Variable estadística

En sus orígenes, la **Estadística** era la "*ciencia del Estado*". El nombre de Estadística alude al enorme interés de esta rama matemática para los asuntos del Estado y su introducción en el mundo científico se debe a la importancia indiscutible para el desarrollo de las ciencias sociales y humanas.

La **Estadística** trata, en primer lugar, de acumular la masa de datos numéricos provenientes de la observación de multitud de fenómenos, procesándolos de forma razonable. Además, se encarga de tabularlos y representarlos en variadas formas de gráficos. Mediante la teoría de la probabilidad analiza y explora la estructura matemática subyacente al fenómeno del que estos datos provienen y, mediante el conocimiento de tal estructura, trata de sacar conclusiones y predicciones que ayuden al mejor aprovechamiento del fenómeno para los fines que de él se pueden pretender.

La tarea de describir y procesar de modo adecuado la masa de datos, provenientes de las observaciones y experimentos, es el objeto de la *EstadísticaDescriptiva*. El análisis de estos datos se realiza mediante la teoría de la *Probabilidad*. Finalmente, el proceso de obtener con confianza conclusiones sobre el fenómeno que se estudia es el objeto de las diversas técnicas existentes en la *Inferencia Estadística*. Si en el experimento interviene una única variable, hablaremos de Estadística *Unidimensional*; en otro caso, la Estadística *Bidimensional o Multidimensional* se ocupa de experimentos en los que se ven involucrados dos o más variables. En esta unidad vamos a tratar los conceptos relativos a la Estadística Descriptiva Unidimensional.

Conceptos básicos

La *Estadística Descriptiva* es una parte de la Estadística cuyo objetivo es examinar a todos los individuos de un conjunto para luego describir e interpretar numéricamente la información obtenida. Sus métodos están basados en la observación y el recuento. Se pretende simplificar los datos observados para obtener de ellos una información lo más completa y concisa posible del total de la población.

La población es el conjunto de todos los individuos sobre que se desea estudiar alguna propiedad o característica.

Cualquier elemento o entidad que sea portador de información sobre alguna propiedad en la cual se está interesado se denomina individuo. (No necesariamente se debe tratar de personas.)

Todo subconjunto finito de la población sobre el que se realice el estudio de la propiedad deseada, es una muestra. Al número de individuos de este subconjunto se le llama tamañode la muestra y se identifica con la letra N.

Ejemplo

Si el experimento consiste en controlar el proceso de fabricación de los rodamientos que produce una cierta máquina, la ***población*** sería el conjunto de todos los rodamientos fabricados por la máquina, los ***individuos*** serían cada uno de los rodamientos y una ***muestra de tamaño N*** estaría formada por ***N*** de esos rodamientos seleccionados mediante algún criterio: al azar, etc.

Para saber más

Conceptos básicos y ejemplos

Variables estadísticas: tipos de variables

Cuando los datos, es decir los resultados de las observaciones, no son magnitudes *medibles numéricamente*, sino cualidades o atributos, se dice que se trata de datos **cualitativos**. Por el contrario, si los datos se pueden cuantificar numéricamente se llaman datos **cuantitativos**, que originan una **variable aleatoria o estadística.** Los datos pueden provenir del estudio de un solo carácter o propiedad (caso **unidimensional**) o de varios simultáneamente (caso **multidimensional**). En esta unidad estudiaremos sólo el caso unidimensional.

Ejemplo

el color de ojos es un **atributo** → no se puede medir con un número

la edad es un **carácter cuantitativo** → se mide con valores numéricos

Una variable estadística o variable aleatoria es el conjunto de valores numéricos que se obtienen al estudiar un carácter cuantitativo de una población o muestra.

Ejemplo

Al estudiar la distribución de edades de los estudiantes de una determinada población, la variable estadística está formada por los valores numéricos que representan todas las edades de dichos estudiantes (población) o por un subconjunto de ellos (muestra), elegido mediante algún criterio.

Las variables estadísticas se denotan por letras: X, Y, Z,....Cada valor concreto de una variable estadística se denota con la misma letra que la variable a la que pertenece, identificada con un subíndice para diferenciarlo del resto de valores de la misma variable.

Ejemplo

X → pesos (en Kg.) de una muestra de 5 neonatos → {3'56,3'50,2'96,3'00,3'85}

x_2 = 3'50 (es el 2º dato de esta variable) x_5 = 3'85 ...

Se llama rango o recorrido de la variable al conjunto de valores que puede tomar.

Ejemplo

Si X mide el nº de hijos de las familias españolas, X puede tomar los valores 0,1, 2, 3, 4, 5,..... (*no existen* familias con 3'141592 hijos)

Si X mide el tiempo de espera para la llegada del autobús, X puede tomar *cualquier valor* dentro del intervalo [0,10] (suponiendo que el tiempo máximo de espera sean 10 minutos) → el autobús puede llegar en ***cualquier momento.***

Existen dos tipos de variables estadísticas, las **discretas** y las **continuas**, que se corresponden con las diferentes situaciones que se pueden presentar al estudiar un carácter cuantitativo.

Una **variable estadística** es **discreta** cuando sólo puede tomar valores aislados y separados entre sí, es decir, su rango es un conjunto discreto de valores.

