Tema 8: Estadística en una variable (unidimensional)

1. Introducción

Se desconocen con exactitud los orígenes de la Estadística. Parece que fueron los chinos, en el 2200 a. C., los primeros en efectuar recuentos de su población. Tanto los egipcios como los griegos y los romanos preveían sus cosechas por medios que podríamos llamar estadísticos y efectuaban censos de población.

En los siglos XVI y XVII la Estadística pasa a tener como principal objetivo el estudio de los asuntos de Estado, de donde deriva el sentido etimológico de la palabra. Desde entonces experimenta una evolución que pasa por varias fases. Inicialmente, la preocupación fundamental era la recogida, clasificación y representación de los datos; más tarde se pasó a la fase de análisis e interpretación de los mismos.

En una primera aproximación, usamos la palabra *estadística* para designar colecciones de datos numéricos de la misma naturaleza, relativos a un determinado fenómeno: estadística de los automóviles vendidos, estadística de las importaciones, estadística de los divorcios, etc. En un sentido más riguroso, la Estadística es un método científico que, a partir del conocimiento de diversos hechos recogidos, hace inferencias que permiten la previsión de nuevos acontecimientos.

Para hablar del objeto de la Estadística, hemos de comenzar por distinguir fenómenos *deterministas* y *aleatorios*.

- **Fenómenos deterministas (o causales)** son los que al repetirlos en idénticas condiciones producen el mismo resultado. Por ejemplo, el tiempo que tarda un móvil, a velocidad constante, en recorrer una distancia dada.
- **Fenómenos aleatorios (de azar o estadísticos)** son los que al repetirlos un gran número de veces, en idénticas condiciones, presentan resultados diferentes, siendo imposible predecir el resultado de cada prueba particular. Por ejemplo, los resultados del lanzamiento de un dado.

El método de trabajo de la Estadística tiene tres vertientes:

- Descripción de los datos observados (Estadística Descriptiva).
- Modelización del comportamiento (Cálculo de Probabilidades).
- Estimación de lo desconocido y generalización (Teoría de Muestras e Inferencia Estadística).

Teniendo en cuenta los métodos de trabajo de la Estadística encontramos sus aplicaciones:

- Descripción.
- Análisis.
- Predicción.

Una *clasificación* más general presenta las técnicas estadísticas en dos grupos con funciones distintas:

- **Estadística Descriptiva.**
    - Reducción y descripción de informaciones voluminosas.

- Recuento, ordenación y clasificación de datos observados.
- Presentación de datos en forma resumida y manejable:
  - Tablas.
  - Gráficas.
  - Cálculo de parámetros estadísticos que caracterizan la distribución de los datos: medias, medianas, cuartiles, percentiles, varianza, desviación típica, ...
- No utiliza el Cálculo de Probabilidades.

- **Estadística Inferencial.**
  - Se apoya en el Cálculo de Probabilidades.
  - Maneja resultados de la Estadística Descriptiva.
  - Plantea y resuelve el problema de establecer previsiones y conclusiones generales sobre una población o colectivo.

Tanto en esta tema como en el siguiente se trabajará la Estadística Descriptiva en una variable (unidimensional) y en dos variables (bidimensional).

2. Primeras definiciones

2.1. Población y muestra

La *población* o *universo* estadístico es el conjunto de elementos que poseen al menos una característica común y sobre los cuales va a incidir el análisis estadístico. El número de elementos de una población es su *tamaño* (que puede ser finito o no). Si la población es finita lo representaremos por N.

No siempre es posible efectuar el estudio de todos los elementos de una población. En este caso, el estudio se puede limitar a una parte de ese todo: a una muestra. Así, una *muestra* es un subconjunto de la población.

Los elementos de la población se llaman *individuos* o *unidades estadísticas*.

Estudiando muestras finitas representativas se obtienen conclusiones que se pueden aplicar a toda la población. Para que una muestra sea representativa de la población es preciso que el *muestreo* sea *aleatorio*, es decir, que cualquier individuo de la población tenga la misma probabilidad de pertenecer a la muestra, en cuyo caso se habla de *muestra aleatoria*.

Ejemplos

a) En un sondeo de opinión realizado por una empresa para conocer la intención de voto de los habitantes de una ciudad, la población está formada por el conjunto de todos los individuos con derecho a voto. De ella se extraerá un conjunto de personas a las que se entrevistará: éstas forman la muestra.

b) Para estudiar la proporción de tornillos defectuosos que produce una fábrica en una semana, se eligen al azar 1000 tornillos. La población la constituyen todos los tornillos fabricados en la semana. La muestra la forman los 1000 tornillos seleccionados.

2.2. Caracteres y modalidades

En relación con cada unidad estadística pueden ser observadas distintas propiedades que permiten clasificar a los individuos de la población: estas propiedades se llaman *caracteres* o *características estadísticas*.

Ejemplo

Consideremos una población formada por N estudiantes. Cada estudiante es un individuo de la población que puede ser estudiado atendiendo a distintos caracteres: sexo, edad, estatura, lugar de nacimiento, nota obtenida en el último examen, color del pelo, ...

Para cada característica, deben estar definidas todas las situaciones posibles en que se puede encontrar una unidad estadística: éstas son las *modalidades*. En cuanto a las modalidades, hemos de cuidarnos no sólo de enumerarlas sino que han de estar bien definidas, de modo que cada individuo pueda pertenecer a una y sólo una de ellas: las modalidades han de ser incompatibles (mutuamente excluyentes) y exhaustivas (cubrir toda la población).

Es posible clasificar los caracteres en *cuantitativos* (o *variables estadísticas*), si son susceptibles de representación numérica, y *cualitativos* (o *atributos*), en caso contrario.

Ejemplo

Consideremos la población formada por todos los alumnos de un Instituto y los siguientes caracteres: sexo, edad, curso y estatura:

- El carácter sexo tiene dos modalidades: hombre y mujer. Este carácter es por tanto cualitativo.
- El carácter edad tiene las siguientes modalidades: {12, 13, 14, 15, 16, 17, 18} si entendemos que la edad se describe por años cumplidos. Este carácter es por tanto cuantitativo y podremos hablar de la variable estadística edad.
- El curso tiene las modalidades 1º ESO, 2º ESO, 3º ESO, 4º ESO, 1º Bachillerato y 2º Bachillerato y por tanto es cualitativo.
- Por último el carácter estatura se puede dividir, por ejemplo, en las siguientes modalidades: $\{(-\infty, 160], (160, 170], (170, 180], (180, +\infty)\}$, donde se está adoptando como unidad de medida los centímetros. Se puede hablar por tanto de la variable estadística estatura.

2.3. Variables estadísticas discretas y continuas

Con respecto a cada unidad estadística o individuo de una población podemos determinar varios caracteres que pueden se cuantitativos o cualitativos, como se ha visto anteriormente. Cada carácter cuantitativo es una variable estadística; dicho de otro modo, una variable estadística es un aspecto medible de la unidad estadística. La medición de la variable de cada individuo de la población permitirá clasificar sus elementos en modalidades.

Las variables suelen representarse por letras mayúsculas: X, Y, ..., y los valores que toma cada una de ellas con las mismas letras que la variable, pero en minúscula y con subíndices: x_1, x_2, x_3, ..., x_k, ..., si nos referimos a la variable X; y_1, y_2, y_3, ..., y_k, ..., si nos referimos a la variable Y.

Diremos que una variable estadística es *discreta* si su campo de variación, esto es, el conjunto de valores que toma la variable, está formado por puntos aislados (en número finito o infinito numerable).

Diremos que una variable estadística es *continua* si su campo de variación es, al menos teóricamente, un intervalo de la recta real. Dados dos valores cualesquiera de los que toma la variable, siempre existe entre ellos una infinidad de valores que puede tomar.

Ejemplo

Son variables estadísticas discretas:

- El número de coches fabricados en un año.
- El número de pacientes atendidos cierto día en un Centro de Salud.
- El número de ordenadores en cada Instituto de la provincia.

Son variables estadísticas continuas:

- El peso de los alumnos de un Instituto.
- La estatura de los mismos alumnos.
- Las temperaturas registradas en un observatorio cada hora.

En la práctica, aunque una variable sea continua, cuando la medimos la estamos haciendo discreta, dada la limitación de los instrumentos de medida. No obstante, al clasificar las variables lo que hacemos es atender a su naturaleza, y no a los resultados obtenidos de la medición.

Atendiendo al número de caracteres cuantitativos que observamos en cada individuo, las variables pueden ser *unidimensionales*, *bidimensionales*, *tridimensionales*, ..., según se estudie en cada individuo de la población uno, dos, tres, ..., caracteres, respectivamente.

