Estadística descriptiva

BENITO J. GONZÁLEZ RODRÍGUEZ (bjglez@ull.es)
DOMINGO HERNÁNDEZ ABREU (dhabreu@ull.es)
MATEO M. JIMÉNEZ PAIZ (mjimenez@ull.es)
M. ISABEL MARRERO RODRÍGUEZ (imarrero@ull.es)
ALEJANDRO SANABRIA GARCÍA (asgarcia@ull.es)

Departamento de Análisis Matemático
Universidad de La Laguna

Índice

1. Introducción y primeras definiciones

En la realización de cualquier experimento cabe distinguir dos fases:

- En una primera fase de *observación* y *análisis* de sucesos, es importante recoger, ordenar y simplificar los datos.
- En una segunda fase se hace necesario *interpretar* los datos recogidos y *extraer conclusiones* a partir de ellos.

En ambas es herramienta útil la *Estadística*, ciencia que se ocupa de recopilar, analizar e interpretar los datos numéricos relativos a un conjunto de individuos (por ejemplo: altura de los alumnos de un colegio, casos de tuberculosis en una región determinada, espectadores de cierto programa de televisión...)

En la primera fase interviene la llamada *Estadística Descriptiva*, que proporciona un conjunto de técnicas y procedimientos para la recogida, clasificación y reducción de los datos a unas pocas medidas representativas. En la segunda fase desempeña un papel relevante la denominada *Estadística Inductiva o Inferencial*, que dota al investigador de un conjunto de métodos para extraer conclusiones de los datos obtenidos.

El propósito de estas notas es dar una breve introducción a la Estadística Descriptiva.

Definición 1.1. *Algunos conceptos generales que conviene tener en cuenta son los siguientes:*

- Población estadística *es el conjunto de elementos sobre el que recae las observaciones.*
- Unidad estadística *o* individuo *es cada uno de los elementos que componen una población.*
- Muestra *es un subconjunto de elementos de la población, a la que sustituye cuando el estudio de todos los individuos de la misma es difícil o costoso.*
- Tamaño *es el número de individuos de la población o muestra.*
- Carácter *es una cualidad o propiedad observable que presenta variación de unos individuos a otros. Se clasifican en* cuantitativos *o* cualitativos*, según que pueda o no asignárseles un valor numérico.*
- Modalidad *es cada uno de los valores de un carácter cualitativo.*
- Variable estadística *es un sinónimo de carácter cuantitativo. Las variables estadísticas se dividen en* continuas *o* discretas*, según que puedan o no tomar todos los valores dentro de un intervalo dado.*

Ejemplo 1.2. *El siguiente cuadro ilustra los conceptos anteriores:*

carácter	*tipo*
estado civil	*cualitativo*
color de ojos	*cualitativo*
color del pelo	*cualitativo*
peso	*cuantitativo (variable estadística continua)*
número de hijos	*cuantitativo (variable estadística discreta)*

2. Ordenación y presentación de los datos

Ahora veremos cómo estructurar los datos obtenidos en la observación de una muestra o población.

2.1. Frecuencias absoluta, relativa y acumuladas

Definición 2.1. *Sea X una variable estadística con valores $x_1 < x_2 < \ldots < x_k$, que pueden aparecer repetidos más de una vez en el conjunto de observaciones realizadas, y sea N el número de tales observaciones.*

- Recorrido *R es la diferencia entre el mayor y el menor de los valores que toma la variable:* $R = x_k - x_1$.
- Frecuencia absoluta n_i *del valor* x_i *es el número de veces que aparece repetido dicho valor en el total de observaciones.*
- Frecuencia relativa f_i *de un valor* x_i *es el cociente entre la frecuencia absoluta* n_i *correspondiente a ese valor y el número N de observaciones:* $f_i = \dfrac{n_i}{N}$.
- Frecuencia absoluta acumulada N_i *en el valor* x_i *es la suma de las frecuencias absolutas de los valores inferiores o iguales a él:* $N_i = \displaystyle\sum_{j=1}^{i} n_j$.
- Frecuencia relativa acumulada F_i *en el punto* x_i *es el cociente entre la frecuencia absoluta acumulada* N_i *en ese valor y el número N de observaciones realizadas:*

$$F_i = \frac{N_i}{N} = \frac{\sum_{j=1}^{i} n_j}{N} = \sum_{j=1}^{i} \frac{n_j}{N} = \sum_{j=1}^{i} f_j.$$

Se tienen las siguientes *propiedades de las frecuencias*:

i) $\sum_{i=1}^{k} n_i = N$.

ii) $\sum_{i=1}^{k} f_i = 1$.

iii) $N_k = N$.

iv) $F_k = 1$.

v) $0 \leq n_i \leq N$.

vi) $0 \leq f_i \leq 1$.

vii) $N_i = N_{i-1} + n_i \Rightarrow n_i = N_i - N_{i-1}$.

viii) El porcentaje correspondiente a un valor x_i de la variable se obtiene multiplicando la frecuencia relativa por 100:

$$(\%) = 100\, f_i.$$

2.2. Tabla de frecuencias

Nos ocupamos a continuación de la tabulación de los datos extraídos de una muestra o población. Con independencia del número de observaciones realizadas, pueden darse dos casos: o bien la variable estadística tiene pocos valores distintos, o, por el contrario, presenta muchos valores diferentes.