Ejemplo

Las variables que miden el nº de hijos de las familias, la calificación de un examen tipo test, el nº de piezas defectuosas que produce una máquina, etc.

Una variable estadística es continua cuando puede tomar, al menos teóricamente, cualquier valor dentro de un intervalo de la recta, es decir, su rango está formado por un intervalo continuo de la recta.

Ejemplo

Las variables que miden el tiempo, la longitud y en general cualquier magnitud que por sus propias características, pueda dividirse de forma indefinida, al menos en teoría.

Representación de datos

En Estadística Descriptiva el material de trabajo lo constituyen los **datos**, que son los resultados de las **observaciones**. Una vez obtenidos los datos hay que ordenarlos y clasificarlos mediante algún criterio racional de modo que sea posible una visión crítica de los mismos.

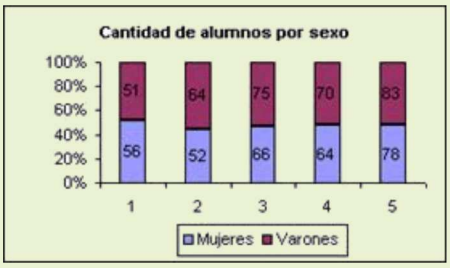
En general, el tratamiento de los datos constará de diversas etapas o fases:

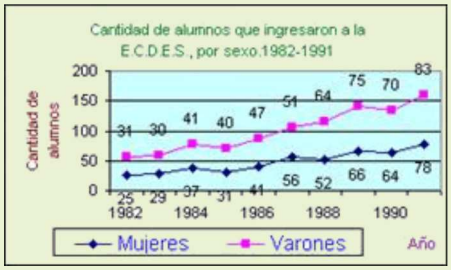
- Construcción de **tablas** para ordenar y clasificar los datos.
- Realización de **gráficos** para representar físicamente los datos.
- Cálculo de algunos estadísticos o **parámetros,** que recojan de forma concisa la información relevante que se encuentre dentro de la muestra.

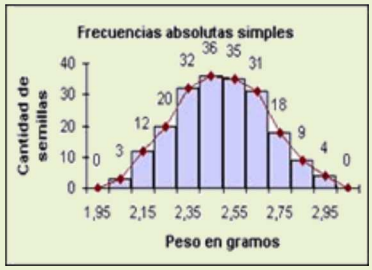
Tablas

Variable xi	Frec. Absoluta	Frec. Relativa	Porcentajes
5	9	0.18	18 %
6	17	0.34	34 %
8	6	0.12	12 %
9	13	0.26	26 %
10	5	0.1	10 %

Gráficos:







Tablas estadísticas

Las tablas estadísticas consisten en masas estructuradas de datos. Deben estar confeccionadas de tal modo que resultan fáciles de leer y de interpretar. Para la construcción de tablas de **datos cuantitativos** pueden tratarse éstos **individualmente** o **agrupándolos** en **clases** o intervalos.

En primer lugar, debemos contar los datos y ordenarlos, para poder trasladar esa información a una tabla. Supondremos que hay **N datos en total** y que hay **k datos distintos**. Cada dato individual se identifica por x_i (el índice **i** varía desde 1 hasta k)

Llamaremos N al tamaño de la muestra con la que estemos trabajando, es decir, al número total de datos que tengamos.

Se llama frecuencia absoluta al número de individuos que toman un determinado valor de una variable estadística (o una modalidad de un atributo). Lo denotaremos con f_i.

(En resumen, f_i es el número de veces que aparece cada valor x_i dentro del total de datos)

Se llama frecuencia absolutaacumulada de un valor a la suma de las frecuencias absolutas de todos los valores menores o iguales que él. Lo denotaremos con F_i

$F_i = f_1 + f_2 + ... + f_i$

Se llama frecuencia relativa al cociente entre la frecuencia absoluta f_i y el número total de datos o tamaño de la muestra N. Lo denotaremos con h_i

$h_i = \frac{f_i}{N}$

Se llama frecuencia relativa acumulada de un valor de una variable estadística a la suma de las frecuencias relativas de todos los valores menores o iguales que él. Lo denotaremos con H_i $H_i = h_1 + h_2 + ... + h_i = \frac{F_i}{N}$

Para variable discreta, o que siendo continua tengamos pocos datos, si tenemos una muestra de tamaño N, la tabla se estructura así:

Variable	Frecuencias	absolutas	Frecuencias relativas	
estadística : Xi	puntuales	acumuladas	puntuales	acumuladas
x1 x2 ..... xk	f1 f2 ..... fk	F1= f1 F2= f1 + f2 ..... Fk= f1 + f2 +..+ fk	h1= f1/N h2=f2/N ..... hk= fk/N	H1= F1/N H2=F2/N ..... Hk= Fk/N

Suponemos que existen *k* valores o clases distintas en nuestra muestra: de x_1 a x_k

Se debe cumplir que la última frecuencia absoluta acumulada debe ser igual a N

$N = f_1 + f_2 + .. + f_k = F_k$

Ejemplo

Las notas de los 20 alumnos de una clase, en un examen han sido: 4, 3, 3, 5, 6, 7, 9, 0, 5, 4, 9, 10, 2, 7, 2, 2, 5, 6, 5, 0

Vamos a formar la tabla:

Variable	Frecuencias	absolutas	Frecuencias relativas	
estadística : xi	puntuales fi	acumuladas Fi	puntuales hi	acumuladas Hi
0	2	2	1/10	1/10
2	3	5	3/20	5/20=1/4
3	2	7	1/10	7/20
4	2	9	1/10	9/20
5	5	14	1/4	14/20=7/10
7	3	17	3/20	17/20
9	3	20	3/20	20/20=1

En la tabla sólo aparecen los valores de x_i que se corresponden con notas que han obtenido los alumnos; así, por ejemplo, no aparece el valor 8 porque ningún alumno ha obtenido un 8 de nota.

