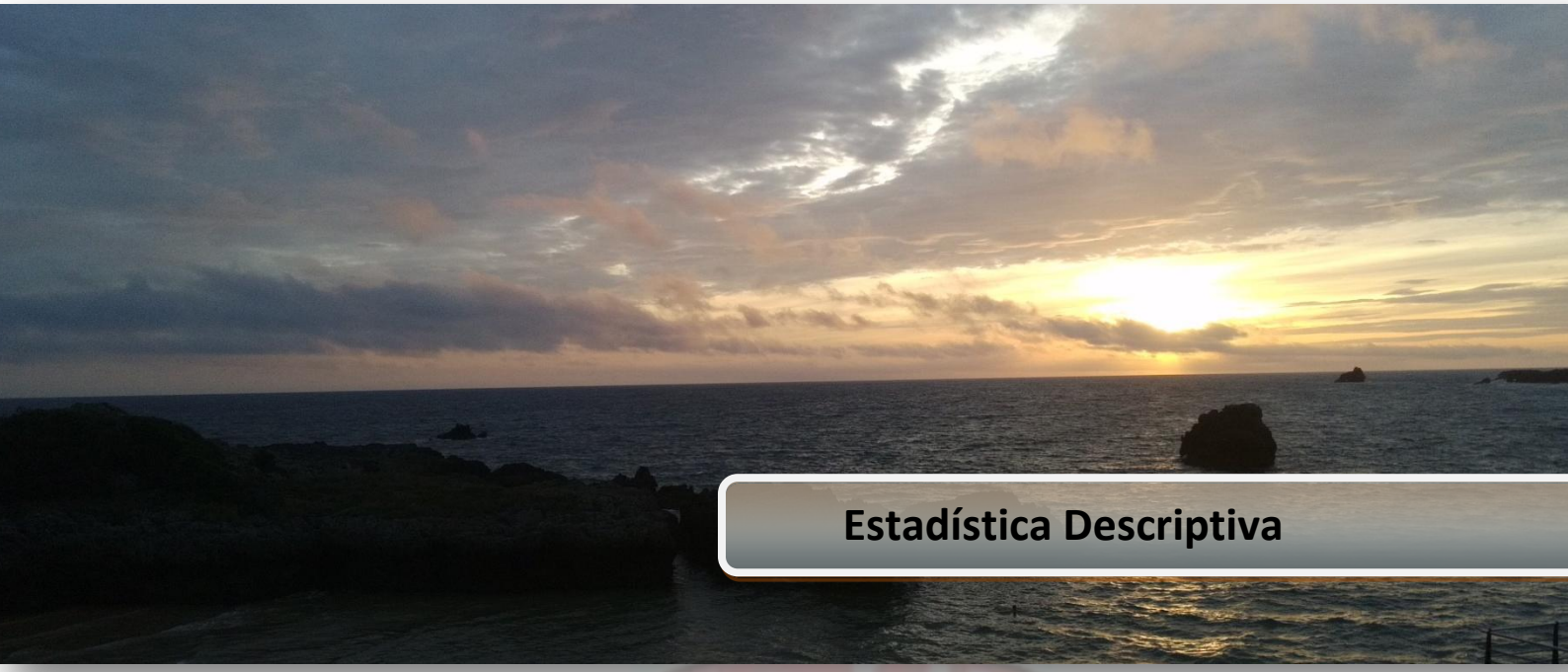


Estadística Descriptiva

1 Elementos básicos de la Estadística

A las características medidas de una muestra se les llama **estadística** muestral, y a las características medidas de una población estadística, o universo, se les llama **parámetros** de la población. En otras palabras las características de una muestra se llaman estadísticas, y las características de una población se llaman parámetros.

En estadística se conoce como ***población*** al agregado de todas la unidades individuales, sean personas, cosas..., que se hallan en una situación determinada, pudiendo ser estas finitas e infinitas. Una ***muestra*** es solo una parte de la población.

1.1 Población

Población estadística, también llamada **universo o colectivo**, es el conjunto de elementos de referencia sobre el que se realizan las observaciones.

El concepto de población en estadística va más allá de lo que comúnmente se conoce como tal. Una población se precisa como un conjunto finito o infinito de personas u objetos que presentan características comunes.

En estadística, población es el conjunto de datos de un problema estadístico determinado.

Algunas de las definiciones más aceptadas son:

"Una población es un conjunto de todos los elementos que estamos estudiando, acerca de los cuales intentamos sacar conclusiones". Levin & Rubin (1996).

"Una población es un conjunto de elementos que presentan una característica común". Cadenas (1974).

Es entonces que cuando tenemos un conjunto muy grande de datos numéricos para analizar decimos que tenemos un Universo o Población de observaciones; tiene como objetivo final descubrir las características y propiedades de aquello que generó los datos. En estadística es representado con N.

Existen distintos tipos de poblaciones:

Población base: es el grupo de personas designadas por las siguientes características: personales, geográficas o temporales, que son elegibles para participar en el estudio.

Población muestreada: es la población base con criterios de viabilidad o posibilidad de realizarse el muestreo.

Muestra estudiada: es el grupo de sujetos en el que se recogen los datos y se realizan las observaciones, siendo realmente un subgrupo de la población muestreada y accesible. El número de muestras que se puede obtener de una población es una o mayor de una.

Población diana: es el grupo de personas a la que va proyectado dicho estudio, la clasificación característica de los mismos, lo cual lo hace modelo de estudio para el proyecto establecido.

1.2 Muestra

Muestra de población, selección de un conjunto de individuos representativos de la totalidad del universo objeto de estudio, reunidos como una representación válida y de interés para la investigación de su comportamiento

Una muestra de población, en estadística, es un conjunto de datos representativos del total de una población o universo.

Los criterios que se utilizan para la selección de muestras pretenden garantizar que el conjunto seleccionado represente con la máxima fidelidad a la totalidad de la que se ha extraído, así como hacer posible la medición de su grado de probabilidad

Otras definiciones altamente aceptadas, son:

"Se llama muestra a una parte de la población a estudiar qué sirve para representarla". Murria R. Spiegel (1991).

"Una muestra es una colección de algunos elementos de la población, pero no de todos". Levin & Rubin (1996).

"Una muestra debe ser definida en base de la población determinada, y las conclusiones que se obtengan de dicha muestra solo podrán referirse a la población en referencia" Cadenas (1974).

La muestra tiene que estar protegida contra el riesgo de resultar sesgada, manipulada u orientada durante el proceso de selección, con la finalidad de proporcionar una base válida a la que se pueda aplicar la teoría de la distribución estadística. A la muestra de una población se le representa en estadística con la letra n.

Es así muestreo probabilístico, consiste en elegir una muestra de una población al azar. Podemos distinguir varios tipos de muestreo.

1.2.1 Muestreo aleatorio simple:

El procedimiento empleado es el siguiente: 1) se asigna un número a cada individuo de la población y 2) a través de algún medio mecánico (bolas dentro de una bolsa, tablas de números aleatorios, números aleatorios generados con una calculadora u ordenador, etc.) se eligen tantos sujetos como sea necesario para completar el tamaño de muestra requerido.

Este procedimiento, atractivo por su simpleza, tiene poca o nula utilidad práctica cuando la población que estamos manejando es muy grande.

1.2.2 Muestreo aleatorio sistemático:

Este procedimiento exige, como el anterior, numerar todos los elementos de la población, pero en lugar de extraer n números aleatorios sólo se extrae uno. Se parte de ese número aleatorio i, que es un número elegido al azar, y los elementos que integran la muestra son los que ocupa los lugares i, i+k, i+2k, i+3k,…, i+(n-1) k, es decir se toman los individuos de k en k, siendo k el resultado de dividir el tamaño de la población entre el tamaño de la muestra: k= N/n. El número i que empleamos como punto de partida será un número al azar entre 1 y k.

El riesgo este tipo de muestreo está en los casos en que se dan periodicidades en la población ya que al elegir a los miembros de la muestra con una periodicidad constante (k) podemos introducir una homogeneidad que no se da en la población.

1.2.3 Muestreo aleatorio estratificado:

Consiste en considerar categorías típicas diferentes entre sí (estratos) que poseen gran homogeneidad respecto a alguna característica. Lo que se pretende con este tipo de muestreo es asegurarse de que todos los estratos de interés estarán representados adecuadamente en la muestra.

Cada estrato funciona independientemente, pudiendo aplicarse dentro de ellos el muestreo aleatorio simple o el estratificado para elegir los elementos concretos que formarán parte de la muestra. La distribución de la muestra en función de los diferentes estratos se denomina afijación, y puede ser de diferentes tipos:

- Afijación Simple: A cada estrato le corresponde igual número de elementos muéstrales.
- Afijación Proporcional: La distribución se hace de acuerdo con el peso (tamaño) de la población en cada estrato.
- Afijación Óptima: Se tiene en cuenta la previsible dispersión de los resultados, de modo que se considera la proporción y la desviación típica.

1.2.4 Muestreo aleatorio por conglomerados:

El muestreo por conglomerados consiste en seleccionar aleatoriamente un cierto número de conglomerados (el necesario para alcanzar el tamaño muestral establecido) y en investigar después todos los elementos pertenecientes a los conglomerados elegidos.

En el muestreo por conglomerados la unidad muestral es un grupo de elementos de la población que forman una unidad, a la que llamamos conglomerado. Las unidades hospitalarias, los departamentos universitarios, una caja de determinado producto, etc.,

son conglomerados naturales. En otras ocasiones se pueden utilizar conglomerados no naturales como, por ejemplo, las urnas electorales. Cuando los conglomerados son áreas geográficas suele hablarse de "**muestreo por áreas**".

Las razones para estudiar muestras en lugar de poblaciones son diversas y entre ellas podemos señalar:

Ahorrar tiempo. Estudiar a menos individuos es evidente que lleva menos tiempo.

Como consecuencia del punto anterior ahorraremos costes.

Estudiar la totalidad de los pacientes o personas con una característica determinada en muchas ocasiones puede ser una tarea inaccesible o imposible de realizar.

Aumentar la calidad del estudio. Al disponer de más tiempo y recursos, las observaciones y mediciones realizadas a un reducido número de individuos pueden ser más exactas y plurales que si las tuviésemos que realizar a una población.

La selección de muestras específicas nos permitirá reducir la heterogeneidad de una población al indicar los criterios de inclusión y/o exclusión.

2 Estadística Descriptiva o Deductiva

Se refiere a la recolección, presentación, descripción, análisis e interpretación de una colección de datos, esencialmente consiste en resumir éstos con uno o dos elementos de información (medidas descriptivas) que caracterizan la totalidad de los mismos.