En este tema nos dedicaremos al estudio de las variables estadísticas unidimensionales y, en el siguiente, a las bidimensionales.

En el siguiente esquema se resumen los conceptos anteriores:

<table>
<tr><td rowspan="3">Caracteres Estadísticos</td><td rowspan="2">Cuantitativos
(Variables Estadísticas)</td><td>Discretas: valores aislados</td></tr>
<tr><td>Continuas: valores en un intervalo de la recta real.</td></tr>
<tr><td>Cualitativos (Atributos)</td><td></td></tr>
</table>

3. Frecuencias y tablas de frecuencias

Consideremos una población o muestra que consta de N individuos. Sea k el número de modalidades definidas para un determinado carácter. Tendremos entonces las modalidades M_1, M_2, ..., M_k.

Se llama *frecuencia absoluta*, n_i, de la modalidad M_i, al número de individuos de la población que pertenecen a dicha modalidad (el número de veces que se repite). Como las modalidades son incompatibles y exhaustivas, se tiene que $\sum_{i=1}^{n} n_i = N$

Se llama *frecuencia relativa*, f_i, de la modalidad M_i, a la proporción de individuos de la población que presentan dicha modalidad. Es decir, si el número total de individuos es N, entonces: $f_i = \frac{n_i}{N}$ y por tanto $0 \leq f_i \leq 1$

Llamaremos *frecuencia absoluta acumulada*, N_i, de la modalidad M_i, a la suma de las frecuencias absolutas hasta la i-ésima modalidad. Es decir:

$$N_i = n_1 + n_2 + ... + n_i = \sum_{r=1}^{i} n_r$$

Llamaremos *frecuencia relativa acumulada*, F_i, de la modalidad M_i, a la suma de las frecuencias relativas hasta la de la i-ésima modalidad. Es decir:

$$F_i = f_1 + f_2 + ... + f_i = \sum_{r=1}^{i} f_r$$

Los datos observados de una población se muestran clasificados y ordenados para dar mayor claridad y ofrecer una visión global del conjunto, que sea interpretable. Las dos formas de representación, que suponen los dos primeros pasos que hay que dar en el tratamiento estadístico de la información, son las *tablas estadísticas* y las *representaciones gráficas*.

Las tablas más simples son las que constan de una primera columna en la que se reflejan las distintas modalidades que presenta el carácter en estudio. Se añaden una o más columnas a su derecha en las que se anotan las respectivas frecuencias y otras más para cálculos posteriores.

El aspecto general de una tabla simple, para un carácter con k modalidades, es la siguiente:

Modalidades M_i	**Frecuencias absolutas ordinarias n_i**	**Frecuencias absolutas acumuladas N_i**	**Frecuencias relativas ordinarias f_i**	**Frecuencias relativas acumuladas F_i**
M_1	n_1	N_1	f_1	F_1
M_2	n_2	N_2	f_2	F_2
...	...	...	...	...
M_i	n_i	N_i	f_i	F_1
...	...	…	…	…
M_k	n_k	$N_k = N$	f_k	$F_k = 1$
	N		**1**	

Observemos que:

- La suma de todas las frecuencias absolutas ordinarias ha de coincidir con el número total de individuos de la población, es decir, con el tamaño N: $\sum_{i=1}^{n} n_i = N$
- La suma de todas las frecuencias relativas ordinarias ha de ser 1, ya que representa la suma de todas las proporciones: $\sum_{i=1}^{n} f_i = 1$

- La última frecuencia absoluta acumulada ha de ser el tamaño, N: $N_k = N$
- La última frecuencia relativa acumulada ha de ser 1: $F_k = 1$

4. Distribuciones de frecuencias

Consideremos una población de tamaño N estudiada según un carácter C que puede ser clasificado en k modalidades $M_1, M_2, ..., M_i, ..., M_k$

Llamamos *distribución de frecuencias* al conjunto de pares ordenados:

$\{(M_1, n_1), (M_2, n_2), ..., (M_i, n_i), ..., (M_k, n_k)\}$ (distribución de frecuencias absolutas)

o bien al conjunto :

$\{(M_1, f_1), (M_2, f_2), ..., (M_i, f_i), ..., (M_k, f_k)\}$ (distribución de frecuencias relativas)

En el caso discreto, las modalidades son los valores numéricos aislados que toma la variable estadística. Entonces, la distribución de frecuencias es:

$\{(x_1, n_1), (x_2, n_2), ..., (x_i, n_i), ..., (x_k, n_k)\}$ (en el caso de frecuencias absolutas)

o bien:

$\{(x_1, f_1), (x_2, f_2), ..., (x_i, f_i), ..., (x_k, f_k)\}$ (en el caso de frecuencias relativas)

4.1. Ejemplo

Un profesor tiene anotadas las calificaciones de los 30 alumnos de un grupo:

5	3	4	1	2	8	9	8	7	6
6	7	9	8	7	7	1	0	1	6
9	9	8	0	8	8	8	9	5	7

Construir la tabla de frecuencias absolutas, absolutas acumuladas, relativas y relativas acumuladas.

x_i	n_i	N_i	f_i	F_i
0	2	2	2/30	2/30
1	3	5	3/30	5/30
2	1	6	1/30	6/30
3	1	7	1/30	7/30
4	1	8	1/30	8/30
5	3	11	3/30	11/30
6	2	13	2/30	13/30
7	5	18	5/30	18/30
8	7	25	7/30	25/30
9	5	30	5/30	30/30
10	0	**30**	0	**30/30 = 1**
	30		**1**	

Se trata de una variable estadística discreta.

4.2. Caso continuo

En el caso continuo, o en el discreto con un gran número de datos, la población se particiona en *clases* o *intervalos*. Es decir, los datos se clasifican en intervalos de la recta real (*"El número de clases debe ser aproximadamente igual a la raíz cuadrada del número de datos"*), dando lugar a datos agrupados en intervalos:

$(e_0, e_1]$	$(e_1, e_2]$	...	$(e_{i-1}, e_i]$	...	$(e_{k-1}, e_k]$
Clase 1ª	Clase 2ª	...	Clase i-ésima	...	Clase última (k-ésima)

En las clases o intervalos tendremos en cuenta los siguientes conceptos:

- *Extremos de clase*: dada la clase i-ésima $(e_{i-1}, e_i]$, a e_{i-1} lo llamaremos *límite inferior* y a e_i *límite superior*.
- *Amplitud de clase*: llamaremos amplitud de la clase i-ésima $(e_{i-1}, e_i]$ a la longitud del intervalo, es decir, al número $a_i = e_i - e_{i-1}$
- *Marcas de clase*: son los puntos medios de las clases o intervalos. En el caso de la clase i-ésima $(e_{i-1}, e_i]$, la marca de clase es $x_i = \dfrac{e_{i-1} + e_i}{2}$

Hemos de tener en cuenta las siguientes observaciones:

- Las amplitudes de las clases no tienen por qué ser iguales. No obstante, si podemos elegir, es cómodo tomar todas las clases con la misma amplitud. Esto habrá que tenerlo muy en cuenta a la hora delas representaciones gráficas: histogramas de frecuencias.
- Más aún, las clases primera y última pueden ser intervalos no acotados, de amplitud infinita. Lo que se pretende con esto es recoger los casos muy extremos, "raros", que se pudieran dar.

En resumen, en el caso de las variables estadísticas continuas, o discretas con datos agrupados (tratamiento continuo por ser muy grande el número de datos), la distribución de frecuencias es un conjunto de la forma:

$$\{(I_1, n_1), (I_2, n_2), ..., (I_i, n_i), ..., (I_k, n_k)\} \text{ (en el caso de frecuencias absolutas)}$$

o bien:

$$\{(I_1, f_1), (I_2, f_2), ..., (I_i, f_i), ..., (I_k, f_k)\} \text{ (en el caso de frecuencias relativas)}$$

donde:

- $I_i = (e_{i-1}, e_i] = \{x_i \in \Re / e_{i-1} < x \leq e_i\}$ es la clase i-ésima.
- Las clases primera y última pueden ser de la forma:
  - $I_1 = (-\infty, e_1] = \{x \in \Re / x \leq e_1\}$
  - $I_k = (e_{k-1}, +\infty) = \{x \in \Re / e_{k-1} < x\}$

4.3. Ejemplo

Las edades de las personas que acuden a un médico a lo largo de un mes son:

3	2	11	13	4	3	2	4	5	6	7	3
4	5	3	2	5	6	27	15	4	21	14	4
3	6	29	13	6	17	6	13	6	5	12	26

Construir la correspondiente tabla de frecuencias agrupando los datos en clases o intervalos de amplitud 5.