Tratamiento de datos agrupados

Vamos a tratar el caso ahora de datos agrupados en intervalos.

Cuando en la población o muestra que estudiamos existen muchos valores diferentes o se trate de una variable continua, es conveniente, aún a costa de perder algo de información, dividir el intervalo de variación de la variable en una serie de subintervalos que cubran el total de datos de la muestra; a cada uno de los intervalos se le llama una clase, a sus extremos, extremos de clase, al punto medio de cada clase, marca de clase y a la diferencia entre sus extremos, amplitud de la clase.

La determinación de los intervalos o clases dependerá del valor de los datos. Es muy importante que los intervalos cubran todos los datos y que no haya intervalos de más.

En estos casos la tabla adopta una estructura como la siguiente:

Intervalos	Marcas de clase	Fr. Absoluta	Fr. Abs. Acumulada	Fr. Relativa	Fr. Rel. Acumulada
$[a_1, b_1)$	x_1	f_1	F_1	$h_1 = f_1/N$	H_1
$[a_2\ b_2)$	x_2	f_2	$F_2 = f_1 + f_2$	$h_2 = f_2/N$	$H_2 = F_2/N$
.....	.....	.....	.....	.....	.....
$[a_k, b_k]$	x_k	f_k	$F_k = f_1 + f_2 + ..+ f_k$	$h_k = f_k/N$	$H_k = F_k/N$

La **marca de clase** x_i es la semisuma de los extremos del intervalo:

$x_i = \frac{a_i + b_i}{2}$ (**punto medio** del intervalo $[a_i, b_i)$

Los demás conceptos son idénticos que en el caso de variable discreta. En cada intervalo, su frecuencia absoluta es el nº de datos que están dentro del intervalo. Los intervalos son cerrados por la izquierda (extremo inferior) y abiertos por la derecha (extremo superior) para que no se solapen los extremos y no haya duplicidad, como se muestra en el siguiente ejemplo.

Ejemplo

Se ha pasado un test de 79 preguntas a 600 personas. El número de respuestas correctas se refleja en la siguiente tabla:

intervalos	Marca de clase xi	F. Absoluta. puntual	F. Absoluta acumulada	F. Relativa. puntual	F. Relativa. acumulada
[0, 10)	5	40	40	1/15	1/15
[10, 20)	15	60	100	1/10	1/6
[20. 30)	25	75	175	1/8	7/24
[30, 40)	35	90	265	3/20	53/120
[40, 50)	45	105	370	7/40	37/60
[50, 60)	55	85	455	17/120	91/120
[60, 70)	65	80	535	2/15	107/120
[70, 80]	75	65	600	13/120	1
		600		1	

El último intervalo debe ser [70, 80) ya que el nº de preguntas era 79 y por tanto no tiene sentido incluir más intervalos, ya que como máximo, se pueden responder 79 preguntas correctas. Por el mismo motivo el primer intervalo es [0, 10) porque no se pueden responder menos de 0 preguntas correctas.

El último intervalo se suele definir cerrado en ambos extremos, para cubrir la eventualidad de que exista un valor de la variable igual al extremo superior de ese intervalo.

Mientras que en el caso del tratamiento individual la tabla quedaba perfectamente determinada por los posibles valores de los datos, en el de clases está claro que no sucede así, pues hay libertad para elegir el número de clase y los extremos de las mismas. Los intervalos, en general, deben tener la **misma amplitud.**

Para decidir el nº de clases que se deben tomar conviene tener en cuenta que si éste es excesivo con respecto al número de datos, pueden aparecer irregularidades accidentales provenientes de pocas observaciones en algunas clases. Sin embargo, si se toma el número de clases demasiado reducido se producirá una pérdida importante de información.

Un criterio *orientativo* para decidir cuántas clases se deben tomar lo proporciona la siguiente fórmula empírica debida a Sturges: **k = nº de intervalos = 1 + 3.32·Log N ,** siendo N el nº total de datos.

Las tablas estadísticas se pueden ampliar con más columnas (por la derecha) para incluir los cálculos parciales necesarios para obtener las medidas de centralización y de dispersión, tales como la media, la varianza y la desviación típica, que veremos más adelante. También se usan para el cálculo de la mediana y de la moda, ya sea en datos aislados (variable discreta) o agrupados en intervalos (variable continua o discreta con muchos valores distintos). Son necesarias así mismo, para las principales representaciones gráficas de datos, como los diagramas de barras o histogramas o los diagramas de sectores, que, por ejemplo, necesitan de la columna de frecuencias relativas. Todo lo anterior demuestra su gran importancia y utilidad.