La Estadística Descriptiva recolecta, describe, analiza, interpreta y presenta los datos de una población en forma de tablas y gráficas

Consiste sobre todo en la presentación de datos en forma de tablas y gráficas; así que se emplea simplemente para resumir de forma numérica o gráfica un conjunto de datos. Esta comprende cualquier actividad relacionada con los datos y está diseñada para resumir o describir los mismos sin factores pertinentes adicionales; esto es, sin intentar inferir nada que vaya más allá de los datos, como tales.

La estadística Descriptiva es el método de obtener de un conjunto de datos conclusiones sobre sí mismos y no sobrepasan el conocimiento proporcionado por éstos. Puede utilizarse para resumir o describir cualquier conjunto ya sea que se trate de una población o de una muestra, cuando en la etapa preliminar de la Inferencia Estadística se conocen los elementos de una muestra.

Así pues, si aplicamos las herramientas ofrecidas por la estadística descriptiva a una muestra, solo nos limitaremos a describir los datos encontrados en dicha muestra, por lo que no se podrá generalizar la información hacia la población.

2.1 Datos en relación al tiempo.

Si se clasifica la Estadística en base al tiempo considerado, tenemos la Estadística Estática (datos de la actualidad) y la Estadística Evolutiva (datos del pasado).

Dentro de la estadística descriptiva se distinguen los datos en función al tiempo en que se encuentra analizada la población; de esta manera, tenemos 2 clasificaciones:

2.1.1 Estadística Estática o Estructural

La estadística estática o estructural, que describe la población en un momento dado empleando datos de la actualidad (por ejemplo la tasa de nacimientos en determinado censo)

2.1.2 Estadística Dinámica o Evolutiva

La estadística dinámica o evolutiva, que describe como va cambiando la población en el tiempo empleando datos del pasado (por ejemplo el aumento anual en la tasa de nacimientos).

2.2 Tipos y clasificación de Datos

Los datos estadísticos son lo que estudiamos en cada individuo de la muestra son las variables (edad, sexo, peso, talla, tensión arterial sistólica, etcétera). Los datos son los valores que toma la variable en cada caso. Lo que vamos a realizar es medir, es decir, asignar valores a las variables incluidas en el estudio. Deberemos además concretar la escala de medida que aplicaremos a cada variable

Los Datos Estadísticos, son aquellos que se estudian en cada elemento de la muestra y son variables que tomaran valores dependiendo del problema.

La naturaleza de las observaciones será de gran importancia a la hora de elegir el método estadístico más apropiado para abordar su análisis. Con este fin, clasificaremos a estos datos estadísticos, a grandes rasgos, en dos tipos: datos cuantitativos o datos cualitativos.

2.2.1 Datos cuantitativos

Son las variables que pueden medirse, cuantificarse o expresarse numéricamente y pueden ser manipulados estadísticamente. Incluyen tabulaciones de frecuencia, porcentajes, medias y promedios. Si entre cada dos datos puede haber una infinidad de ellos, se llaman continuos, y si entre un dato y otro siempre hay un hueco o salto, se llaman discretos.

Las Datos Cuantitativos son aquellos que se pueden expresar mediante valores numéricos, y se dividen en continuos (enteros y decimales) y discretos (sólo enteros):

Datos cuantitativos continuos: si admiten tomar cualquier valor dentro de un rango numérico determinado, es decir, que pueden expresarse con números decimales o fraccionarios. (Densidad de un líquido, la fuerza de un muelle, edad, peso, talla).

Datos cuantitativos discretos: si no admiten todos los valores intermedios en un rango. Suelen tomar solamente valores enteros (Nota de un examen, número de hijos, número de partos, número de hermanos, etc.).

2.2.2 Datos cualitativos.

Son datos que no se pueden expresar numéricamente, debido a que suponen cualidades, opiniones, sentimientos entre otros, y se dividen en nominales (categorías que no mantiene relación de orden) y los jerarquizados (escalas utilizadas bajo un orden).

Datos que expresan cualidades, como opiniones, sentimientos, observaciones y cambios en el comportamiento que clasifica a cada caso en una de varias categorías. La situación más sencilla es aquella en la que se clasifica cada caso en uno de dos grupos (hombre/mujer, enfermo/sano, fumador/no fumador).

Son datos dicotómicos o binarios. Como resulta obvio, en muchas ocasiones este tipo de clasificación no es suficiente y se requiere de un mayor número de categorías (color de los ojos, grupo sanguíneo, profesión, etcétera).

En el proceso de medición de estas variables, se pueden utilizar dos escalas:

Escalas nominales: ésta es una forma de observar o medir en la que los datos se ajustan por categorías que no mantienen una relación de orden entre sí (color de los ojos, sexo, profesión, presencia o ausencia de un factor de riesgo o enfermedad, etcétera).

Escalas ordinales o jerarquizados: en las escalas utilizadas, existe un cierto orden o jerarquía entre las categorías (grados de disnea, estadiaje de un tumor, etcétera).

2.3 Variables Estadísticas

Una variable estadística es cada una de las características o cualidades que poseen los individuos de la población que estamos interesados en estudiar. Se pueden clasificar en función a la Medición o a la influencia.

2.3.1 VARIABLES CUALITATIVAS Y CUANTITATIVAS

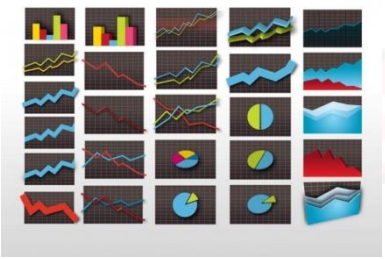
Las variables cualitativas Son las variables que expresan distintas cualidades, características o modalidad. Cada modalidad que se presenta se denomina atributo o categoría y la medición consiste en una clasificación de dichos atributos. Las variables cualitativas pueden ser dicotómicas cuando sólo pueden tomar dos valores posibles como sí y no, hombre y mujer o son politómicas cuando pueden adquirir tres o más valores. Podemos distinguir dos tipos:

Variable cualitativa nominal: presenta modalidades no numéricas que no admiten un criterio de orden.

Por ejemplo: El estado civil, con las siguientes modalidades: soltero, casado, separado, divorciado y viudo.

Variable cualitativa ordinal o variable cuasicuantitativa: La variable puede tomar distintos valores ordenados siguiendo una escala establecida, aunque no es necesario que el intervalo entre mediciones sea uniforme, por ejemplo: leve, moderado, fuerte; o la nota en un examen: suspenso, aprobado, notable, sobresaliente.

Una variable cuantitativa es la que se expresa mediante un número, por tanto se pueden realizar operaciones aritméticas con ella. Podemos distinguir dos tipos:



Variable discreta: Una variable discreta es aquella que toma valores aislados, es decir no admite valores intermedios entre dos valores específicos. Es decir, sólo puede ser expresado con números enteros.

Por ejemplo: El número de hermanos de 5 amigos: 2, 1, 0, 1, 3.

Variable continua: Una variable continua es aquella que puede tomar valores comprendidos entre dos números por lo cual tiene un número infinito de valores posibles. Es decir, puede ser expresada con números decimales o fraccionarios.

Por ejemplo: La altura de los 5 amigos: 1.73, 1.82, 1.77, 1.69, 1.75.

En la práctica medimos la altura con dos decimales, pero también se podría dar con tres decimales.

2.3.2 Variables dependientes e independientes

Variables independientes son las que el investigador escoge para establecer agrupaciones en el estudio, clasificando intrínsecamente a los casos del mismo. Un tipo especial son las variables de control, que modifican al resto de las variables

independientes y que de no tenerse en cuenta adecuadamente pueden alterar los resultados por medio de un sesgo.

Es aquella característica o propiedad que se supone ser la causa del fenómeno estudiado. En investigación experimental se llama así a la variable que el investigador manipula.

Variables dependientes son las variables de respuesta que se observan en el estudio y que podrían estar influenciadas por los valores de las variables independientes. Hayman la define como propiedad o característica que se trata de cambiar mediante la manipulación de la variable independiente. La variable dependiente es el factor que es observado y medido para determinar el efecto de la variable independiente

3 Estadística descriptiva con R

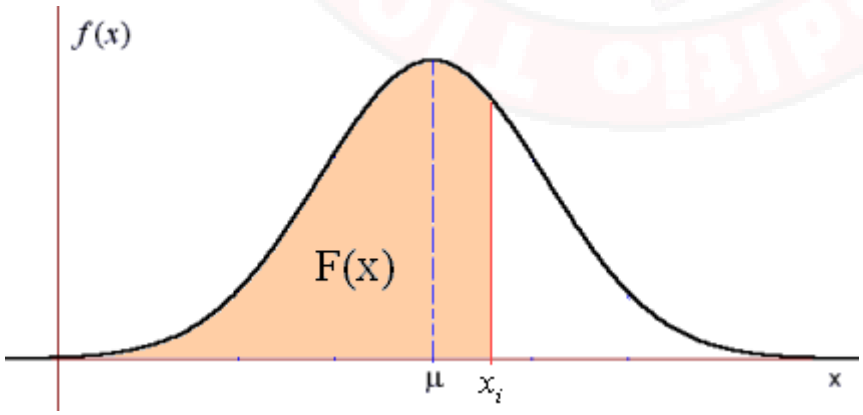
3.1 Distribución de probabilidad y Función de densidad de una v.a.

Una variable aleatoria puede tomarse como una cantidad cuyo valor no es fijo pero puede tomar diferentes valores; una distribución de probabilidad se usa para describir la probabilidad de que se den los diferentes valores (se denota usualmente por F(x)).

$Fx\ (x) = P\ (X \leq x)$

La distribución de probabilidad de una v.a. describe teóricamente la forma en que varían los resultados de un experimento aleatorio. Intuitivamente se trataría de una lista de los resultados posibles de un experimento con las probabilidades que se esperarían ver asociadas con cada resultado.

La función de densidad de probabilidad, función de densidad, o, simplemente, densidad de una variable aleatoria continua es una función, usualmente denominada f(x) que describe la densidad de la probabilidad en cada punto del espacio de tal manera que la probabilidad de que la variable aleatoria tome un valor dentro de un determinado conjunto sea la integral de la función de densidad sobre dicho conjunto.



Función de densidad de una distribución

3.2 Parámetros y estadísticos

Parámetro: Es una cantidad numérica calculada sobre una población.

