INTRODUCCIÓN

¿QUÉ ES LA ESTADÍSTICA?

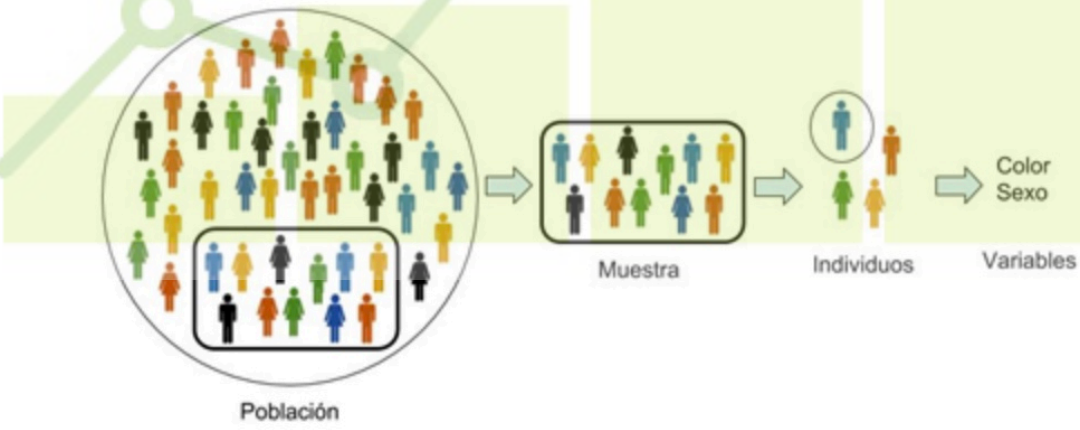
La estadística es un proceso complejo que abarca la recopilación, análisis, presentación e interpretación de datos para mejorar la toma de decisiones en diversas áreas.

Sus tres ramas principales son:

- La **estadística descriptiva**, que resume características de conjuntos de datos mediante tablas y gráficos
- La **teoría de la probabilidad**, que estudia fenómenos aleatorios y controla la aleatoriedad mediante cálculos de probabilidades
- La **inferencia estadística**, que extrae conclusiones sobre poblaciones a partir de muestras recogidas.

CONCEPTOS BÁSICOS

Para introducir algunos de los conceptos principales que deben tenerse claros para poder avanzar, mostramos el siguiente esquema:



- La **población** es el conjunto total sobre el cual se llevará a cabo el estudio, compuesto por todos los elementos o unidades estadísticas que podrían ser objeto de análisis.
    - Dado que analizar cada elemento de la población es impracticable en la mayoría de los casos debido a limitaciones como tiempo, costo o acceso, se recurre a trabajar con una muestra representativa de la población.
- La **muestra** consiste en una selección de elementos de la población a los cuales se puede acceder y sobre los cuales se realizará el estudio, ya que es más factible llegar a ellos en términos de tiempo y recursos.
    - Se plantean preguntas sobre cómo seleccionar la muestra de manera adecuada, lo cual implica considerar técnicas de muestreo para garantizar la representatividad y validez de los resultados.
- Las **variables**, también conocidas como caracteres estadísticos de interés, son las características o cualidades que poseen los individuos de una población y que son objeto de estudio.
- La recolección de datos implica recopilar, analizar y resumir información para su presentación e interpretación en un conjunto de datos para el estudio.

TIPOS DE VARIABLES

Según el tipo de valores que toman las variables, distinguimos diferentes tipos de variables. Las variables esencialmente pueden ser de dos tipos: **cualitativas** y **cuantitativas**.

- Las **variables cualitativas** (o **atributos**) son aquellas que no aparecen en forma numérica, sino como categorías o atributos. Es decir, los valores son categorías y dichos valores son diferentes por una cualidad, no por una cantidad. Por ejemplo: partido político al que votó un individuo; región en que vive; sexo; estado civil; marca de coche que conduce, etc. Nótese que con estas variables no se pueden hacer operaciones algebraicas con ellas.