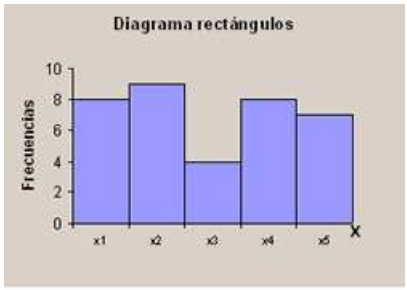
Gráficos estadísticos

El siguiente paso, después de haber recogido y ordenado los datos en una tabla, suele ser la representación gráfica de los mismos, usando alguno de los diversos tipos de gráficos estadísticos. La representación gráfica debe ser lo suficientemente clara y precisa para que de un vistazo obtengamos información relevante acerca de la distribución de los datos.

Existen diversos tipos de gráficos y sería muy prolijo enumerarlos a todos. Vamos a tratar los más usuales y pondremos algún ejemplo de los demás.

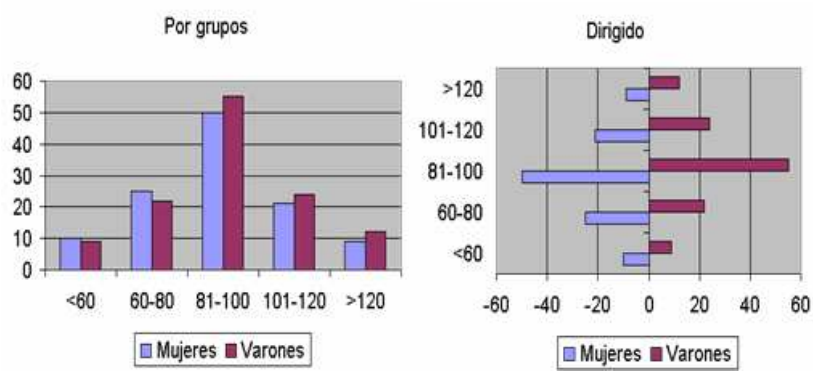
- **Diagrama de barras**: se usa en variable discreta, cuando los datos están separados entre sí. Consiste en colocar en el eje OX los valores de la variable estadística y sobre cada uno de ellos levantar una línea o barra, cuya altura sea igual a la frecuencia absoluta de ese valor.

- **Histograma**: es equivalente al diagrama de barras, pero para variable continua o cuando los datos están agrupados en intervalos. Sobre el eje OX se colocan los distintos intervalos o clases y sobre cada uno de ellos se levanta un rectángulo de altura igual a la frecuencia absoluta del intervalo:

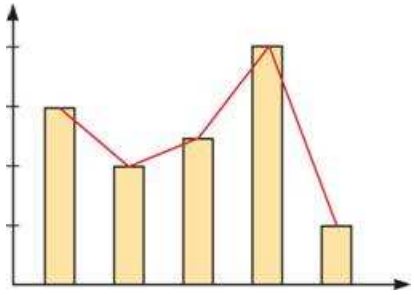


También podemos elaborar ambos tipos de gráficos para las frecuencias acumuladas, obteniendo gráficos en escalera, como en el siguiente ejemplo (se llaman en escalera porque al ir acumulando las frecuencias absolutas, cada rectángulo es mayor que el anterior y se obtiene algo parecido a una serie de escalones):

O representar dos histogramas de la misma variable en dos situaciones distintas:

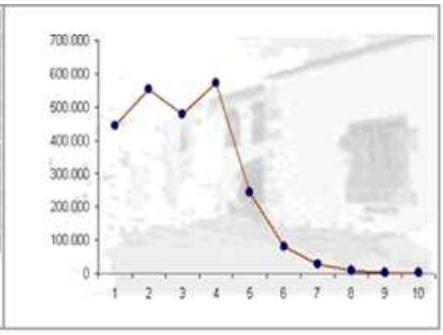


- **Polígono de frecuencias**: son líneas poligonales que unen los vértices superiores de las barras de un diagrama de barras o de los rectángulos en un histograma:



También se pueden presentar sin barras ni rectángulos, en este caso, cada vértice da la línea poligonal corresponde con una frecuencia absoluta:

Nº Residentes	Viviendas
1 persona	444.390
2 personas	551.618
3 personas	477.622
4 personas	573.254
5 personas	244.544
6 personas	81.973
7 personas	26.793
8 personas	9.989
9 personas	3.712
10 o más personas	3.284



- **Diagrama de sectores:** es un tipo de gráfico muy adecuado para representar cualquier tipo de variable. Consiste en un círculo dividido en sectores circulares, que se corresponden con los distintos datos o intervalos de la variable, de forma que el área o número de grados de cada sector es proporcional a la frecuencia absoluta de cada dato o clase. Pueden estar en 2 o 3 dimensiones.

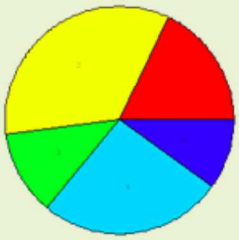
Para calcular la amplitud de cada sector circular, debemos multiplicar la frecuencia relativa de cada dato o clase por 360º y así obtendremos el nº de grados que debe tener cada sector.

