Tema 3.

Cómo descubrir la variabilidad de un conjunto de datos

Tema 3:
Cómo descubrir la variabilidad de un conjunto de datos

COLECTIVOS O POBLACIONES Y CARACTERÍSTICAS O VARIABLES

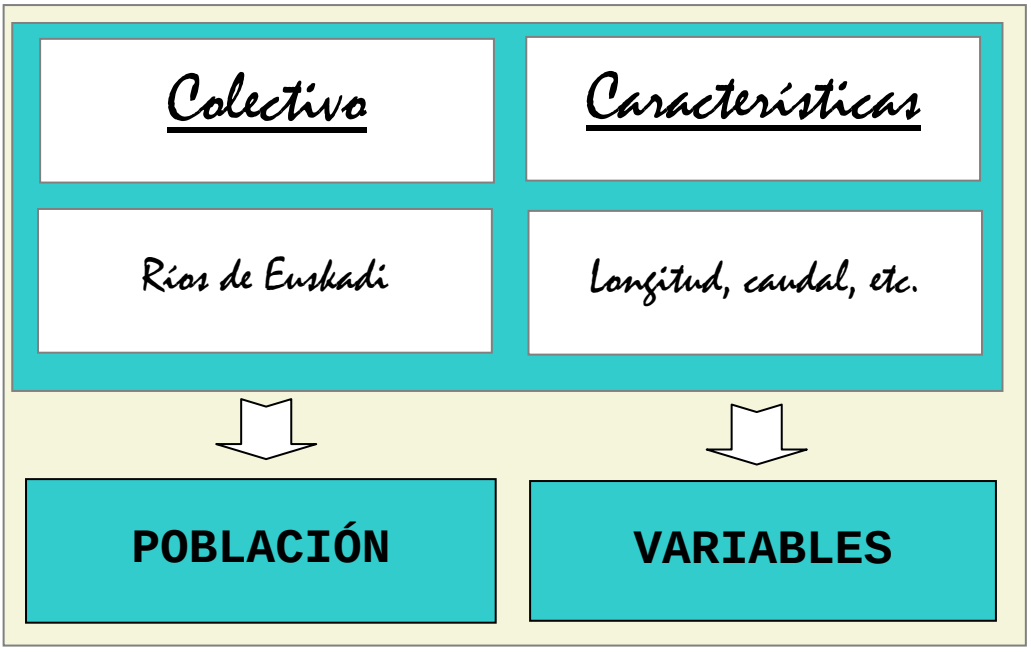
En el tema anterior hemos podido ver que es muy frecuente estudiar la característica de algún colectivo cuyo comportamiento nos interesa conocer. Hablábamos, entre otros ejemplos, de un posible colectivo de ciudades sobre el cual se estudiaba el número de emigrantes que habían fijado su residencia en ellas. Otro colectivo que mencionábamos era el de productos y servicios, del cual nos interesaba estudiar el encarecimiento que habían sufrido tras el cambio de moneda al euro.

Lo que nos interesa ahora es presentar dos conceptos cuya comprensión es esencial en el razonamiento estadístico. No hay que olvidar que en ocasiones resulta difícil entender los procedimientos estadísticos simplemente porque antes no se han entendido bien algunos de los conceptos fundamentales del lenguaje estadístico. Son conceptos sencillos pero, si no se ha asimilado su significado, pueden bloquear la comprensión y el aprendizaje posterior.

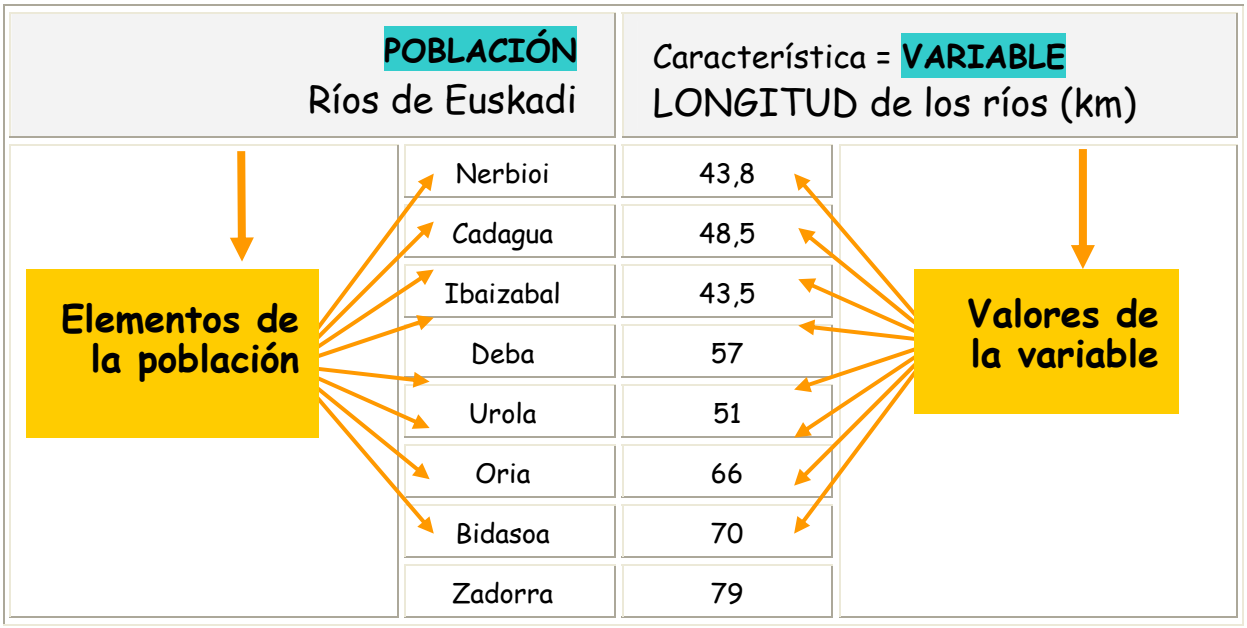
Supongamos que nos han encargado un estudio sobre los ríos de Euskadi. El estudio se enmarca en un proyecto de investigación sobre la calidad ambiental de distintos elementos y características del territorio perteneciente a la comunidad autónoma. Se nos ha pedido que hagamos una caracterización de los rasgos que incluya la longitud de los ríos, el caudal medio e indicadores de la calidad ambiental del agua de los ríos y de los bosques de ribera.

El colectivo cuyas características debemos estudiar está integrado en este caso por los ríos que tienen su recorrido dentro del territorio de la comunidad autónoma. Para referirnos a este colectivo utilizaremos en adelante el término correcto que es POBLACIÓN. Una población en Estadística es el conjunto de elementos sobre los que se quiere estudiar y conocer algo. A los integrantes de la población (colectivo de estudio) se les denomina ELEMENTOS DE LA POBLACIÓN.

De cada uno de los ríos estudiaremos algunas de sus características. Estudiaremos su longitud, el caudal máximo, la calidad de sus aguas y el régimen de alimentación. Sabemos que la longitud varía de unos ríos a otros, lo mismo que la cantidad de agua que llevan o el nivel de calidad de sus aguas. A estas características que se estudian para cada uno de los elementos de la población se les conoce, en el lenguaje estadístico, con el nombre de VARIABLES.



Un conjunto de elementos = UNA POBLACIÓN	Características de los elementos = VARIABLES
Los ríos de Euskadi	Longitud
	Caudal máximo
	Calidad del agua en la desembocadura
	Régimen de alimentación



En el ejemplo que vemos en la tabla, en la columna izquierda tenemos los datos correspondientes a la población, que estaría compuesta por todos y cada uno de los ríos de Euskadi. En la columna derecha tenemos los datos correspondientes a los valores de una de las variables que, en este caso, sería la longitud de los ríos.

Veamos ahora algunos detalles de las cuatro características o variables que vamos a estudiar para cada uno de los ríos:

- La longitud del río, que se expresa generalmente en km.

- El caudal máximo de un río se refiere al mayor caudal registrado en un punto dado del río, dentro de un período de tiempo determinado. El caudal se mide en litros o en metros cúbicos por segundo.

- La calidad del agua se suele definir por medio de calificativos que indican niveles diferentes dentro de una escala de valoración. Los calificativos surgen generalmente de una valoración global elaborada a partir de varias medidas de distintos aspectos relacionados con la calidad del agua.[1] En el

[1] Gobierno Vasco (2004-2008): *Red de seguimiento del estado ecológico de los ríos de la CAPV*. Se trata de un un estudio exhaustivo desarrollado durante varios años por la empresa Ondotek II para la Agencia Vasca del Agua. Los resultados, divididos en 25 tomos, pueden descargarse en la web de la Agencia Vasca del

estudio sobre el estado ecológico de los ríos de la CAPV utilizaron para la calificación una escala de cinco niveles de calidad: muy bueno, bueno, aceptable, deficiente y malo.

- El régimen de alimentación hace referencia al origen de las aguas que alimentan al río. El origen del agua puede ser pluvial o nival; puede ser también mixto, con preferencia de uno u otro tipo (agua de lluvia o agua procedente de la fusión de nieve). El régimen se define entonces mediante términos como pluvial nival, pluvio-nival o nivo-pluvial.

El tipo de medida o de datos que corresponde a cada una de las variables es bien diferente. Supongamos que hemos obtenido los datos de las cuatro variables correspondientes a uno de los ríos y que los resultados son los siguientes:

Río	Longitud	Caudal máximo	Calidad del agua en la desembocadura	Régimen de alimentación
Adibide	72 kms	220 m3/s	Buena	pluvial

Los datos fictícios de la tabla que hemos elaborado nos sirven para explicar en qué difieren el tipo de datos asociados a cada variable:

- Las variables longitud y caudal son del mismo tipo en tanto que se definen mediante un valor, un número que expresa la cantidad de kilómetros o la de metros cúbicos por segundo.

- La variable sobre la calidad del agua no se define mediante un valor sino mediante un calificativo. Ahora bien, aunque no se trata de un valor sí expresa una cuantía, que puede ser mayor o menor. A diferencia de lo que ocurre con el caudal o la longitud, no es una medida precisa. Si la calidad del agua de un río tiene una calificación de muy buena y la de otro río solamente es buena, no podemos saber de forma precisa cuál es la medida de la diferencia entre ambos niveles de calidad. Diremos, en definitiva, que

---

Agua (Ver bibliografía). Aunque son informes muy voluminosos y también muy técnicos, contienen tablas y gráficos de resumen que permiten ver qué aspectos se han tenido en cuenta y se han medido para otorgar una calificación de calidad a las aguas del río.