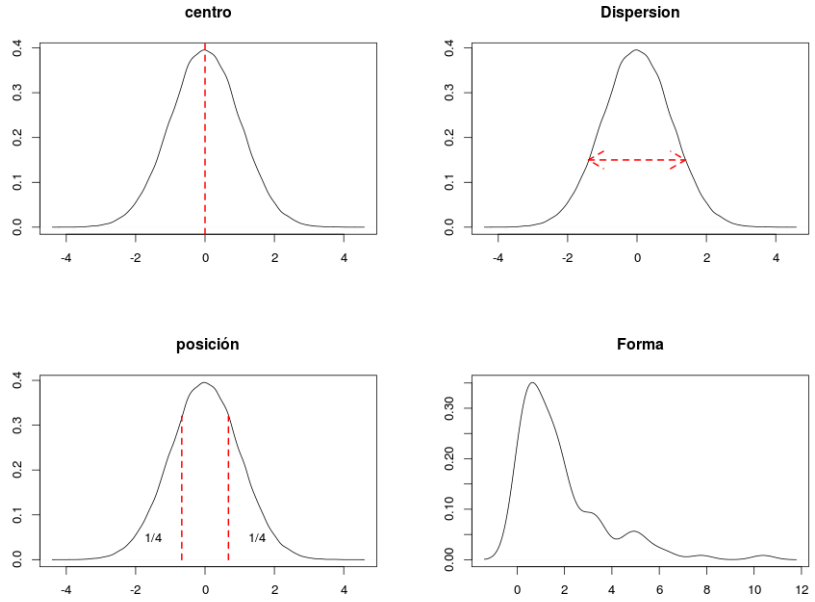
- La altura media de los individuos de un país.

Estadístico: Es una cantidad numérica calculada sobre una muestra.

- La altura media de los que estamos en este aula.

Si un estadístico se usa para aproximar un parámetro también se le suele llamar estimador.

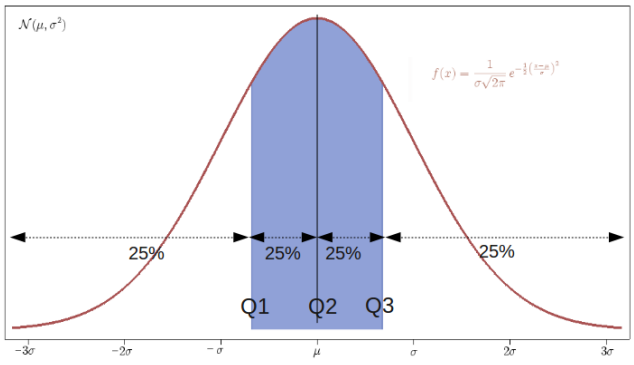
Los estadísticos se calculan, y estos estiman parámetros. Hay diferentes tipos según las cosas que queramos saber de la distribución de una variable.



Tipos de estadísticos

3.2.1 Medidas de posición.

Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos. Las más populares: Cuantiles, percentiles, cuartiles, deciles.



El cuantil de orden *p* de una distribución (con $0 < p < 1$) es el valor de la variable x_p que marca un corte de modo que una proporción p de valores de la población es menor o igual que x_p. Por ejemplo, el cuantil de orden 0,3 dejaría un 30% de valores por debajo y el cuantil de orden 0,50 se corresponde con la **mediana** de la distribución.

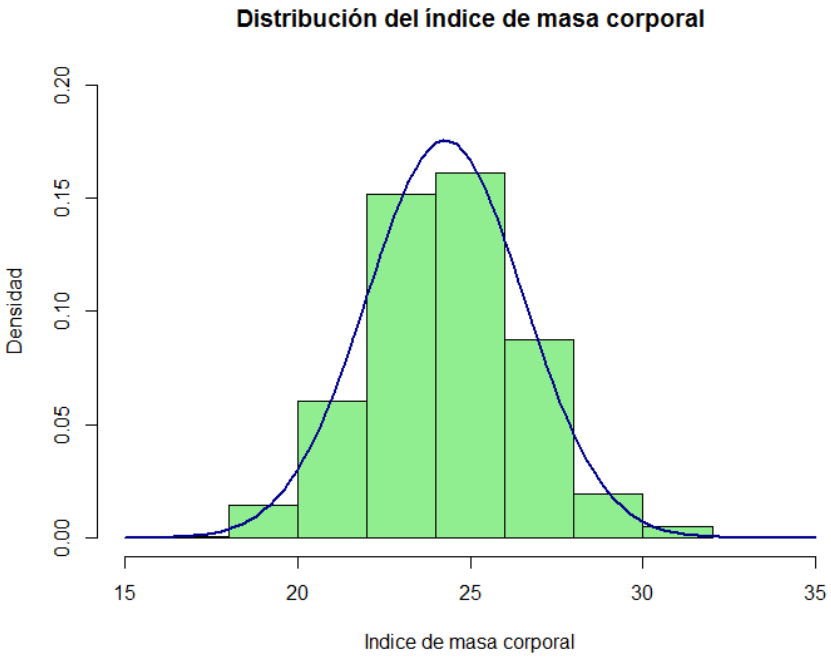
Los cuantiles suelen usarse por grupos que dividen la distribución en partes iguales; entendidas estas como intervalos que comprenden la misma proporción de valores. Los más usados son:

- Los **cuartiles**, que dividen a la distribución en cuatro partes (corresponden a los cuantiles 0,25; 0,50 y 0,75);

- Los **quintiles**, que dividen a la distribución en cinco partes (corresponden a los cuantiles 0,20; 0,40; 0,60 y 0,80);
- Los **deciles**, que dividen a la distribución en diez partes;
- Los **percentiles**, que dividen a la distribución en cien partes.

Ejemplo: El 5% de los españoles se consideran que tienen infrapeso con riesgo de anoráxia. ¿Qué Indice de Masa Corporal se considera "demasiado bajo" o infrapeso? Percentil 5 o cuantil 0,05.

```
>BMI<-rnorm(n=1000, m=24.2, sd=2.2)
>quantile(BMI, 0.05)
 5%
20.41249
```



Representación gráfica de la función de distribución

```
>hist(BMI, freq=FALSE, xlab="Indice de masa corporal", ylab="Densidad",
 + main="Distribución del índice de masa corporal", col="lightgreen",
 + xlim=c(15,35),  ylim=c(0, .20),breaks=10)
>curve(dnorm(x, mean=mean(BMI), sd=sd(BMI)), add=TRUE, col="darkblue", lwd=2)
```

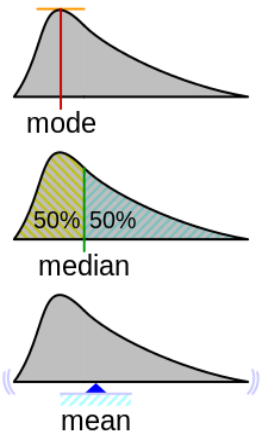
Medidas de centralización o tendencia central:

Indican valores con respecto a los que los datos "parecen" agruparse.

La media, moda y mediana son parámetros característicos de una distribución de probabilidad La media se confunde a veces con la mediana o moda.; sin embargo, para las distribuciones con sesgo, la media no es necesariamente el mismo valor que la mediana o que la moda

Media aritmética

Según la Real Academia Española (2001) *«[…] resulta al efectuar una serie determinada de operaciones con un conjunto de números y que, en determinadas condiciones, puede representar por sí solo a todo el conjunto»*. Existen distintos tipos de medias, tales como la media geométrica, la media ponderada y la media armónica aunque en el lenguaje común, el término se refiere generalmente a la media aritmética.



La media aritmética es el promedio de un conjunto de valores, o su distribución que a menudo se denomina "promedio".

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

La media aritmética "mean" es la suma de los valores dividido por el tamaño muestral.

```
>x  <- c(1,  2,  3,  4,5  )
>mean(x)
3
>media <-  (1  +  2  +3  +  4  +  5)/5
>media
3
```

La media aritmética se trata de un parámetro conveniente cuando los datos se concentran simétricamente con respecto a ese valor. Muy sensible a valores extremos (en estos casos hay otras 'medias', menos intuitivas, pero que pueden ser útiles: media aritmética, geométrica, ponderada...)

Mediana

Representa el valor de la variable de posición central en un conjunto de datos ordenados.

Mediana("median"): Es un valor que divide a las observaciones en dos grupos con el mismo número de individuos (percentil 50). Si el número de datos es par, se elige la media de los dos datos centrales.

Mediana de 1,2,4,5,6,6,8 es 5.

Mediana de 1,2,4,5,6,6,8,9 es 5.5

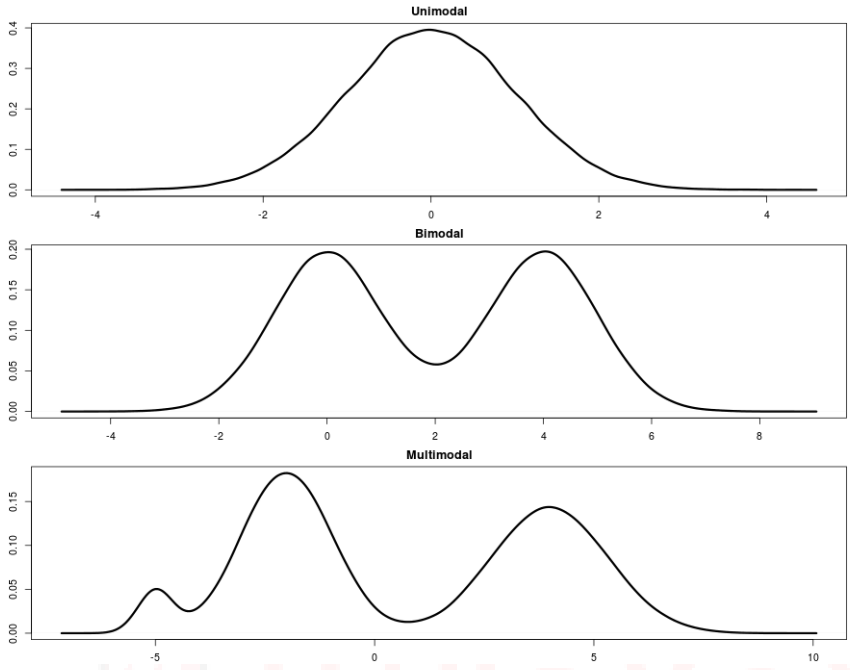
Es conveniente cuando los datos son asimétricos. No es sensible a valores extremos.

Mediana de 1,2,4,5,6,6, 800 es 5. (¡La media es 117,7!)

```
> y <- c(1, 2, 4, 5, 6, 6, 8, 9)
> z <- c(1, 2, 4, 5, 6, 6, 800)
> median(y)
[1] 5.5
> median(z)
[1] 5
```

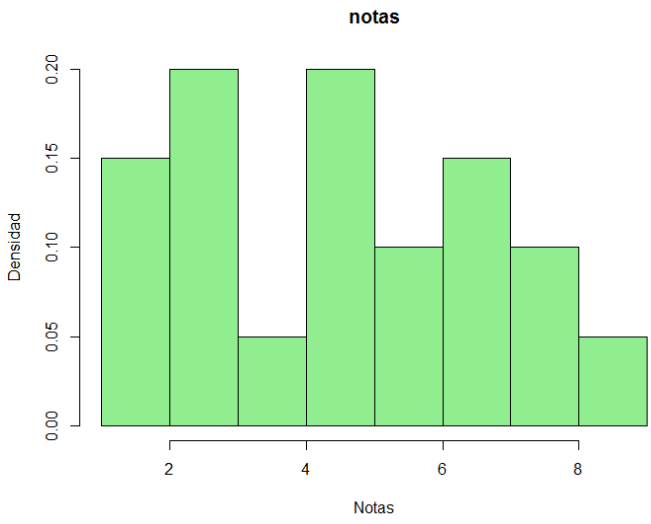
Moda

La moda es el valor con una mayor frecuencia en una distribución de datos. Se hablará de una distribución bimodal de los datos cuando encontremos dos modas, es decir, dos datos que tengan la misma frecuencia absoluta máxima.



