Universidad Francisco de Vitoria
UFV Madrid

Estadística descriptiva

TEMA 1

(PARTE I)

2º CURSO DE INGENERÍA INFORMÁTICA
ESCUELA POLITÉCNICA SUPERIOR

Objeto de la estadística descriptiva

La aplicación del tratamiento estadístico tiene dos fases fundamentales:

- Organización y análisis inicial de los datos recogidos.
- Extracción de conclusiones válidas y toma de decisiones razonables a partir de ellos.

- Los objetivos de la Estadística Descriptiva son los que se abordan en la primera de estas fases. Es decir, **su misión es ordenar, describir y sintetizar la información recogida.**
- En este proceso será **necesario establecer medidas cuantitativas** que reduzcan a un número manejable de parámetros el conjunto (en general grande) de datos obtenidos.
- La **realización de gráficas (visualización de los datos en diagramas)** también forma parte de la Estadística Descriptiva dado que proporciona una manera visual directa de organizar la información
- La finalidad de la Estadística Descriptiva no es, entonces, extraer conclusiones generales sobre el fenómeno que ha producido los datos bajo estudio, sino solamente su descripción

Variable estadística

- Variable: Es una representación numérica de una característica u hecho sometido a medición.

- Variable aleatoria o estadística: Se entiende por variable estadística al símbolo que representa al dato o carácter objeto de nuestro estudio de los elementos de la muestra y que puede tomar un conjunto de valores. (al menos dos)

- En el caso de que estemos tratando con caracteres cuantitativos, la variables estadísticas pueden clasificarse en:

  - **Discretas**, cuando solo pueden tomar una cantidad (finita o infinita) numerable de valores.
  - **Continuas**, cuando pueden tomar teóricamente infinitos valores entre dos valores dados. Es la diferencia básica que existe entre contar y medir.

El número de electrones de un átomo es una variable discreta. La velocidad o la altura de un móvil son variables continuas.

Variable estadística

- Variable: Es una representación numérica de una característica u hecho sometido a medición.

- Variable aleatoria o estadística: Se entiende por variable estadística al símbolo que representa al dato o carácter objeto de nuestro estudio de los elementos de la muestra y que puede tomar un conjunto de valores. (al menos dos)

- En el caso de que estemos tratando con caracteres cuantitativos, la variables estadísticas pueden clasificarse en:

  - **Discretas**, cuando solo pueden tomar una cantidad (finita o infinita) numerable de valores.
  - **Continuas**, cuando pueden tomar teóricamente infinitos valores entre dos valores dados. Es la diferencia básica que existe entre contar y medir.

El número de electrones de un átomo es una variable discreta. La velocidad o la altura de un móvil son variables continuas.

Variable estadística

Por otra parte, las variables se pueden asimismo clasificar en **unidimensionales**, cuando solo se mida un carácter o dato de los elementos de la muestra, o **bidimensionales, tridimensionales**, y en general n–dimensionales, cuando se estudien simultáneamente varios caracteres de cada elemento.

La temperatura o la presión atmosférica (por separado), son variables monodimensionales. La temperatura y la presión atmosférica (estudiadas conjuntamente), o la longitud y el peso de una barra conductora, son ejemplos de variables bidimensionales. La velocidad, carga eléctrica y masa de un ión es tridimensional.

Por otra parte, las variables también se pueden clasificar en **categóricas** (por ejemplo, sexo), en este caso los posibles valores distintos serán pocos (hombre, mujer) y cada uno de ellos se repetirá varias veces (pues los resultados serán hombre o mujer).
Por el contrario si la variable es **cuantitativa** (por ejemplo altura) habrá muy pocas repeticiones o ninguna especialmente si se hace con la adecuada dispersión)

Distribuciones de frecuencias

El **primer paso** para el estudio estadístico de una muestra es **su ordenación y presentación en una tabla de frecuencias**.

- Supongamos que tenemos una muestra de tamaño *N*, donde la variable estadística *x* toma valores distintos $x_1, x_2, \ldots, x_k$

- En primer lugar hay que ordenar los diferentes valores que toma la variable estadística en orden (normalmente creciente). La diferencia entre el **valor mayor y menor** que toma la variable se conoce como **recorrido, o rango**

Valores de la variable estadística x_i	Frecuencias absolutas n_i	Frecuencias relativas f_i	Frecuencias absolutas acumuladas N_i	Frecuencias relativas acumuladas 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

En la primera columna de esta tabla se escriben los distintos valores de la variable, ordenados de mayor a menor.
En una tabla de frecuencias de una variable cualitativa o categórica se escribirán en la primera columna las diferentes cualidades o atributos que puede tomar la variable.

Frecuencia absoluta y relativa

- **Frecuencia absoluta *ni***: Definida como el número de veces que aparece repetido el valor en cuestión de la variable estadística en el conjunto de las observaciones realizadas. Si *N* es el número de observaciones(o tamaño de la muestra), las frecuencias absolutas cumplen las propiedades

$$0 \leq n_i \leq N \quad ; \quad \sum_{i=1}^{k} n_i = N.$$

La frecuencia absoluta, aunque nos dice el número de veces que se repite un dato, no nos informa de la importancia de éste.