2.2.1. Variables con pocos valores distintos

En este caso la tabla de frecuencias se confecciona ordenando los valores de menor a mayor y disponiéndolos en una columna. En las demás columnas vamos anotando las correspondientes frecuencias n_i, f_i, N_i, F_i:

x_i	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	N_k	F_k

Ejemplo 2.2. *Se lanzan* 5 *monedas* 1000 *veces. La siguiente tabla recoge el número de lanzamientos en los que han salido* 0*,* 1*,* 2*,* 3*,* 4 *y* 5 *caras. Elaborar una tabla de frecuencias.*

número de caras	0	1	2	3	4	5
número de tiradas	38	144	342	287	164	25

RESOLUCIÓN.

x_i	n_i	f_i	$f_i\,(\%)$	N_i	F_i	$F_i\,(\%)$
0	38	0.038	3.8	38	0.038	3.8
1	144	0.144	14.4	182	0.182	18.2
2	342	0.342	34.2	524	0.524	52.4
3	287	0.287	28.7	811	0.811	81.1
4	164	0.164	16.4	975	0.975	97.5
5	25	0.025	2.5	1000	1.000	100.0

□

2.2.2. Variables con muchos valores distintos

En este caso se agrupan los valores de la variable en *intervalos de clase*. El punto medio de cada intervalo de clase se denomina *marca de clase* y es el valor que representa la información contenida en el intervalo. Los extremos de los intervalos de clase se llaman *límites de la clase*. Las *frecuencias absolutas de clase* se obtienen contando el número de datos que caen en el intervalo correspondiente. A partir de éstas se calculan las *frecuencias absolutas acumuladas y las relativas, acumuladas o no, de clase*. La tabla de frecuencias incorporará los intervalos de clase, las correspondientes marcas de clase, y las frecuencias (absolutas y relativas, acumuladas o no) de clase.

La elección de los intervalos de clase, tanto en número como en amplitud (constante o variable), es una cuestión subjetiva del investigador, aunque hay una serie de procedimientos que podemos tener en cuenta:

- El *número de intervalos n* suele oscilar entre 5 y 20.
- La *amplitud de cada intervalo* suele ser fija y se calcula redondeando por exceso el cociente

$$\ell = \frac{R}{n}$$

(donde R es el recorrido de la variable y n el número de intervalos que queremos formar) al mismo número de cifras decimales que los datos.

- Los intervalos suelen elegirse *semiabiertos por la derecha*, esto es, de la forma $[a,b)$, de tal manera que se solapen en los extremos, tomándose tantos cuantos sean necesarios para cubrir todo el recorrido de la variable. Nótese que $a \in [a,b)$, pero $b \notin [a,b)$.

En ocasiones los datos se presentan agrupados en intervalos no solapados. En tal caso es aconsejable (principalmente a efectos de su representación gráfica) reemplazarlos por otros del tipo anterior, *cuidando de no modificar las frecuencias*; esto se logra sustituyendo los extremos de los intervalos originales por los puntos medios de los extremos derecho e izquierdo de cada dos intervalos contiguos. Los nuevos extremos reciben el nombre de *límites reales de clase*.

Ejemplo 2.3. *Convertir los siguientes intervalos de clase en intervalos con límites reales de clase. Hallar las marcas de clase. Elaborar la tabla de frecuencias.*

intervalos	$130-139$	$140-149$	$150-159$	$160-169$
n_i	25	32	15	17

RESOLUCIÓN.

intervalos de clase	marcas de clase x_i	n_i	f_i	$f_i\,(\%)$	N_i	F_i	$F_i\,(\%)$
$[129.5, 139.5)$	134.5	25	0.28	28	25	0.28	28
$[139.5, 149.5)$	144.5	32	0.36	36	57	0.64	64
$[149.5, 159.5)$	154.5	15	0.17	17	72	0.81	81
$[159.5, 169.5)$	164.5	17	0.19	19	89	1.00	100

□

3. Representaciones gráficas

A continuación describiremos algunas representaciones gráficas que ayudan a visualizar la información recogida.

3.1. Caracteres cuantitativos

3.1.1. Diagramas de barras

Los diagramas de barras pueden trazarse verticales u horizontales, sin más que intercambiar el papel de los ejes.