  Dentro de las variables cualitativas podemos distinguir entre:

  - **Variables ordinales**. Si sus valores se pueden ordenar. Por ejemplo: clase social (baja, media, alta); opinión sobre una propuesta política (muy en contra, más bien en contra, indiferente, más bien a favor, muy a favor); y el grado de satisfacción en el trato con el personal sanitario (muy satisfecho, satisfecho, poco satisfecho).
  - **Variables nominales**. Si sus valores no se pueden ordenar. Por ejemplo: sexo (hombre, mujer); fuma (Sí, No); y estado civil (soltero, casado, separado, viudo).

- Las **variables cuantitativas** son aquellas que pueden expresarse numéricamente como el peso, el número de goles de un partido de fútbol, la temperatura, los ingresos anuales, nota en un examen, número de años de educación, kilómetros de distancia entre trabajo y residencia...

  Dentro de las variables cuantitativas podemos distinguir entre:

  - **Variables discretas**. Si toma valores en el conjunto de los números enteros. Por ejemplo: número de hermanos (1, 2, 3...., etc., pero nunca podrá ser 3,45); no de monedas que una persona lleva en el bolsillo (0, 1, 2, ...); edad de una persona: (1,2,3 años…).
  - **Variables continuas**: Si toma cualquier valor dentro de un intervalo real. Por ejemplo: la velocidad de un vehículo (80,3 km/h, 94,57 km/h...); altura de las personas (1,65m, 1,58m...); peso de las personas: (55,3kg, 68,2kh...)

Las variables estadísticas también se pueden clasificar en:

- **Variables unidimensionales**: sólo recogen información sobre una característica (por ejemplo: edad de los alumnos de una clase).
- **Variables bidimensionales**: recogen, a la vez y sobre el mismo individuo, información sobre dos características de la población, que pueden o no estar relacionadas, (por ejemplo: edad y altura de los alumnos de una clase).
- **Variables multidimensionales**: recogen, a la vez y sobre el mismo individuo, información sobre tres o más características de la población, que pueden o no estar relacionadas (por ejemplo: edad, altura y peso de los alumnos de una clase).

EJERCICIO:

Clasifica las siguientes variables estadísticas:

- Número de aprobados en un curso.
- Color de las manzanas de una frutería.
- Libros leídos por un grupo de alumnos.
- Número de pulsaciones por minuto.
- Número de compañeros de clase.
- Estado civil.
- Medida de la palma de la mano.
- Temperaturas mínimas en una semana.
- Género de cine preferido.
- Veces por semana que se come pescado.
- Nacionalidad.
- Edad.
- Color de ojos.
- Peso de los recién nacidos.

TABULACIÓN DE LOS DATOS

Para comenzar nuestros cálculos, en primer lugar, vamos a aprender a ordenarlas para posteriormente realizar un sencillo recuento de las observaciones de las que disponemos separadas por clases, valores o intervalos numéricos, según el tipo de variable que estemos tratando. Es lo que conocemos mediante tablas de distribución de frecuencias.

Valores	Frecuencias absolutas	Frecuencias relativas	Frecuencias absolutas acumuladas	Frecuencias relativas acumuladas
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
⋮	⋮	⋮	⋮	⋮
x_k	n_k	f_k	$N_k = N$	$F_k = 1$
Total	N	1		

EJEMPLO:

Un ejemplo de una variable estadística discreta es el siguiente. Supongamos que hemos medido la variable número de hijos para un conjunto de 15 familias y se han obtenido los resultados: 1, 2, 1, 3, 2, 2, 4, 1, 1, 1, 0, 0, 2, 0 y 5