Ejemplo

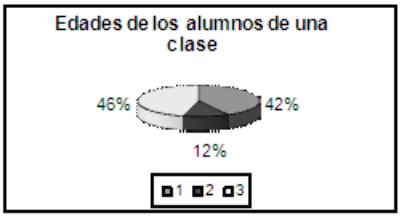
Variable xi	Frec. Absoluta	Frec. Ab. Acum.	Frec. Relativa
1	9	9	0.18
2	17	26	0.34
3	6	32	0.12
4	13	45	0.26
5	5	N = 50	0.1

El diagrama de sectores quedaría:

Variable x_i	Nº grados
1	64.8º
2	122.4º
3	43.2º
4	93.6º
5	36º



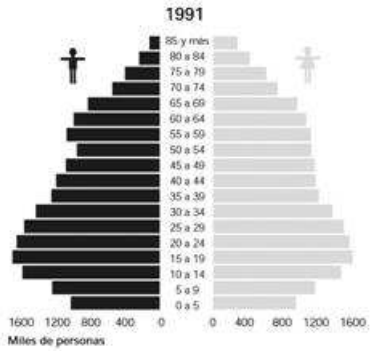
Un diagrama de sectores en 3D podría ser como el siguiente, en el que se representan, en porcentajes, las distintas edades de los alumnos de una clase:

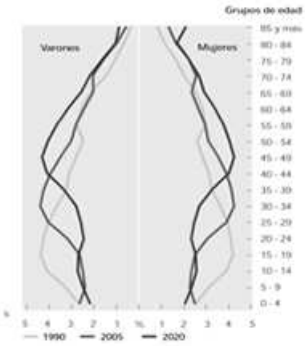


Diagramas de sectores o circulares 3D

- **Otros gráficos estadísticos**: además de todos los anteriores, se suelen usar otros gráficos, tales como:

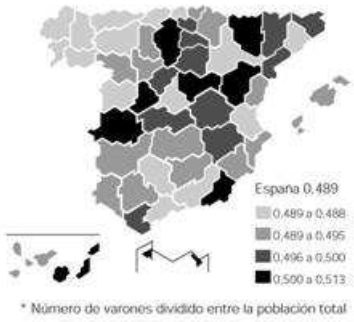
Pirámides de población



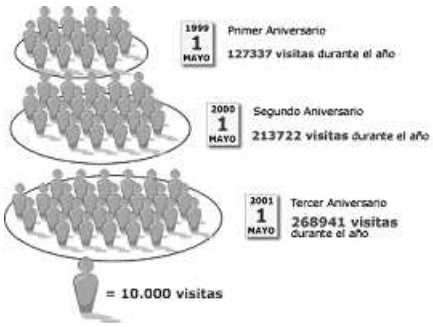


Pirámide de población proyectada

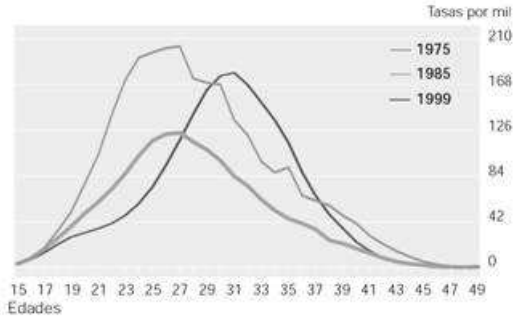
Cartogramas



Pictogramas



Series cronológicas o temporales



Para saber más

- Representación grafica del análisis de datos
- Los distintos gráficos estadísticos con explicaciones y ejemplos
- Excelentes gráficos estadísticos
- Aplicaciones interactivas de estadística básica

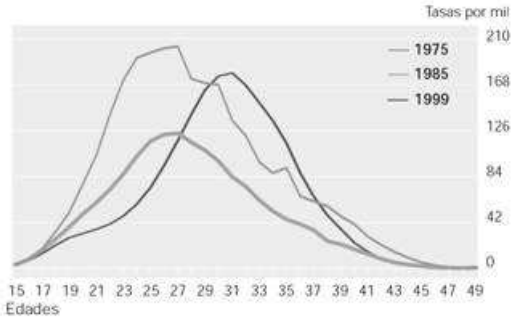
- Curso completo de estadística con gráficos y ejemplos
- Herramienta interactiva en castellano para realizar diagramas de barras
- Tutorial para realizar gráficos y todo tipo de cálculos estadisticos

Medidas de centralización

También llamados **promedios** o **medidas de tendencia central**. Son valores típicos o representativos de un conjunto de datos. Pretenden resumir todos los datos en un único valor.

Definimos tres medidas de tendencia central:

- media
- mediana
- moda



Media aritmética

Es el parámetro que se obtiene al sumar todos los valores de la variable y dividirlos por el número de datos:

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{N}$$

Si los datos están tabulados en una tabla estadística, conocemos la frecuencia absoluta f_i de cada valor x_i y entonces la media se puede calcular más rápidamente con la 2ª fórmula.

Ejemplo

Si la variable X toma los valores → 2,3,3,5,7,3,1,6,2,3,1,4,5,4,2,8,6,3,4.5 entonces la media se puede calcular:

$$\overline{x} = \frac{1\cdot2+2\cdot3+3\cdot5+4\cdot3+5\cdot3+6\cdot2+7\cdot1+8\cdot1}{20} = \frac{77}{20} = 3'85$$

Como se observa, la media aritmética no tiene porqué coincidir con ningún valor de la variable.