- Para *variables no agrupadas en intervalos de clase*, se considera un sistema de ejes cartesianos, se colocan en abscisas los distintos valores de la variable, y sobre cada uno de ellos se levanta una línea perpendicular cuya altura es la frecuencia (absoluta o relativa) de dicho valor (Figura 3.1).

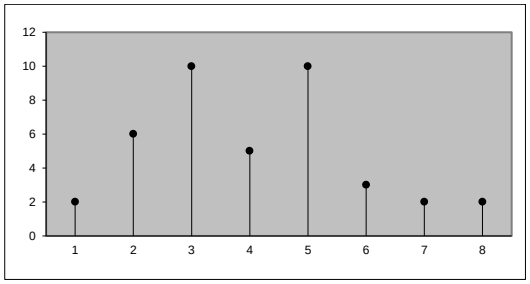

Figura 3.1. Diagrama de barras verticales para frecuencias absolutas (variables no agrupadas).

- Para una *variable agrupada en intervalos de clase*, sobre el eje de abscisas de un sistema de referencia cartesiano se disponen los intervalos de clase uno a continuación del otro, y sobre cada intervalo se construye un rectángulo de área proporcional a la frecuencia absoluta o relativa correspondiente. Este tipo de diagrama se denomina *histograma* (Figura 3.2).

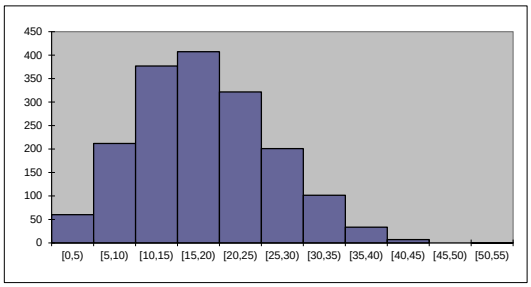

Figura 3.2. Histograma (variables agrupadas).

3.1.2. Diagramas de trazos

Representan *frecuencias no acumuladas*.

- Si la *variable no está agrupada en intervalos* el diagrama de trazos se obtiene uniendo mediante una poligonal los extremos superiores de las barras en el diagrama de barras (Figura 3.3).

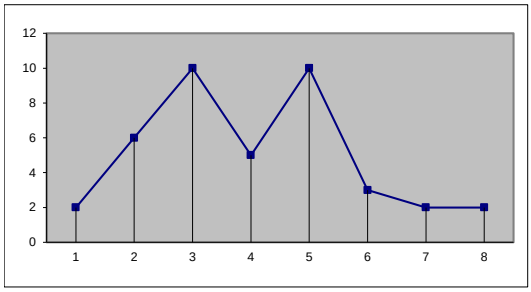

Figura 3.3. Diagrama de trazos para frecuencias no acumuladas (variables no agrupadas).

- Si la *variable está agrupada en clases*, el diagrama de trazos se forma uniendo los puntos medios de las bases superiores de los rectángulos que conforman el histograma (Figura 3.4).

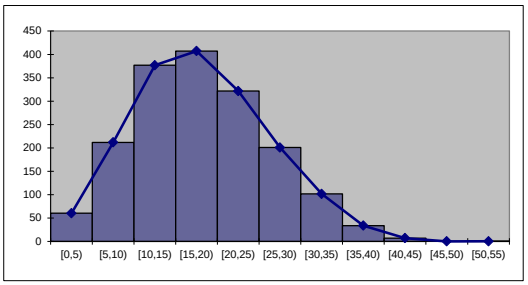

Figura 3.4. Diagrama de trazos para frecuencias no acumuladas (variables agrupadas).

3.1.3. Diagramas de barras acumulativos y polígonos de frecuencias acumuladas

- Los *diagramas de frecuencias acumuladas* o *diagramas de barras acumulativos* para *variables no agrupadas* son diagramas de barras construidos a partir de las frecuencias (absolutas o relativas) acumuladas, en los que se trazan segmentos horizontales desde el extremo superior de cada barra hasta la barra situada inmediatamente a la derecha (Figura 3.5).

- Los *polígonos de frecuencias acumuladas* para una *variable agrupada* se obtienen representando en abscisas los distintos intervalos de clase, levantando sobre el extremo derecho de cada intervalo una línea vertical de longitud equivalente a la frecuencia (absoluta o relativa) acumulada del mismo y uniendo los extremos superiores del diagrama de barras creciente que resulta (Figura 3.6).