Clases I_i	**Marcas de clase x_i**	**n_i**	**N_i**	**f_i**	**F_i**
(0, 5]	2,5	17	17	17/36	17/36
(5, 10]	7,5	7	24	7/36	24/36
(10, 15]	12,5	7	31	7/36	31/36
(15, 20]	17,5	1	32	1/36	32/36
(20, 25]	22,5	1	33	1/36	33/36
(25, 30]	27,5	3	**36**	3/36	36/36=**1**
		N = 36			

Observemos que se trata de una variable estadística discreta a la que, por haber un número grande de datos, se trata como continua agrupando los datos en intervalos.

5. Representaciones gráficas

Aunque las tablas de frecuencias contienen información suficiente para permitir el análisis de los datos, comúnmente se recurre a su representación gráfica con el objetivo de obtener una mejor idea del comportamiento de los datos.

Según sea el carácter estudiado, se emplean distintos tipos de representaciones gráficas o diagramas:

Carácter cualitativo (atributo)	• Diagrama rectangular. • Diagrama de sectores. • Pictogramas. • Cartogramas. • Pirámides de población.	
Carácter cuantitativo (variable estadística)	Variable discreta	• Diagrama de barras. • Función de distribución.
	Variable continua	• Histograma. • Función de distribución.

En este tema veremos los diagramas rectangulares y de sectores para caracteres cualitativos y los diagramas de barras e histogramas para los cuantitativos.

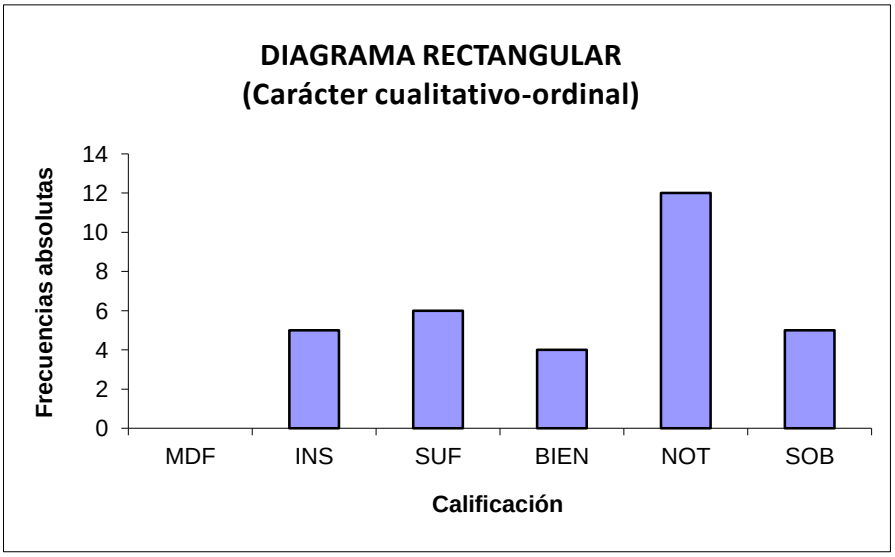
5.1. Diagrama rectangular (carácter cualitativo)

Están constituidos por varios rectángulos de base constante, una por cada modalidad, y con altura proporcional a la frecuencia absoluta (sin más que cambiar la escala del eje de ordenadas se tendría la misma gráfica para las frecuencias relativas.

Por ejemplo, consideremos que las calificaciones obtenidas por los 32 alumnos de una clase en la asignatura de matemáticas vienen dadas en la siguiente tabla:

M_i	n_i	N_i	f_i	F_i
Muy Deficiente	0	0	0/32	0/32
Insuficiente	5	5	5/32	5/32
Suficiente	6	11	6/32	11/32
Bien	4	15	4/32	15/32
Notable	12	27	12/32	27/32
Sobresaliente	5	**32**	5/32	32/32 = **1**
	N = 32		32/32 = **1**	

Un diagrama rectangular sería el siguiente:



5.2. Diagrama de sectores (carácter cualitativo)

Consiste en hacer corresponder un círculo a la frecuencia total (preferentemente relativa, expresada en términos porcentuales) y hacer corresponder a cada modalidad M_i un sector circular de amplitud proporcional a la frecuencia correspondiente. Para ello se recurre a cualquiera de las reglas de tres simples que tienes a continuación:

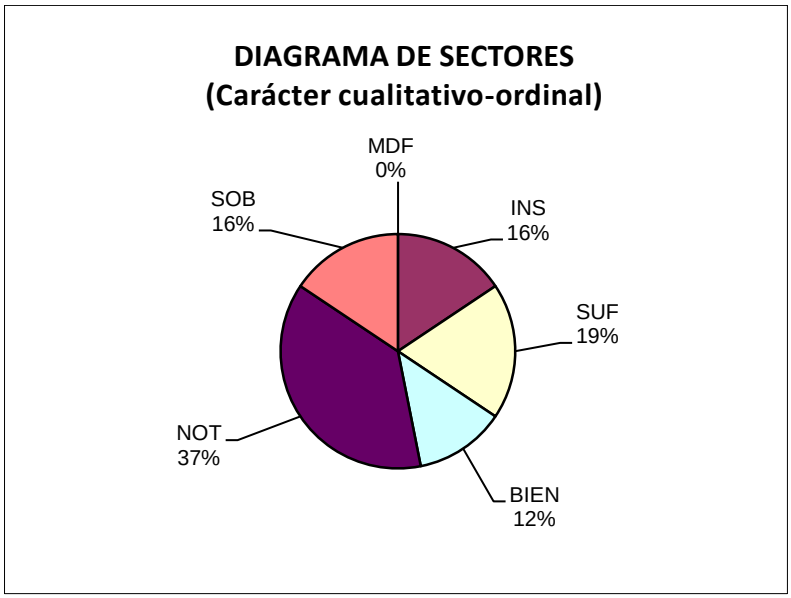
$$\begin{array}{c} N \longrightarrow 360^o \\ n_i \longrightarrow \alpha_i \\ \dfrac{N}{n_i} = \dfrac{360^o}{\alpha_i} \end{array} \qquad \begin{array}{c} 1 \longrightarrow 360^o \\ f_i \longrightarrow \alpha_i \\ \dfrac{1}{f_i} = \dfrac{360^o}{\alpha_i} \end{array} \qquad \begin{array}{c} 100 \longrightarrow 360^o \\ p_i\,(100\cdot f_i) \longrightarrow \alpha_i \\ \dfrac{100}{p_i} = \dfrac{360^o}{\alpha_i} \end{array}$$

De donde:

$$\alpha_i = \frac{n_i}{N}\cdot 360^o \qquad \alpha_i = f_i \cdot 360^o \qquad \alpha_i = \frac{p_i}{100}\cdot 360^o$$

Para el ejemplo anterior se tendría:

M_i	n_i	f_i	p_i (%)	α_i (º)
Muy Deficiente	0	0/32 = 0,0000	0,00	0,00
Insuficiente	5	5/32 = 0,15625	15,625	56,25
Suficiente	6	6/32 = 0,1875	18,75	67,50
Bien	4	4/32 = 0,1250	12,50	45,00
Notable	12	12/32 = 0,3750	37,50	135,00
Sobresaliente	5	5/32 = 0,15625	15,625	56,25
	N = 32	32/32 = **1**	100,00	360,00



5.3. Diagrama de barras (variable estadística discreta)

Se llama así la representación gráfica de frecuencias de una variable estadística discreta (carácter cuantitativo discreto) obtenida de la forma siguiente:

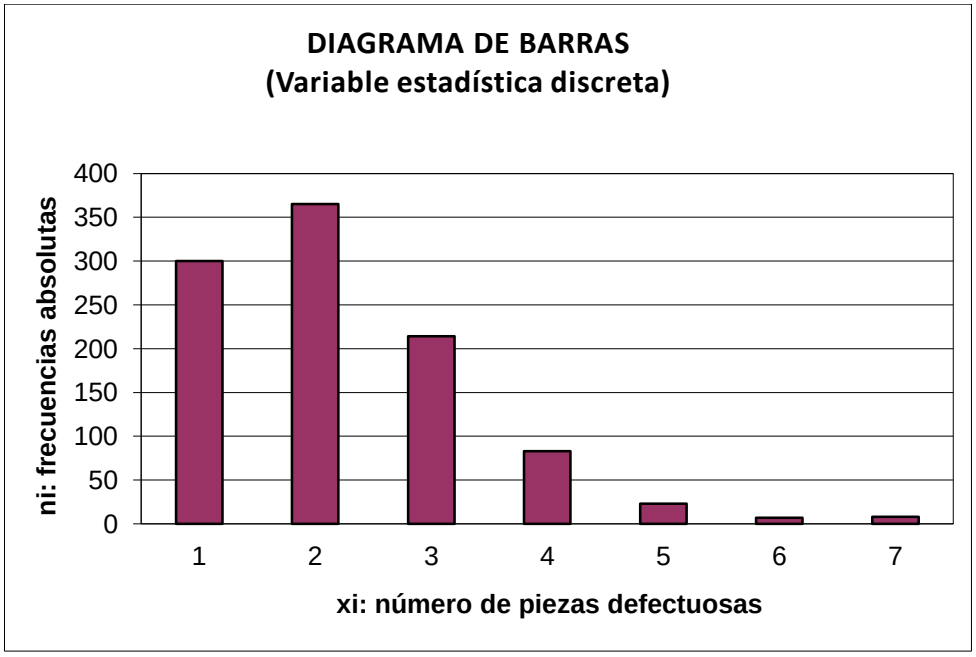
- Sobre el eje de abscisas se marca cada uno de los valores de la variable en una escala aritmética (divisiones iguales).
- Sobre el eje de ordenadas se lleva a cabo una graduación aritmética que permita representar las frecuencias absolutas o relativas (si se van a hacer comparaciones mejor relativas).
- Sobre cada punto del eje de abscisas, correspondiente a un valor de la variable, se levanta una barra de altura proporcional a la frecuencia de dicho valor.

Es un diagrama similar al diagrama rectangular para caracteres cualitativos.

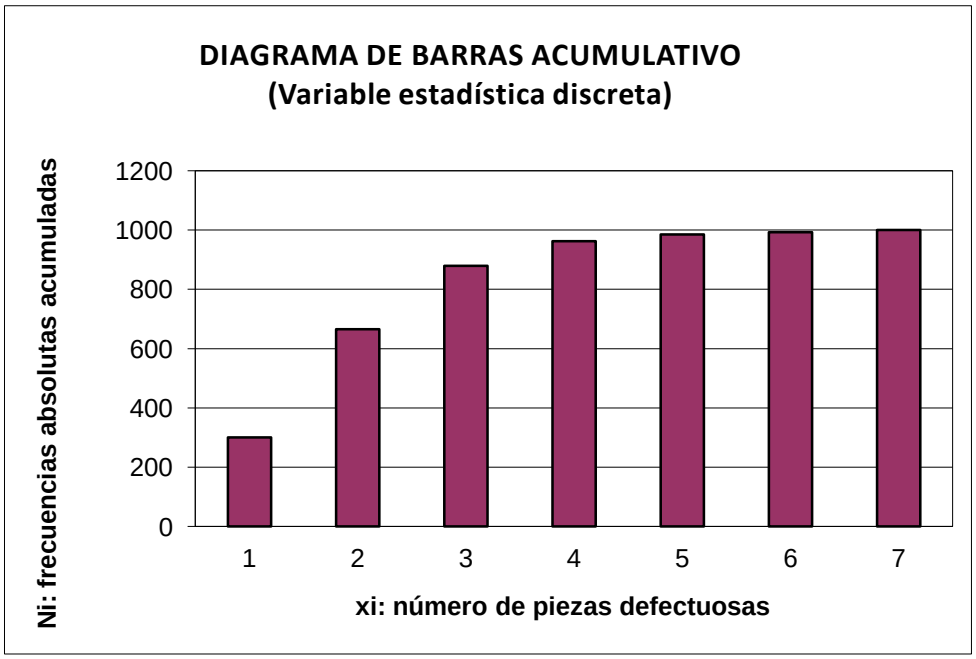
Por ejemplo, consideremos una población formada por 1000 lotes de ciertas piezas mecánicas. El carácter (cuantitativo) que se observa en cada unidad estadística es el número de piezas defectuosas que contiene: 0, 1, 2, 3, 4, 5 ó 6 (estas son las modalidades, los valores de la variable discreta en cuestión).

Las frecuencias vienen dadas en la siguiente tabla:

Número de piezas defectuosas por lote	x_i	0	1	2	3	4	5	6	
Número de lotes con x_i piezas defectuosas	n_i	300	365	214	83	23	7	8	**1000**
Frecuencias acumuladas	N_i	300	665	879	962	985	992	**1000**	



Cambiando frecuencias absolutas ordinarias, n_i, por frecuencias absolutas acumuladas N_i, tendríamos el diagrama de barras acumulativo.



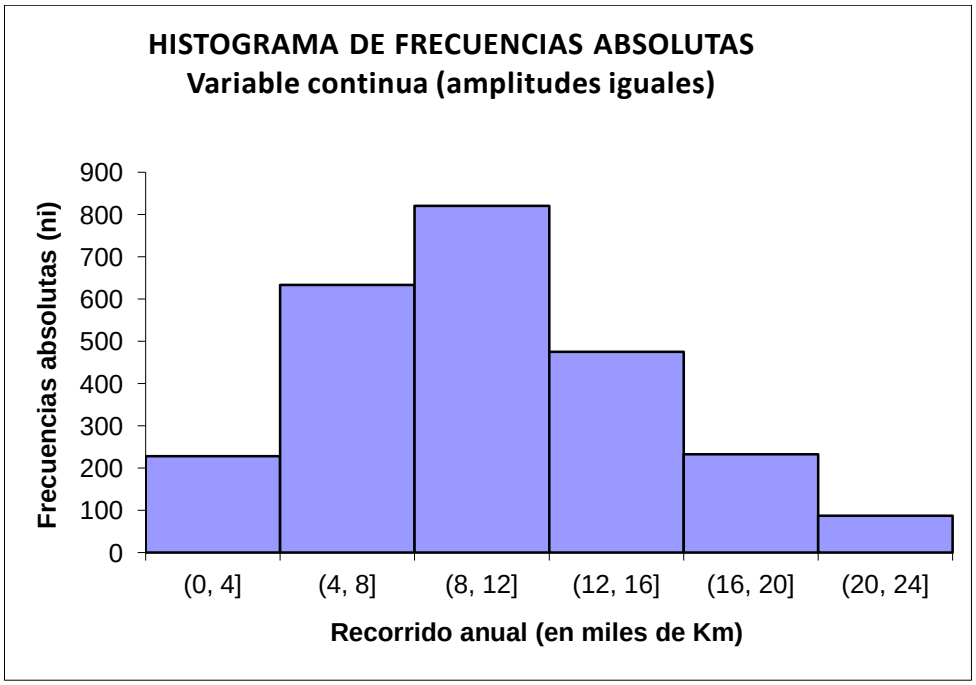
5.4. Histograma de frecuencias (variable continua)

Se llama *histograma* a la representación gráfica de las frecuencias de una distribución estadística de una variable continua cuyas observaciones están agrupadas en clases. Se construye de la forma siguientes:

- Sobre el eje de abscisas, graduado según una escala aritmética, se marcan los extremos de las clases sucesivas.
- Sobre el eje de ordenadas se marcarán las frecuencias.
- Sobre cada intervalo o clase se dibuja un rectángulo de modo que las áreas de loas rectángulos sean proporcionales a las frecuencias.

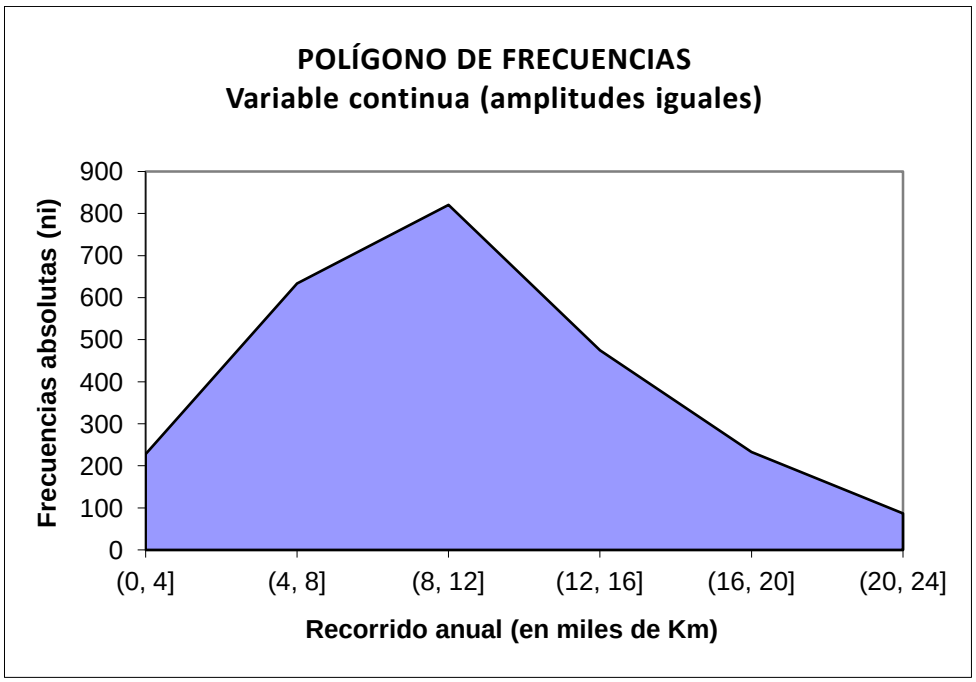
Por ejemplo, consideremos un parque automovilístico de 2478 coches clasificados según el número de kilómetros recorridos en un año:

Kilometraje anual (en miles de kilómetros)	**Número de vehículos**
$(e_{i-1}, e_i]$	**n_i**
(0, 4]	228
(4, 8]	634
(8, 12]	821
(12, 16]	475
(16, 20]	233
(20, 24]	87
	N = 2478



Observemos que todos los intervalos tienen la misma amplitud. Entonces, para la construcción del histograma, podemos asignar como altura de cada rectángulo la frecuencia absoluta del intervalo correspondiente.

Uniendo el vértice superior izquierdo o los puntos medios de los techos de los rectángulos, se obtiene una línea poligonal que encierra sobre el eje X un área igual a la que encierran los rectángulos. Tal línea es el *polígono de frecuencias*.



6. Reducción numérica de los datos

Hasta ahora hemos tratado y representado gráficamente las distribuciones de frecuencias según un carácter. Con ello tenemos una primera aproximación al conocimiento de las mismas.

Ahora daremos un conjunto de medidas descriptivas que resuman cuantitativamente, de modo sucinto y significativo, las características más importantes de una distribución. Esto nos permitirá comparar distintas distribuciones.

Por ejemplo, si se desea comparar las temperaturas de Granada y Ciudad Real a lo largo de un año, sería mejor disponer de unos pocos números que representaran de forma resumida a cada una de las provincias que comparar las temperaturas de todos y cada uno de los días del año. Lo único que hay que hacer es tomar esos números de modo que sean representativos de todo el grupo; es decir, unos valores que representen o resuman a toda la población. Estos números se llaman *parámetros estadísticos* o, simplemente, *estadísticos*. En nuestro ejemplo se suele recurrir a la temperatura media de las máximas y a la temperatura media de las mínimas.

6.1. Medidas de centralización

Las medidas o estadísticos de centralización, o de tendencia central, nos indican los punto en torno a los cuales se encuentran los valores de la variable estadística en estudio; es decir, nos indican los puntos centrales de una distribución. Representan el conjunto de los datos mediante un solo valor numérico, tratando de resumir y sintetizar la distribución de frecuencias. Las medidas de posición central más utilizadas son la mediana, la moda y la media aritmética.

- **Mediana**

Sea X una variable estadística (carácter cuantitativo) de una población o muestra con N individuos.

Se llama *mediana* a un valor, representado por Me, tal que, ordenados los N valores de X en orden creciente, el 50% de ellos son menores o iguales que Me y el 50% restante son mayores o iguales que Me.

Para determinar la mediana los haremos en el caso discreto y continuo.

✓ **Caso discreto**

Consideraremos la siguiente distribución de frecuencias que nos servirá de ejemplo:

x_i	n_i	N_i	f_i	F_i
3	1	1	1/9	1/9
4	2	3	2/9	3/9
5	1	4	1/9	4/9
6	1	5	1/9	5/9
7	3	8	3/9	8/9
8	0	8	0	8/9
9	0	8	0	8/9
10	1	**9**	1/9	9/9 = **1**
	N = 9		**1**	

Podemos proceder de dos formas:

- **Directamente sobre los datos**: ordenamos los datos sin agrupar; es decir, repitiendo cada uno tantas veces como indique su frecuencias absoluta.

3	4	4	5	**6** **Me**	7	7	7	10

En este caso, N = 9 es impar y la mediana es el valor central: Me = 6 deja a la mitad de individuos por encima y a la otra mitad por debajo.

- **A partir de la tabla de frecuencias**: observamos en la columna de las frecuencias absolutas acumuladas donde se encuentra el valor N/2. Este dejará por encima la frecuencia absoluta acumulada N_i y por debajo la frecuencia absoluta acumulada N_{i+1}. La mediana es el valor de la variable que se encuentra inmediatamente por debajo de esta posición, es decir, x_{i+1}. En nuestro ejemplo N/2 = 4,5 y por tanto Me = 6. Observa la tabla:

x_i	n_i	N_i	f_i	F_i
3	1	1	1/9	1/9
4	2	3	2/9	3/9
5 = x_i	1	4 = N_i	1/9	4/9
N/2 →		**4,5**		
← **Me = x_{i+1}** 6	1	5 = N_{i+1}	1/9	5/9
7	3	8	3/9	8/9
8	0	8	0	8/9
9	0	8	0	8/9
10	1	**9**	1/9	9/9 = **1**
	N = 9		**1**	

Puede ocurrir que N/2 coincida con algún valor de N_i. Entonces la mediana es el valor medio entre x_i y x_{i+1}: $Me = \frac{x_i + x_{i+1}}{2}$

Por ejemplo, consideremos ahora la siguiente distribución de frecuencias. En este caso N/2 = 5, que coincide con uno de los valores de N_i. Por tanto $Me = \frac{x_i + x_{i+1}}{2} = \frac{6+7}{2} = 6{,}5$.

N/2

x_i	n_i	N_i	f_i	F_i
3	1	1	1/10	1/10
4	2	3	2/10	3/10
5	1	4	1/10	4/10
⑥	1	⑤	1/10	5/10
⑦	3	8	3/10	8/10
8	0	8	0	8/10
9	0	8	0	8/10
10	2	**10**	2/10	10/10 = **1**
	N = 10		**1**	

$Me = \frac{6+7}{2} = 6{,}5$

Observa que si calculamos la mediana directamente sobre los datos, al ser ahora N par, quedan dos valores centrales. La mediana es el valor medio de estos:

3 4 4 5 [6 7] 7 7 10 10

$Me = \frac{6+7}{2} = 6{,}5$

✓ **Caso continuo**

Para este caso tomaremos el ejemplo de clases de igual amplitud de la página 134: consideremos un parque automovilístico de 2478 coches clasificados según el número de kilómetros recorridos en un año:

Kilometraje anual (en miles de kilómetros)	**Número de vehículos**	**Frecuencias absolutas acumuladas**
$(e_{i-1}, e_i]$	**n_i**	**N_i**
(0, 4]	228	228
(4, 8]	634	862
	1239 ← **N/2**	
(8, 12] ← **Intervalo mediano**	821	1683
(12, 16]	475	2158
(16, 20]	233	2391
(20, 24]	87	**2478**
	N = 2478	

En este caso, la primera clase cuya frecuencia absoluta acumulada es mayor o igual que N/2 es el *intervalo mediano* o *clase mediana* de la distribución: que los llamaremos $I_i = (e_{i-1}, e_i]$ En nuestro ejemplo $I_i = (8, 12]$. Para obtener la mediana se recurre a la siguiente fórmula:

$$Me = e_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}} \cdot a_i$$

donde e_{i-1} es el límite inferior del intervalo mediano, a_i es la amplitud del intervalo mediano, N_{i-1} es la frecuencia absoluta acumulada que se encuentra inmediatamente por encima del intervalo mediano, N_i es la frecuencia absoluta acumulada correspondiente al intervalo mediano y N es el número de individuos de la población.

En nuestro ejemplo: $Me = e_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}} \cdot a_i = 8 + \frac{\frac{2478}{2} - 862}{1683 - 862} \cdot 4 = 9{,}84$

- **Moda**

✓ **Caso discreto**

Dada una variable estadística discreta X con distribución de frecuencias

$$\{(x_1, n_1), (x_2, n_2), ..., (x_i, n_i), ..., (x_k, n_k)\}$$

se llama *moda*, y se representa por Mo, a la modalidad que presenta una frecuencia máxima. En el diagrama de barras es la modalidad a la que corresponde la barra más alta. Una distribución puede tener, pues, más de una moda, en el caso de que la frecuencia más alta corresponda a más de una modalidad.