- La **frecuencia absoluta** de un valor (o modalidad) es el número de observaciones que presenta ese valor (o modalidad). Lo vamos a denotar por n_i . Dado que el número total de observaciones es N, se verifica que :

$$n_1 + n_2 + \ldots + n_k = N.$$

- La **frecuencia relativa** de un valor (o modalidad) es la proporción de observaciones que presenta ese valor (o modalidad). Lo vamos a denotar por f_i. Es decir que:

$$f_i = n_i/N$$

La suma de todas las frecuencias relativas debe ser igual a 1. Es decir que:

$$\sum f_i = 1$$

A veces la frecuencia relativa se expresa en forma de porcentaje:

$$p_i = f_1 x 100$$

- La **frecuencia absoluta acumulada** de un valor x_i es el número de observaciones que son menores o iguales que ese valor (o modalidad). Si la denotamos por N_i se tienen que cumplir:

$$N_i = n_1 + n_2 + \ldots + n_i$$

- La **frecuencia relativa acumulada** de un valor x_i es la proporción de observaciones que son menores o iguales que x_i. Si la denotamos por F_i

se verifica:

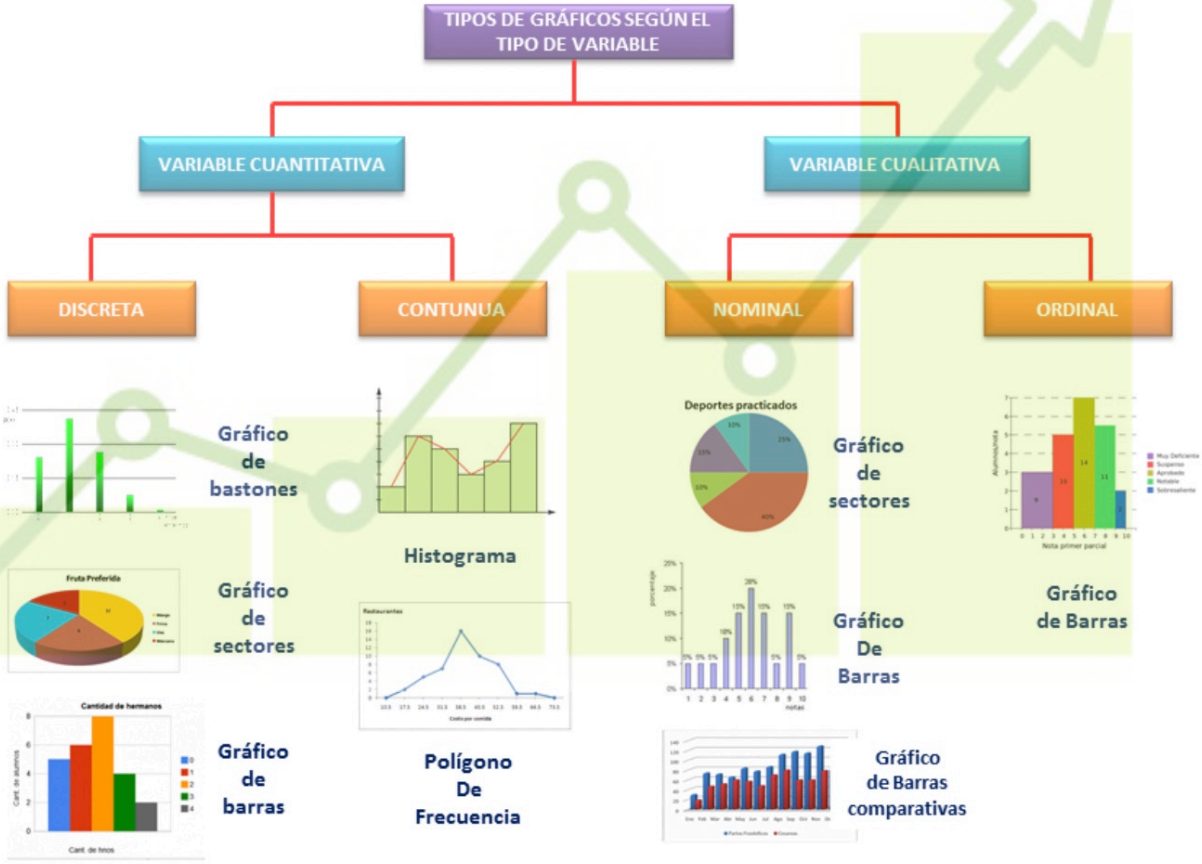
$$F_i = f_1 + f_2 + \ldots + f_i = N_i/N$$