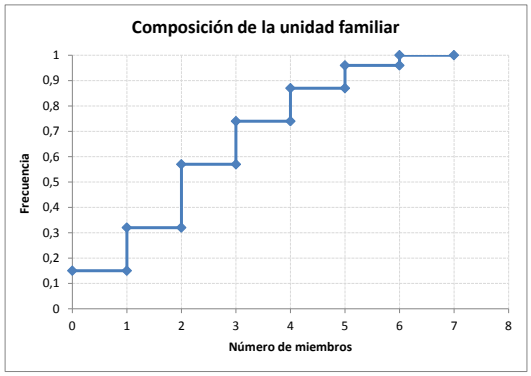

Figura 3.5. Diagrama de barras acumulativo para frecuencias relativas (variables no agrupadas).

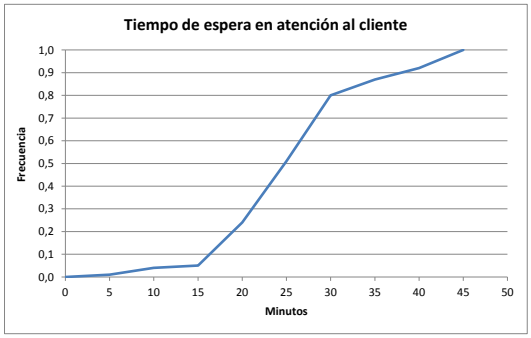

Figura 3.6. Polígono de frecuencias relativas acumuladas (variables agrupadas).

3.2. Caracteres cualitativos

3.2.1. Diagrama de rectángulos

Es similar a los diagramas de barras de variables no agrupadas para frecuencias no acumuladas (Figura 3.7).

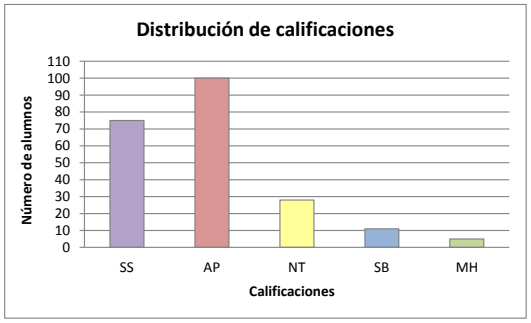

Figura 3.7. Diagrama de rectángulos.

3.2.2. Diagrama de sectores

En un círculo se asigna un sector circular a cada una de las modalidades de un carácter, siendo la amplitud del sector proporcional a la frecuencia de la modalidad (Figura 3.8).

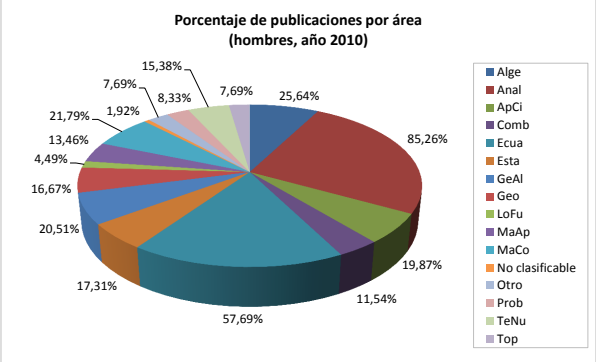

Figura 3.8. Diagrama de sectores.

3.2.3. Perfiles ortogonales y radiales

Se construyen por un procedimiento similar a los diagramas de trazos de variables no agrupadas para frecuencias no acumuladas. Las barras pueden ser ortogonales a los ejes de un sistema de referencia cartesiano (*perfil ortogonal*, Figura 3.9) o presentar una disposición radial (*perfil radial*, Figura 3.10).

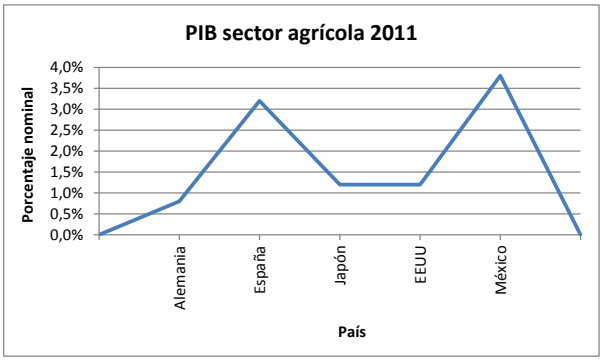

Figura 3.9. Perfil ortogonal.

3.2.4. Pictograma

Cada modalidad se representa con un dibujo alusivo de tamaño proporcional a la frecuencia de la misma, o que se repite un número de veces proporcional a dicha frecuencia (Figura 3.11).

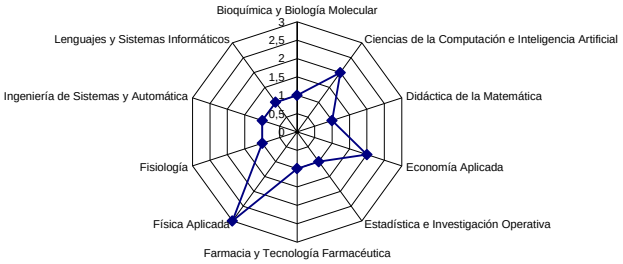

Figura 3.10. Perfil radial.