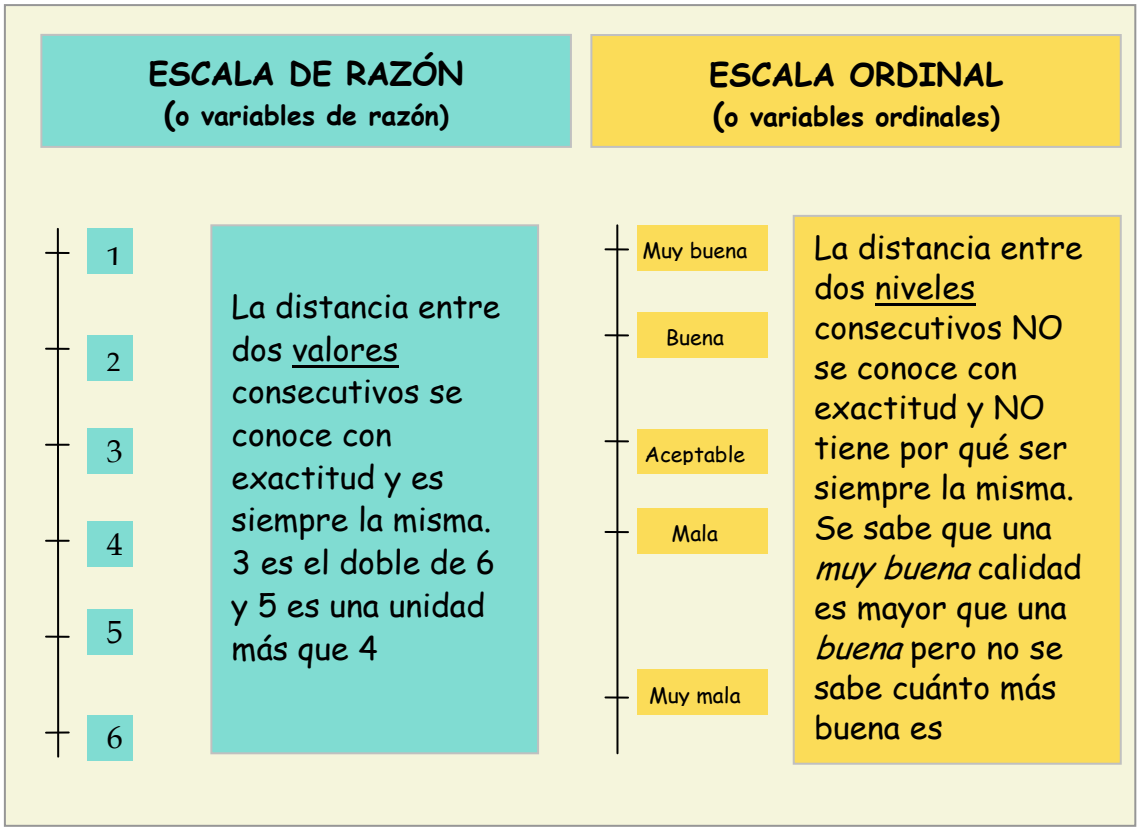
no se puede operar matemáticamente con las mediciones de este tipo.

- Por último, la variable sobre el régimen de alimentación es una característica más de los ríos, que se define mediante un término o una expresión que no alude a cantidad ninguna. El hecho de que la alimentación de un río sea pluvial y la de otro sea nival significa únicamente que son diferentes. El régimen de alimentación de los ríos es una característica variable, es decir, hay variaciones de un río a otro, pero no son de cantidad.

El hecho de que la variable o característica de los elementos de una población se defina de una forma u otra, es decir, mediante un número alusivo a la cantidad o una palabra o expresión, repercute en el tipo de procedimientos que se pueden utilizar para estudiarla. Por este motivo, antes de abordar cualquier tipo de análisis, es importante conocer los distintos tipos de variables y aprender a diferenciarlas entre sí:

Variable	Río Adibide	Río Ibai	TIPO DE VARIABLE O ESCALAS DE MEDICIÓN
Longitud	72 km	48 km	**DE RAZÓN:** el valor de la variable se establece mediante medidas de precisión estándar, al igual que la diferencia entre valores.
Caudal	220 m3/s	110 m3/s	
Calidad del agua	Buena	Mala	**ORDINAL:** las medidas de la variable son niveles dentro de una escala en la que no se conoce la posición exacata de cada nivel
Régimen de alimentación	Pluvial	Nival	**NOMINAL:** las medidas de la variable no implican cantidad. En lugar de hablar de valores de la variable se habla de modalidades de la variable.

Se verá todavía más claro si comparamos los tipos de variable o escalas de medición ordinal y de razón mediante una ilustración de la escala:



Como decíamos más arriba, el hecho de que las variables sean de un tipo u otro implica que el tratamiento que se puede hacer con ellas es también distinto. Cuando la escala de medición de la variable es ***de razón***, se puede hacer todo tipo de operaciones matemátias con los valores. Se pueden sumar, restar, multiplicar, dividir, etc. Si la escala de medición es ***ordinal*** los distintos niveles sólo se pueden ordenar de mayor a menor o al revés. Por último, cuando la variable es de tipo ***nominal***, es decir, se define mediante una característica o atributo, ni siquiera se pueden ordenar; se habla entonces, no de valores ni de niveles, sino de modalidades de la variable.

Aunque es cierto que el tipo de manipulación y de estudio que se puede hacer con variables es distinto según el tipo o la escala de medición, también lo es el hecho de que la Estadística ofrece posibilidades para el análisis de todas ellas. Lo importante en este caso es saber los procedimientos de análisis son diferentes en función del tipo de variables. A este respecto, hay que aclarar, por último, que en los materiales que aquí presentamos nos ocuparemos exclusivamente de los procedimientos para tratar con variables que pueden ser medidas en escala de razón.

Volviendo al tema de la población objeto de estudio en nuestro ejemplo, todavía queda un interrogante:

- Cuáles son realmente los ríos de Euskadi que debemos analizar. ¿Tenemos que tener en cuenta todos los arroyos y riachuelos o sólo los ríos principales? Y si sólo debemos encargarnos de estos últimos, ¿cómo sabremos cuáles son los ríos principales?. La duda puede surgir, por ejemplo, porque existen algunos riachuelos de pequeña entidad que nacen cerca de la costa y tienen recorridos muy cortos. Nos preguntamos si estos también debemos incluirlos en el estudio.

Sea cual sea la respuesta a nuestra pregunta, lo que queremos destacar aquí es que lo importante en cualquier estudio es disponer de una buena definición de la población que se va a analizar. Definir la población significa fijar o concretar con claridad y precisión las características que debe tener un elemento para ser considerado parte de la población o, en caso contrario, para excluirlo.

La definición nos debe permitir resolver cualquier duda a la hora de decidir si un elemento formará o no parte de la población. En el caso de los ríos, en el supuesto de que no hubiera que estudiarlos todos, se podría utilizar, por ejemplo, un límite mínimo de longitud por debajo del cual el río no formara parte de la población.

DESCUBRIENDO LA VARIABILIDAD: EL ANÁLISIS PRELIMINAR DE LOS DATOS

Una vez que hemos definido de forma correcta y sin ambigüedad la población que estudiaremos, de modo que hemos podido identificar, sin ninguna duda, todos los integrantes de dicha población y una vez, también, que hemos recopilado los datos pertenecientes a la variable la sigiente tarea consiste en **descubrir** la variabilidad de los datos. Para ello se realiza lo que se denomina un análisis preliminar de los datos.

El análisis preliminar de los datos consiste en ordenarlos y elaborar una representación gráfica de los mismos. El objetivo es detectar y corregir posibles errores y anomalías en los datos y obtener el máximo posible de inforrmación sobre su variabilidad, es decir, sobre las diferencias que hay entre los elementos de la población en relación a la característica o variable que queremos estudiar. Si la población fueran los ríos de Euskadi y la variable su longitud, analizaríamos en qué medida son diferentes las longitudes de dichos ríos. En esta primera fase de trabajo no se utilizan fórmulas, no se manipulan matemáticamente los datos; simplemente se observan siguiendo una serie de criterios.

El primer paso en cualquier trabajo de análisis es ordenar los datos en función de los valores de la variable. Podemos ordenarlos en orden creciente, del valor más pequeño al más grande, o decreciente. A continuación localizamos el valor más alto y el más bajo y calculamos la diferencia. A esta diferencia entre los valores extremos de la variable se le denomina **rango**. El cálculo del rango es una tarea sencilla (sobre todo si la realizamos mediante un programa de ordenador) pero muy provechoso porque nos proporciona una primera idea sobre la variabilidad de los datos que estamos analizando.

A continuación debemos hacer una revisión cuidadosa de los datos con el fin de detectar y corregir posibles errores y anomalías.

No debemos olvidar que todas la conclusiones que podamos obtener sobre la población que estamos estudiando tienen su punto de origen en los datos que hemos recogido. Entenderemos así que es preciso tener un cuidado exquisito a la hora de comprobar los posibles errores en los datos que puedan comprometer la validez de los resultados.

Primera aproximación a la variabilidad: el rango de la variable

Tal como hemos indicado, la primera operación que realizamos para tener una idea del grado de variabilidad es calcular la diferencia entre el valor más alto y el más bajo de nuestra serie de datos, es decir, el rango. Para calcular el rango utilizaremos, en este caso, los datos sobre la renta personal disponible en los municipios de Euskadi, correspondientes al año 2001[2]

Explicado de forma resumida, podemos decir que la renta personal disponible es la cantidad de dinero anual de la que una persona dispone para gastar o ahorrar. Esta cantidad disponible surge de restar a la cantidad de dinero que una persona obtiene mediante todo tipo de ingresos la cantidad que destina al pago de impuestos (directos) y de las cuotas obligatorias de la Seguridad Social[3].

En nuestro ejemplo lo que nos interesa es conocer en qué medida varía la renta de un municipio a otro. Nos interesa, por ejemplo, conocer si la renta de los habitantes de los distintos municipios es similar o, si por el contrario, existen municipios con una renta sensiblemente inferior o superior a la de otros.

Lanestosa	7.192
Valle de Carranza	7.275
Lapuebla de Labarca	7.800
Trucios	7.813
Kripan	8.131
.....	
Zigoitia	14.294
Gordexola	14.335
Sukarrieta	15.016
Leintz-Gatzaga	15.346
Laukiz	20.627

La tabla muestra los municipios que encabezan la lista con los valores de renta media personal disponible y los municipios que están al final de la lista, con los valores más bajos.

[2] Ver la serie completa de datos en la tabla nº 1 del Anexo.

[3] Una visión sencilla sobre la importancia de los estudios sobre la renta personal en Larrañaga Sarriegi, M. (2006) Distribución de la renta..

Observamos que son los municipios de Lanestosa con 7.192 euros anuales de renta per cápita y Laukiz, con 20.627, los que presentan los valores máximo y mínimo. Hacemos la resta y vemos que el rango de la variable renta personal per capita de los municipios de Euskadi en 2001 fue de 13.435 euros

Tenemos ya una primera idea sobre la variabilidad que estamos tratando de analizar. En este caso, sólo con el dato del rango, podemos afirmar que en Euskadi, las diferencias de renta entre los municipios son muy grandes, o lo que es lo mismo, que la variabilidad entre municipios es muy elevada: Laukiz, el municipio con mayor renta de Euskadi, posee una renta anual casi tres veces superior a la de Lanestosa.

Detección de posibles errores y anomalías

Cuando las diferencias entre los valores máximo y mínimo de la variable son elevadas, como en el caso de la renta personal disponible de los municipios de Euskadi en 2001, lo primero que hay que preguntarse, y comprobar, es si la diferencia observada es real o si podría tratarse de un error de los datos. Nos preguntamos en este caso si es real que los habitantes de Laukiz hayan podido disponer, de media, en 2001 de una cantidad de dinero tan superior a la del resto de habitantes y municipios.