MÉTODOS GRÁFICOS

Consiste en crear gráficos que sintetizan el comportamiento de la variable. En una representación gráfica siempre debe haber constancia en el gráfico de las variables que estamos representando. Un resumen gráfico es un método complementario al resumen numérico. Por sí solo no nos da información. Los gráficos se basarán en el sistema cartesiano. Van acompañados siempre de resúmenes numéricos. Se dividen en dos grupos:

1. Cualitativos.

2. Cuantitativos.

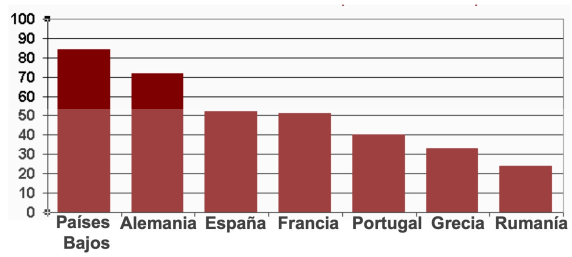


MÉTODOS GRÁFICOS PARA DATOS CUALITATIVOS

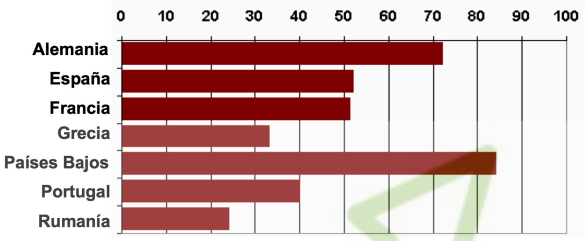
- **Diagramas de barras (no histogramas)**

Se construye colocando las distintas modalidades de la variable cualitativa sobre el eje de abscisas y sobre cada una de ellas se levanta un rectángulo de igual base y altura igual a su frecuencia (absoluta o relativa).

Orientación vertical y orden por frecuencias



Orientación horizontal y orden alfabético



- **Diagramas de sectores**

Sirve para variables cualitativas no agrupadas. Se construye repartiendo el área del círculo en sectores de tamaño proporcional a la frecuencia de cada modalidad. Hay tantos sectores como valores de la variable y los ángulos se calculan de forma proporcional a las frecuencias relativas de cada sector:

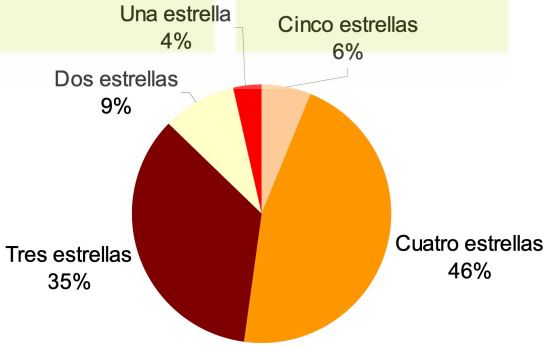
$\alpha = fi \times 360^{o}$

Alojamientos Turísticos. 2009

Categoría	Número de viajeros
Total	69.152.754
Cinco estrellas	4.216.253
Cuatro estrellas	31.960.442
Tres estrellas	24.079.125
Dos estrellas	6.331.715
Una estrella	2.565.219