Para calcular la moda de una distribución tenemos que utilizar un paquete específico. En este caso utilizamos el paquete "modeest"

```
> library(modeest)
> notas<-c(3,3,5,2,7,8,5,7,6,4,5,6,5,3,9,7,8,3,1,1)
> mfv(notas)
[1] 3 5
> hist(notas, freq=FALSE, xlab="Notas", ylab="Densidad",
+ main="notas", col="lightgreen",breaks=11)
```



3.2.2 Medidas de dispersión

Las medidas de posición resumen la distribución de datos, pero resultan insuficientes y simplifican excesivamente la información. Estas medidas adquieren verdadero significado cuando van acompañadas de otras que informen sobre la heterogeneidad de los datos.

Los parámetros de dispersión indican, de un modo bien definido, lo homogéneos que estos datos son. Hay medidas de dispersión absolutas, entre las cuales se encuentran la varianza, la desviación típica o la desviación media, aunque también existen otras menos utilizadas como los recorridos o la meda; y medidas de dispersión relativas, como el coeficiente de variación, el coeficiente de apertura o los recorridos relativos.

Rango

Rango de una variable estadística es la diferencia entre el mayor y el menor valor que toma la misma. Es la medida de dispersión más sencilla de calcular, aunque es algo burda porque sólo toma en consideración un par de observaciones. Basta con que uno de estos dos datos varíe para que el parámetro también lo haga, aunque el resto de la distribución siga siendo, esencialmente, la misma.

Existen otros parámetros dentro de esta categoría, como los recorridos o rangos intercuantílicos, que tienen en cuenta más datos y, por tanto, permiten afinar en la dispersión. Entre los más usados está el **rango intercuartílico**, que se define como la diferencia entre el cuartil tercero y el cuartil primero. En ese rango están, por la propia definición de los cuartiles, el 50% de las observaciones. Este tipo de medidas también se usa para determinar valores atípicos. En el diagrama de caja que aparece a la derecha se marcan como valores atípicos todos aquellos que caen fuera del intervalo $[L_i, L_s] = [Q_1 - 1{,}5 \cdot R_s, Q_3 + 1{,}5 \cdot R_s]$, donde Q_1 y Q_3 son los cuartiles 1º y 3º, respectivamente, y R_s representa la mitad del recorrido o rango intercuartílico, también conocido como **recorrido semiintercuartílico**

Desviaciones medias

Dada una variable estadística X y un parámetro de tendencia central, c, se llama desviación de un valor de la variable, xi, respecto de c, al número $|xi - c|$. Este número mide lo lejos que está cada dato del valor central c, por lo que una media de esas medidas podría resumir el conjunto de desviaciones de todos los datos.

Así pues, se denomina desviación media de la variable X respecto de c a la media aritmética de las desviaciones de los valores de la variable respecto de c, esto es, si

$$X = x_1, x_2, \ldots, x_n, \text{ entonces } DM_c = \frac{\sum_{i=1}^{n} |x_i - c|}{n}$$

De este modo se definen la desviación media respecto de la media (c = $\overline{x}$) o la desviación media respecto de la mediana (c = $\overline{Me}$), cuya interpretación es sencilla en virtud del significado de la media aritmética

Sin embargo, el uso de valores absolutos impide determinados cálculos algebraicos que obligan a desechar estos parámetros,

Varianza y desviación típica

La suma de todas las desviaciones respecto al parámetro más utilizado, la media aritmética, es cero. Por tanto si se desea una medida de la dispersión sin los inconvenientes para el cálculo que tienen las desviaciones medias, una solución es elevar al cuadrado tales desviaciones antes de calcular el promedio.

Se define la **varianza** como:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n},$$

o sea, la media de los cuadrados de las desviaciones respecto de la media.

La desviación típica, σ, se define como la raíz cuadrada de la varianza, esto es,

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

Tiene las mismas unidades que la variable.

Coeficiente de variación de Pearson

Es la razón entre la desviación típica y la media. Se define como:

$$C_V = \frac{\sigma}{\bar{x}}$$

Donde σ es la desviación típica y $\bar{x}$ es la media aritmética.

Se interpreta como el número de veces que la media está contenida en la desviación típica. Suele darse su valor en tanto por ciento, multiplicando el resultado anterior por 100. De este modo se obtiene un porcentaje de la variabilidad. Mide la desviación típica en forma de "qué tamaño tiene con respecto a la media". También se la denomina variabilidad relativa. Es una cantidad adimensional. Interesante para comparar la variabilidad de diferentes variables.

Su principal inconveniente es que en el caso de distribuciones cuya media se acerca a cero, su valor tiende a infinito e incluso resulta imposible de calcular cuando la media es cero. Por ello no puede usarse para variables tipificadas. P.e. Si el peso tiene CV=30 % y la altura tiene CV=10 %, los individuos presentan más dispersión en peso que en altura.

```
> pesos <- rnorm(1000,  3,  0.8)
> range(pesos)
[1] 0.4445757 5.5031138
> IQR  <- (quantile(pesos,  0.75,  names  =  F) - quantile(pesos,
+ 0.25,  names  =  F))
> IQR
[1] 1.088238
> var(pesos)
[1] 0.6307161
> sd(pesos)
[1] 0.7941764
```

```
> CV <- sd(pesos)/mean(pesos)
> CV
[1] 0.2634968
```

El **índice de Gini** o **coeficiente de Gini** es un parámetro de dispersión usado para medir desigualdades entre los datos de una variable o la mayor o menor concentración de los mismos.

Este coeficiente mide de qué forma está distribuida la suma total de los valores de la variable. Se suele usar para describir salarios. Los casos extremos de *concentración* serían aquel en los que una sola persona acapara el total del dinero disponible para salarios y aquel en el que este total está igualmente repartido entre todos los asalariados.

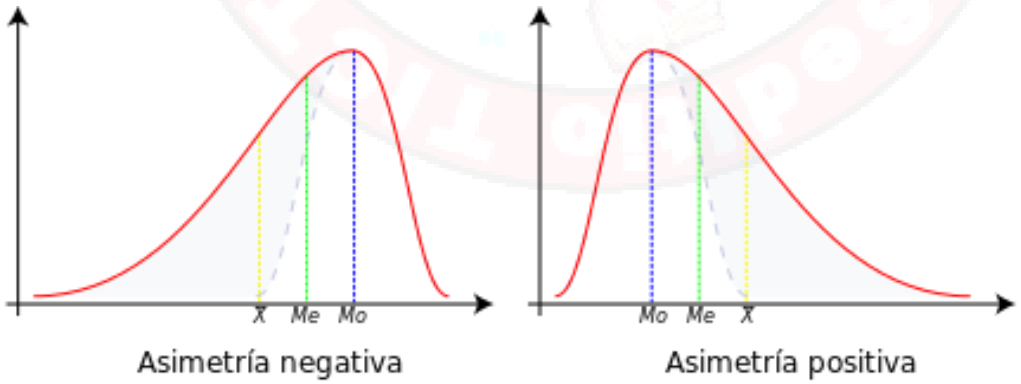
3.2.3 Medidas de forma

Las medidas de forma caracterizan la forma de la gráfica de una distribución de datos estadísticos. La mayoría de estos parámetros tiene un valor que suele compararse con la campana de Gauss, esto es, la gráfica de la distribución normal, una de las que con más frecuencia se ajusta a fenómenos reales.

Medidas de asimetría

Se dice que una distribución de datos estadísticos es simétrica cuando la línea vertical que pasa por su media, divide a su representación gráfica en dos partes simétricas. Ello equivale a decir que los valores equidistantes de la media, a uno u otro lado, presentan la misma frecuencia.

En las distribuciones simétricas los parámetros media, mediana y moda coinciden, mientras que si una distribución presenta cierta asimetría, de un tipo o de otro, los parámetros se sitúan como muestra el siguiente gráfico:



Asimetría negativa Asimetría positiva

La posición relativa de los parámetros de centralización pueden servir como una primera medida de la simetría de una distribución.

La asimetría resulta útil en muchos campos. Muchos modelos simplistas asumen una distribución normal, esto es, simétrica en torno a la media. La distribución normal tiene una asimetría cero. Pero en realidad, los valores no son nunca perfectamente simétricos y la asimetría de la distribución proporciona una idea sobre si las desviaciones de la media son

positivas o negativas. Una asimetría positiva implica que hay más valores distintos a la derecha de la media.

Las medidas de asimetría, sobre todo el coeficiente de asimetría de Fisher, junto con las medidas de apuntamiento o curtosis se utilizan para contrastar si se puede aceptar que una distribución estadística sigue la distribución normal. Esto es necesario para realizar numerosos contrastes estadísticos en la teoría de inferencia estadística.

Coeficiente de asimetría de Pearson

Sólo se puede utilizar en distribuciones uniformes, unimodales y moderadamente asimétricas. Se basa en que en distribuciones simétricas la media de la distribución es igual a la moda.

$$A_p = \frac{\mu - moda}{\sigma},$$

Donde μ es el momento central de orden 1, que corresponde a la media aritmética de la variable X.

Si la distribución es simétrica, μ=moda y $A_p = 0$. Si la distribución es asimétrica positiva la media se sitúa por encima de la moda y, por tanto, $A_p > 0$.

Coeficiente de asimetría de Bowley

Está basado en la posición de los cuártiles y la mediana, y utiliza la siguiente expresión:

$$A_B = \frac{Q_{3/4} + Q_{1/4} - 2Me}{Q_{3/4} - Q_{1/4}}$$

En una distribución simétrica el tercer cuartil estará a la misma distancia de la mediana que el primer cuartil. Por tanto A_B=0.

Si la distribución es positiva o a la derecha, $A_B > 0$.

Medidas de apuntamiento o curtosis

Con estos parámetros se pretende medir cómo se reparten las frecuencias relativas de los datos entre el centro y los extremos, tomando como comparación la campana de Gauss.