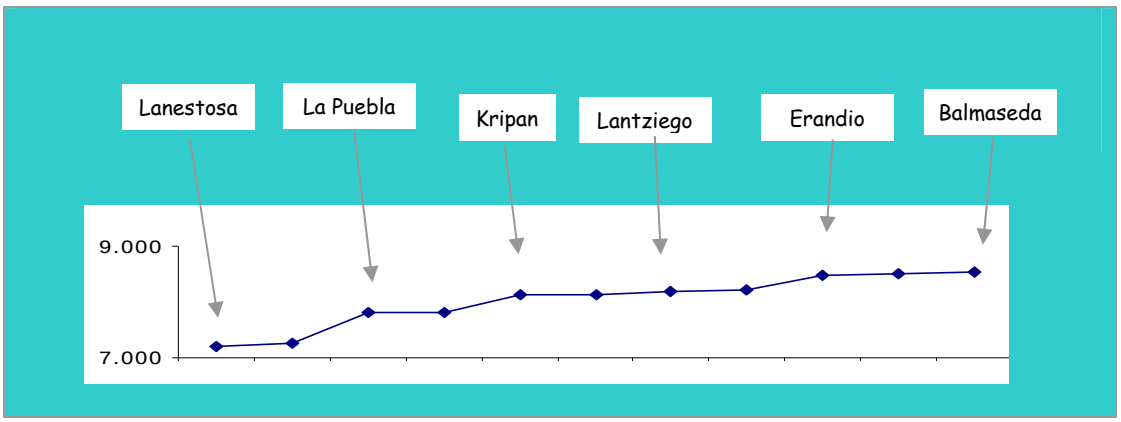
Para una persona recién iniciada en el análisis de datos la duda será, quizás, qué criterio utilizar para decidir que una serie de datos es sopechosa de contener errores. La tarea requiere algo de práctica pero es más sencilla de lo que puede parecer. Veremos cómo proceder con los datos de renta.

Si uno de los dos valores extremos de la serie fuera erróneo, es decir, si fuera anormalmente alto o bajo a consecuencia de un error producido en la fases previas de recogida de datos o de informatización, sería muy diferente del resto de los datos. Sería un valor discordante.

Para comprobar si se trata de valores raros, es decir, si los valores de Lanestosa y Laukiz son una excepción o no, lo que hacemos es mirar en el listado de datos si hay otros municipios que tienen valores de renta similares a los de dichos municipios:

Municipios con los valores de renta más bajos	
Lanestosa	7.192
Valle de Carranza	7.275
Lapuebla de Labarca	7.800
Trucios	7.813
Kripan	8.131
Ekora	8.138
Lantziego	8.185
Valle de Arana	8.218
Erandio	8.492
Elantxobe	8.507
Balmaseda	8.547

En la tabla vemos los 10 municipios que siguen en la lista a Lanestosa y que tienen, por tanto, rentas progresivamente superiores a la suya. Hay, por ejemplo, otros tres municipios más cuya renta no llega a los 8.000 euros. Vemos también que los aumentos de renta de cada municipio al siguiente forman una progresión que asciende suavemente. El gráfico siguiente proporciona una imagen visual de la progresión:



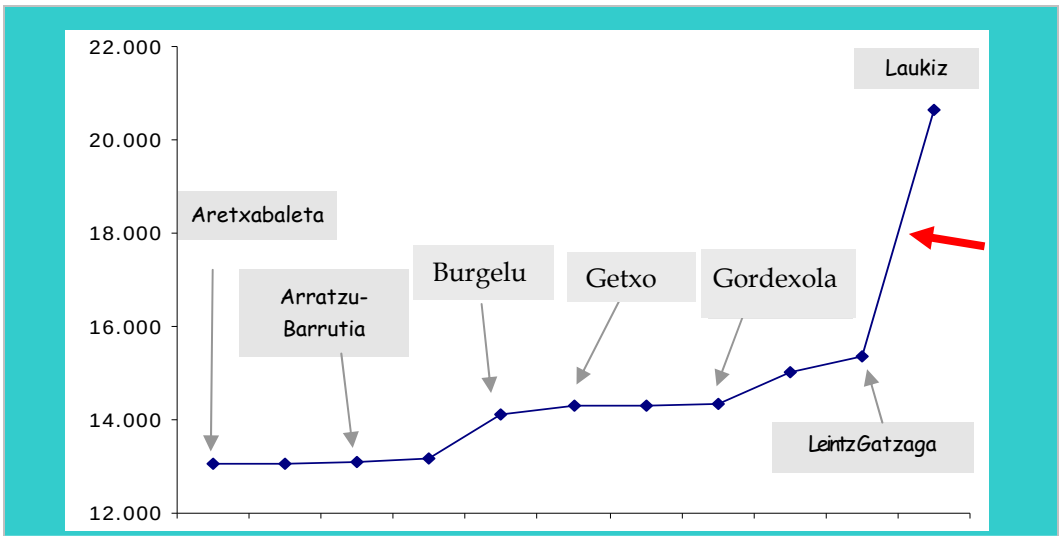
La pendiente suave de la gráfica muestra que las diferencias de valores de un municipio a otro son pequeñas, y, sobre todo, que conforman una progresión muy regular (las diferencias entre municipios son bastante similares)

A continuación haremos lo mismo con los once municipios cuyas rentas son las mayores de la comunidad autónoma:

Municipios con los valores de renta más altos	
Aretxabaleta	13.044
Legutio	13.069
Arratzu-Ubarrundia	13.100
Lezama	13.157
Burgelu	14.105
Getxo	14.291
Zigoitia	14.294
Gordexola	14.335
Sukarrieta	15.016
Leintz Gatzaga	15.346
Laukiz	20.627

La diferencia entre la renta de Leintz Gatzaga y Laukiz es de 5.281 €

En este caso, el aumento en el valor de la renta de un municipio a otro es también progesivo hasta que llegamos a los dos últimos municipios, entre los que existe una diferencia muy importante en sus valores de renta: 5.281 euros separan las rentas de Leintz-Gatzaga y Laukiz.



En esta ocasión el gráfico refleja claramente el salto que se produce entre los valores de renta de Leintz-Gatzaga y Laukiz.

Ya estamos en situación de afirmar que el valor de renta de Laukiz es un elemento raro, ya que no existe otro municipio que tenga un valor similar. El valor de Lanestosa, sin embargo, no es un elemento raro ya que existen otros municipios con valores cercanos.

El hecho de que los valores raros que encontramos a veces en los datos resulten sospechosos no significa que necesariamente tengan que ser errores. A veces ocurre, sin más, que determinados elementos de la población que analizamos tienen valores mucho más altos o bajos que el resto de los elementos. Sea como fuere, lo que debemos hacer es intentar comprobar la veracidad de los datos que nos resulten dudosos. En el ejemplo que estamos analizando se trata de intentar averiguar si el dato de Laukiz se corresponde con una situación real, es decir, si verdaderamente este municipio disfrutó en 2001 de unas rentas muy superiores a las del resto o si, por el contrario, se trata de un dato erróneo. Para intentar resolver la duda no nos queda otro remedio que revisar todo el proceso de recogida y manipulación de los datos, hasta llegar a la fuente original de los mismos.

Los datos de renta que estamos utilizando los hemos obtenido directamente de de la página Web del Instituto Vasco de Estadística (EUSTAT). Una segunda inspección de los listados de datos que ofrece la entidad en Internet nos ha permitido comprobar que no se ha producido ningún error y que, por tanto, el dato que tenemos de Laukiz, hasta donde nosotros podemos comprobar, es correcto.

Una vez comprobado que no se ha producido ningún error en el proceso de recogida y manipulación de los datos todavía puede ocurrir que, por el conocimiento que tenemos del tema que estamos estudiando o por alguna otra razón, sigamos pensando que se trata de un valor erróneo que se ha "colado" en los listados originales de la fuente de datos. Ante una situación similar tenemos dos opciones:

- Eliminar el dato que seguimos considerando sospechoso y realizar todos los análisis posteriores sin dicho dato, o

- Realizar un doble análisis de los datos incluyendo el supuesto dato erróneo en uno de los análisis y excluyéndolo en el otro.

En cualquier caso, sea cual fuere la decisión que tomemos, todo lo ocurrido debe quedar reflejado en el informe final del estudio.

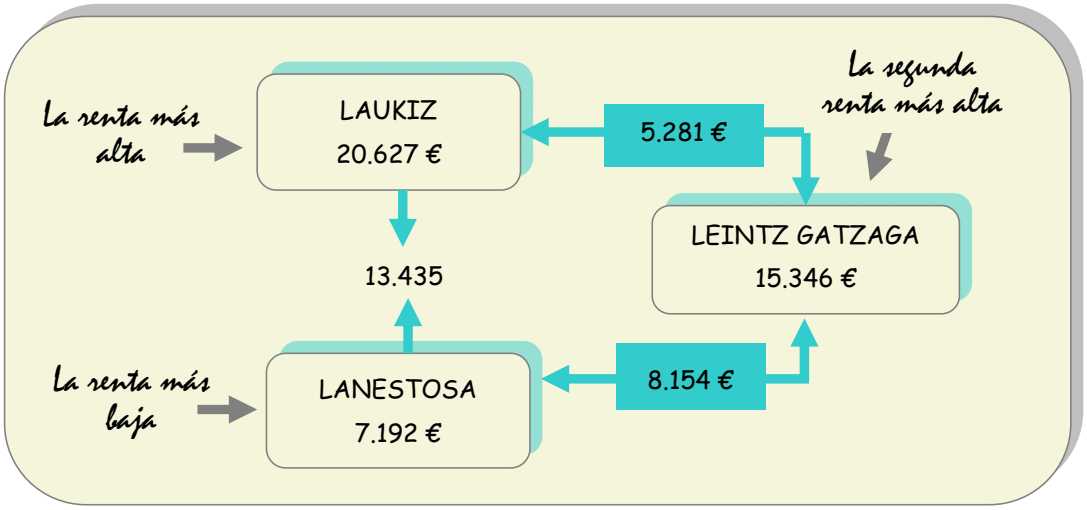
En ocasiones, todavía es posible encontrar recursos alternativos para seguir indagando. Cuando el tema que analizamos es de los que suscitan algún tipo de interés en la sociedad, puede ocurrir que, de haberse producido alguna situación particularmente llamativa, haya sido recogida por la prensa. Esto es lo que ocurrió con la renta de Laukiz, de cuya situación particular se hizo eco la prensa de Álava en 2007:

> PUEBLO RICO, PUEBLO POBRE
>
> Vacas gordas, vacas flacas
>
> A Laukiz y al Valle de Carranza les separan 15.000 euros de renta per cápita. A su pesar, son los municipios más rico y más pobre de Vizcaya. «Ni tanto ni tan poco», matizan sus lugareños. (…)
>
> Los vecinos de Laukiz no es la primera vez que escuchan la misma canción. Jóvenes o mayores, la primera reacción se reduce a una carcajada. «Ya estamos. Pregúntale a los de allí arriba, que son los que elevan nuestro nivel», sostiene Íñigo Barrena. No existe una distinción terminológica, pero Laukiz, salvando las distancias, es como uno de esos municipios de anuncio televisivo que compiten en fiestas populares: los de arriba y los de abajo. Pues en las alturas, residen los propietarios de una de las urbanizaciones más lujosas de Vizcaya, Unbe-Mendi, integrada por 127 chalés. Por eso, no es extraño toparse con turismos de marcas prohibitivas kilómetros antes de llegar a sus inmediaciones. «
>
> El correo, 12 de mayo de 2007

Es cierto que la noticia de prensa es de 2007 y nuestros datos son de 2001. No se trata de que esta noticia avale el dato de 2001; sirve solamente para confirmar que, en algunas ocasiones, la renta de Laukiz destaca por su diferencia con la del resto de municipios y para orientarnos sobre las razones de la misma.