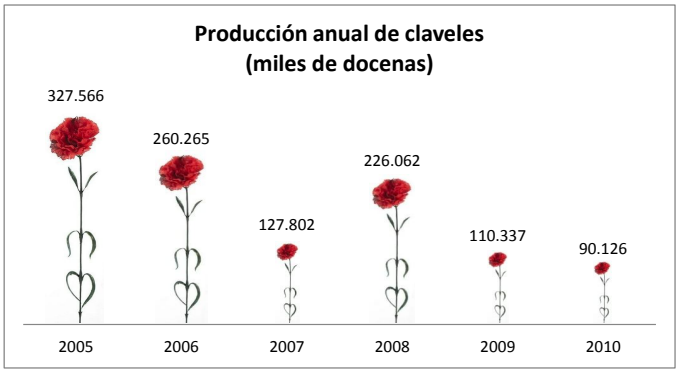

Figura 3.11. Pictograma.

3.2.5. Cartograma

Es la representación sobre un mapa del carácter estudiado. Las distintas modalidades se visualizan mediante sombreados de distinta intensidad (Figura 3.12).

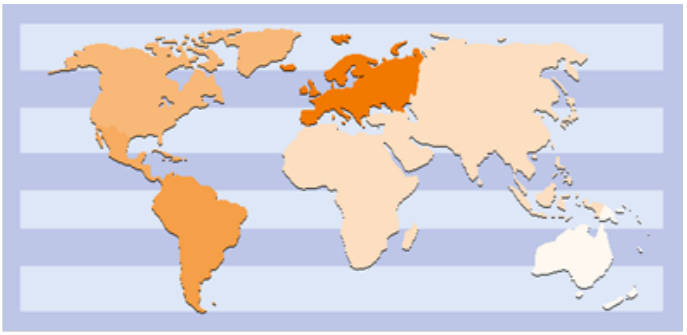

Figura 3.12. Cartograma.

4. Medidas descriptivas

Distinguimos dos tipos de *medidas descriptivas*: las de *centralización* y las de *dispersión*.

4.1. Medidas de centralización

Es conveniente reducir la información recopilada a unos pocos valores, con el fin de poder comparar muestras o poblaciones. Estos valores que centralizan la información se llaman *medidas de centralización* o *de tendencia central*. Entre ellos se encuentran la *media*, la *mediana*, la *moda* y los *cuantiles* (*cuartiles*, *deciles* y *centiles* o *percentiles*). A continuación veremos cómo se definen y computan.

4.1.1. Media aritmética

Definición 4.1. *La* media aritmética *se denota* $\bar{x}$ *y se calcula mediante la fórmula*

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \sum_{i=1}^{k} x_i f_i,$$

donde $x_1, x_2, \ldots, x_k$ *son las marcas de clase o los valores de la variable, según que ésta esté o no agrupada en intervalos, y* $n_1, n_2, \ldots, n_k$ *son las frecuencias absolutas correspondientes.*

Ejemplo 4.2. *Hallar la media aritmética en los Ejemplos 2.2 y 2.3.*

RESOLUCIÓN. En el Ejemplo 2.2:

$$\bar{x} = \frac{1}{1000}[(0 \cdot 38) + (1 \cdot 144) + (2 \cdot 342) + (3 \cdot 287) + (4 \cdot 164) + (5 \cdot 25)] = 2.47 \simeq 2.5.$$

En el Ejemplo 2.3:

$$\bar{x} = \frac{1}{89}[(134.5 \cdot 25) + (144.5 \cdot 32) + (154.5 \cdot 15) + (164.5 \cdot 17)] \simeq 147.2.$$

□

4.1.2. Mediana

Definición 4.3. *La* mediana M_e *es el valor de la variable que ocupa el punto central cuando los datos están ordenados de forma creciente; es decir, es el valor de la variable que deja por detrás la mitad de las observaciones, y por delante la otra mitad.*

Veamos cómo se calcula la mediana cuando conocemos una distribución de frecuencias.

4.1.2.1. Cálculo de la mediana: variables no agrupadas

1. Se divide el número de observaciones N entre 2.
2. Se comprueba si el número $N/2$ obtenido figura en la columna de frecuencias absolutas acumuladas N_i.
   - Si no está, estará comprendido entre dos números (N_i, N_{i+1}) de la citada columna, con lo cual la mediana será aquel valor de la variable que corresponde al mayor de ambos números:

$$M_e = x_{i+1}.$$

   - Si el valor $N/2$ está en la columna de las N_i es que coincide con la frecuencia absoluta acumulada de algún valor x_i. En tal caso la mediana es el punto medio del intervalo (x_i, x_{i+1}):

$$M_e = \frac{x_i + x_{i+1}}{2}.$$

La Figura 4.1 ilustra el cálculo de la mediana para variables no agrupadas en intervalos de clase usando el diagrama de barras acumulativo de frecuencias absolutas. A la izquierda y a la derecha de la gráfica se representan situaciones en las que $M_e = x_{i+1}$ y $M_e = (x_i + x_{i+1})/2$, respectivamente.