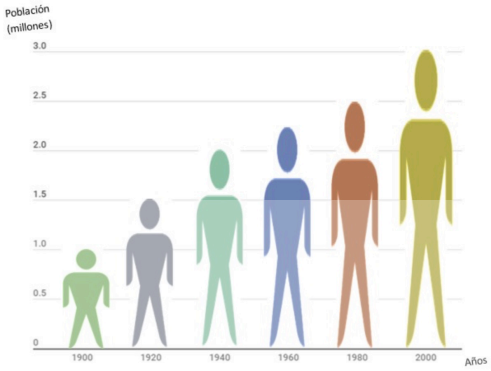
Fuente: Encuesta de Ocupación en Alojamientos Turísticos

Viajeros hospedados en hoteles españoles por categoría del establecimiento. 2009



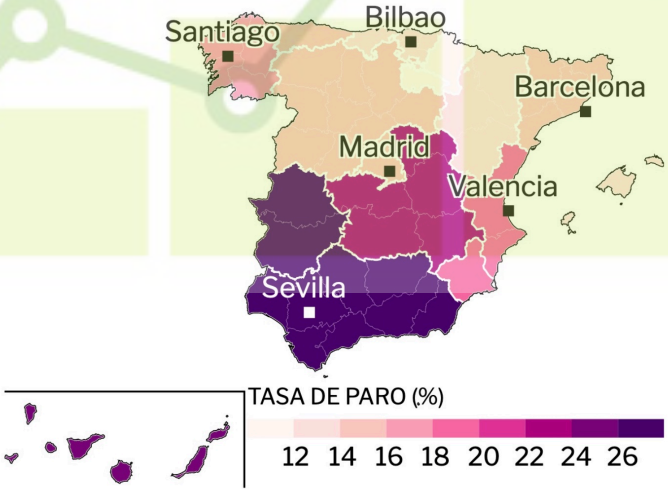
- **Pictograma**

Un pictograma es un tipo de gráfico, que en lugar de barras, utiliza figuras proporcionales a la frecuencia. Generalmente se emplea para representar variables cualitativas. Este tipo de gráfico no permite buenas comparaciones.



- **Cartograma**

Un cartograma es un mapa o diagrama que muestra datos cuantitativos asociados a sus respectivas áreas.



MÉTODOS GRÁFICOS PARA DATOS CUATITATIVOS

Datos cuantitativos sin agrupar:

- **Diagrama de dispersión o nube de puntos**

El gráfico de dispersión es uno de los más importantes en el ámbito de la Estadística. Consiste en la representación de dos variables del mismo individuo o hecho en un sistema de ejes cartesianos. Con ello se obtiene una nube de puntos que en ocasiones, si existe relación entre las dos variables representadas, adopta una forma más o menos definida. El objetivo del gráfico de dispersión es, precisamente, estudiar la posible relación entre las dos variables.

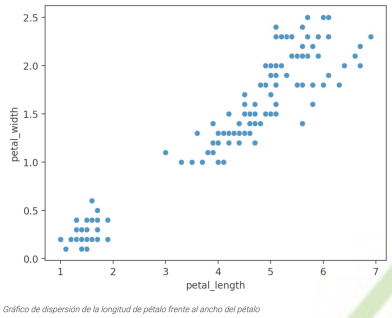


Gráfico de dispersión de la longitud de pétalo frente al ancho del pétalo

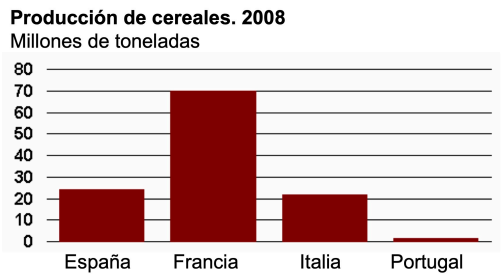
- **Diagrama de barras (para tablas de frecuencias no agrupadas)**

Un gráfico de barras es una representación gráfica en un eje cartesiano de las frecuencias de una variable cualitativa o discreta. En uno de los ejes se posicionan las distintas categorías o modalidades de la variable cualitativa o discreta y en el otro el valor o frecuencia de cada categoría en una determinada escala.

