

	<p>Temas de Estadística Práctica Antonio Roldán Martínez Proyecto http://www.hojamat.es/</p>
	<p>Tema 1: Recogida de datos Resumen teórico</p>

Recogida, tabulación y organización de datos

<u>Tipos de medida</u>	<u>Constantes y variables</u>	<u>Recogida de datos</u>	<u>Organización de datos</u>	<u>Agrupación de datos</u>
--	---	--	--	--

Tipos de medida

Características y modalidades

Llamamos *característica* a cualquier propiedad de objetos o personas que deseamos estudiar en Estadística. Las distintas formas de presentarse esta característica se llaman *modalidades*. Por ejemplo, 1,82 y 1,65 cm. son dos modalidades de la característica *altura*, y varón y mujer dos modalidades de la característica *sexo*.

Si una característica sólo tiene dos modalidades la llamaremos *dicotómica*.

Medida

Es la operación de asignar un número a cada una de las modalidades de una característica, convirtiendo algunas relaciones entre modalidades en sus correspondientes relaciones entre los números que representan su medida. Por ejemplo, los ciudadanos españoles se corresponden con su DNI, su peso con los kg que da la balanza, y el sexo se puede corresponder con los símbolos V y M, etc.

Escala de medida

Es un conjunto básico de modalidades y números (considerados como sus medidas) a partir del cual se construye un procedimiento para medir las restantes modalidades. Así, la escala centígrada de temperaturas se basa en asignar 0º a la temperatura de fusión del agua y 100º a la de ebullición. La medida de la dureza de los minerales se basa en un conjunto determinado de ellos, desde el talco hasta el diamante.

En Estadística consideramos cuatro tipos básicos de escalas, desde la *nominal*, que apenas permite análisis, hasta la de *razón*, que es la más completa.

Se diferencian unas de otras esencialmente en las operaciones que permiten.

Escala nominal

Una escala se llama *nominal* si la única relación que tiene en cuenta es la de *igualdad* (y su contraria la desigualdad). Suele estar formada por nombres, códigos o números considerados como etiquetas (como el DNI). Así, son nominales los apellidos, la Comunidad Autónoma, el distrito postal, etc.

Si una escala nominal se construye con números (como los distritos postales), sólo se admitirán entre ellos las relaciones de igualdad y desigualdad, pero no operaciones como suma o producto. No podemos sumar dos Comunidades Autónomas. Tampoco tendría utilidad multiplicar dos números de teléfono.

Escala ordinal

La escala *ordinal* añade a la nominal la posibilidad de ordenar los datos, es decir, considera las relaciones de *mayor* y *menor*, aunque no se plantea una distancia entre unas medidas y otras. La escala de Insuficiente, Suficiente, Bien, Notable y Sobresaliente es ordinal. No se considera si entre Bien y Notable existe la misma diferencia que entre Notable y Sobresaliente. Son ordinales muchas de las medidas en Psicología o Ciencias de la Educación.

Escala de intervalos

Se introduce una medida tipo (o patrón) llamada *unidad* y se tiene en cuenta cuantas unidades están comprendidas entre dos medidas distintas. Tienen sentido, además de la igualdad y el orden, las *diferencias* entre dos medidas. Podemos sumar y restar medidas, pero no tienen sentido sus cocientes. Son de intervalo la gran mayoría de las escalas de las ciencias experimentales: temperatura, peso, velocidad, intensidad de la corriente eléctrica, etc.

Escala de razón

En esta escala se le da también un sentido a las *razones* entre dos medidas, es decir, las veces que una medida contiene a la otra. Fue la medida por excelencia de la Geometría griega y se ha trasladado a todas las Ciencias Sociales y de la Naturaleza. Se distingue también por la existencia de un *cero verdadero*, no convencional. Así, la escala centígrada de temperatura es sólo de intervalo y la Kelvin es de razón.

Podemos dividir medidas, pero sólo para su comparación o razón.

Resumen

- En escala nominal sólo distinguiremos la igualdad o desigualdad entre dos modalidades.
- La escala ordinal añade la posibilidad de establecer un orden.
- Si se usa una unidad y tienen sentido las diferencias, se trata de una escala de intervalo.
- Por último, si se pueden comparar dos medidas mediante un cociente o razón, la escala es de razón.