La media aritmética es el valor que mejor representa a la distribución de los valores de la variable. Si hubiera que elegir un único valor que represente a toda la variable, ése sería la media. Además, es el centro de gravedad de la variable, es decir, el valor que está equidistante de todos los valores, teniendo en cuenta los respectivos pesos o frecuencias absolutas de cada valor individual.

Para datos agrupados en intervalos, tomaremos como x_i la marca de clase de cada intervalo, siendo f_i la respectiva frecuencia absoluta y aplicaremos la misma fórmula.

Ejemplo

Dada la siguiente tabla:

Variable xi	Marca de clase	Frec. Absoluta	Frec. Relativa
[0,5)	2'5	5	5
[5,10)	7'5	2	7
[10,15)	12'5	9	16
[15,20)	17'5	4	20
[20,25]	22'5	3	23

$$\overline{x} = \frac{2'5 \cdot 5 + 7'5 \cdot 2 + 12'5 \cdot 9 + 17'5 \cdot 4 + 22'5 \cdot 3}{23} = \frac{250}{23} = 10'869$$

Mediana

Es el valor de la variable que ocupa el lugar central de la distribución, es decir el valor de la variable que deja el 50% de observaciones a su izquierda y el 50% a su derecha.

Para poder hallar la mediana, lo primero que hay que hacer es ordenar los valores de la variable de forma creciente, y escribir los valores de las frecuencias absolutas acumuladas F_i.

Distinguiremos dos casos, datos no agrupados y datos agrupados en intervalos.

Para **datos no agrupados**, se calcula primero el 50% de la población, N/2 y se lleva ese valor a la columna de frecuencias absolutas acumuladas. Se pueden dar entonces 2 situaciones:

- si el valor **no está** en la columna de frecuencias absolutas acumuladas, se toma como valor de la **mediana** el **primer valor de la variable** cuya frecuencia absoluta acumulada supere a N/2. .
- si el valor **sí está** en la columna de frecuencias absolutas acumuladas, se toma como **mediana** la **media aritmética de ese valor** de la variable y del **siguiente**.

Ejemplos

Dada la siguiente tabla de valores

Variable Xi	Frec. Absoluta	Frec. Ab. Acumulada
3	3	3
4	6	9
5	8	17
6	4	21
7	4	25

N = 25 → N/2 = 12'5 → el primer valor cuya frecuencia absoluta acumulada es 5 → la **mediana** es **5**

b) dada la tabla de valores

Variable Xi	Frec. Absoluta	Frec. Ab. Acumulada
1	2	2

2	6	8
3	4	12
4	8	20
5	4	24

N = 24 → N/2 = ***12*** → el valor ***3*** tiene una frecuencia absoluta acumulada igual a ***12*** → la **mediana** de esta variable es

$$\frac{3+4}{2} = 3'5$$

Para datos agrupados en intervalos, se calcula como antes la mitad del número de datos N/2 y se lleva ese valor a la columna de frecuencias absolutas acumuladas.

- Si el valor **no está** en la columna, se toma como intervalo mediano el primer intervalo cuya frecuencia absoluta acumulada supere a N/2, y después de situarnos en el intervalo por la hipótesis de uniformidad hacemos una proporción entre la amplitud del intervalo, los elementos que tiene y la amplitud que correspondería a la diferencia entre N/2 y la frecuencia acumulada anterior, valor que añadiríamos al extremo inferior del intervalo mediano.

$$\text{mediana} = A_i + \left(\frac{\frac{N}{2} - \left(\sum f_i\right)}{f\_\text{mediana}} \right) \cdot c \quad \text{donde:}$$

A_i = extremo inferior del intervalo mediano

N = número de datos

$\left(\sum f_i\right)$ = suma de las frecuencias de los intervalos anteriores al intervalo mediano

f_mediana = frecuencia absoluta del intervalo mediano

c = anchura del intervalo

Ejemplo

Si la tabla es

Clases	Frec. Absoluta	Frec. Ab. Acumulada
[0,10)	9	9
[10,20)	13	22
[20,30)	5	27
[30,40]	8	35

El intervalo mediano es [10,20). Dentro de él, vamos a calcular la mediana

Mediana = $10 + \left(\frac{17'5 - 9}{13} \right) \cdot 10 \approx 16'54$

- Si el valor **sí está** en la columna de frecuencias acumuladas, se toma como mediana el extremo superior del intervalo correspondiente.

Moda

La moda es el valor de la variable que más se repite, es decir, el que tenga mayor frecuencia absoluta.

Una variable puede tener una o más modas: en este caso se dice bimodal, trimodal, etc...

Ejemplo

Si la variable X toma los valores → 2,3,3,5,7,3,1,6,2,3,1,4,5,4,2,8,6,3,4.5

Formamos la tabla con las frecuencias absolutas: N = 20

Variable xi	Frec. Absoluta
1	2
2	3
3	5 **
4	3
5	3
6	2
7	1
8	1

La moda es **3**, ya que su frecuencia absoluta es 5, la mayor de todas → el 3 es el valor que más se repite.