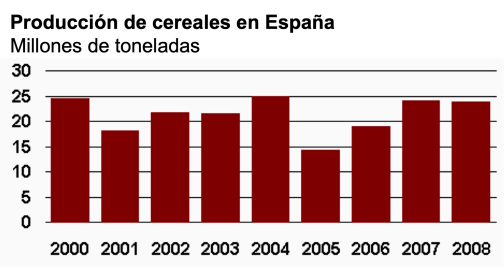
Gráfico de barras

Se suelen usar para:

• **Comparar magnitudes** de varias categorías.

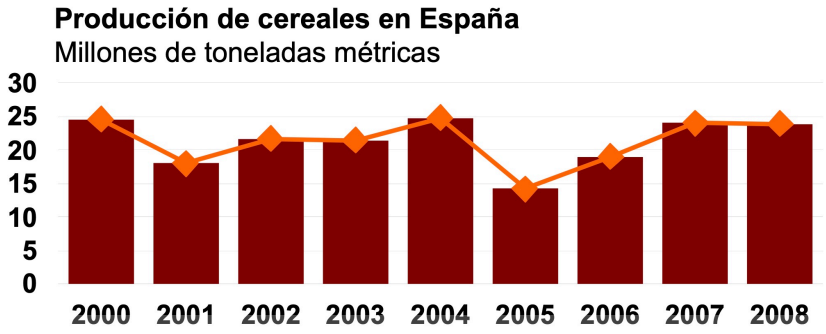


• Ver la **evolución en el tiempo** de una magnitud concreta.



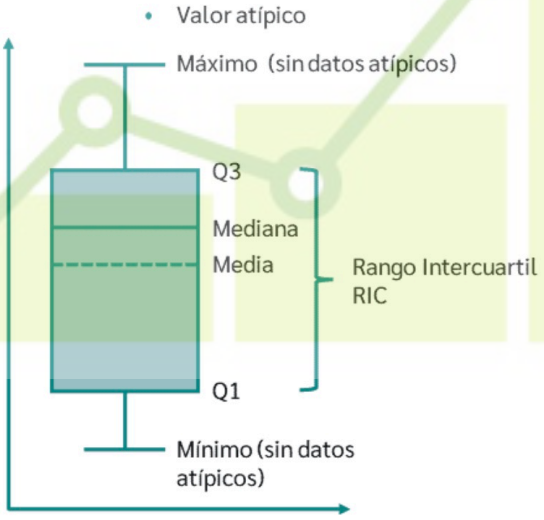
- **Poligono de frecuencias**

Si se unen los puntos medios de las bases superiores de las barras en los gráficos de barra se obtiene el polígono de frecuencias.



- **Box Plot o diagrama de caja**

Un diagrama de caja es un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles. De esta manera, se muestran a simple vista la mediana y los cuartiles de los datos, y también pueden representarse sus valores atípicos.



La caja indica el intervalo en el que se encuentra el 50% de los datos.

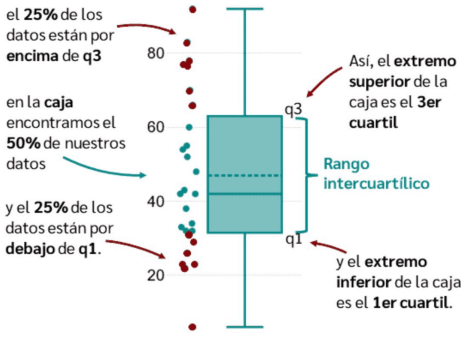
Por lo tanto, el extremo inferior de la caja es el 1er cuartil y el extremo superior es el 3er cuartil.