LA VARIABILIDAD INTERNA DE LOS VALORES. LA DISTRIBUCIÓN

Para ahora ya tenemos claro que las diferencias de renta entre los municipios de la C.A. no son tan exageradas como parecía en principio. Es cierto que la diferencia de 13.435 euros entre los municipios con mayor y menor renta (Laukiz y Lanestosa) es real, pero también es cierto que hay un único municipio con un valor de renta tan elevado. Si comparamos, por ejemplo, la renta de Lanestosa con la del segundo municipio "más rico", la diferencia se reduce a 8.154 euros.



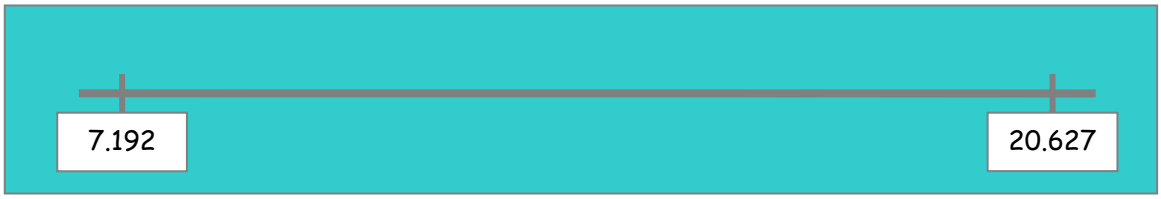
Tal como decíamos anteriormente, el rango nos proporciona una primera idea de las diferencias que presentan los valores máximo y mínimo de la variable, pero no nos dice nada sobre el resto de los valores. En el caso de la renta nos preguntamos, por ejemplo, si son muchos los municipios que tienen rentas próximas a las más bajas o si son mayoría los que tienen valores próximos a los de Leintz-Gatzaga. Podría ocurrir también que una mayoría de los municipios tuviera valores de renta a mitad de camino entre los de Laukiz y Leintz-Gatzaga…

El concepto de distribución de los valores de la variable

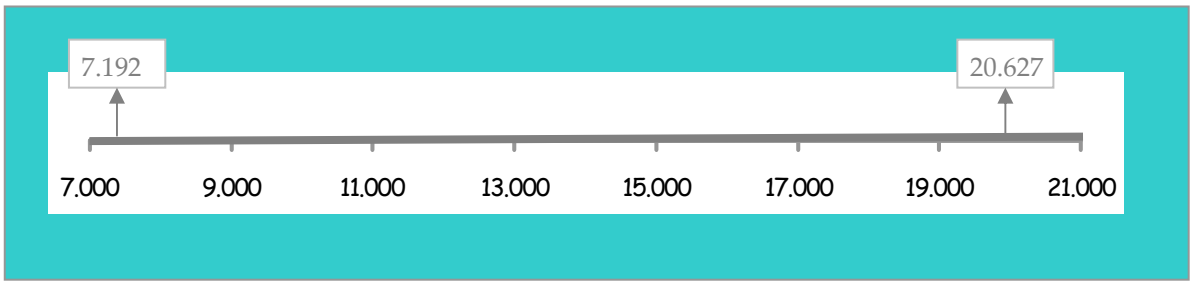
Para visualizar la similitud o diferencia de los datos entre sí, pero de toda la serie, utilizaremos un gráfico que, tras un vistazo, nos permitirá responder a preguntas como las que hemos formulado más arriba.

Iniciaremos el gráfico trazando un segmento de recta mediante el cual representaremos el rango de la variable, es decir, la distancia entre el valor más

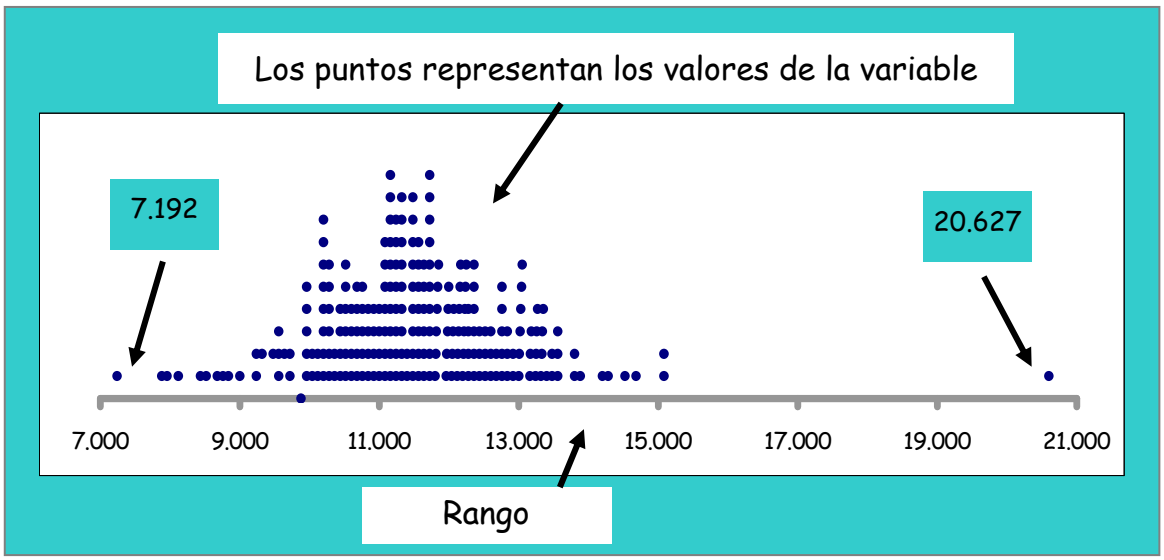
alto y el más bajo. En puntos próximos a los extremos del segmento anotamos los valores extremos del rango.



A continuación dividimos el segmento en partes iguales que representan tramos de incremento del valor de renta. Tenemos una escala graduada:



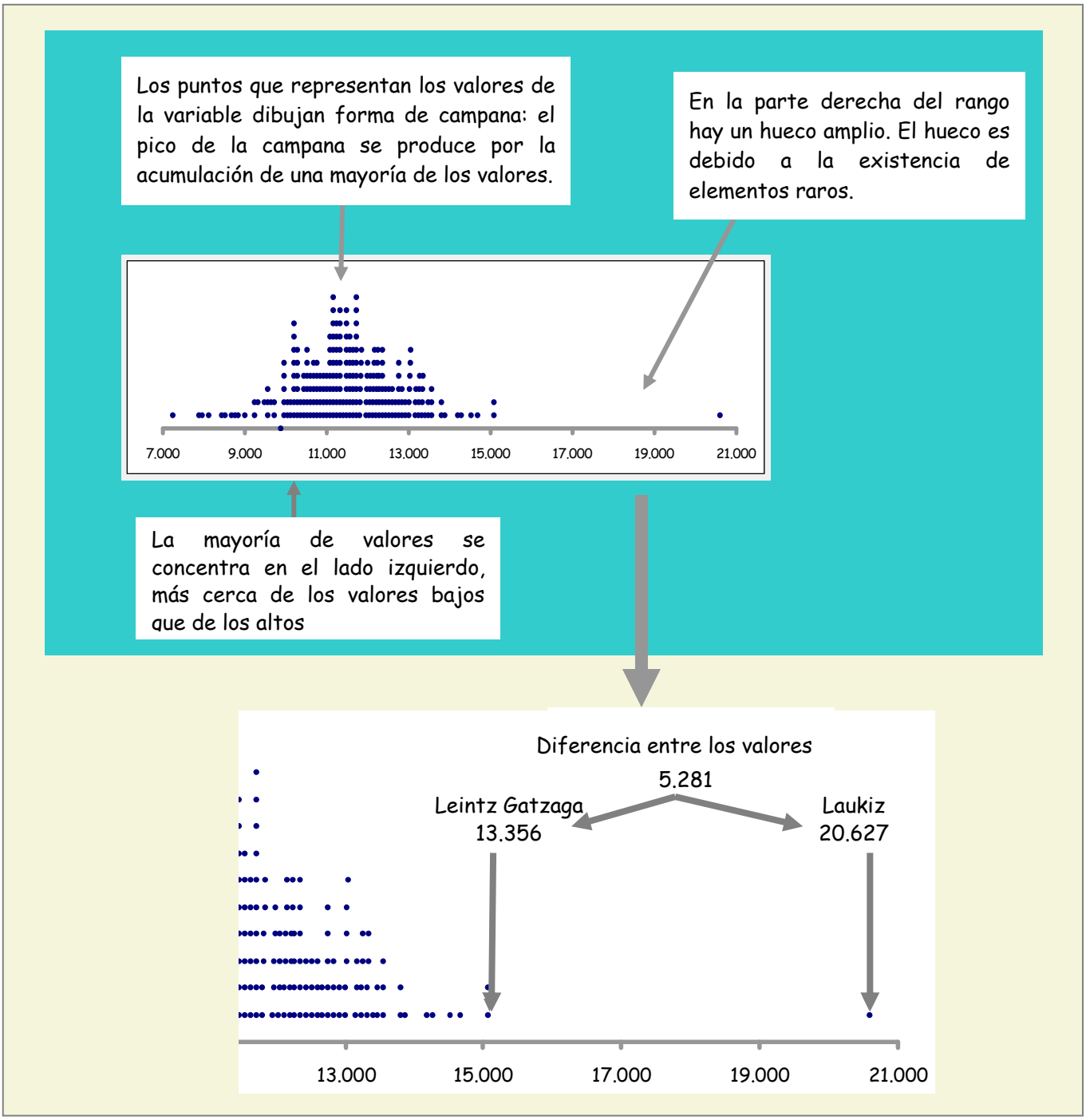
Sobre la recta podemos "colocar" ahora los distintos valores de la variable. Cada dato de la serie lo representamos mediante un pequeño símbolo situado en la posición que ocuparía por su valor. Cuando el espacio en el que debemos situar un punto está ya ocupado, colocamos el nuevo punto encima del anterior.



El resultado final es un gráfico que muestra la distribución de los valores de la variable en el interior del rango. Por ello, para referirnos a la información que nos proporciona el gráfico, se utiliza habitualmente la expresión ***distribución de los valores de la variable.***

El gráfico es una simplificación, ya que hemos sustituido los valores por puntos situados en una escala. La variabilidad la percibimos ahora de forma visual, fijándonos en la posición de los valores dentro del rango y en la distancia que hay entre ellos. Aunque sólo nos proporciona una idea aproximada, su ventaja, con respecto a la serie original de datos, consiste en la facilidad con la que podemos visualizar la distribución y las distancias que hay entre los valores, sin necesidad de hacer ningún cálculo. Un vistazo rápido es suficiente para percibir la variabilidad, ya que la mayor o menor proximidad de los valores en la recta es el indicativo de la misma.