Para variable continua o datos agrupados, hablamos de **intervalo modal** → el que tenga mayor frecuencia absoluta.

Autoevaluación

Calcula la media, mediana y la moda del siguiente conjunto de datos:

3,5,2,7,5,6,3,4,5,2,3,1,6,3,4,5,5,6,3,2,1

a)

$\overline{x} =$ 4'21 Me = 5 Mo = 5

b)

$\overline{x} =$ 3'857 Me = 4 Mo =3 y 5 (bimodal)

c)

$\overline{x} =$ 1'45 Me = 4 Mo = 3

Comprobar

Calcula la media y mediana de:

X_i	f_i
[5,9)	5
[9,13)	9
[13,17)	6
[17,21]	10

a)

$\overline{x} =$ 4'21 Me = 13

b)

$\overline{x} =$ 35'7 Me = 15'01

c)

$\overline{x} =$ 13'8 Me = 13'667

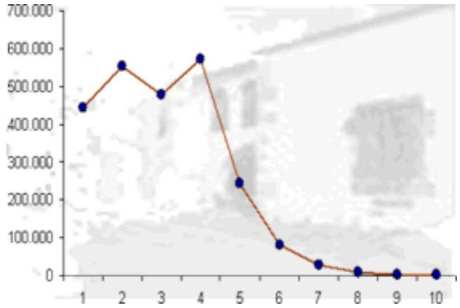
Comprobar

Parámetros de dispersión

Las medidas de dispersión nos indican la mayor o menor separación de los valores de una variable respecto a un promedio. Acompañando a un promedio debe ir una medida de dispersión que nos indica la mayor o menor representatividad de ese promedio.

Definimos tres parámetros o medidas de dispersión:

- la varianza
- la desviación típica
- el coeficiente de variación



Varianza

La varianza es el primer parámetro que mide la dispersión de los datos, es decir, lo que se separan respecto de la media. Es un promedio cuadrático de la separación de cada uno de ellos x_i respecto del valor medio $\overline{x}$. La varianza siempre es positiva, ya que por definición es una suma de cuadrados (positivos) multiplicada por valores positivos (f_i). Por tanto, cuanto mayor sea la varianza, mayor será la dispersión de los datos.

$$s_x^2 = \frac{\sum_{i=1}^{k} \left(x_i - \overline{x} \right)^2 \cdot f_i}{N} = \frac{\sum_{i=1}^{k} x_i^2 \cdot f_i}{N} - \overline{x}^2$$

La segunda fórmula del cálculo de la varianza es la más usada, por ser la más concisa. Se puede demostrar fácilmente que ambas fórmulas son equivalentes, sin más que desarrollar en la primera de ellas el cuadrado de la diferencia $\left(x_i - \overline{x} \right)^2$ y operar convenientemente.

Ejemplo

Calculemos la varianza de la siguiente serie de datos

1 2 3 1 2 4 0 2 0 1 3 2 1 2 3 1 2 2 3 4

xi	fi	xi · fi	xi2 · fi

0	2	0	0
1	5	5	5
2	7	14	28
3	4	12	36
4	2	8	32

$$\bar{x} = \frac{39}{20} = 1'95 \qquad s_x^2 = \frac{101}{20} - 1'95^2 = 5'05 - 3'8025 = 1'2475$$

Propiedades:

- La varianza siempre es mayor o igual que cero. Tan solo hay un caso en que es cero y es cuando todos los valores de la variable son iguales, porque en este caso todos los valores de la variable coinciden con la media $\bar{x}$ y cada paréntesis de la fórmula es igual a 0.
- Si a los valores de la variable se les suma una constante, la varianza de la nueva variable es la misma que la anterior.

Es decir: si $x_i' = x_i + K \rightarrow S^2_{x'} = S^2_x$

Demostración:

$$S_x'^2 = \frac{\sum (x_i' - \bar{x}')^2 \cdot f_i}{N} = \frac{\sum ((x_i + k) - (\bar{x} + k))^2 \cdot f_i}{N} = S_x^2$$

- Si a los valores de la variable se les multiplica por una constante, la varianza de la nueva variable queda multiplicada por el cuadrado de la constante.

Es decir si $x_i' = k \cdot x_i$ entonces $S^2_{x'} = k^2 \cdot S^2_x$ (siendo $k \neq 0$)

Existe un problema con la varianza y es que al ser una medida cuadrática (valores al cuadrado), sus unidades serían el cuadrado de las unidades de la variable en cuestión. Para resolverlo, lo más sencillo sería extraer la *raíz cuadrada* de la varianza, que es la definición del siguiente parámetro de dispersión, la desviación típica.

Desviación típica

La desviación típica es la raíz cuadrada positiva de la varianza:

$$s_x = +\sqrt{s_x^2}$$

Es la medida de dispersión más utilizada. Sus unidades son las mismas que las de la media y que las de la variable.