Si consideramos el ejemplo de las páginas 132 y 133:

Número de piezas defectuosas por lote	$\mathbf{x_i}$	0	**1**	2	3	4	5	6	
Número de lotes con x_i piezas defectuosas	$\mathbf{n_i}$	300	**365**	214	83	23	7	8	**1000**

El valor que se presenta con más frecuencia es el 1 (365 veces). Por tanto Mo = 1.

✓ **Caso continuo**

Dada una variable estadística continua X con distribución de frecuencias

$$\{(I_1, n_1), (I_2, n_2), ..., (I_i, n_i), ..., (I_k, n_k)\}$$

se llama *clase o intervalo modal* al intervalo que presenta una "mayor densidad de frecuencia". En el histograma es al que le corresponde el rectángulo de mayor altura.

En el ejemplo de la página 134:

Kilometraje anual (en miles de kilómetros)	**Número de vehículos**
$\mathbf{(e_{i-1}, e_i]}$	$\mathbf{n_i}$
(0, 4]	228
(4, 8]	634
(8, 12]	821
(12, 16]	475
(16, 20]	233
(20, 24]	87
	N = 2478

La clase o intervalo modal es, en este caso, (8, 12] pues es la que se presenta en un mayor número de ocasiones (821).

Si queremos especificar más concretamente a que valor de la variable le atribuimos el papel de moda, aplicaremos la siguiente fórmula:

$$Mo = e_{i-1} + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} \cdot a_i$$

donde e_{i-1} es el límite inferior de la clase modal, n_i es la frecuencia absoluta correspondiente al intervalo modal, n_{i-1} es la frecuencia absoluta inmediatamente anterior a n_i, n_{i+1} es la frecuencia absoluta inmediatamente posterior a n_i y a_i es la amplitud de la clase modal.

En nuestro ejemplo $Mo = 8 + \frac{821-634}{(821-634)+(821-475)} \cdot 4 = 9{,}04$

Si llamamos $\Delta_1 = n_i - n_{i-1}$ (exceso de la clase modal sobre la clase contigua anterior) y $\Delta_2 = n_i - n_{i+1}$ (exceso de la clase modal sobre la clase contigua posterior), la fórmula anterior se convierte en:

$$Mo = e_{i-1} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot a_i$$

En el ejemplo $\Delta_1 = 821 - 634 = 187$ y $\Delta_2 = 821 - 475 = 346$, y entonces se tiene que $Mo = 8 + \frac{187}{187+346} \cdot 4 = 9{,}04$

✓ **Observaciones:**

- Cuando una distribución presenta varios máximos locales, bien en el diagrama de barras (caso discreto) o bien en el histograma (caso continuo), se habla de una distribución multimodal.
- Cuando la clase modal sea una clase extrema, la primera o la última, se supone que la clase anterior o la posterior, respectivamente, es de frecuencia nula.

- **Media aritmética**

✓ **Caso discreto**

Sea X una variable estadística discreta de una población finita de tamaño N y sean x_1, x_2, ..., x_N los N valores observados de X.

La *media aritmética*, o simplemente *media*, de esos N valores es:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Si de esos N valores sólo hay k distintos x_1, x_2, ..., x_k, que se repiten, respectivamente, n_1, n_2, ..., n_k veces (sus frecuencias absolutas), entonces podemos escribir:

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \ldots + n_kx_k}{N} = \frac{\sum_{i=1}^{k} n_ix_i}{N}$$

o bien, si empleamos frecuencias relativas:

$$\bar{x} = f_1x_1 + f_2x_2 + \ldots + f_kx_k = \sum_{i=1}^{k} f_ix_i$$

Usemos uno de los ejemplos anteriores para ver cómo se ordenan los cálculos:

x_i	n_i	n_ix_i	f_i	f_ix_i
3	1	3	1/10	3/10
4	2	8	2/10	8/10
5	1	5	1/10	5/10
6	1	6	1/10	6/10
7	3	21	3/10	21/10
8	0	0	0	0
9	0	0	0	0
10	2	20	2/10	20/10
	$\Sigma n_i = N = 10$	**$\Sigma n_ix_i = 63$**	**$\Sigma f_i = 1$**	**$\Sigma f_ix_i = 63/10$**

$$\bar{x} = \frac{\sum_{i=1}^{8} n_ix_i}{N} = \frac{63}{10} = 6{,}3$$

o bien, si preferimos trabajar con frecuencias relativas:

$$\bar{x} = \sum_{i=1}^{8} f_ix_i = \frac{63}{10} = 6{,}3$$

✓ **Caso continuo**

En este caso, reemplazamos las clases por sus marcas x_i (lo que equivale a suponer que todos los puntos del intervalo están concentrados en su punto medio). Se trata de una especie de "discretización" de la variable. Las fórmulas para el calculo de la media son las mismas de antes.

Por ejemplo:

Clase	**Marca de clase**	**Frecuencias absolutas**	
$(e_{i-1}, e_i]$	**x_i**	**n_i**	**x_in_i**
(0, 150]	75	120	9000
(150, 300]	225	159	35775
(300, 350]	325	89	28925
(350, 400]	375	78	29250
(400, 500]	450	66	29700
(500, $\rightarrow$)	550	52	28600
		N = 564	**161250**

Para la clase extrema (500, →) se podrían adoptar diversos convenios. Hemos adoptado el de asignarle la misma amplitud que a la anterior.

La media es, por tanto: $\bar{x} = \dfrac{\sum_{i=1}^{6} n_i x_i}{N} = \dfrac{161250}{564} = 285{,}9$

6.2. Medidas de posición

Son una generalización de la mediana. En general, sirven para determinar en qué posición de la distribución se encuentra un individuo, supuestos ordenas en orden creciente.

Sea X una variable estadística (discreta o continua) sobre una población finita de tamaño N, y sea t un número real tal que $0 < t < 1$.

Se llama *cuantil de orden* t al valor C_t tal que $t \cdot N$ individuos de la población son tales que $X \leq C_t$ y los $(1 - t) \cdot N$ individuos restantes son tales que $X \geq C_t$. Dicho de otro modo, el 100t % de los individuos se encuentra por debajo del cuantil C_t y el 100(1 –t) % de individuos restante se encuentra por encima del cuantil C_t.

Si t = 0,5, entonces $C_{0,5}$ = Me (la mediana). Si para un individuo ocurre que $X \leq Me$, tal individuo está en la primera mitad de la población ordenada.

La interpretación de los cuantiles y las circunstancias que se pueden dar en su determinación, según los casos, son exactamente las mismas que para la mediana.

En el caso discreto, bien a partir de los datos sin agrupar o bien a partir de la distribución de frecuencias absolutas tomando como referencia el valor tN para mirar en la columna de frecuencias absolutas acumuladas. En el ejemplo de variable discreta al final de este apartado se verá con toda claridad.

Para el caso continuo, con los datos agrupados en intervalos, existe una fórmula análoga a la de la mediana para el cuantil de orden t:

$$C_t = e_{i-1} + \frac{tN - N_{i-1}}{N_i - N_{i-1}} \cdot a_i$$

Los cuantiles se estudian en grupos que dividen a la población en un cierto número de partes iguales, ordenados los individuos por el valor de la variable en orden creciente.

Según el número de partes en que dividen a la población reciben distintos nombre:

- **Cuartiles**

Dividen a la población en cuatro partes, cada una de las cuales contiene al 25% de las observaciones. Los cuartiles son:

Primer cuartil:	$Q_1 = C_{1/4}$	$(t = 1/4 = 0{,}25)$
Segundo cuartil:	$Q_2 = C_{1/2} = Me$	$(t = 1/2 = 0{,}5)$
Tercer cuartil:	$Q_3 = C_{3/4}$	$(t = 3/4 = 0{,}75)$

En el caso continuo, una vez determinado el intervalo $(e_{i-1}, e_i]$ que contiene a Q_k, de frecuencia absoluta acumulada N_k, las fórmulas para los tres cuartiles son:

$$Q_1 = e_{i-1} + \frac{\frac{1}{4}N - N_{i-1}}{N_i - N_{i-1}} \cdot a_i$$

$$Q_2 = e_{i-1} + \frac{\frac{1}{2}N - N_{i-1}}{N_i - N_{i-1}} \cdot a_i = Me$$

$$Q_3 = e_{i-1} + \frac{\frac{3}{4}N - N_{i-1}}{N_i - N_{i-1}} \cdot a_i$$

Es conveniente observar que los cuartiles no tienen por qué estar unos a la misma distancia de otros: lo que han de verificar es que entre cada dos consecutivos esté el 25% de la población:

25 %	25 %	25 %	25 %

e_0 Q_1 $Q_2 = Me$ Q_3 e_k

- **Deciles**

Dividen a la población en diez partes, cada una de las cuales contiene al 10% de las observaciones. Los deciles son:

Primer decil: $D_1 = C_{1/10}$ $(t = 0{,}10)$

Segundo decil: $D_2 = C_{2/10}$ $(t = 0{,}20)$

……

Quinto decil: $D_5 = C_{5/10} = Q_2 = Me$ $(t = 0{,}50)$

……

Noveno decil: $D_9 = C_{9/10}$ $(t = 0{,}90)$

La forma de calcularlos es la misma de antes:

$$D_\lambda = e_{i-1} + \frac{\frac{\lambda}{10}N - N_{i-1}}{N_i - N_{i-1}} \cdot a_i$$

- **Centiles o percentiles**

Dividen a la población en cien partes, cada una de las cuales contiene al 1% de ella. Los percentiles son:

$P_1 = C_{1/100}$ $(t = 0{,}01)$

……..