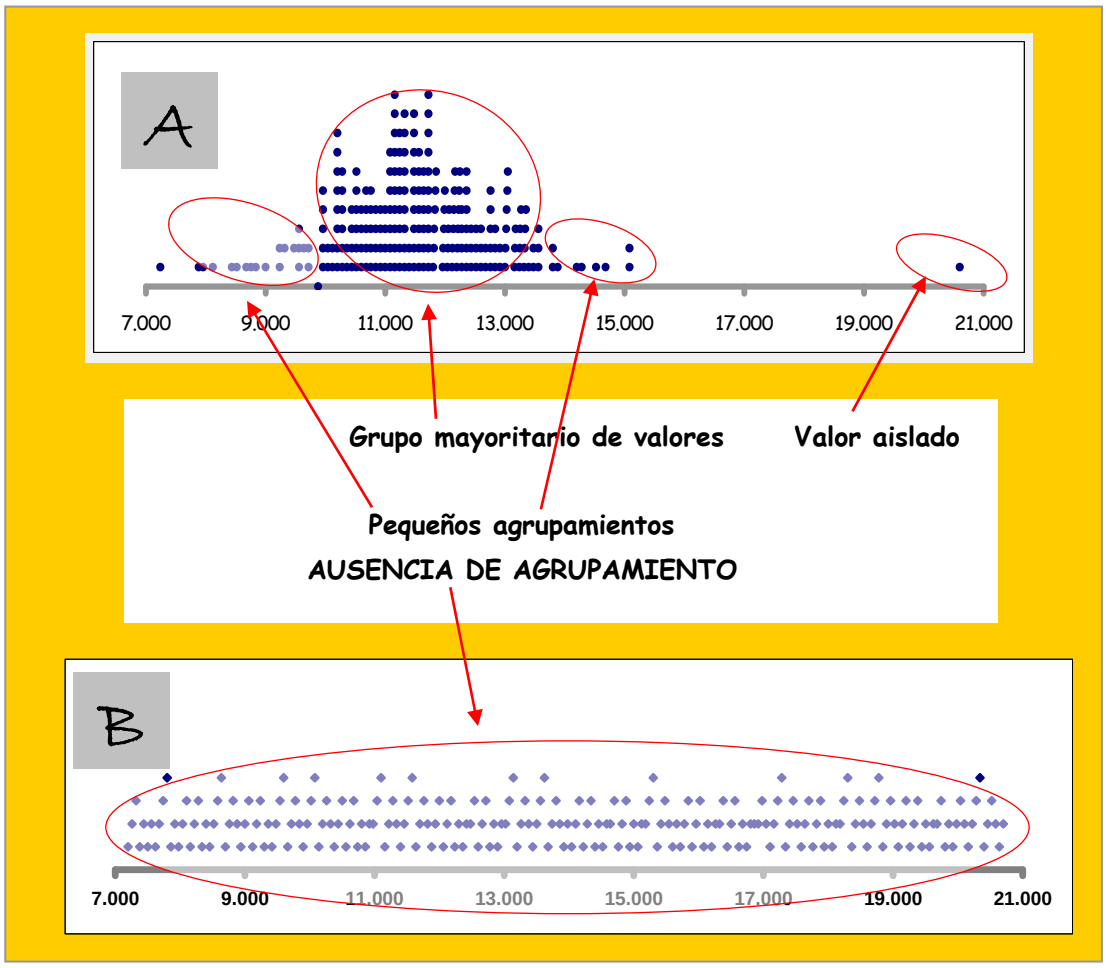
En este punto, el gráfico resulta una herramienta adecuada para preguntarnos sobre la información que se puede obtener a través de la observación de la distribución de valores.


Los puntos que representan los valores de la variable dibujan forma de campana: el pico de la campana se produce por la acumulación de una mayoría de los valores.
En la parte derecha del rango hay un hueco amplio. El hueco es debido a la existencia de elementos raros.
7.000
9.000
11.000
13.000
15.000
17.000
19.000
21.000
La mayoría de valores se concentra en el lado izquierdo, más cerca de los valores bajos que de los altos
Diferencia entre los valores
5.281
Leintz Gatzaga
13.356
Laukiz
20.627
13.000
15.000
17.000
19.000
21.000

Cuando analizamos una serie de datos, una de las características de ésta que más nos interesa es el modo en que se distribuyen los valores dentro del rango, ya que dicha distribución es un indicador directo de la variabilidad. Cuanto mayor es el agrupamiento de los valores, menor es la variabilidad y viceversa.

En la ilustración de la página anterior, elaborada con los datos de renta disponible de 2001, el gráfico nos muestra que existe un agrupamiento notable de los valores. Aunque hay algún valor raro, muy alejado del resto, la mayoría de los valores se encuentran agrupados en un sector del rango

Ahora veremos la comparación entre el gráfico de renta disponible **(A)** y otro que representa una situación completamente diferente **(B):**



En este último gráfico **(B)**, los valores no se agrupan en un sector del intervalo, sino que se distribuyen de forma regular a lo largo del mismo. Si los puntos del gráfico se correspondieran con valores de renta, podríamos decir que la

variabilidad entre los municipios es total; que no existe grupo alguno de municipio cuyos valores de renta sean más parecidos entre sí que los del resto.

Análisis de la distribución

El gráfico que hemos utilizado en el apartado anterior para representar el rango y la distribución de valores resulta muy útil para comprender los conceptos implicados, pero no es el tipo de gráficos que se utilizan habitualmente para analizar o describir la variabilidad. Se trata únicamente de una herramienta didáctica que no resulta válida para el análisis de la distribución porque:

- No resultaría ni factible ni práctico intentar representar mediante puntitos todos los valores de la variable sobre la recta del rango.

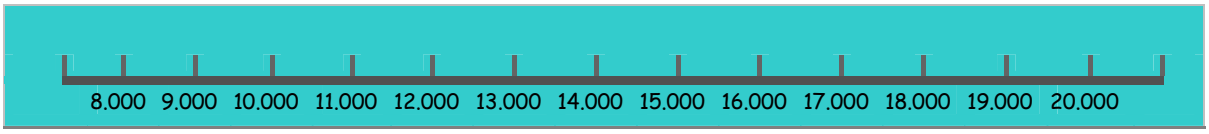
- Aunque nos proporciona una imagen de la distribución, le falta la precisión necesaria para sustentar una descripción. Por ejemplo, aunque en el gráfico veíamos que la densidad de puntos aumentaba de forma progresiva hacia la zona central del rango no podríamos precisar en qué punto del rango empieza a aumentar la densidad de valores.

- No cumple con un requisito que resulta esencial para avanzar en el anáisis de los datos: reducir el número inicial de datos de forma que sea más manejable.

Puesto que el gráfico que hemos utilizado hasta ahora sólo tiene un valor didáctico, presentaremos a continuación el tipo de gráficos y procedimientos habituales para el análisis de la distribución de valores.

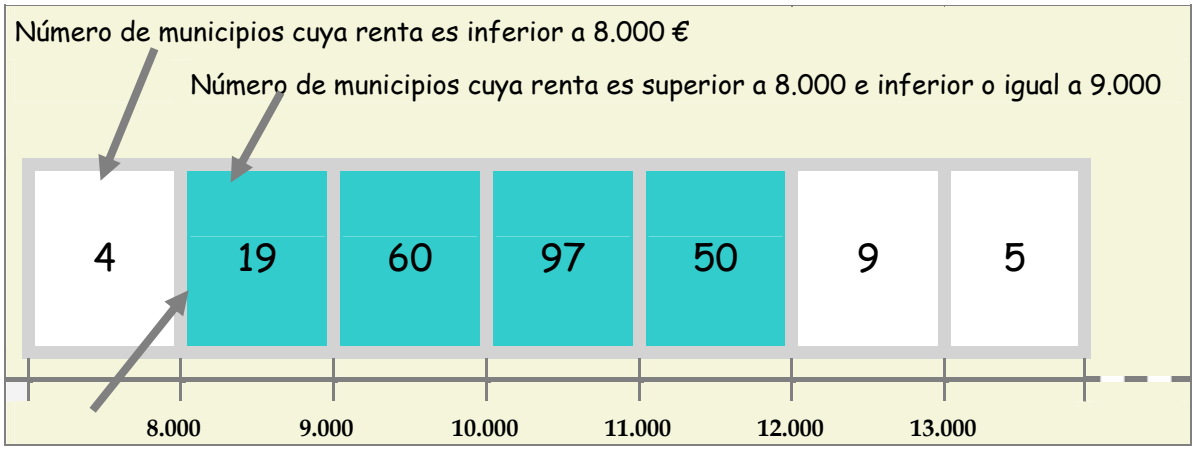
Cuando analizamos la distribución interna de los valores de una variable nuestro objetivo es conocer y caracterizar el comportamiento de los datos como colectivo. Queremos saber si existen agrupamientos de datos y, si los hay, en que sectores del rango se encuentran. Para obtener este tipo de información lo mejor es analizar el rango por sectores y ver cuántos valores de la variable hay en cada sector. Para comprender el procedimiento haremos este análisis con los datos de la renta personal disponible de los municipios de Euskadi en 2001.

Empezaremos dividiendo el rango de la variable en sectores:



A los sectores en que dividimos el rango se les denomina intervalos o clases

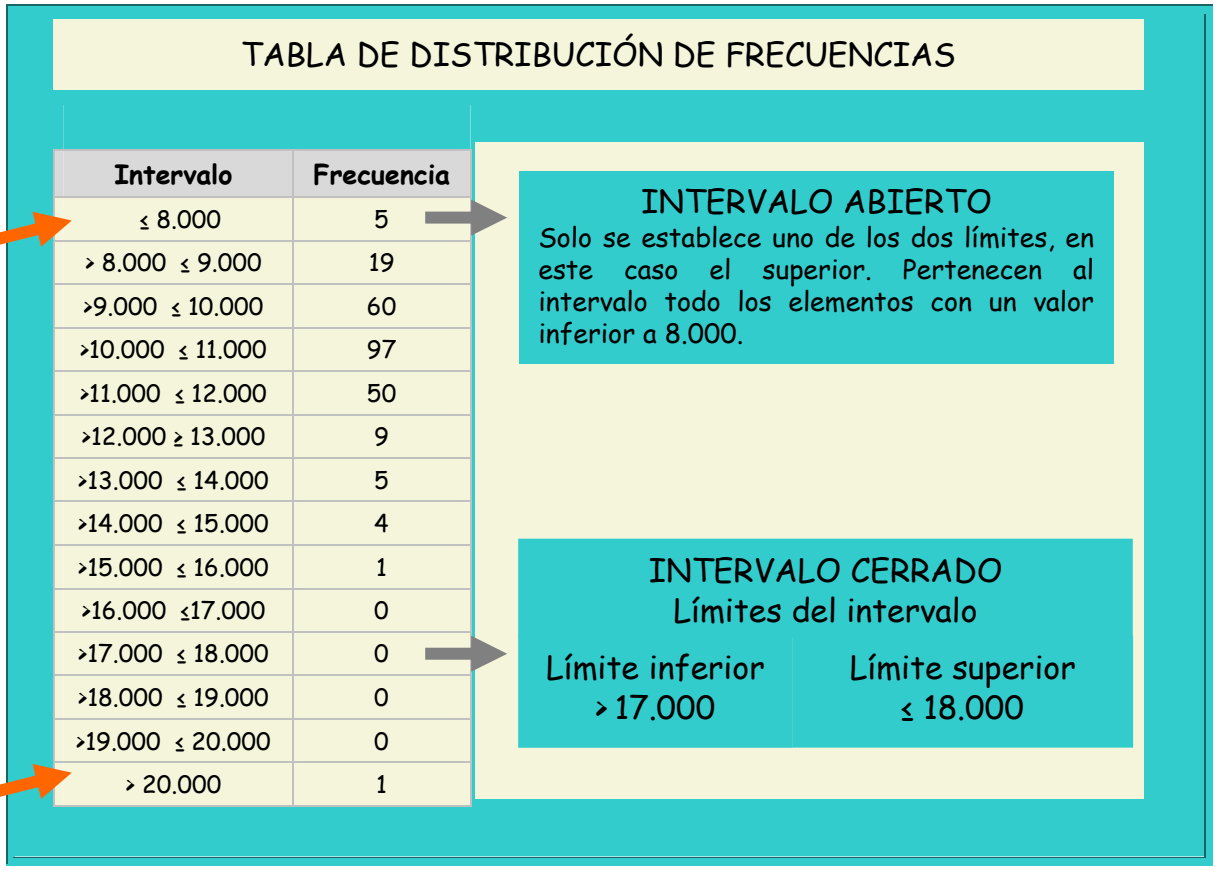
A continuación contaremos el número de valores que *caen* dentro de cada uno de los sectores (clases):