Ejemplo

Siguiendo con el ejemplo anterior, si $s_x^2 = 1'2475 \rightarrow s_x = 1'117$

El uso combinado de la media y la desviación típica permite, en distribuciones bastante simétricas, determinar una serie de intervalos en los que se encuentran los valores de la distribución con ciertos porcentajes:

en el intervalo ($\bar{x} - s_x, \bar{x} + s_x$) se encuentra el **68%** de los valores de la variable

en el intervalo ($\overline{x} - 2 \cdot s_x , \overline{x} + 2 \cdot s_x$) se encuentra el **95%** de los valores de la variable

en el intervalo ($\overline{x} - 3 \cdot s_x , \overline{x} + 3 \cdot s_x$) se encuentra el **99%** de los valores de la variable

Los porcentajes asociados a cada intervalo pueden variar en función del tipo de distribución. Los incluidos aquí corresponden a una distribución estándar unimodal y simétrica.

Para saber más

- Curso completo estadística Unidimensional
- Curso estadística
- Estadística descriptiva
- Curso Estadística

Autoevaluación

Calcula la varianza y la desviación típica del siguiente conjunto de datos:

3 6 7 6 3 4 6 7 7 6 4 7 5 8 6 7 3 5 7 3 8 5 4 3 5 6 5 7 7 6

a) s_x = 2'542
b) s_x = 1'589
c) s_x = 1'543

Comprobar

Calcula la desviación típica de:

Variable x_i	[0,6)	[6,12)	[12,18)	[18,24]
Frecuencias f_i	4	8	3	4

a) s_x= 0'04
b) s_x = 6'22
c) s_x = 1'677

Comprobar

Coeficiente de variación

El uso conjunto de la media y la desviación típica también permite comparar dos variables aleatorias entre sí para determinar cuál de ellas es más dispersa.

Si las dos variables tienen la **misma media**, será **más dispersa** aquella que tenga **mayor desviación típica**.

Ejemplo

Si las variables X e Y cumplen que:

X → $\overline{x} = 5'25 \quad s_x = 1'56$

Y → $\overline{y} = 5'25 \quad s_y = 2'05$

Vemos que la variable **Y es más dispersa que la variable X**, ya que tiene la misma media (5'25) pero su desviación típica es mayor (2'05 > 1'56) → **X** es **más homogénea** que **Y**

El problema surge al comparar variables aleatorias que tienen **distinta media:** necesitamos una medida para comparar la dispersión de este tipo de variables. Ese valor es el coeficiente de variación.

El coeficiente de variación CV(X) de una variable X es el cociente entre la desviación típica, s_x y la media de la variable $\overline{x}$:

$$CV(X) = \frac{s_x}{\overline{x}} = \frac{\text{desviación típica}}{\text{media}} \rightarrow (\text{si } \overline{x} \neq 0)$$

El coeficiente de variación es un parámetro adimensional, es decir, que no tiene unidades, ya que es el cociente de dos magnitudes expresadas en las mismas unidades (al dividirlas, éstas se simplifican). Por ello sirve para comparar las dispersiones de aquellas variables que tienen medias distintas o cuyos datos vienen expresados en unidades distintas.

Ejemplo

Si los alumnos de un grupo de 4º de ESO tienen una talla media $\overline{x}$ = 1'75 m. y una desviación típica s_x = 0'05 m. y la nota media en un examen de Matemáticas ha sido $\overline{y}$ = 5'35 ptos. con una desviación típica de s_y = 1'25 ptos. podemos compararlas a través del coeficiente de variación:

$$CV(X) = \frac{0'05}{1'75} \approx 0'03 \rightarrow 3\%$$

$$CV(Y) = \frac{1'25}{5'35} \approx 0'25 \rightarrow 25\%$$

Podemos concluir que las notas de este grupo están más dispersas que las tallas.

Para saber más

- Pagina con herramientas interactivas de estadística
- Curso básico de estadística: conceptos y ejemplos
- Sitio que recopila diversos enlaces a software de estadística de uso libre
- Sitio con descargas de software estadístico de libre distribución
- Software gratis de Matemáticas y Estadística (I)
- Software gratis de Matemáticas y Estadística (II)

En los siguientes recursos se muestran dos ejercicios de tablas estadísticas completas, con gráficos incluidos:

- Variable discreta
- Variable continua

Autoevaluación

Halla la dispersión absoluta y relativa de los siguientes valores de una variable aleatoria:

X à 5,2,3,4,2,4,3,5,2,2,5,6,1,3,2

a) s_x= 1'44 CV = 0'44
b) s_x = 1'76 CV = 0 '447
c) s_x = 2'09 CV = 0'245

Comprobar

La siguiente tabla muestra las calificaciones de María y Paco obtenidas en 10 controles de Matemáticas:

X (Notas de María	4	5	5	4	6	7	8	9	3	9
Y (Notas de Paco	5	6	6	6	7	7	6	5	7	5

Halla las medias y desviaciones típicas. ¿Quién ha sido más regular?

a)

$\overline{x}$ = 5.16 sx = 1.678 CV(X) = 0.325

$\overline{y}$ = 6 sy = 0.45 CV(Y) = 0.075 à el más regular ha sido María

b)

$\overline{x}$ = 4.56 sx = 1.24 CV(X) = 0.272

$\overline{y}$ = 3.67 sy = 2 CV(Y) = 0.545 à los dos han sido igual de regulares

c)

$\overline{x}$ = 6 sx = 2.049 CV(X) = 0.3415

$\overline{y}$ = 6 sy = 0.774 CV(Y) = 0.129 à el más regular ha sido Paco

Comprobar