$P_{25} = C_{25/100} = Q_1$ $(t = 0{,}25)$

……..

$P_{50} = C_{50/100} = Q_2 = Me$ $(t = 0{,}50)$

……..

$P_{75} = C_{75/100} = Q_3$ $(t = 0{,}75)$

……..

$P_{99} = C_{99/100}$ $(t = 0{,}99)$

Veamos dos ejemplos (uno de variable discreta y otro de variable continua) en los que se aprecie el cálculo de los distintos parámetros:

• **Ejemplo 1**: consideremos el ejemplo de las páginas 132 y 133: una población formada por 1000 lotes de ciertas piezas mecánicas. El carácter que se observa es el número de piezas defectuosas que contiene: 0, 1, 2, 3, 4, 5 ó 6.

Las frecuencias vienen dadas en la siguiente tabla:

x_i	0	1	2	3	4	5	6	
n_i	300	365	214	83	23	7	8	**1000**
N_i	300	665	879	962	985	992	**1000**	
n_ix_i	0	365	428	249	92	35	48	**1217**

Calcular la media, la moda, la mediana, los tres cuartiles, los deciles sexto y séptimo, y los percentiles P_{40} y P_{95}

Es claro que $\bar{x} = \dfrac{1217}{1000} = 1{,}217$ y que Mo = 1. Para determinar los demás parámetros miraremos en la fila de frecuencias absolutas acumuladas.

✓ **Mediana**: la primera frecuencia absoluta acumulada que es mayor que N/2 = 500 es $N_2 = 665$. Por tanto Me = 1

✓ **Cuartiles**: la primera frecuencia absoluta acumulada que es mayor o igual que N/4 = 250 es $N_1 = 300$. Entonces $Q_1 = 0$ (el primer 25% de los lotes observados, ordenados por orden creciente de piezas defectuosas, tiene 0 piezas defectuosas). Por otro lado $Q_2 = Me = 1$ (el segundo 25% de los lotes observados tienen 0 ó 1 pieza defectuosas). Por último, la primera frecuencia absoluta acumulada que es mayor que 3N/4 = 750 es $N_3 = 879$. Entonces $Q_3 = 2$ (el tercer 25% de la población tiene 0, 1 ó 2 piezas defectuosas).

✓ **Sexto y séptimo deciles**: la primera frecuencia absoluta acumulada que es mayor que 6N/10 = 600 es $N_2 = 665$. Por tanto $D_6 = 1$ (es decir, el primer 60% de los lotes observados tienen 0 ó 1 piezas defectuosas). De forma similar, como 7N/10 = 700, la primera frecuencia acumulada que es mayor que tal valor es $N_3 = 879$ y entonces $D_7 = 2$ (lo que quiere decir que el 70% de los lotes tienen 0, 1 ó 2 piezas defectuosas).

✓ **Percentiles P_{40} y P_{95}**: la primera frecuencia absoluta acumulada que es mayor que 40N/100 = 400 es $N_2 = 665$. Entonces $P_{40} = 1$ (el 40% de los lotes tiene 0 ó 1 pieza defectuosa). Finalmente, como la primera frecuencia absoluta acumulada que es mayor que 95N/100 = 950 es $N_4 = 962$, tenemos que $P_{95} = 3$ (el 95% de los lotes tienen 0, ,1, 2 ó 3 piezas defectuosas).

• **Ejemplo 2**: los pesos en kg. de 100 alumnos de un colegio vienen dados por la tabla

I_i	n_i	x_i	N_i	n_ix_i
(40, 48]	8	44	8	352
(48, 56]	22	52	30	1144
(56, 64]	29	60	59	1740
(64, 72]	21	68	80	1428
(72, 80]	20	76	**100**	1520
	100			**6184**

Calcular la media, la moda, la mediana, los el tercer cuartil y el percentil P_{35}

✓ **Media**: $\bar{x} = \frac{6184}{100} = 61{,}84$

✓ **Moda**: el intervalo modal es (56, 64], y entonces Mo = $e_{i-1} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot a_i$ = $= 56 + \frac{7}{7+8} \cdot 8 = 59{,}73$

✓ **Mediana**: el primer intervalo cuya frecuencia absoluta acumulada es mayor que N/2 = 50 es (56, 64]. Por tanto Me = $e_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}} \cdot a_i = 56 + \frac{50-30}{59-30} \cdot 8 = 61{,}52$

✓ **Q_3**: el primer intervalo cuya frecuencia absoluta acumulada es mayor que 3N/4 = 75 es (64, 72]. Por tanto $Q_3 = e_{i-1} + \frac{\frac{3}{4}N - N_{i-1}}{N_i - N_{i-1}} \cdot a_i = 64 + \frac{75-59}{80-59} \cdot 8 = 70{,}095$. Esto quiere decir que el 75% de los alumnos tienen un peso inferior a 70,095 kg.

✓ **P_{35}**: el primer intervalo cuya frecuencia absoluta acumulada es mayor que 35N/100 = 35 es (56, 64]. Entonces $P_{35} = e_{i-1} + \frac{\frac{35}{100}N - N_{i-1}}{N_i - N_{i-1}} \cdot a_i = 56 + \frac{35-30}{59-30} \cdot 8 =$ $= 57{,}38$. Es decir, el 35% de los alumnos tienen un peso inferior a 57,38 kg.

6.3. Medidas de dispersión

Las medidas de centralización sintetizan la información: representan la totalidad del conjunto de datos mediante unos valores centrales. Sin embargo, un promedio no es suficiente. Es preciso añadir también una medida de cómo de representativo es dicho promedio.

Consideremos las siguientes distribuciones:

A:	20	22	24	26	28
B:	10	10	20	35	45

que podrían representar los pesos de dos grupos de niños. Observamos que los dos grupos tienen el mismo peso medio: $\bar{x} = 24$, siendo, no obstante, muy diferentes en cuanto a concentración-dispersión de sus valores. En el grupo A los valores se encuentran próximos a la media, luego tienen poca dispersión. En el grupo B, los valores están alejados de la media, estando formado por valores más dispersos.

Al grado en que los datos numéricos tienden a extenderse alrededor de un valor promedio (estadístico de centralización como la media o mediana, por ejemplo) se le llama *variación* o *dispersión* de los datos.

Se utilizan distintas medidas de dispersión. Las más empleadas son: rango o recorrido, recorrido intercuartílico, desviación absoluta media, varianza, desviación típica y coeficiente de variación. Nosotros nos ceñiremos a las más importantes: varianza, desviación típica y coeficiente de variación.

- **Varianza y desviación típica**

La más empleada de las medidas de dispersión es la *varianza*, que se define como la media de los cuadrados de las desviaciones respecto a la media; esto es, la media de la variable : $\left(X-\overline{X}\right)^2$:

$$\text{Var}(X)=\sigma^2=\overline{\left(X-\overline{X}\right)^2}=\frac{\sum_{i=1}^{k}n_i(x_i-\overline{x})^2}{N}=\sum_{i=1}^{k}f_i(x_i-\overline{x})^2$$

Puesto que la varianza de X no viene dada en las mismas unidades de X (si, por ejemplo, la variable viene dada en metros, la varianza resulta en metros cuadrados), en su lugar se emplea la *desviación típica*, σ, definida como

$$\sigma=+\sqrt{\text{Var}(X)}=+\sqrt{\sigma^2}$$

En la medida en que la varianza o la desviación típica tomen valores más o menos grandes, esto indicará el grado de dispersión o alejamiento de los datos respecto de la media. En el caso trivial de que todos los valores de la variable estén concentrados en un punto (que coincidirá con la media), estos estadísticos de dispersión se anularán.