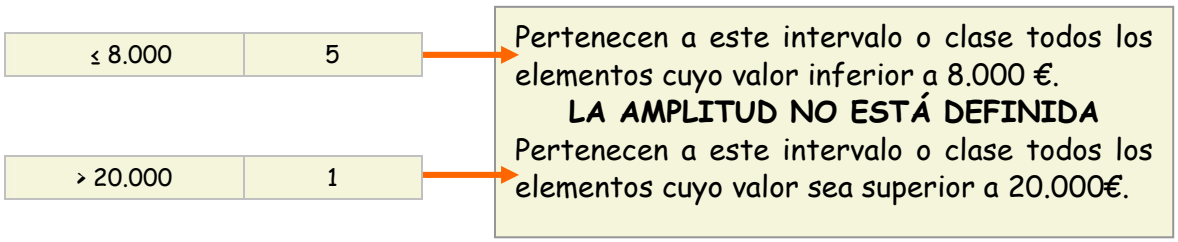
Ahora podemos precisar que en el punto del rango correspondiente a 8.000 € se produce un cambio importante en el agrupamiento (frecuencia) de valores: sólo son 5 los municipios con rentas inferiores a 8.000 y sube hasta 19 el número de municipios con rentas entre 9.000 y 10.000 €. El agrupamiento cambia también bruscamente a partir de los 12.000 €

Cuando agrupamos los valores de la variable en segmentos del rango –en intervalos- obtenemos los datos para elaborar una tabla de frecuencias de la variable. Es una tabla de dos columnas: en la columna de la izquierda indicamos los segmentos del rango o intervalos y en la de la derecha el número de valores de la variable que caen dentro de cada intervalo (frecuencia).

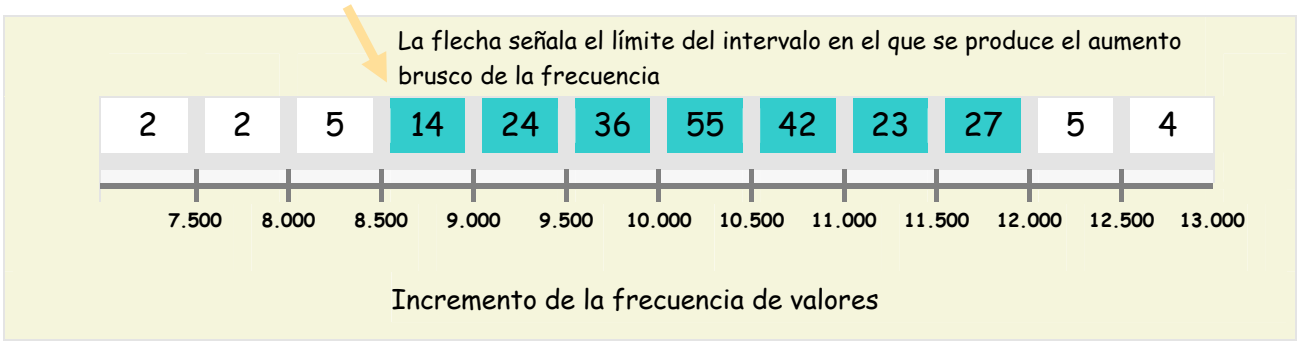
Los intervalos o clases están formados por dos valores, que son sus límites. Al valor más pequeño se le denomina límite inferior y al valor mayor se le denomina límite superior.



Hemos marcado mediante flechas rojas el primer y el último intervalo. En ambos casos podemos ver que sólo se indica uno de los dos límites. En estos casos, al faltar uno de los límites, se dice que el intervalo es abierto. El significado de esto parece evidente:



En la tabla superior podemos ver la distribución de frecuencias resultante de dividir el rango en intervalos de 1.000 €. Pero podríamos haberlo dividido el rango en intervalos o fragmentos más pequeños, como se muestra en el gráfico siguiente.



En esta ocasión hemos dividido el rango en intervalos de 500 €. La distribución que obtenemos es distinta de la anterior, no sólo porque los datos estén más desglosados, sino porque el punto del rango en el que la frecuencia de valores aumenta bruscamente es en el paso del intervalo >8.000 - ≤8.500 al intervalo >8.500 - ≤9.000.

Intervalo	Frecuencia
<7.500	2
≥7.500 ≤ 8.000	2
≥ 8.000 ≤ 8.500	5
≥8.500 ≤ 9.000	14
≥ 9.000 ≤ 9.500	24
≥ 9.500 ≤ 10.000	36
≥ 10.000 ≤ 10.500	55
≥ 10.500 ≤ 11.000	42
≥ 11.000 ≤ 11.500	23
≥11.500 ≤ 12.000	27
≥ 12.000 ≤ 12.500	5
≥ 12.500 ≤ 13.000	4
≥ 13.000 ≤ 13.500	5
≥ 13.500 ≤ 14.000	0
≥ 14.000 ≤ 14.500	4
≥ 14.500 ≤ 15.000	0
≥ 15.000 ≤ 15.500	2
≥ 15.500 ≤ 16.000	0
≥ 16.000 ≤ 16.500	0
≥ 16.500 ≤ 17.000	0
≥ 17.000 ≤ 17.500	0
≥ 17.500 ≤ 18.000	0
> 18.000	1

A la hora de elaborar tablas de distribución de frecuencias de los valores de la variable tenemos que tener claros unos cuantos principios fundamentales:

- Las tablas de distribución de frecuencias constituyen una simplificación con respecto a la serie de datos inicial. Como hemos visto en el ejemplo de la renta personal disponible, de una larga lista de valores, de la que es muy difícil extraer conclusiones, pasamos a una tabla de dimensiones mucho más reducidas. La ventaja de la tabla de distribución de frecuencias es que nos permite, de un simple vistazo, identificar los intervalos que acumulan un mayor número de valores y los que contienen pocos o, incluso, están vacíos. La conclusión parece evidente: a cambio de la ventaja que nos proporciona la observación de los datos agrupados en intervalos perdemos información; una vez agrupados los datos ya no vemos los valores individuales ni a que elemento de la población pertenece cada uno de ellos.

Serie original de datos (fragmento)	
Elementos de la población	**Valor**
Lanestosa	7.192
Karrantza Harana	7.275
Lapuebla de Labarca	7.800
Turtzioz	7.813
Kripan	8.131
Ekora	8.138
Lantziego	8.185
Harana	8.218
Erandio	8.492

Tabla de distribución de frecuencias	
Intervalo	**Frecuencia**
< 7.500	2
≥7.500 ≤ 8.000	2
≥ 8.000 ≤ 8.500	5

Al pasar de la serie inicial de datos a la tabla de distribución de frecuencias, las nueve filas de la serie inicial se han convertido en tres

PÉRDIDA DE INFORMACIÓN: sabemos que hay dos municipios con rentas inferiores a 7.500 €, pero no sabemos ni cuáles son ni sus valores concretos de renta.

- El tamaño o amplitud de los intervalos lo elegimos nosotros. Los intervalos o clases puede ser más grandes o más pequeños, como en los dos ejemplos que hemos mostrado. La cuestión es entonces cuántos grupos, clases o intervalos hacer. Por el momento no debe preocuparnos. Para decidir la amplitud de los intervalos lo más conveniente es hacer diferentes pruebas hasta encontrar una clasificación que muestre de la forma más adecuada posible la distribución de los valores de la variable en el rango.

Una clasificación de los valores de la variable en intervalos será adecuada cuando refleje las dicontinuidades naturales de la distribución de valores, es decir, cuando los límites de los intervalos coincidan con puntos del rango en los que la frecuencia aumenta o disminuye significativamente. En función de este criterio, de las dos pruebas que hemos realizado anteriormente con los valores de renta disponible sería la que corresponde a la tabla con intervalos de menor tamaño la que mejor refleja las discontinuidades naturales de la distribución