El parámetro usado con más frecuencia para esta medida es el **coeficiente de curtosis de Fisher**, definido como:

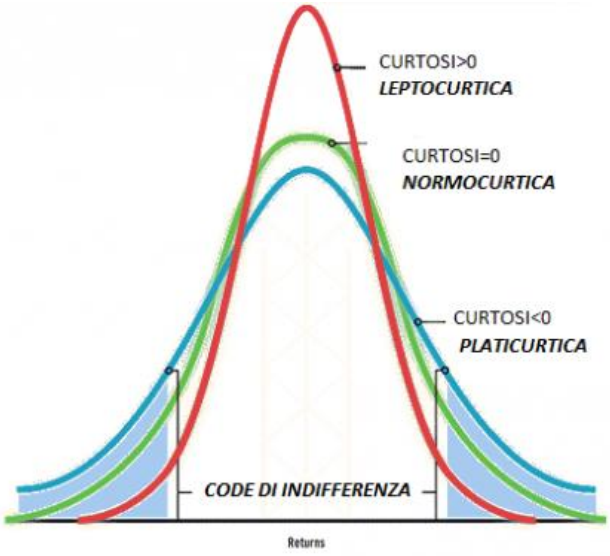
$$\gamma_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n\sigma^4} - 3,$$

aunque hay otros como el **coeficiente de curtosis de Kelley** o el **coeficiente de curtosis percentílico**.

La comparación con la distribución normal permite hablar de distribuciones **platicúrticas** o más aplastadas que la normal; distribuciones **mesocráticas**, con igual apuntamiento que la normal; y distribuciones **leptocúrticas**, esto es, más apuntadas que la normal.

- Platicúrtica (aplanada): curtosis < 0

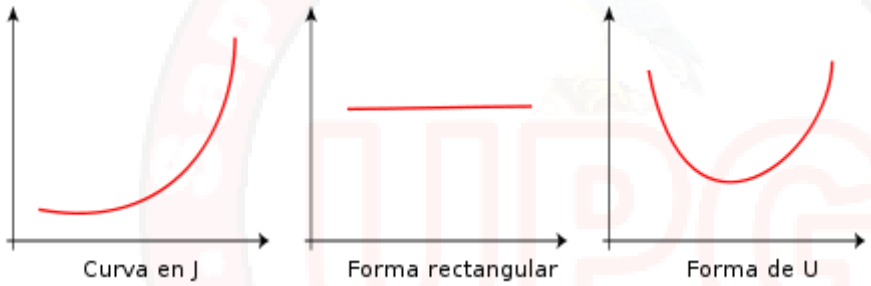
- Mesocúrtica (como la normal): curtosis = 0
- Leptocúrtica (apuntada): curtosis > 0



Regla aproximativa (para ambos estadísticos).

- Curtosis y/o coeficiente de asimetría entre -1 y 1, es generalmente considerada una muy ligera desviación de la normalidad.
- Entre -2 y 2 tampoco es malo del todo, según el caso.

Por último, existen otras medidas para decidir sobre la forma de una distribución con ajuste a modelos menos usuales como los que se muestran en las siguientes gráficas:



3.2.4 Otros parámetros

Proporción

La proporción de un dato estadístico es el número de veces que se presenta ese dato respecto al total de datos. Se conoce también como frecuencia relativa y es uno de los parámetros de cálculo más sencillo. Tiene la ventaja de que puede calcularse para variables cualitativas.

El dato con mayor proporción se conoce como moda.

En inferencia estadística existen intervalos de confianza para la estimación de este parámetro.

Número índice

Un número índice es una medida estadística que permite estudiar las fluctuaciones o variaciones de una magnitud o de más de una en relación al tiempo o al espacio. Los índices más habituales son los que realizan las comparaciones en el tiempo. Algunos ejemplos de uso cotidiano de este parámetro son el índice de precios o el IPC

Tasa

La tasa es un coeficiente que expresa la relación entre la cantidad y la frecuencia de un fenómeno o un grupo de fenómenos. Se utiliza para indicar la presencia de una situación que no puede ser medida en forma directa.31 Esta razón se utiliza en ámbitos variados, como la demografía o la economía, donde se hace referencia a la tasa de interés.

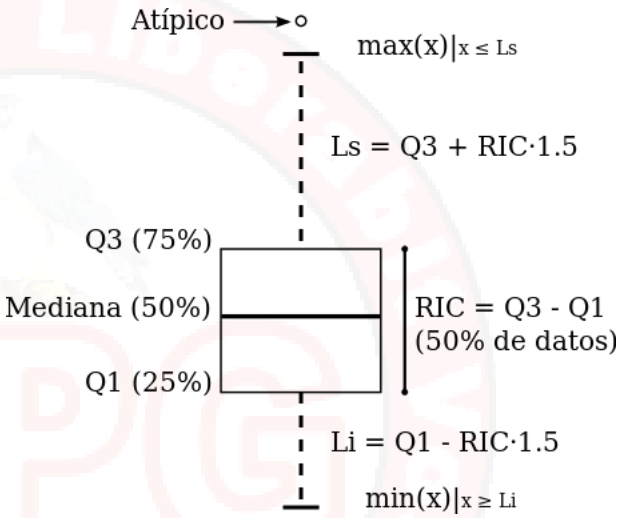
Algunos de los más usados son: tasa de natalidad, tasa de mortalidad, tasa de crecimiento demográfico, tasa de fertilidad o tasa de desempleo.

4 Gráficos en estadística descriptiva

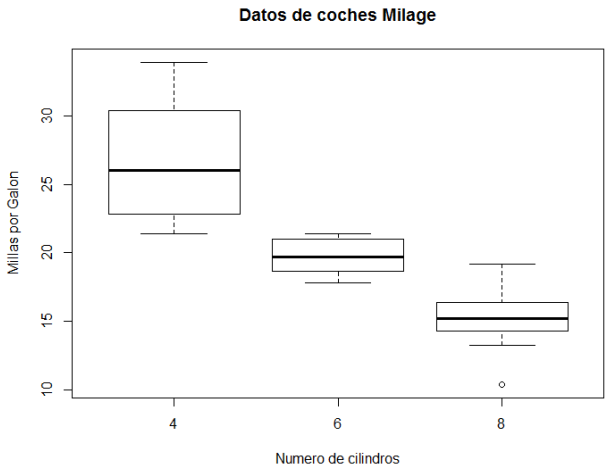
4.1 Diagrama de cajas o Boxplot

Un diagrama de caja, John Tukey (1977), es un gráfico, basado en cuartiles, mediante el cual se visualiza un conjunto de datos. Está compuesto por un rectángulo (la caja) y dos brazos (los bigotes).

También llamados 'diagramas de cajas y bigotes'.



Boxplot o diagrama de caja y bigotes



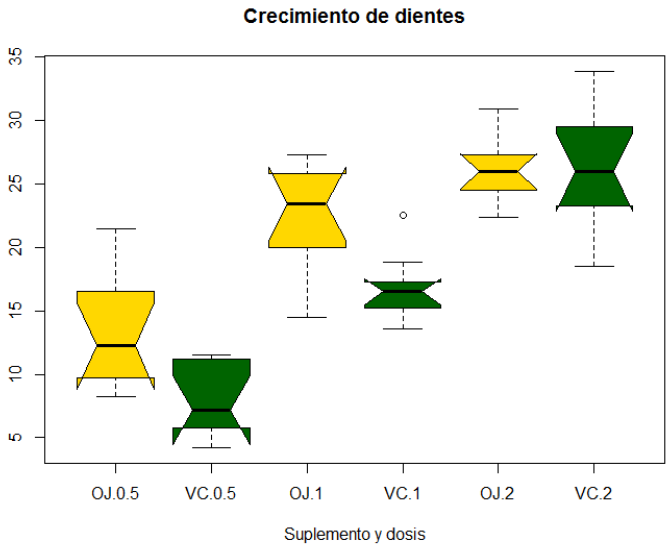
Ejemplo de Boxplot

Es un gráfico que suministra información sobre los valores mínimo y máximo, los cuartiles Q1, Q2 o mediana y Q3, y sobre la existencia de valores atípicos y simetría de la distribución.

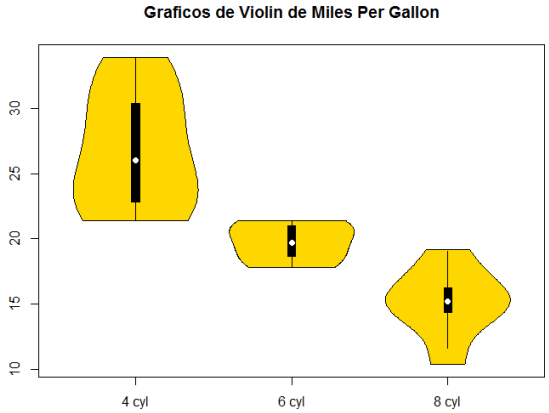
Los valores atípicos son los inferiores a Li y los superiores a Ls. Proporcionan una visión general de la simetría de la distribución de los datos, si la media no está en el centro del rectángulo, la distribución no es simétrica. Son útiles para ver la presencia de valores atípicos. Muy útiles para comparar distribuciones.

```
# Boxplot de MPG
>boxplot(mpg~cyl,data=mtcars, main="Datos de coches Milage",
       + xlab="Numero de cilindros", ylab="Millas por Galon")
```

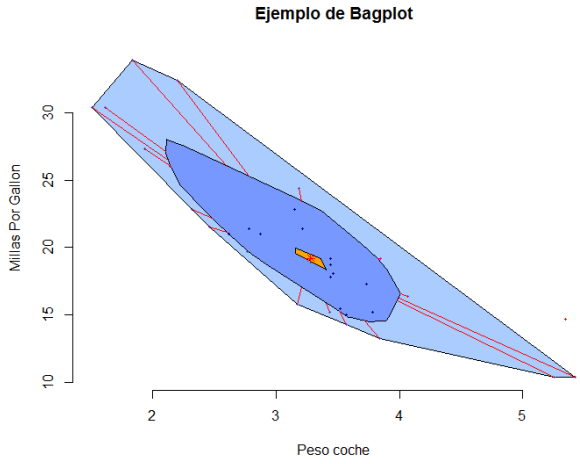
```
# Boxplot de Tooth Growth con dos factores
# cajas coloreadas para mejor interpretación
> boxplot(len~supp*dose, data=ToothGrowth, notch=TRUE,
+ col=(c("gold","darkgreen")),
+ main="Crecimiento de dientes", xlab="Suplemento y dosis")
```




```
# Gráficos de tipo Violin
> library(vioplot)
> x1 <- mtcars$mpg[mtcars$cyl==4]
> x2 <- mtcars$mpg[mtcars$cyl==6]
> x3 <- mtcars$mpg[mtcars$cyl==8]
> vioplot(x1, x2, x3, names=c("4 cyl", "6 cyl", "8 cyl"),
+ col="gold")
> title("Graficos de Violin de Miles Per Gallon")
```

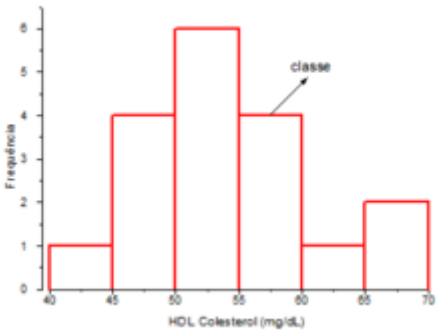


```
# Ejemplo de Bagplot
> library(aplpack)
> attach(mtcars)
> bagplot(wt,mpg, xlab="Peso coche", ylab="Millas Por Gallon",
+ main="Ejemplo de Bagplot")
```



4.2 Histograma

Un histograma es una representación gráfica de una variable en forma de barras, donde la altura de cada barra es proporcional a la frecuencia de los valores representados, ya sea en forma diferencial o acumulada. Sirven para obtener una "primera vista" general, o panorama, de la distribución de la población, o la muestra, respecto a una característica, cuantitativa y continua, de la misma y que es de interés para el observador.