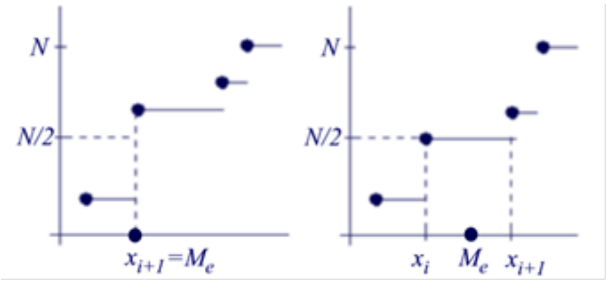

Figura 4.1. Cálculo de la mediana para variables no agrupadas.

Un procedimiento alternativo es el siguiente. Si en la tabla de frecuencias se han calculado las frecuencias relativas acumuladas en porcentaje, se comprueba si entre ellas figura el 50 %. En caso negativo, la mediana es el valor de la variable al que corresponde el porcentaje inmediatamente superior al 50 %. En caso afirmativo, la mediana se obtiene promediando el valor de la variable que acumula un porcentaje del 50 % con el valor de la variable inmediatamente posterior.

Ejemplo 4.4. *Calcular la mediana si los resultados de un experimento son:*

a) $\{1,3,7,9,10\}$.

b) $\{1,3,7,9,9,10\}$.

RESOLUCIÓN.

a) La tabla de frecuencias en este caso es:

x_i	n_i	N_i	f_i	F_i
1	1	1	20 %	20 %
3	1	2	20 %	40 %
7	1	3	20 %	60 %
9	1	4	20 %	80 %
10	1	5	20 %	100 %

- *Forma 1*: *Utilizando la columna de frecuencias absolutas acumuladas.* La mitad de las observaciones es $N/2 = 5/2 = 2.5$. Esta cifra no figura en la columna de las N_i; la inmediatamente superior que aparece es 3; el x_i que acumula una N_i de 3 es 7; por tanto, $M_e = 7$.
- *Forma 2*: *Utilizando la columna de frecuencias relativas acumuladas.* El 50 % no figura en la columna de las F_i, por lo que buscamos el porcentaje inmediatamente superior que aparece. Este porcentaje es el 60 %, y el x_i que acumula una F_i del 60 % es 7. Por tanto, $M_e = 7$.

b) Ahora la tabla de frecuencias es:

x_i	n_i	N_i	f_i	F_i
1	1	1	16.7 %	16.7 %
3	1	2	16.7 %	33.3 %
7	1	3	16.7 %	50.0 %
9	2	5	33.3 %	83.3 %
10	1	6	16.7 %	100.0 %

- *Forma 1*: *Utilizando la columna de frecuencias absolutas acumuladas.* La mitad de las observaciones es $N/2 = 6/2 = 3$, cifra que figura en la columna de las N_i. El valor de la variable que acumula una N_i de 3 es 7, y el inmediatamente superior, 9; por tanto, $M_e = (7+9)/2 = 8$.
- *Forma 2*: *Utilizando la columna de frecuencias relativas acumuladas.* El 50 % figura en la columna de las F_i, y el x_i que acumula una F_i del 50 % es 7. Por tanto, $M_e = (7+9)/2 = 8$.

□

4.1.2.2. Cálculo de la mediana: variables agrupadas

1. Se divide el número de observaciones N entre 2.
2. Se traslada el valor $N/2$ a la columna de frecuencias absolutas acumuladas N_i.
   - Si el valor está en la tabla, es la frecuencia absoluta acumulada de cierto intervalo de clase $[a_i, a_{i+1})$, y por tanto la mediana es el extremo superior del mismo: $M_e = a_{i+1}$.
   - Si $N/2$ está comprendido entre N_i y N_{i+1}, que corresponden a las frecuencias absolutas acumuladas de dos intervalos $[a_{i-1}, a_i)$ y $[a_i, a_{i+1})$, respectivamente, la mediana se encuentra en el intervalo $[a_i, a_{i+1})$ y su valor se calcula mediante interpolación lineal:

$$\frac{N/2 - N_i}{N_{i+1} - N_i} = \frac{x}{a_{i+1} - a_i} \quad \Rightarrow \quad M_e = a_i + x.$$

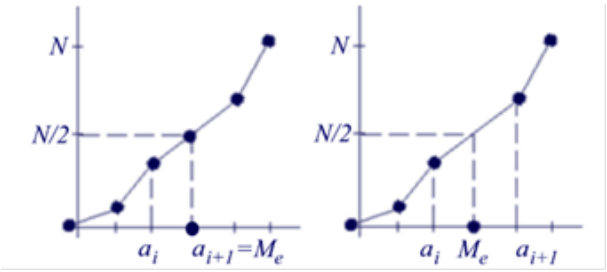

Figura 4.2. Cálculo de la mediana para variables agrupadas.