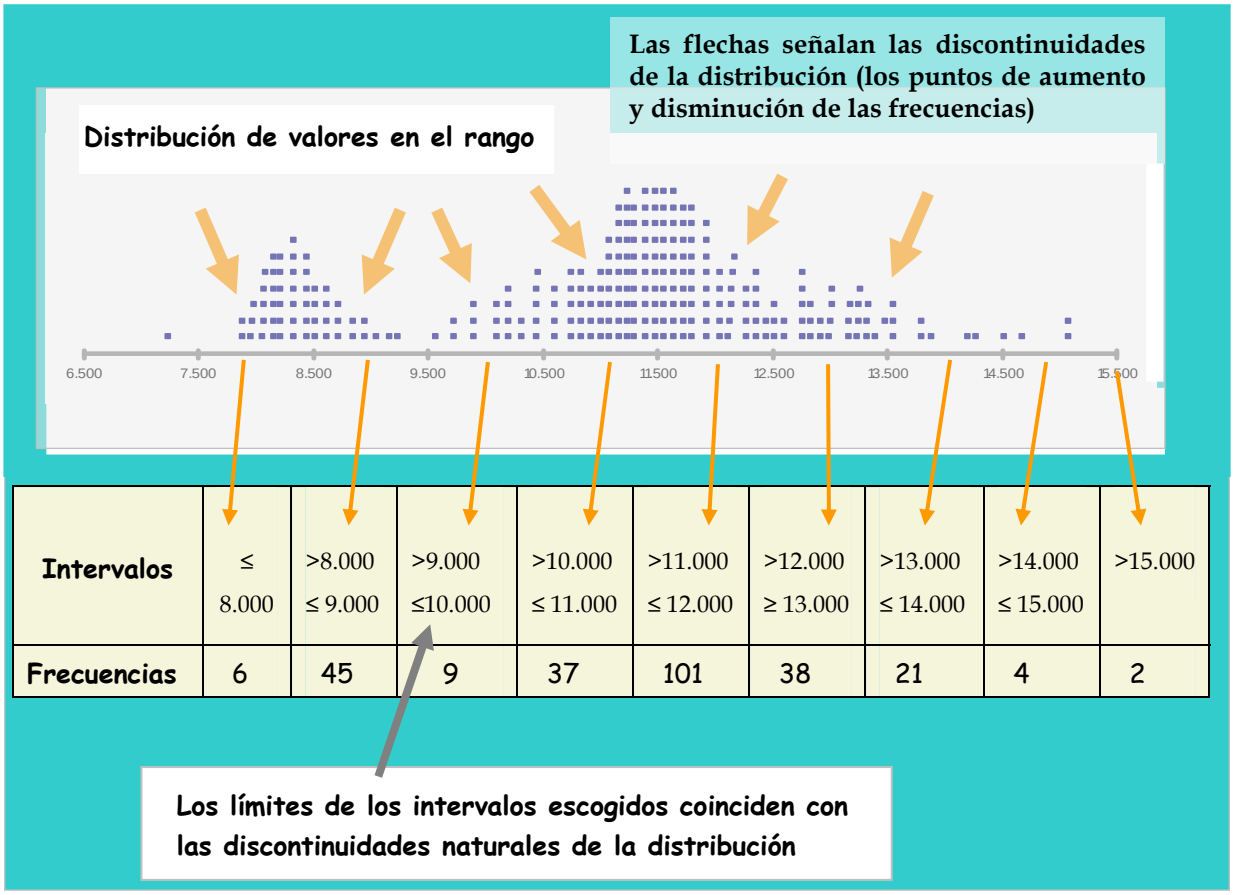
Veremos a continuación un gráfico que representa una distribución imaginaria. En dicho gráfico indicaremos cuáles son las discontinuidades naturales de la distribución, es decir, los puntos del rango en los que deberíamos situar, en la medida de lo posible, los límites de los intervalos. Trataremos, por último, de adecuar la distribución en intervalos de la tabla de frecuencias a la distribución propia de los valores de la variable.

El objetivo es doble:

- Obtener una tabla cuyo tamaño o número de clases permita visualizar de forma rápida la distribución de valores de la variable.

- Descubrir en qué punto o sector del rango se produce un aumento o descenso brusco de la frecuencia de valores.

Distribución de los valores en el rango



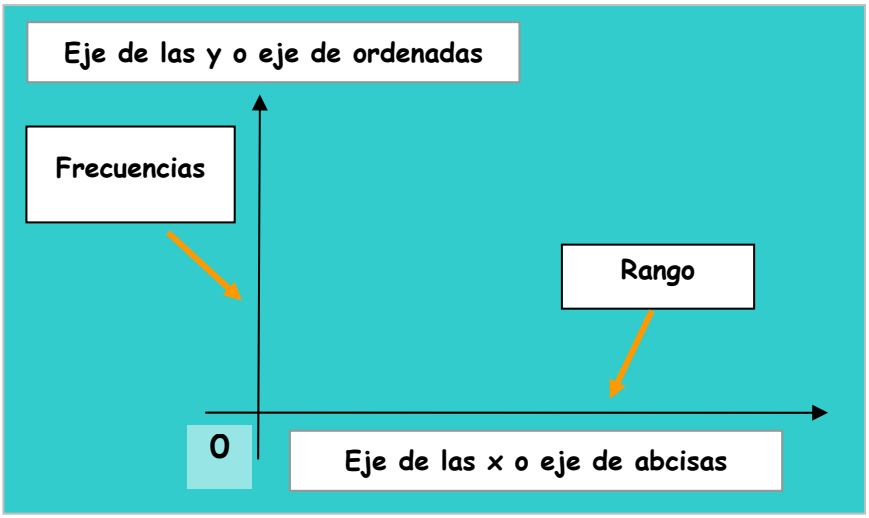
Intervalos	≤ 8.000	>8.000 ≤ 9.000	>9.000 ≤10.000	>10.000 ≤ 11.000	>11.000 ≤ 12.000	>12.000 ≥ 13.000	>13.000 ≤ 14.000	>14.000 ≤ 15.000	>15.000
Frecuencias	6	45	9	37	101	38	21	4	2

Los límites de los intervalos escogidos coinciden con las discontinuidades naturales de la distribución

Representación gráfica de la distribución: el histograma

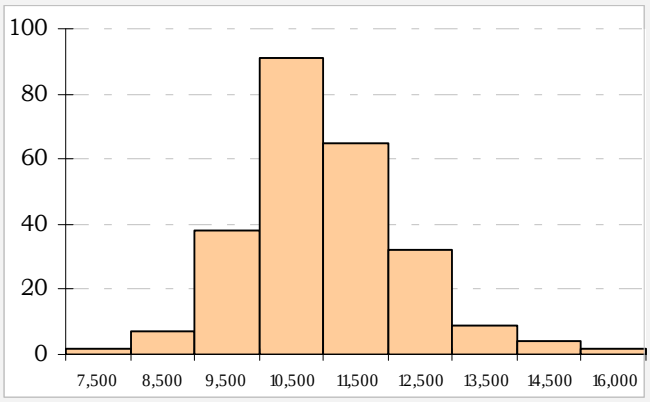
El histograma es un tipo de gráfico que se utiliza para analizar y mostrar la distribución interna de los valores de la variable. Se trata de una herramienta gráfica extremadamente útil para representar de modo adecuado las características más destacables de la distribución de valores de una variable. En definitiva, mediante el histograma se busca una lectura ágil, lo más visual posible, de los rasgos de la distribución.

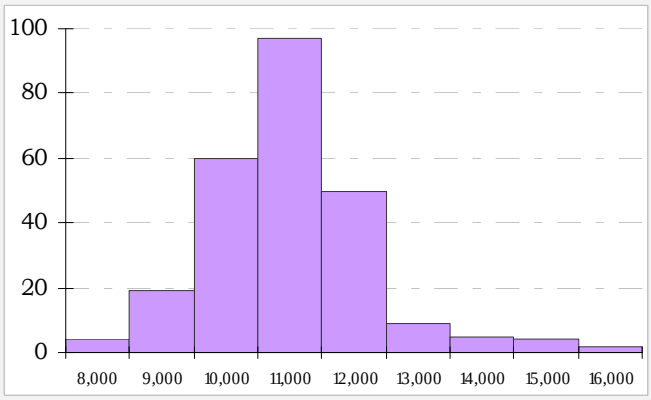
Como hemos dicho anteriormente, para elaborar el histograma se utilizan las tablas de distribución de frecuencias. Los valores de la tabla se representan en un eje de coordenadas cartesianas:



Se puede afirmar que elaborar un histograma es una tarea fácil siempre y cuando, al hacer la tabla de distribución de frecuencias, se hayan tomado las decisiones adecuadas. Como hemos visto anteriormente, para mostrar de la forma más adecuada posible las características de la distribución, lo más importante es una buena elección del número de clases o intervalos y de los límites de éstos. Repetiremos que nos corresponde a nosotros tomar estas decisiones y que, en función de las elecciones realizadas, la imagen de la distribución que ofrece el histograma puede ser distinta. De hecho, hemos visto los distintos resultados que obteníamos para los valores de renta personal disponible al elegir intervalos de diferente amplitud. A continuación veremos los distintos histogramas que se pueden realizar con los mismos datos.

Renta personal disponible de los municipios de Euskadi. 2001

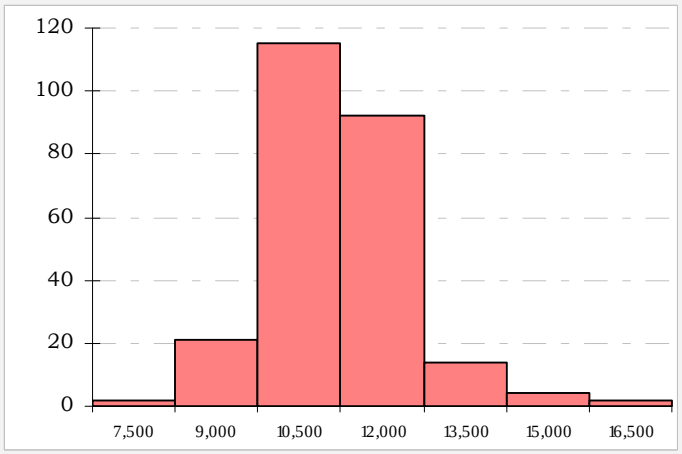


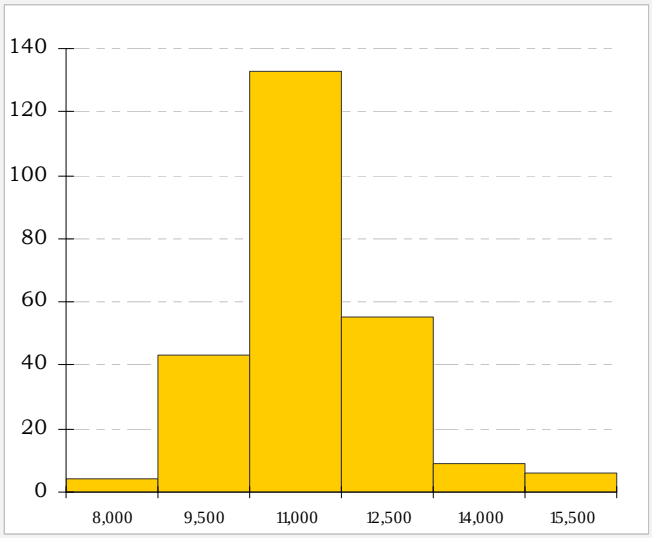


Amplitud de los intervalos: 1.000

Histograma nº 1. Primer intervalo: 6.500 -7.500 € Histograma nº 2. Primer intervalo: 7.000- 8.000 €

Al cambiar el punto de inicio (límite inferior del primer intervalo) la imagen que ofrece el histograma cambia: en el de la izquierda la sensación es que se acumula un mayor número de valores en la parte derecha del gráfico; en el histograma de la derecha ocurre lo contrario.

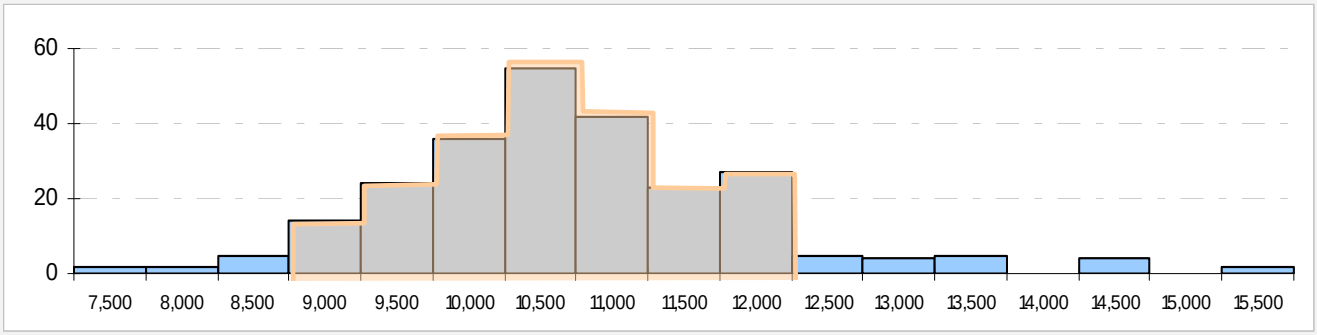




Amplitud de los intervalos: 1.500

Histograma nº 3. Primer intervalo: 6.000- 7.500 € Histograma nº 4. Primer intervalo: 6.500- 8.000 €

En este caso, al haber aumentado la amplitud de los intervalos, se obtiene la sensación de que hay una mayor acumulación de valores en la parte central del gráfico. Esta sensación es todavía más acentuada en el histograma de la izquierda, en el que las dos barras centrales, muy destacadas sobre el resto, acaparan la atención.