En el eje vertical se representan las frecuencias, es decir, la cantidad de población o la muestra, según sea el caso, que se ubica en un determinado valor o subrango de valores de la característica que toma la característica de interés.

Así pues, podemos evidenciar comportamientos, observar el grado de homogeneidad, acuerdo o concisión entre los valores de todas las partes que componen la población o la muestra, o, en contraposición, poder observar el grado de variabilidad, y por ende, la dispersión de todos los valores que toman las partes, también es posible no evidenciar ninguna tendencia y obtener que cada miembro de la población toma por su lado y adquiere un valor de la característica aleatoriamente sin mostrar ninguna preferencia o tendencia, entre otras cosas.

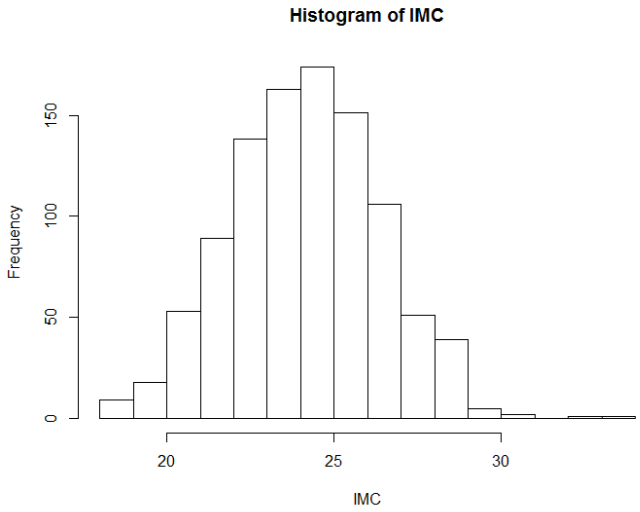
En general se utilizan para relacionar variables cuantitativas continuas, pero también se suele usar para variables cuantitativas discretas, en cuyo caso es común llamarlo diagrama de frecuencias y sus barras están separadas, esto es porque en el "x" ya no se representa un espectro continuo de valores, sino valores cuantitativos específicos como

ocurre en un diagrama de barras cuando la característica que se representa es cualitativa o categórica. Su utilidad se hace más evidente cuando se cuenta con un gran número de datos cuantitativos y que se han agrupado en intervalos de clase.

Representar histogramas en R es tan sencillo como crear un objeto hist, con la función hist().

Ejemplo. Vamos a representar el Indice de Mas Corporal (IMC) que se comporta como una distribución normal de media 24.2 y desviación típica de 2.2

```
## Representación de Histogramas
## Ejemplo Indice de Masa Corporal
> IMC<-rnorm(n=1000, m=24.2, sd=2.2)
> hist(IMC)
```



Histograma de Indice de Masa Corporal

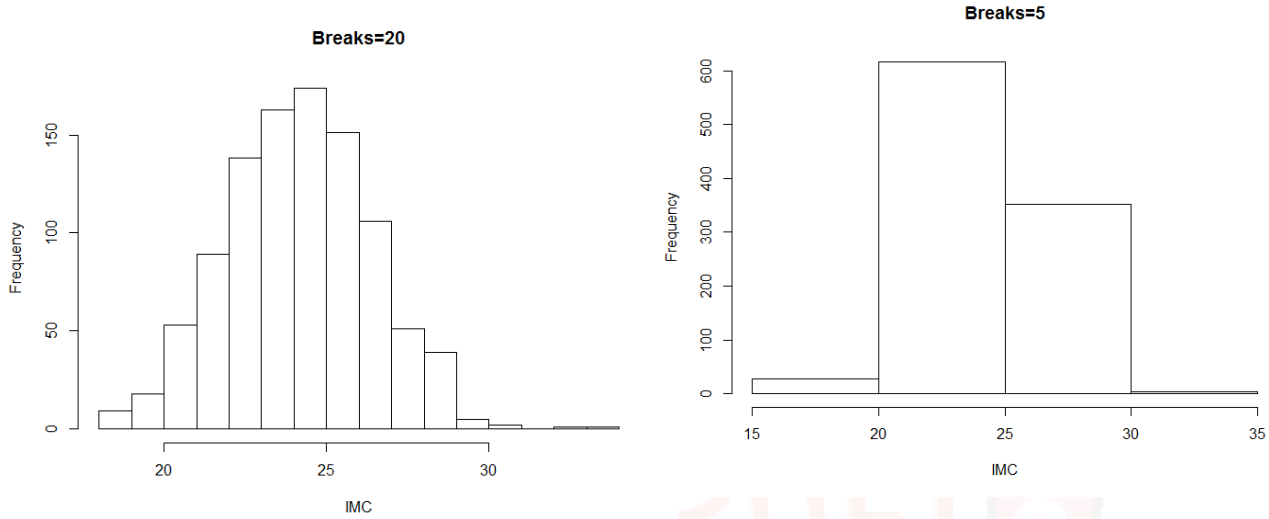
Si queremos obtener información sobre cualquier histograma, basta con poner lo siguiente:

```
#Información del Histograma
> histinfo<-hist(IMC)
> histinfo
$breaks
 [1] 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
$counts
 [1]   9  18  53  89 138 163 174 151 106  51  39   5   2   0   1   1
$density
 [1] 0.009 0.018 0.053 0.089 0.138 0.163 0.174 0.151 0.106 0.051 0.039 0.005
[13] 0.002 0.000 0.001 0.001
$mids
 [1] 18.5 19.5 20.5 21.5 22.5 23.5 24.5 25.5 26.5 27.5 28.5 29.5 30.5 31.5 32.5
[16] 33.5
$xname
```

```
[1] "IMC"
$equidist
[1] TRUE
attr(,"class")
[1] "histogram"
```

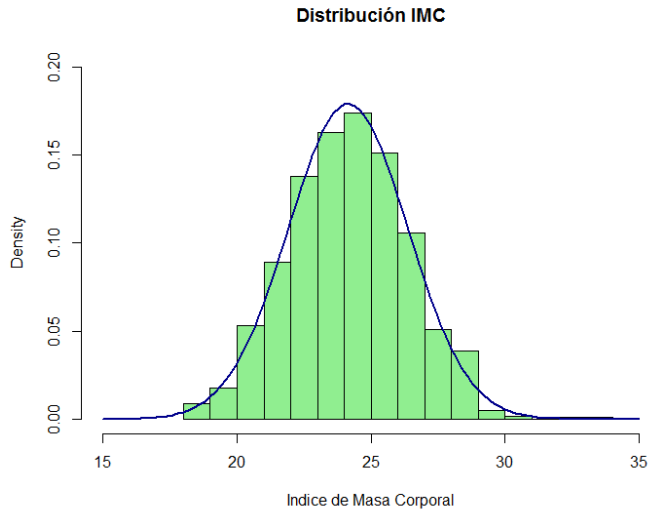
Podemos cambiar el numero de clases con el parámetro break()

```
>hist(IMC, breaks=20, main="Breaks=20")
>hist(IMC, breaks=5, main="Breaks=5")
```



Podemos añadir titulo, nombre de ejes y curva de distribución:

```
> hist(IMC, freq=FALSE, xlab="Indice de Masa Corporal",
+ main="Distribución IMC", col="lightgreen", xlim=c(15,35),  ylim=c(0, .20))
> curve(dnorm(x, mean=mean(IMC), sd=sd(IMC)), add=TRUE, col="darkblue", lwd=2)
```



Función de distribución del Indice de Masa Corporal (IMC)

4.3 Gráfica de tallos y hojas

El diagrama "tallo y hojas" (Stem-and-Leaf Diagram) permite obtener simultáneamente una distribución de frecuencias de la variable y su representación gráfica. Para construirlo basta separar en cada dato el último dígito de la derecha (que constituye la hoja) del bloque de cifras restantes (que formará el tallo).

Esta representación de los datos es semejante a la de un histograma pero además de ser fáciles de elaborar, presentan más información que estos

Permite la descripción de los datos agrupados en filas y columnas donde recuenta la frecuencia hasta la fila donde se encuentra la mediana, señalada por medio de paréntesis ():

```
> stem.leaf(AAA)
1 | 2: represents 12
 leaf unit: 1
            n: 52
   1    2. | 6
   4    3* | 122
  10    3. | 899999
  18    4* | 01112344
  (9)   4. | 566667789
  25    5* | 011333344
  16    5. | 55788
  11    6* | 000011134
   2    6. | 59
```

4.4 Diagrama de dispersión

Un diagrama de dispersión o gráfica de dispersión o gráfico de dispersión es un tipo de diagrama matemático que utiliza las coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos. Los datos se muestran como un conjunto de puntos, cada uno con el valor de una variable que determina la posición en el eje horizontal (x) y el valor de la otra variable determinado por la posición en el eje vertical (y) Muestra conjuntamente datos de dos variables (en X y en Y) para ver su correlación, y permite considerar grupos (niveles de un factor)

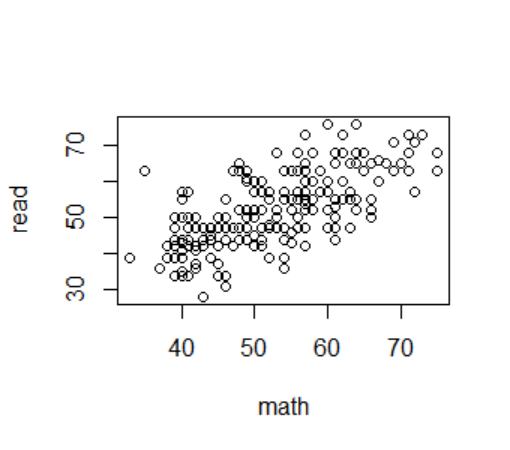
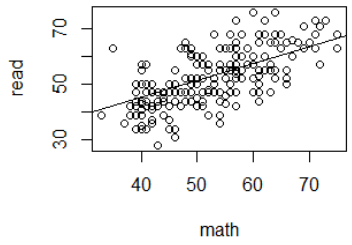
Se emplea cuando una variable está bajo el control del experimentador. Si existe un parámetro que se incrementa o disminuye de forma sistemática por el experimentador, se le denomina parámetro de control o variable independiente y habitualmente se representa a lo largo del eje horizontal (eje de las abscisas). La variable medida o dependiente usualmente se representa a lo largo del eje vertical (eje de las ordenadas). Si no existe una variable dependiente, cualquier variable se puede representar en cada eje

