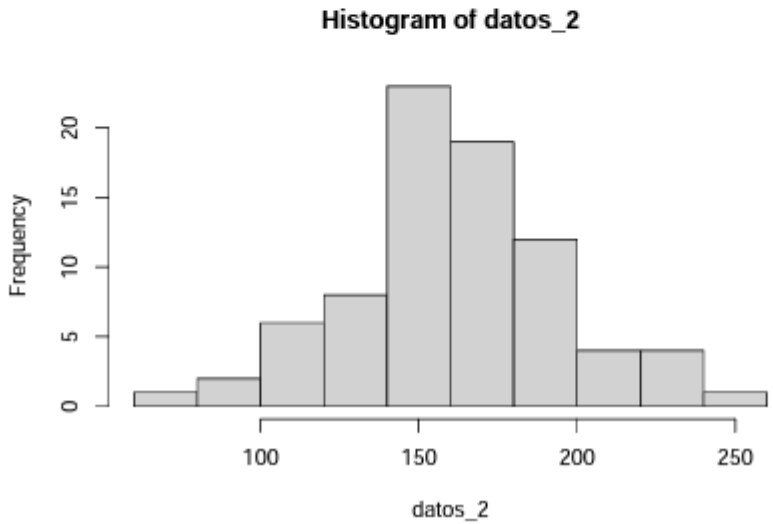
```
mpg.grp <- cut(mpg, breaks, right=TRUE)#
ni <- table(mpg.grp)
fi <- table(mpg.grp)/length(mpg.grp)
Ni <- cumsum(ni) Fi <- cumsum(fi)
Tabla_Frec_grp = cbind(ni,fi,Ni,Fi); Tabla_Frec_grp
```

Haz el histograma del desplazamiento agrupando los valores en intervalos de 50 unidades comenzando por 50. ¿Cuántas muestran tienen desplazamiento entre 200 y 250?

Histograma

Es la gráfica adecuada para representar variables cuantitativas con un gran número de valores distintos. Los datos se agrupan en intervalos y se representan gráficamente por rectángulos yuxtapuestos cuyas bases descansan sobre el eje horizontal y cuyas alturas son tales que el área de cada rectángulo es proporcional a la frecuencia de cada intervalo. Si todos los intervalos tienen igual longitud, entonces la altura de cada rectángulo es proporcional a la frecuencia del intervalo. Para evitar confusiones, la diferencia principal con el gráfico de barras es la inexistencia de espacios entre rectángulos. La función hist() permite hacer el histograma de unos datos y, además, modificar la longitud de los intervalos, si se desea. A diferencia del gráfico de barras, la función calcula automáticamente la frecuencia del intervalo. El histograma del ejemplo 2 se genera de la siguiente forma:

```
h=hist(datos_2)
```



Haz el histograma del desplazamiento agrupando los valores en intervalos de 50 unidades comenzando por 50. ¿Cuántas muestran tienen desplazamiento entre 200 y 250?

```
new_breaks = seq(50,500,by=50)
h2.wind <- hist(Datos$Displacement,breaks=new_breaks)
```

```
h2.wind$counts[4]
```

```
bp = boxplot(Datos[-c(1,2)])
```

```
mean(Datos$MPG[Datos$Origin=="USA"],na.rm = TRUE)
```

```
median(Datos$Weigth[Datos$Cylinders==8],na.rm = TRUE)
```

```
quantile(Datos$Displacement,0.25)
```

```
hp <- na.omit(Datos$Horsepower)
sd.hp.sin <- sqrt((length(hp)-1)/length(hp))*sd(hp)
```

```
pie(table(Datos$Cylinders))
```