Histograma nº 5. Amplitud de los intervalos: 500

En los histogramas realizados hemos utilizado distinto número de intervalos y, por tanto, distintas amplitudes. Hemos variado también el límite inferior del primer intervalo, empezando en algunos en 6.000 € y en otros en 6.500 €. Es evidente que la impresión visual resultante varía de unos a otros. A la vista de las diferencias cabe preguntarse cuál de los histogramas resultaría más adecuado para representar gráficamente la distribución que estamos analizando; nos podemos preguntar también qué es mejor, si utilizar pocos o muchos intervalos. Lo cierto es que no hay una respuesta única válida para todas las ocasiones. Lo único que se puede decir, a modo de principio general, es que se debería optar por una división en intervalos que, siendo lo más sencilla posible, no enmascare los rasgos propios distribución. Aplicando este principio a los histogramas que acabamos de realizar podríamos concluir lo siguiente:

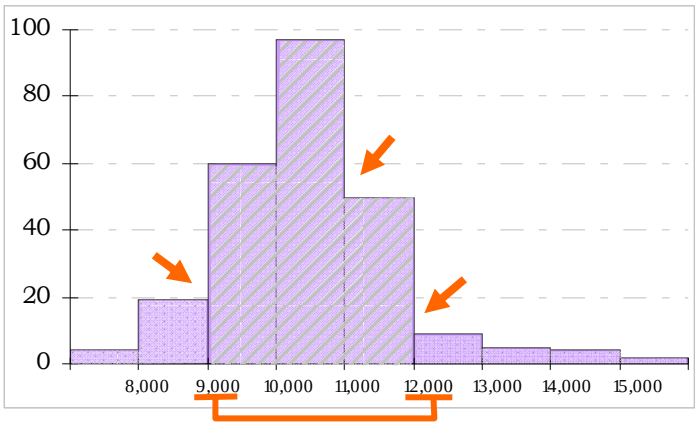
- El histograma realizado con intervalos de amplitud de 500 muestra de forma impecable las discontinuidades propias de la distribución (los puntos en los que se produce un aumento o un descenso notable de las frecuencias). Nos permite ver de forma clara que la renta de la mayoría de los municipios se sitúa entre los 8.500 y los 12.000 € y que hay un mayor número de valores en la parte izquierda del histograma (mayor número de municipios del lado de las rentas menores) Pese a todo, revisaremos el resto de gráficos para ver si existe la posibilidad de acortar el número intervalos sin perder información esencial.

- El histograma nº 2, (amplitud de intervalo 1.000 €) mucho más simple que el

anterior, muestra también las características esenciales de la distribución de frecuencias, fundamentalmente el brusco descenso que se produce a partir de 12.000 €. Es igualmente perceptible la mayor acumulación de municipios en el lado de las rentas más bajas.

- Los intervalos de los histogramas nº 3 y 4 son excesivamente amplios, de modo que distorsionan los rasgos propios de la distribución.

Finalmente pensamos que el histograma nº 2 proporciona una imagen idónea de la distribución de los valores de renta pesonal media disponible en los municipios de Euskadi, en el año 2001. Permite observar de forma clara:



- Que la renta media de los habitantes de los municipios de Euskadi oscila en un rango de valores muy amplio

- Que pese a la gran variabilidad general, existe una mayoría de municipios con rentas más similares entre ellos (el área rallada), ya que son muy pocos los que tienen rentas muy bajas o muy altas (áreas punteadas).

- Que en la mayoría de municipios la renta media de sus habitantes se encuentra entre 9.000 y 12.00 euros. Podemos entonces hablar de que las rentas se concentran fundamentalmente en este intervalo.

- Que hay más municipios con rentas en la mitad inferior del rango que en la mitad superior.

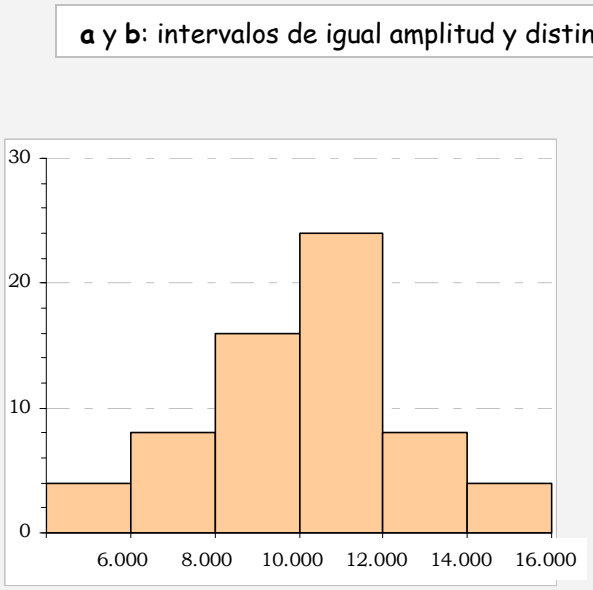
- Una buena aproximación a las "discontinuidades" naturales: los puntos en los que se produce un descenso o aumento brusco de las frecuencias. Las flechas rojas en el gráfico señalan los límites de los intervalos escogidos que

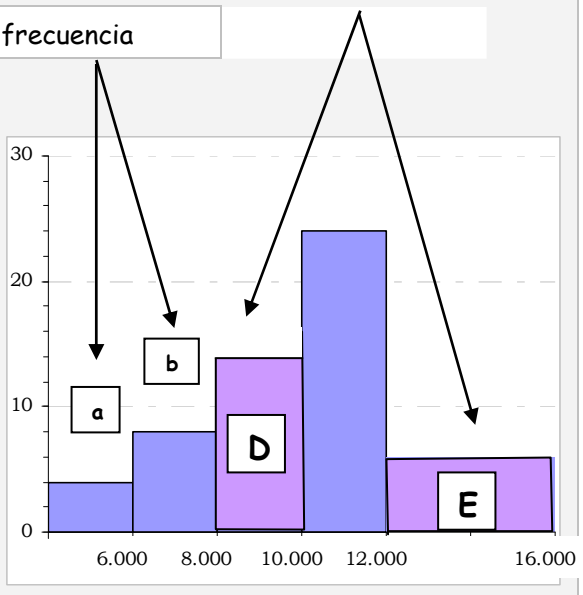
coinciden con las discontinuidades naturales que podemos observar en el gráfico de la página 27.

Los cinco histogramas que hemos realizado tienen una característica en común: los intervalos o clases de cada uno de ellos son de igual amplitud. Esto no es una condición obligatoria pero sí facilita la lectura e interpretación del gráfico debido al hecho de que en los histogramas la frecuencia de cada clase viene representada por el área del rectángulo y no por su altura. Cuando todos los intervalos son de idéntica amplitud para comparar las distintas frecuencias de cada uno basta con mirar la altura de los rectángulos o barras. Ante intervalos de diferente amplitud hay que fijarse no en las alturas sino en las áreas. Una ilustración permitirá comprender mejor esta idea:

D y E: intervalos de distinta amplitud, igual frecuencia, **misma área**, **distinta altura**

a y b: intervalos de igual amplitud y distinta frecuencia





Intervalo	Frecuencia
> 4.000 ≤ 6.000	4
> 6.000 ≤ 8.000	8
>8.000 ≤ 10.000	16
>10.000 ≤ 12.000	24
>12.000 ≤ 14.000	8
>14.000 ≤ 16.000	4

Intervalo	Frecuencia
> 4.000 ≤ 6.000	4
> 6.000 ≤ 8.000	8
>8.000 ≤ 10.000	12
>10.000 ≤ 12.000	24
>12.000 ≤ 16.000	12

En el histograma de la izquierda está representada una distribución mediante intervalos de igual amplitud. Siendo así, nos fijaremos solamente en las alturas, que son proporcionales a la frecuencia: el segundo intervalo agrupa dos veces más valores que el primero y, por tanto, su altura es el doble; el tercer intervalo, también con una frecuencia doble a la del segundo es dos veces más alto que éste.

En el histograma de la derecha observamos un intervalo de amplitud diferente a la del resto (E, de 12.000 a 16.000). Vemos también que este intervalo distinto, de doble amplitud, tiene la misma frecuencia que el intervalo D. Si nos fijamos en las áreas de los intervalos D y E veremos que son iguales, pero si atendemos a la altura comprobaremos que son diferentes.

Aunque la lectura del histograma es más ágil y directa cuando utilizamos intervalos de igual amplitud, en algunas ocasiones se aconseja realizar intervalos distintos. Este sería el caso, por ejemplo, de las distribuciones en las que muchos de los valores se concentran en una pequeña parte del rango y sólo unos pocos valores se diseminan en el resto. Ante distribuciones de este tipo, puede resultar adecuado utilizar intervalos amplios para las zonas del rango con valores dispersos y utilizar intervalos más pequeños para las zonas que acumulan gran número de valores.

Al presentar las tablas de distribución de frecuencias hemos hablado de los intervalos abiertos y cerrados. Hemos visto que es posible, y frecuente también, realizar tablas en las que el primer y el último intervalo se dejan abiertos. Ahora bien, en el caso del histograma todos los intervalos tienen que ser cerrados. Esto es así porque, como acabamos de ver, el área de cada uno de los rectángulos que forman el histograma se define en función de la amplitud del intervalo y de la frecuencia de los valores que acumula. Está claro entonces que sólo se puede calcular el área para los intervalos cerrados.

Para finalizar el tema haremos un cuadro resumen de las principales características del histograma:

- El histograma es una herramienta gráfica extremadamente útil no sólo en el momento de presentar los resultados del análisis. Resulta de gran ayuda también durante la fase de exploración de los datos, ya que nos permite identificar la estructura natural de distribución de los valores de la variable.

- Dependiendo de las decisiones que se tomen en relación al número y amplitud de los intervalos del histograma, la imagen que éste ofrece sobre la distribución de los valores de la variable, puede ser muy diferente.

- No existe ninguna norma que permita decidir cuál es el número y la amplitud idónea de los intervalos de un histograma. Son decisiones que debe tomar el autor del histograma. Lo importante es reflejar de la manera más sencilla posible la estructura propia de los datos.

- Cuando se elabora un histograma con el fin de presentar una imagen gráfica de la distribución, es conveniente hacer varios ensayos, analizar las diferencias y escoger el que consideremos más idóneo.