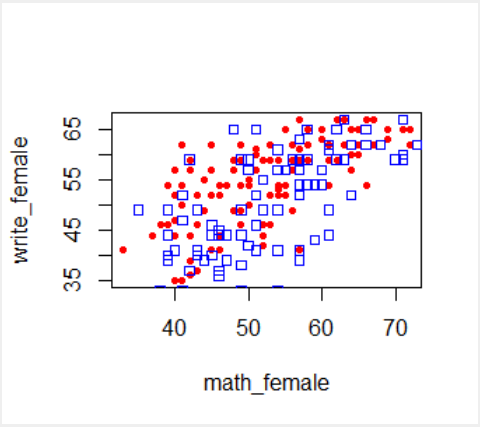
y el diagrama de dispersión mostrará el grado de correlación (no causalidad) entre las dos variables

Para hacer las pruebas vamos a utilizar datos genéricos provenientes de "University of California, Los Angeles"

'http://www.ats.ucla.edu/stat/r/modules/hsb2.csv'

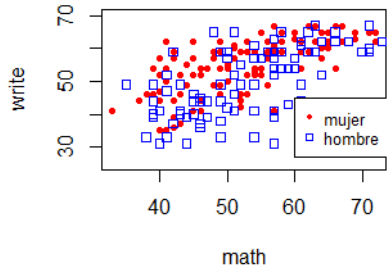
```
hsb2 <- read.table('http://www.ats.ucla.edu/stat/r/modules/hsb2.csv', header=T,
sep=",")
attach(hsb2)
```

	`plot(math, read)`
	**`plot(math, read)`** **`abline(lsfit(read, math))`**



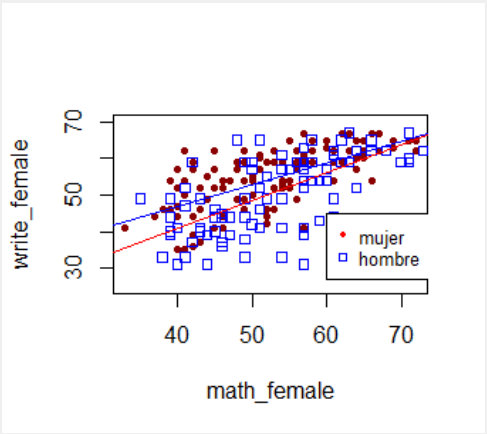

```
math_male<-hsb2$math[female==0]
write_male<-hsb2$write[female==0]
math_female<-hsb2$math[female==1]
write_female<-hsb2$write[female==1]

plot(math_female, write_female,
type="p",
     pch=20, col="red")
points(math_male, write_male,
pch=22, col="blue")
```




```
Añadiendo leyenda

hsb2_female<-hsb2[female==1,]
> hsb2_male<-hsb2[female==0,]
> with(hsb2_female, plot(math, write,
pch=20,
+ col="red", ylim=c(25, 70)))
> with(hsb2_male, points(math, write,
pch=22,
+ col="blue"))
> legend(60, 45, c("mujer", "hombre"),
pch=c(20, 22),
+ cex=.8, col=c("red", "blue"))
```




```
math_male<-hsb2$math[female==0]
> write_male<-hsb2$write[female==0]
> math_female<-hsb2$math[female==1]
> write_female<-hsb2$write[female==1]
> plot(math_female, write_female,
type="p", pch=20,
+ col="darkred", ylim=c(25, 70))
> points(math_male, write_male,
pch=22, col="blue")
> abline(lsfit(write_female,
math_female), col="red")
> abline(lsfit(write_male, math_male),
col="blue")
> legend(60, 45, c("mujer", "hombre"),
pch=c(20, 22),
+ cex=.8, col=c("red", "blue"))
```

4.5 Gráfica de sectores

Se trata de un diagrama en forma de círculo dividido en tantos sectores como datos distintos haya, en el que el ángulo de cada sector es proporcional a la frecuencia relativa del correspondiente dato.

Esta representación gráfica se denomina diagrama de sectores o diagrama de tarta. También puede usarse para datos cuantitativos agrupados en clases, y en tales casos, cada sector corresponde a una clase.

La función para diseñar diagramas de sectores en R es: **pie()**

Por ejemplo, la encuesta de población activa elaborada por el Instituto Nacional de Estadística referente al cuarto trimestre de 1970, presenta para el número de empleados por rama de actividad los siguientes datos:

Rama de Actividad	Miles de Empleados
Agricultura, caza y pesca	3706.3
Fabriles	3437.8
Construcción	1096.3
Comercio	1388.3
Transporte	648.7
Otros servicios	2454.8

Para almacenarlos en R:

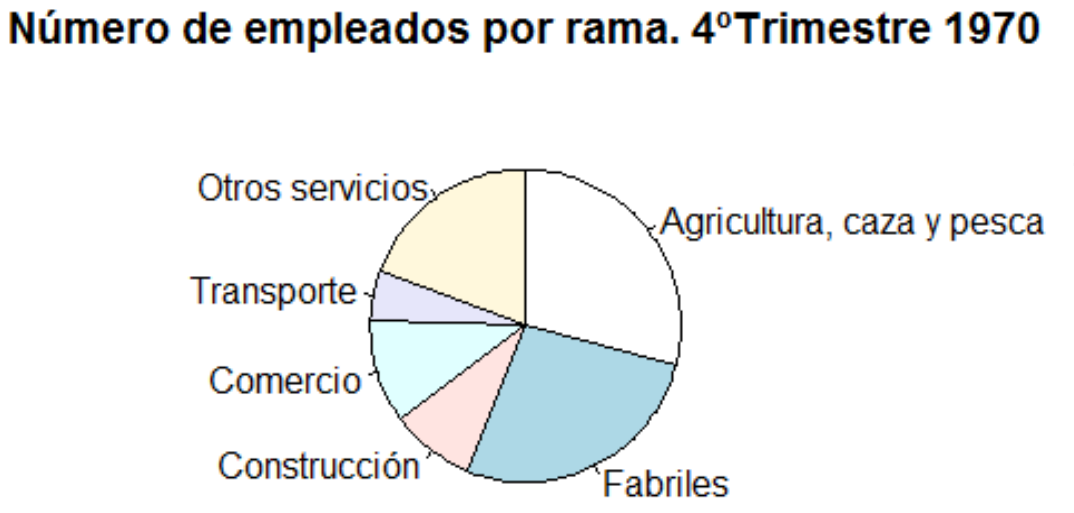
```
> Sector <- c(3706.3, 3437.8, 1096.3, 1388.3, 648.7, 2454.8)
```

Y posteriormente, le asignaremos la Rama de Actividad al vector Sector mediante la función names():

```
>names(Sector) <- c("Agricultura, caza y pesca", "Fabriles", "Construcción",
"Comercio", "Transporte", "Otros servicios")
```

Si usamos ahora la función pie() con los datos anteriores obtenemos:

```
pie(Sector, clockwise=TRUE, main="Número de empleados por rama. 4ºTrimestre 1970",
col=c(2,3,4,5,6,7))
```



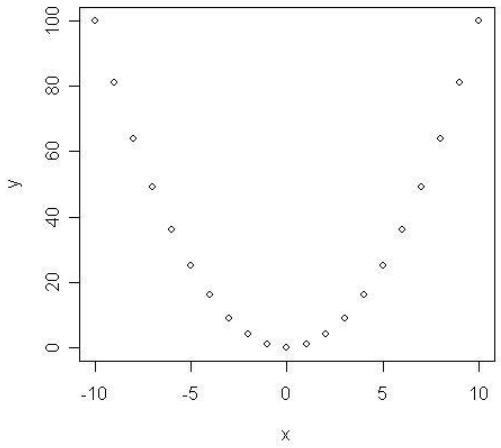
4.6 Gráfica XY:

Permite comparar datos de dos variables cuantitativas

La función plot()

El comando plot se utiliza para crear una nueva figura.

```
> x = seq(-10,10)        # Generamos los números -10, -9,....,9, 10
> y = x^2                # Generamos los cuadrados de dichos números
> plot(x,y)              # Graficamos
```



Dicha función admite bastantes argumentos. Vamos a destacar los más importantes:

---

axes=F **Suprime la generación de los ejes**

---

log="x" **Hace que alguno de los ejes se tome en escala logarítmica**
log="y"
log="xy"

---

type="p" **Dibuja puntos individuales (opción por defecto)**
type="l" **Dibuja líneas**
type="b" **Dibuja puntos y líneas**
type="o" **Dibuja puntos atravesados por líneas**
type="h" **Dibuja con líneas verticales**
type="s" **Dibuja a base de funciones escalera**
type="S" **Casi lo mismo**
type="n" **No dibuja nada. Pero deja marcados los puntos para manejos posteriores**

---

xlab="cadena" **Etiqueta para el eje de las x**
ylab="cadena" **Etiqueta para el eje de las y**
main="cadena" **Título del gráfico**
sub="cadena" **Subtítulo del gráfico**

---

pch="simbolo" **Se dibuja con el simbolo especificado. Por ejemplo:**
pch=18

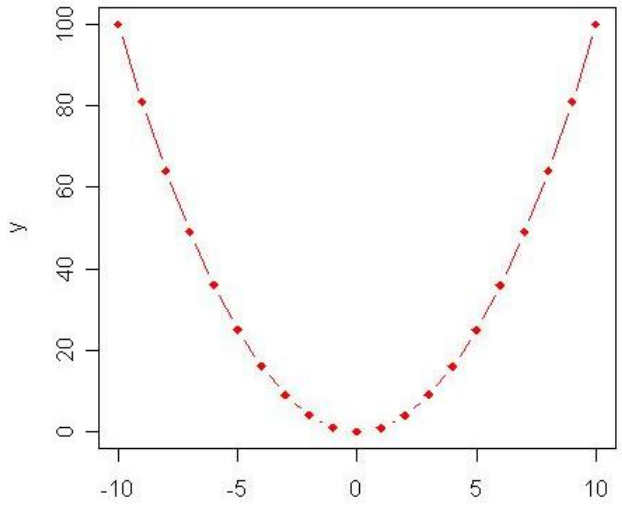
pch="x"
pch="P"

---

col= numero entero Color para dibujar
col=2 Color rojo
col=3 Color verde

Algunos ejemplos:

```
> x = seq(-10,10)
> y = x^2
> plot(x,y,type="l",xlab="eje de las x",ylab="eje de las y", main="Parabola")
> plot(x,y,type="h",xlab="eje de las x",ylab="eje de las y", main="Parabola",axes=F)
> plot(x,y,pch=18,col=2,type="b")
```