Constantes y variables

Llamaremos *constante* a una característica que sólo admite una modalidad, por ejemplo la constante de gravitación universal. Por el contrario, *variable* es aquella que admite varias modalidades, a las que también llamaremos *datos* o *valores*.

Tipos de variables

Una variable se llama ***cualitativa*** si sólo admite una medida nominal. Son cualitativas la localidad de nacimiento de cada persona, el color de su piel o su domicilio.

Llamaremos ***casi cuantitativa*** a aquella que admite como máximo una medida ordinal, como podría ser la motivación en el estudio, el grado de extraversión o la valoración de una prueba de gimnasia rítmica.

Por último, llamaremos ***cuantitativa*** a aquella variable que admita medidas de intervalo o de razón. Si entre cada dos valores pueden existir infinitos otros, la llamaremos ***continua***, como el peso, la estatura, etc. y si sólo admite un número finito de valores entre cada dos, recibirá el nombre de ***discreta*** (edades medidas en años, número de hermanos, etc.). A causa de la falta de precisión en las medidas, muchas variables continuas se pueden tratar como discretas, y lo contrario sucede cuando existen tantos valores distintos que puede ser útil tratar como continua una variable discreta.

Los datos de cualquier tipo de variable pueden ser simples, de un solo valor, y se llaman en este caso ***unidimensionales***. Llamaremos ***bidimensionales*** a los datos compuestos de dos valores, como un resultado de Baloncesto (89 a 76). Existen además datos ***tridimensionales*** y, en general, multidimensionales.

Recogida de los datos

Los datos se recogen de conjuntos reales, por lo que debemos hacer algunas distinciones:

Población y muestra

Llamaremos **población** a un conjunto bien definido por ciertas características que deseamos estudiar: La población de una Comunidad Autónoma, los aprobados de 2º de Bachillerato en mi Centro, los profesores de E.S.O. en la Delegación Norte, etc.

Como las poblaciones pueden contener muchos elementos distintos, elegiremos una **muestra**, o subconjunto de ellas, que sea más fácil de estudiar que la población. Existe toda una ciencia para conseguir muestras representativas de la población.

Un estudio estadístico de la población recibe el nombre de **censo** y el de la elección y estudio de una muestra, **muestreo**.

Un número que caracterice o describa una población recibe el nombre de **parámetro**. La estatura media de los alumnos y alumnas de 16 años es un parámetro de esa población, o la Renta per cápita de la población española. Si ese mismo número lo calculamos en una muestra, recibe el nombre de **estadístico**. Si mido la estatura media de los alumnos de mi clase de Bachillerato estaré calculando un estadístico. Cuando se calcula un parámetro a partir de un estadístico, estaremos realizando una **estimación**. Un caso representativo es el de los sondeos antes de unas elecciones.

Recuento

Los censos y muestreos necesitan siempre un **recuento de datos**.

Si los recuentos de datos se efectúan manualmente, se suelen representar mediante trazos verticales:

María ||||

José Mari |||||||

aunque son más útiles configuraciones de cinco en cinco o de diez en diez.

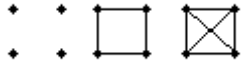
☑ ☑ |||

Así, esta configuración de recogida representa el número 13 mediante cuadrados con diagonal que representan el número 5

HHH

También es útil representar el número 5 con cuatro barras y la quinta tachando a las demás.

Si se desea contar el número diez, se puede seguir esta secuencia de puntos y rectas:



(cuatro puntos más cuatro rectas más dos diagonales: 10)

Organización de los datos

Una vez efectuado el recuento de datos dispondremos de unos números llamados **frecuencias** asignados a los distintos valores de las variables. A la operación de construir ese conjunto se le suele llamar coloquialmente **confeccionar una estadística**.

Frecuencias

El número de veces que se repite un valor concreto en una recogida de datos se llama **frecuencia absoluta** o simplemente frecuencia. Se representa por la letra **n** o por la **f**, según los distintos textos. Aquí usaremos **n**. La suma de todas las frecuencias coincide con el número total de elementos estudiados, al que representaremos por **N**.