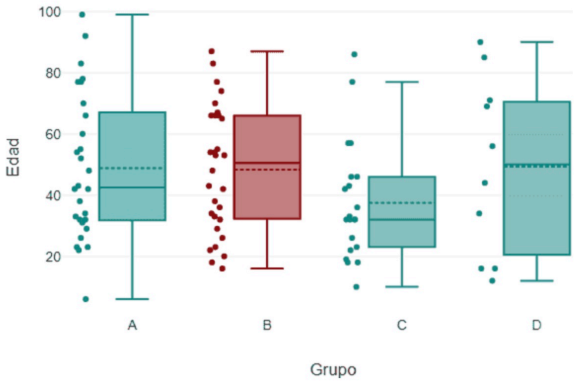
Entre Q1 y Q3, está el rango intercuartil (RIC)

En el diagrama de caja, la línea continua indica la mediana y la línea discontinua, la media.

Los bigotes en forma de T se extienden hasta los valores máximo y mínimo que siguen estando dentro de 1,5 veces el rango intercuartílico (RIC).

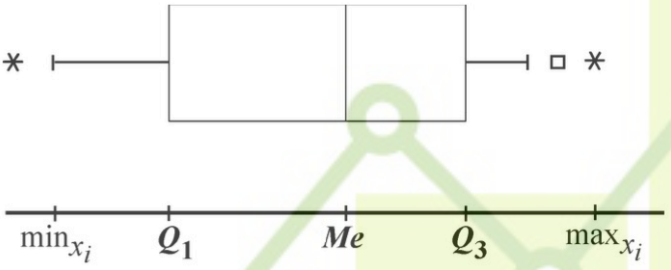
Los puntos que están aún más alejados se consideran valores atípicos.



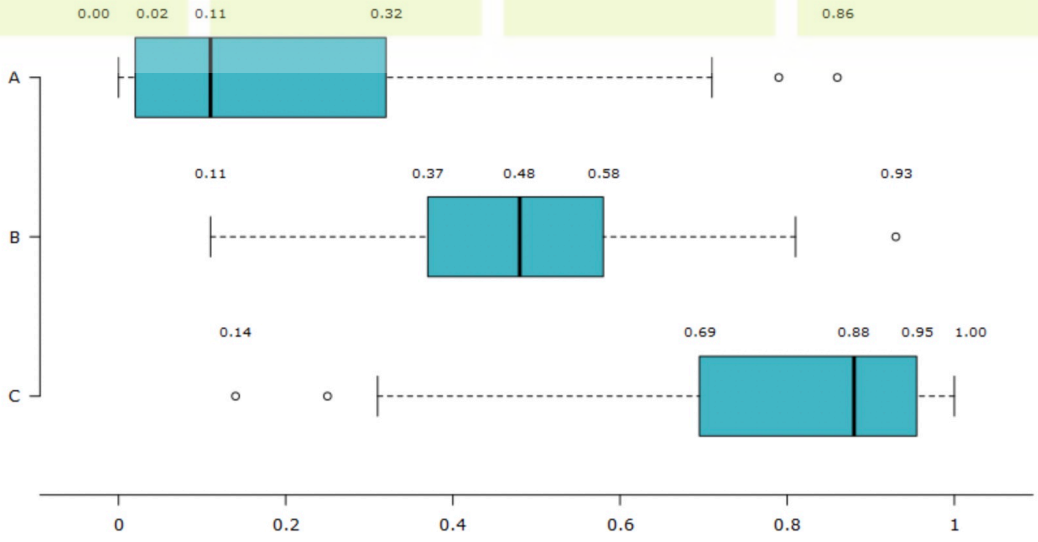


Proceso de construcción del boxplot:

- Se ordenan los datos de menor a mayor y se calculan los cuartiles y la mediana. La caja central queda limitada por Q1 y Q3
- Se calculan los límites inferior y superior (LI y LS)
- Se consideran atípicos los datos fuera del intervalo [LI, LS]
- Se dibuja una línea desde cada extremo de la caja central hasta el valor más lejano no atípico
- Identificamos los datos fuera de [LI, LS], señalándolos como atípicos.
    - Limite Superior: $LS = Q_3 + 1{,}5(Q_3 - Q_1)$
    - Limite Inferior: $LI = Q_1 - 1{,}5(Q_3 - Q_1)$