5 Funciones útiles de R para estadísticos descriptivos

Summary

Hemos presentado a lo largo del texto muchas funciones para obtener descriptivos, la más recurrida es summary(), pero hay muchas más y algunas no están en los paquetes por defecto. Veamos algunos ejemplos.

```
> p1   <- rnorm(1000,  3,  0.8)
> p2   <- rnorm(1000,  2,  0.5)
> p  <- c(p1,  p2)
> altura  <- c(rnorm(1000,  87,  7),  rnorm(1000,  97,  6))
> #  length(p); hist(p)
> grupo <- c(rep("M",  1000),  rep("H",  1000))
> df <- data.frame(p, altura, grupo)
> head(df)
         p   altura grupo
1 3.631115 99.81936     M
```

```
2 3.135630 89.78738     M
3 3.808861 97.43998     M
4 3.533223 92.98578     M
5 1.602240 87.92617     M
6 3.293906 91.16983     M
> str(df)
'data.frame':       2000 obs. of  3 variables:
 $ p     : num  3.63 3.14 3.81 3.53 1.6 ...
 $ altura: num  99.8 89.8 97.4 93 87.9 ...
 $ grupo : Factor w/ 2 levels "H","M": 2 2 2 2 2 2 2 2 2 ...
```

```
> summary(df)
       p                altura        grupo
 Min.   :0.08224   Min.   : 65.55   H:1000
 1st Qu.:1.90634   1st Qu.: 86.96   M:1000
 Median :2.38038   Median : 92.51
 Mean   :2.51208   Mean   : 92.30
 3rd Qu.:3.06516   3rd Qu.: 98.11
 Max.   :6.04453   Max.   :116.60
```

También podemos asignar dataframes a cada grupo para simplificar la sintaxis

```
> df.M <-  df[which(df$grupo == "M"),]
> summary(df.M)
       p                altura        grupo
 Min.   :0.08224   Min.   : 65.55   H:   0
 1st Qu.:2.48552   1st Qu.: 82.94   M:1000
 Median :3.01158   Median : 87.68
 Mean   :3.02616   Mean   : 87.56
 3rd Qu.:3.56867   3rd Qu.: 92.14
 Max.   :6.04453   Max.   :106.88
```

```
> df.H <-  df[which(df$grupo == "H"),]
> summary(df.H)
       p                altura        grupo
 Min.   :0.2053   Min.   : 77.77   H:1000
 1st Qu.:1.6572   1st Qu.: 92.81   M:   0
 Median :2.0126   Median : 97.00
 Mean   :1.9980   Mean   : 97.05
 3rd Qu.:2.3391   3rd Qu.:101.15
 Max.   :3.3470   Max.   :116.60
```

stat.desc()

La función stat.desc() del paquete pastecs, tiene varias opciones muy interesantes:

```
> library("pastecs")
> stat.desc(df)
                         p       altura grupo
nbr.val      2.000000e+03 2.000000e+03    NA
nbr.null     0.000000e+00 0.000000e+00    NA
nbr.na       0.000000e+00 0.000000e+00    NA
min          8.224415e-02 6.554759e+01    NA
max          6.044528e+00 1.166041e+02    NA
range        5.962284e+00 5.105648e+01    NA
sum          5.024165e+03 1.846078e+05    NA
median       2.380383e+00 9.251240e+01    NA
mean         2.512083e+00 9.230389e+01    NA
SE.mean      1.909287e-02 1.805699e-01    NA
CI.mean.0.95 3.744401e-02 3.541249e-01    NA
var          7.290755e-01 6.521097e+01    NA
std.dev      8.538592e-01 8.075331e+00    NA
coef.var     3.399009e-01 8.748635e-02    NA
```

```
> stat.desc(df[-3],  norm   =  TRUE)
                         p        altura
nbr.val      2.000000e+03  2.000000e+03
nbr.null     0.000000e+00  0.000000e+00
nbr.na       0.000000e+00  0.000000e+00
min          8.224415e-02  6.554759e+01
max          6.044528e+00  1.166041e+02
range        5.962284e+00  5.105648e+01
sum          5.024165e+03  1.846078e+05
median       2.380383e+00  9.251240e+01
mean         2.512083e+00  9.230389e+01
SE.mean      1.909287e-02  1.805699e-01
CI.mean.0.95 3.744401e-02  3.541249e-01
var          7.290755e-01  6.521097e+01
std.dev      8.538592e-01  8.075331e+00
coef.var     3.399009e-01  8.748635e-02
skewness     5.411580e-01 -2.106689e-01
skew.2SE     4.943776e+00 -1.924576e+00
kurtosis     1.160297e-01 -1.720119e-01
```

```
kurt.2SE     5.302614e-01 -7.861029e-01
normtest.W   9.797425e-01  9.962903e-01
normtest.p   3.059961e-16  8.206427e-05
```

```
> stat.desc(df.M[-3],  basic =  FALSE,  norm   =  TRUE)
                      p      altura
median       3.01157743 87.68085348
mean         3.02615949 87.55634911
SE.mean      0.02606603  0.22067863
CI.mean.0.95 0.05115045  0.43304682
var          0.67943789 48.69905598
std.dev      0.82428023  6.97847089
coef.var     0.27238493  0.07970263
skewness     0.05906331 -0.12030911
skew.2SE     0.38182307 -0.77775520
kurtosis     0.03840118 -0.11264143
kurt.2SE     0.12424810 -0.36445455
normtest.W   0.99874020  0.99774431
normtest.p   0.71550603  0.19133830
```

Hmisc

Hay más funciones que ofrecen descriptivos, por ejemplo Hmisc (y muchas más).

```
install.packages('Hmisc')
library("Hmisc")
describe(df$p)
df$p
      n missing  unique    Info    Mean    .05    .10    .25     .50     .75
   2000       0    2000       1   2.512   1.305   1.518  1.906   2.380   3.065
    .90     .95
  3.695   4.072
lowest : 0.08224 0.20532 0.47317 0.47469 0.47567
highest: 5.35935 5.41162 5.43792 5.68467 6.04453
```

```
> head(df$p)
[1] 3.631115 3.135630 3.808861 3.533223 1.602240 3.293906
```

```
> describe(df$altura)
df$altura
      n missing  unique    Info    Mean     .05     .10     .25     .50     .75
   2000       0    2000       1    92.3   78.26   81.51   86.96   92.51   98.11
    .90     .95
 102.32  104.74
```

```
lowest :  65.55  66.32  66.69  68.34  68.85
highest: 113.27 113.64 113.82 114.71 116.60
```

```
> head(df$altura,  25)
 [1] 99.81936 89.78738 97.43998 92.98578 87.92617 91.16983 89.54132 92.63484
 [9] 82.03902 91.20508 89.67584 80.88947 89.40407 94.56393 88.35981 99.46379
[17] 81.09473 85.02808 73.52344 82.53400 95.56726 89.98571 87.20829 98.14927
[25] 95.12412
```

Función tapply()

Con la función tapply nos podemos construir fácilmente nuestras tablas de descriptivos de una forma muy elegante.

```
> tapply(df$p,  df$g,  mean)
       H        M
1.998006 3.026159
```

```
> m <- tapply(df$p,  df$g, mean)
> s <- tapply(df$p, df$g, sd)
> m2 <- tapply(df$p, df$g, median)
> n <- tapply(df$p,df$g,length) cbind(media = m, sd = s, mediana = m2,n)
> n <- tapply(df$p,df$g,length)
> cbind(media  =  m, sd  =  s, mediana =  m2,   n)
     media        sd  mediana    n
H 1.998006 0.5003633 2.012560 1000
M 3.026159 0.8242802 3.011577 1000
```

Tablas de frecuencias y probabilidades

En estadística, se le llama distribución de frecuencias a la agrupación de datos en categorías mutuamente excluyentes que indican el número de observaciones en cada categoría. Esto proporciona un valor añadido a la agrupación de datos. La distribución de frecuencias presenta las observaciones clasificadas de modo que se pueda ver el número existente en cada clase.

```
> pais <- c( "ES", "ES", "ES", "US", "US","UK" )
> sexo <- c( "F", "F", "M", "F", "M","M" )
> t <- table( pais, sexo )  # tabla de frecuencias absolutas
> t
    sexo
pais F M
  ES 2 1
  UK 0 1
  US 1 1
> # frec relativas
```

```
> prop.table( t )  # porcentajes totales
    sexo
pais         F         M
  ES 0.3333333 0.1666667
  UK 0.0000000 0.1666667
  US 0.1666667 0.1666667
> prop.table( t ) * 100
    sexo
pais        F        M
  ES 33.33333 16.66667
  UK  0.00000 16.66667
  US 16.66667 16.66667
```

```
> # porcentajes por filas
> prop.table( t, 1 )
    sexo
pais         F         M
  ES 0.6666667 0.3333333
  UK 0.0000000 1.0000000
  US 0.5000000 0.5000000
> # porcentajes por columnas
> prop.table( t, 2 )
    sexo
pais         F         M
  ES 0.6666667 0.3333333
  UK 0.0000000 0.3333333
  US 0.3333333 0.3333333
```

6 Creación de Funciones en R

Para crear funciones en R empleamos **function(<parametro1>,…<parametroN>)** y para llamarla hacemos lo mismo que hacemos con las funciones habituales. Esta es la forma de programar con R. Del mismo modo si deseamos medir el coeficiente de curtosis (momento de orden 4) para medir la asimetría hemos de crear la función:

```
> kurtosis=function(x) {
+ m4=mean((x-mean(x))^4)
 + kurt=m4/(sd(x)^4)-3
+ kurt}
> kurtosis(alturas)
[1] -0.9660813
```

Con todo lo visto anteriormente podemos crear una función que nos haga un pequeño análisis descriptivo de un vector:

```
> descriptivos<-function(x){
 + desc<-c(mean(x),varianza(x),min(x),max(x),quantile(x),kurtosis(x))
+ nom<-c("Media","Varianza","Mínimo","Máximo","Cuantil 0","Cuantil 25","Cuantil
50","Cuantil 75",
+ "Cuantil 100", "Kurtosis")
+ names(desc)<-nom
+ desc}
> descriptivos(alturas)
  Media Varianza Mínimo Máximo Cuantil 0 Cuantil 25 Cuantil 50 Cuantil 75 Cuantil
100 Kurtosis
1.77090909 0.01200826 1.54000000 1.90000000 1.54000000 1.71000000 1.76000000
1.88000000 1.90000000 -0.96608127
```

Creamos la función descriptivos que recibirá un parámetro vector. Obtenemos algunas medidas descriptivas que almacenamos en otro vector y asignamos los nombres de los valores con la función names, por último simplemente vemos el vector.