Representaremos esto así

$$\sum n = N$$

Para poder comparar distintos conjuntos de datos es más útil el uso de las **frecuencias relativas o proporciones**, que son los cocientes de dividir cada frecuencia absoluta entre el total de valores **N**. Se representan por **f** (así lo haremos nosotros) o por **h**.

$$f = \frac{n}{N}$$

La suma de todas las frecuencias relativas es igual a uno:

$$\sum f = 1$$

La frecuencia relativa es una razón (o cociente). Por tanto se puede convertir en **porcentaje** multiplicándola por 100, y así representa el tanto por ciento del total que representa cada dato. Lo representamos por **p**.

Por tanto se cumplirá que $p = f \times 100$

y que $\sum p = 100$

Frecuencias acumuladas

Cuando la variable está medida al menos a nivel ordinal permite la acumulación de frecuencias.

La frecuencia acumulada de un valor es el número de datos del conjunto que son *menores o iguales* a él. Por tanto, se calculará sumando todas las frecuencias de datos menores o iguales al mismo. Podemos acumular las frecuencias absolutas y también las relativas y los porcentajes.

Las frecuencias acumuladas serán crecientes y la última absoluta coincidirá con N, la última relativa con 1 y el porcentaje último con 100.

Distribución de frecuencias

El conjunto formado por los valores de la variable y sus frecuencias (hasta seis columnas) constituye la **distribución de frecuencias** de la población o muestra, y se representa en las **tablas de frecuencias**, primer paso obligado de un estudio estadístico.

Tabla de frecuencias

Dato X	n_i	f_i	p_i
2	3	0,06	6%
3	7	0,14	14%
4	12	0,24	24%
5	18	0,36	36%
6	7	0,14	14%
7	3	0,06	6%
Total	50	1	100%

Agrupación de datos

Si la variable que se estudia es continua, o discreta con muchos valores distintos, se organizarán sus datos en forma de intervalos. Para ello se fija un valor mínimo y otro máximo, de forma que todos los datos estén comprendidos entre ellos (a veces esto

no se garantiza y quedan intervalos abiertos). La diferencia entre ambos se denomina **rango** de los datos y posteriormente se divide en un número de *intervalos* mediante valores intermedios.

Esto se construye para después situar cada dato en su intervalo correspondiente y hacer el recuento, con lo que cada intervalo poseerá una *frecuencia*.

Para representar cada intervalo disponemos de

Extremo inferior: Es el valor mínimo que puede tener un valor incluido en ese intervalo.

Extremo superior: Es el valor máximo posible. Se considera no alcanzable. Así si un intervalo comprende desde 5 hasta 10, incluiremos en el mismo los valores comprendidos entre estos dos, incluyendo el 5 y sin incluir el 10.

Marca de clase: Promedio entre los dos extremos (o punto medio del intervalo), que se elige como representante de todos los valores comprendidos. Esto constituye una pérdida de exactitud, compensada por el mejor manejo de los datos agrupados.

Amplitud: Es la diferencia entre los dos valores extremos.

Tabla de datos agrupados

Extremo inf.	Extremo sup.	Marca de clase	frecuencia
120	130	125	4
130	140	135	23
140	150	145	31
150	160	155	7

En la figura podemos ver una tabla con cuatro intervalos de amplitud 10, en los que se representan los dos extremos y las marcas de clase, junto a sus frecuencias.

El número aconsejado de intervalos suele estar entre 5 y 15, aunque se pueden elegir con libertad según las características del estudio. Una regla empírica nos aconseja elegir como número de intervalos **la raíz cuadrada entera del número de observaciones**. También se suelen manejar intervalos todos iguales, aunque a veces se altera la amplitud de algunos para destacar algo de interés en la distribución.

En algunas cuestiones se puede suponer que los datos incluidos en un intervalo se distribuyen en el mismo de manera uniforme, pero otras veces es mejor suponer que todos coinciden con la marca de clase. Ambas hipótesis son falsas, lo que supone que la agrupación en intervalos supone siempre una pérdida de información.