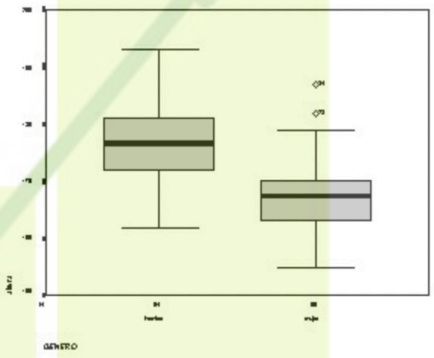
Los boxplots son útiles para comparar la distribución de una variable entre diferentes poblaciones.



- **Diagrama de tallo y hojas o Stem-and-Leaf Diagram**

El diagrama de tallo y hojas (Stem-and-Leaf Diagram) es un semigráfico que permite presentar la distribución de una variable cuantitativa. Consiste en separar cada dato en el último dígito (que se denomina hoja) y las cifras delanteras restantes (que forman el tallo).

```
altura Stem-and-Leaf Plot for
GENERO= hombre

 Frequency    Stem &  Leaf

      2.00       16 .  23
      2.00       16 .  59
     15.00       17 .  000000122223444
     15.00       17 .  555555567788888
     12.00       18 .  000000122233
      5.00       18 .  55568
      3.00       19 .  023

  Stem width:    10
  Each leaf:        1 case(s)

altura Stem-and-Leaf Plot for
GENERO= mujer

 Frequency    Stem &  Leaf

      .00       15 .
     4.00       15 .  5799
    16.00       16 .  0000122233334444
    20.00       16 .  55556667778888888999
    12.00       17 .  000000000134
     6.00       17 .  588889
     2.00 Extremes    (>=182)

  Stem width:    10
  Each leaf:        1 case(s)
```

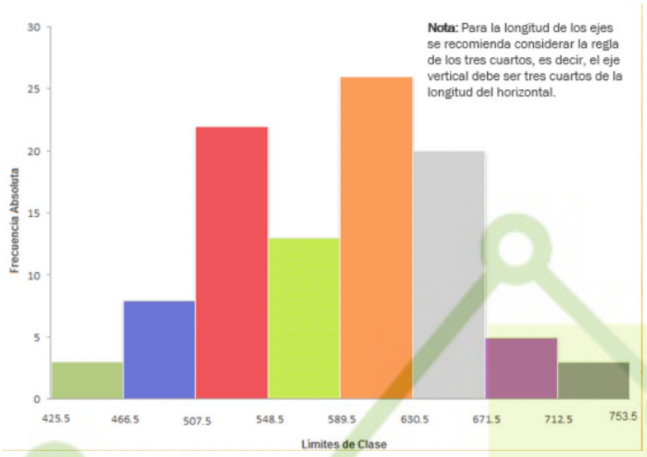


Datos cuantitativos agrupados:

- **Histograma**

Un histograma es una representación gráfica de una variable en forma de barras, teniendo en cuenta que la superficie de cada barra es proporcional a la frecuencia de los valores representados. Un histograma nos permite ver cómo se distribuyen los valores de la variable en estudio.

Usamos los histogramas cuando analizamos variables continuas, o cuando trabajamos con variables discretas que toman un gran número de valores y son agrupadas en intervalos. Recordemos que cuando tenemos variables cualitativas, se emplean los diagramas de barras.



- **Polígono de frecuencias**

El polígono de frecuencias es una línea que se obtiene uniendo los puntos medios de las bases superiores (los techos) de cada rectángulo en el histograma. De forma que empiece y acabe sobre el eje de abscisas, en el punto medio del que sería el intervalo anterior al primero y posterior al último respectivamente.