La Figura 4.2 ilustra el cálculo de la mediana para variables agrupadas en intervalos de clase usando el polígono de frecuencias absolutas acumuladas. A la izquierda y a la derecha de la gráfica se representan situaciones en las que

$$M_e = a_{i+1} \quad \text{y} \quad M_e = a_i + \frac{(a_{i+1} - a_i)(N/2 - N_i)}{N_{i+1} - N_i},$$

respectivamente.

Ejemplo 4.5. *Dadas las siguientes tablas de frecuencias, calcular la mediana.*

a)

intervalos de clase	n_i	N_i
$[7.5, 9.0)$	3	3
$[9.0, 10.5)$	8	11
$[10.5, 12.0)$	6	17
$[12.0, 13.5)$	14	31
$[13.5, 15.0)$	1	32
$[15.0, 16.5)$	2	34

b)

intervalos de clase	n_i	N_i
$[7.5, 9.0)$	3	3
$[9.0, 10.5)$	8	11
$[10.5, 12.0)$	10	21
$[12.0, 13.5)$	10	31
$[13.5, 15.0)$	1	32
$[15.0, 16.5)$	2	34

RESOLUCIÓN.

a) En este caso $N/2 = 34/2 = 17$ es la frecuencia absoluta que acumula el intervalo $[10.5, 12.0)$, por lo que $M_e = 12.0$.

b) Al ser $N/2 = 34/2 = 17$ y $11 < 17 < 21$ la mediana estará en el intervalo $[10.5, 12.0)$, y aplicando la fórmula anterior se tendrá

$$M_e = 10.5 + \frac{(12.0 - 10.5)(17 - 11)}{21 - 11} = 11.4.$$

□

4.1.3. Moda

Definición 4.6. *La* moda*, que denotaremos M_o, es el valor (no necesariamente único) de la variable que tiene la máxima frecuencia. Si hay dos modas, la distribución se llama* bimodal*; si tres,* trimodal*; etc.*

Cuando la variable viene agrupada en intervalos de clase hablaremos de intervalo modal*, que es el intervalo o intervalos con mayor frecuencia absoluta de clase.*

El intervalo modal se reconoce fácilmente en el histograma por ser aquél al que corresponde el rectángulo de mayor área por unidad de base. Razones geométricas justifican la siguiente fórmula para calcular la posición de la moda en el caso de intervalos de igual amplitud:

$$M_o = a_i + c\,\frac{n_i - n_{i-1}}{2n_i - n_{i+1} - n_{i-1}},$$

donde $c = a_{i+1} - a_i$ representa la longitud del intervalo modal $[a_i, a_{i+1})$, al que le corresponde la máxima frecuencia absoluta n_i. En la práctica omitiremos frecuentemente este cálculo, contentándonos con indicar el intervalo o intervalos modales; no obstante, a continuación lo ilustraremos por completitud.

Ejemplo 4.7. *Dadas las siguientes tablas de frecuencias, calcular la moda.*

a)

x_i	n_i	N_i
1	3	3
3	8	11
4	6	17
7	14	31
9	1	32
10	2	34

b)

intervalos de clase	n_i	N_i
$[7.5, 9.0)$	3	3
$[9.0, 10.5)$	8	11
$[10.5, 12.0)$	9	20
$[12.0, 13.5)$	2	22
$[13.5, 15.0)$	9	31
$[15.0, 16.5)$	3	34

c)

intervalos de clase	n_i	N_i
$[7.5, 9.0)$	3	3
$[9.0, 10.5)$	8	11
$[10.5, 12.0)$	10	21
$[12.0, 13.5)$	10	31
$[13.5, 15.0)$	1	32
$[15.0, 16.5)$	2	34

RESOLUCIÓN.

a) En este caso $M_o = 7$, ya que 7 es el valor de la variable que presenta la mayor frecuencia absoluta, a saber, 14.

b) Ahora tenemos dos intervalos modales con frecuencia absoluta máxima de 9, que son $[10.5, 12.0)$ y $[13.5, 15.0)$. Por tanto la distribución es bimodal, siendo las modas:

$$M_{o_1} = 10.5 + 1.5\,\frac{9-8}{18-2-8} = 10.7 \quad \text{y} \quad M_{o_2} = 13.5 + 1.5\,\frac{9-2}{18-3-2} = 14.3.$$

c) Aquí se presenta un caso de distribución bimodal, ya que tanto el intervalo $[10.5, 12.0)$ como el $[12.0, 13.5)$ tienen frecuencia absoluta máxima de 10; deberíamos aplicar, por tanto, para cada uno de los dos intervalos la fórmula anterior, determinando así las dos modas de la distribución. No obstante, aparece una peculiaridad adicional, y es que ambos intervalos modales son contiguos. En esta situación se considera que la distribución es unimodal, y se elige como moda el extremo común: $M_o = 12.0$.