Hay una fórmula que se obtiene del desarrollo de la expresión de la varianza que permite calcular ésta de manera simplificada. Es la siguiente:

$$\text{Var}(X)=\sigma^2=\frac{\sum_{i=1}^{k}n_i x_i^{\,2}}{N}-\overline{x}^{\,2}$$

- **Coeficiente de variación**

Las medidas de dispersión estudiadas hasta ahora se expresan en la misma unidad de medida que la variable estadística, designando medidas de dispersión absoluta. Esto presenta algunos problemas técnicos:

- ✓ ¿Cómo hacer comparaciones entre dos distribuciones de naturaleza diferente (alturas y pesos) o, aun siendo de la misma naturaleza, expresadas en unidades diferentes (metros y pulgadas)?
- ✓ Por otro lado, una variación de 100 € en una serie de compras cuyo precio medio es de 1000 € tiene una repercusión muy diferente que la misma variación de 100 € en una serie de compras cuyo precio medio es de 1000000 €.

Para resolver estos problemas recurrimos a una medida de dispersión relativa, que recibe el nombre de *coeficiente de dispersión o de variación de Pearson*:

$$CV=\frac{\sigma}{\overline{x}}$$

Esta es una medida abstracta que no tiene dimensiones. Tiene las siguientes propiedades:

- ✓ Suele expresarse en %: $CV=\frac{\sigma}{\overline{x}}\cdot 100$

- ✓ Cuanto menor es el coeficiente de variación más homogénea respecto a la media es la distribución.
- ✓ Cuanto más cerca de 0 esté, más representativa de la distribución es la media.
- ✓ A medida que se aleja de 0, la media es menos representativa.
- ✓ Al ser una medida relativa, permite comparar distribuciones del mismo tipo aunque tengan distinto tamaño.
- ✓ Tiene el inconveniente de que deja de ser útil cuando $\bar{x}$ está próxima a 0.
- ✓ Es independiente de las unidades utilizadas.

- **Ejemplo:** volvamos sobre el ejemplo de la página 143 en el que se daban los pesos en kg. de 100 alumnos de un colegio. Calculemos la varianza, la desviación típica y el coeficiente de variación. Para ello vamos a diseñar la tabla de manera que nos sea útil para realizar los cálculos

I_i	n_i	x_i	x_i^2	$n_i x_i$	$n_i x_i^2$
(40, 48]	8	44	1936	352	15488
(48, 56]	22	52	2704	1144	59488
(56, 64]	29	60	3600	1740	104400
(64, 72]	21	68	4624	1428	97104
(72, 80]	20	76	5776	1520	115520
	100			**6184**	**392000**

Recordemos que $\bar{x} = \frac{6184}{100} = 61{,}84$. Calcularemos la varianza con la fórmula simplificada: $\sigma^2 = \frac{\sum_{i=1}^{k} n_i x_i^2}{N} - \bar{x}^2 = \frac{392000}{100} - 61{,}84^2 = 95{,}8144$. Por tanto la desviación típica será: $\sigma = +\sqrt{\sigma^2} = +\sqrt{95{,}8144} \cong 9{,}788$.

El coeficiente de variación es pues: $CV = \frac{\sigma}{\bar{x}} = \frac{9{,}788}{61{,}84} \cong 0{,}158$, es decir la desviación típica es el 16,2 % de la media; por tanto, la media es muy representativa de la población.

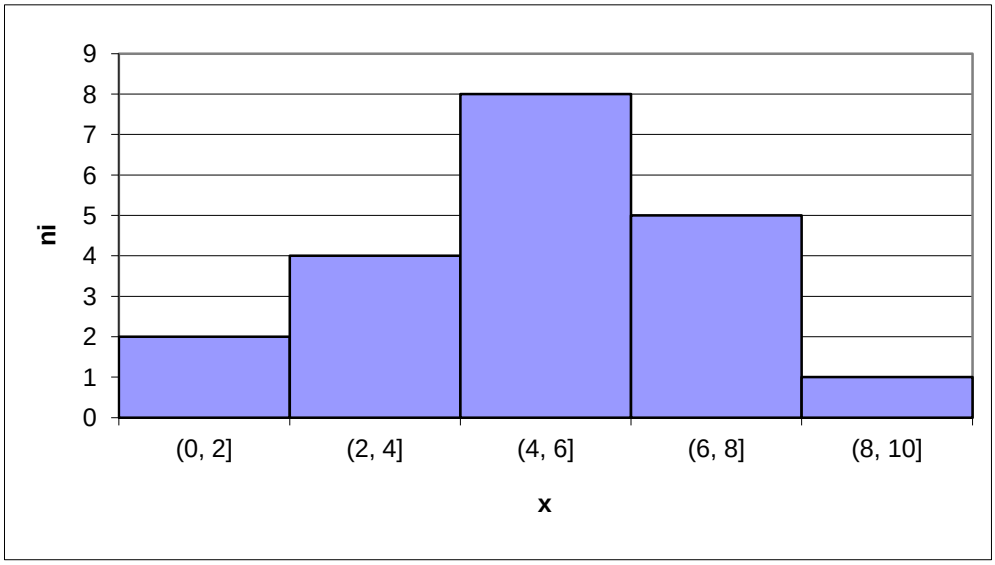
Ejercicios y problemas

1. Completar los datos que faltan en la siguiente tabla estadística, donde (como se debe saber) n_i, N_i, f_i y F_i son las frecuencias absolutas, absolutas acumuladas, relativas y relativas acumuladas.

x_i	n_i	N_i	f_i	F_i	$n_i x_i$	x_i^2	$n_i x_i^2$
1	4		0,08				
2	4						
3		16	0,16				
4	7		0,14				
5	5	28					
6		38					
7	7	45	0,14				
8							
	N =						

Calcular la moda, media, varianza y desviación típica. Calcular el coeficiente de variación e interpretarlo. Calcular la mediana y los cuartiles.

2. Las puntuaciones obtenidas por 20 personas en una prueba quedan reflejadas en el siguiente histograma de frecuencias absolutas. Calcular la moda, media, varianza y desviación típica. Calcular el coeficiente de variación e interpretarlo.



3. Las calificaciones de dos grupos de diez alumnos en la Primera Evaluación en una cierta asignatura se recogen en la siguiente tabla:

Grupo A	0	1	1	3	5	5	6	8	8	9
Grupo B	2	2	4	4	4	5	5	6	6	8

Contestar razonadamente a las siguientes cuestiones:

a) ¿Cuál de los dos grupos obtuvo mejores resultados?

b) ¿Qué grupo es más homogéneo?

4. La siguiente tabla recoge los minutos de retraso en la incorporación al trabajo de los empleados de una empresa:

Retraso en minutos	(0, 4]	(4, 8]	(8, 12]	(12, 16]	(16, 20]
Número de empleados	5	15	18	10	4

a) Representar los datos mediante un histograma.

b) Calcular el retraso medio y la desviación típica.

c) Calcular la mediana y los cuartiles y explicar qué miden estos parámetros.

5. En un estudio sobre el sueldo en euros de 50 personas se han obtenido los siguientes datos:

Sueldo	(500, 700]	(700, 900]	(900, 1300]	(1300, 1500]	(1500, 2100]
Nº de personas	10	10	20	9	1

a) Construir el histograma (nótese que las clases tienen amplitudes desiguales).

b) Calcular la media, la varianza, la desviación típica y el coeficiente de variación y explicar el significado de estos parámetros.

6. Los pesos en kilogramos de 50 personas vienen dados por la tabla:

Peso	(50, 60]	(60, 70]	(70, 80]	(80, 90]	(90, 100]
Número de empleados	10	15	20	4	1

Calcular el peso medio, los cuartiles y la desviación típica. Interpreta los resultados. ¿Se puede decir que es un grupo homogéneo?

7. La tabla de frecuencias que se da a continuación corresponde a la variable estadística X = "Posición en la liga de un cierto equipo", medida durante quince años consecutivos:

X	1º	2º	3º	4º	5º o peor
Número de veces	2	1	4	2	6

a) Indicar de qué tipo de variable se trata.

b) Representar gráficamente la distribución en diagramas rectangular y en otro de sectores.

c) Dar una medida de posición central y otra de dispersión adecuadas al experimento. Explicar por qué lo son, así como su significado.

8. Hacer un estudio estadístico completo (diagramas, medidas de centralización, de posición, de dispersión e interpretación de los resultados) del ejemplo 4.1 de la página 128 y del ejemplo 4.3 de la página 130.