□

4.1.4. Cuantiles

Reciben esta denominación genérica los *cuartiles*, *deciles* y *centiles* (o *percentiles*), medidas de centralización definidas como se indica seguidamente.

Definición 4.8. *Los* cuartiles *son los tres valores de la variable que dividen las observaciones en cuatro partes iguales. Más precisamente:*

- Primer cuartil $P_{1/4}$ *es el valor de la variable que deja la cuarta parte de las observaciones menores o iguales que él y las tres cuartas partes superiores a él. Se calcula como la mediana, pero en lugar de* $N/2$*, se toma* $N/4$.
- Segundo cuartil $P_{2/4}$ *es el valor de la variable que deja inferiores o iguales a él las dos cuartas partes de las observaciones; coincide, por tanto, con la mediana.*
- Tercer cuartil $P_{3/4}$ *es el valor de la variable que deja inferiores o iguales a él las tres cuartas partes de las observaciones y la cuarta parte de éstas superiores a él. Se calcula igual que la mediana, tomando* $3N/4$.

Ejemplo 4.9. *Supuesta la siguiente distribución de frecuencias para una variable discreta, calcular los cuartiles.*

x_i	n_i	N_i
1	3	3
5	2	5
6	5	10
8	3	13
10	3	16
11	8	24
12	11	35
14	7	42

RESOLUCIÓN. Se tiene:

- Primer cuartil: $N/4 = 42/4 = 10.5 \to 10 < 10.5 < 13 \Rightarrow P_{1/4} = 8$.
- Segundo cuartil: $N/2 = 42/2 = 21 \to 16 < 21 < 24 \Rightarrow P_{2/4} = 11 = M_e$.
- Tercer cuartil: $3N/4 = 3 \cdot 42/4 = 31.5 \to 24 < 31.5 < 35 \Rightarrow P_{3/4} = 12$.

□

Definición 4.10. *Se define el* decil *k*-ésimo D_k *como el valor de la variable que deja inferiores o iguales a él a las* $k/10$ *partes de las observaciones, con* $k = 1,2,3,\ldots,9$. *Para calcular* D_k *se emplean las mismas técnicas anteriores, partiendo de* $kN/10$.

Definición 4.11. *El* centil *(o* percentil*)* P_k *es el valor de la variable que deja inferior o iguales a él las* $k/100$ *partes de las observaciones, con* $k = 1,2,\ldots,99$. *El cálculo de* P_k *es análogo al de los anteriores cuantiles, a partir de* $kN/100$.

En la sección de ejercicios resueltos se ilustra el cómputo de los deciles y centiles.

4.2. Medidas de dispersión (o de concentración)

Las medidas de tendencia central reducen la información de una muestra a unos pocos valores, pero en algunos casos estos valores estarán más próximos a la realidad de las observaciones que en otros. La *varianza* y la *desviación típica* son dos medidas descriptivas que cuantifican la representatividad de la media aritmética.

4.2.1. Varianza

Definición 4.12. *La* varianza*, que denotaremos* σ^2*, es la media aritmética de los cuadrados de las desviaciones de los datos con respecto a la media aritmética (o bien, la media aritmética de los cuadrados menos el cuadrado de la media aritmética):*

$$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i-\bar{x})^2 n_i}{N} = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} - \bar{x}^2.$$

4.2.2. Desviación típica

Definición 4.13. *La* desviación típica *es la raíz cuadrada de la varianza:*

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2 n_i}{N}} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 n_i}{N} - \bar{x}^2}.$$

4.3. Teorema de Chebyshev

Este resultado debe su nombre al matemático ruso Pafnuti Lvóvich Chebyshev (1821-1894).

Teorema 4.14 (Chebyshev)**.** *En el intervalo* $(\bar{x}-k\sigma,\bar{x}+k\sigma)$ *está como mínimo el* $\left(1-\frac{1}{k^2}\right)\cdot 100\,\%$ *de las observaciones, y fuera de este intervalo está como máximo el* $\frac{1}{k^2}\cdot 100\,\%$ *de ellas.*

La sección de ejercicios resueltos contiene ejemplos sobre el cálculo de la varianza y la desviación típica, así como aplicaciones del Teorema de Chebyshev.