- **Frecuencia relativa *fi***: Cociente entre la frecuencia absoluta y el número de observaciones realizadas N. Es decir

$$f_i = \frac{n_i}{N},$$

Cumpliéndose las siguientes propiedades

$$0 \leq f_i \leq 1 \quad ; \quad \sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{n_i}{N} = \frac{\sum_{i=1}^{k} n_i}{N} = 1.$$

Esta frecuencia relativa se puede expresar también en tanto por ciento del tamaño de la muestra, para lo cual basta con multiplicar por 100:

$$(\%)_{x_i} = 100 \mathrm{x} f_i.$$

Frecuencia acumulada absoluta y relativa

- **Frecuencia absoluta acumulada *Ni***: Frecuencia absoluta acumulada Ni: Suma de las frecuencias absolutas de los valores inferiores o igual a *xi*, o número de medidas por debajo, o igual, que *xi*. Evidentemente la frecuencia absoluta acumulada de un valor se puede calcular a partir de la correspondiente al anterior como

$$N_i = N_{i-1} + n_i \qquad \text{y} \qquad N_1 = n_1.$$

Además la frecuencia absoluta acumulada del ultimo valor será: $N_k = N.$

- **Frecuencia relativa acumulada *Fi*:** Cociente entre la frecuencia absoluta acumulada y el número de observaciones. Coincide además con la suma de las frecuencias relativas de los valores inferiores o iguales a *xi*

$$F_i = \frac{N_i}{N} = \frac{\sum_{j=1}^{i} n_j}{N} = \sum_{j=1}^{i} \frac{n_j}{N} = \sum_{j=1}^{i} f_i,$$

y la frecuencia relativa acumulada del último valor es 1

$$F_k = 1.$$

Se puede expresar asimismo como un porcentaje (multiplicando por 100) y su significado será el tanto por ciento de medidas con valores por debajo o igual que xi.

Frecuencia acumulada absoluta y relativa

Supongamos que el número de hijos de una muestra de 20 familias es el siguiente:

2	1	1	3	1	2	5	1	2	3
4	2	3	2	1	4	2	3	2	1

El tamaño de la muestra es $N = 20$, el número de valores posibles $k = 5$, y el recorrido es $5 - 1 = 4$.

x_i	n_i	f_i $n_i/20$	N_i $\sum_1^i n_j$	F_i $\sum_1^i f_j$
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00

Agrupamiento en intervalos de clase

- Cuando el número de valores distintos que toma la variable estadística es demasiado grande o la variable es continua no es útil elaborar una tabla de frecuencias como la vista anteriormente. En estos casos se realiza un **agrupamiento de los datos en intervalos y se hace un recuento del número de observaciones que caen dentro de cada uno de ellos.**

- Dichos intervalos se denominan **intervalos de clase**, y al valor de la variable en el centro de cada intervalo se le llama **marca de clase**. De esta forma se sustituye cada medida por la marca de clase del intervalo a que corresponda.

- A la diferencia entre el extremo superior e inferior de cada intervalo se le llama amplitud del intervalo. Normalmente se trabajara con intervalos de **amplitud constante**. La tabla de frecuencias resultante es similar a la vista anteriormente.

Agrupamiento en intervalos de clase

En el caso de una distribución en k intervalos esta sería:

Intervalos de clase	Marcas de clase	Frecuencias absolutas	Frecuencias relativas	Frecuencias absolutas acumuladas	Frecuencias relativas acumuladas
$a_i - a_{i+1}$	c_i	n_i	$f_i = n_i/N$	N_i	$F_i = N_i/N$
$a_1 - a_2$	c_1	n_1	f_1	N_1	F_1
$a_2 - a_3$	c_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_k - a_{k+1}$	c_k	n_k	f_k	N_k	F_k

Agrupamiento en intervalos de clase

El realizar el estudio mediante el agrupamiento en intervalos de clase simplifica el trabajo, pero también supone una pérdida de información, ya que no se tiene en cuenta cómo se distribuyen los datos dentro de cada intervalo. Para que dicha pérdida sea mínima es necesario elegir con cuidado los intervalos. Aunque no existen ningunas reglas estrictas para la elección de los intervalos, los pasos a seguir son:

1. Determinar el **recorrido, o rango, de los datos**. Esto es, la diferencia entre el mayor y el menor de los valores que toma la variable.
2. **Decidir el número k de intervalos** de clase en que se van a agrupar los datos. Dicho número se debe situar normalmente entre 5 y 20, dependiendo del caso. En general el número será más grande cuanto más datos tenga la muestra. Una regla que a veces se sigue es elegir k como el entero más próximo a raíz de N.
3. **Dividir el recorrido entre el número de intervalos para determinar la amplitud (constante) de cada intervalo**. Dicha amplitud no es necesario que sea exactamente el resultado de esa división sino que normalmente se puede redondear hacia un número algo mayor.
4. **Determinar los extremos de los intervalos de clase**. Evidentemente el extremo superior de cada intervalo ha de coincidir con el extremo inferior del siguiente. Es importante que ninguna observación coincida con alguno de los extremos, para evitar así una ambigüedad en la clasificación de este dato. Una forma de conseguir esto es asignar a los extremos de los intervalos una cifra decimal más que las medidas de la muestra. Por ejemplo, si la variable estadística toma valores enteros: 10, 11, 12, . . ., los intervalos se podrían elegir: 9.5 − 11.5, 11.5 − 13.5, ...
5. **Calcular las marcas de clase** de cada intervalo como el valor medio entre los limites inferior y superior de cada intervalo de clase. Otra consideración a tomar en cuenta a la hora de elegir los intervalos es intentar que las marcas de clase coincidan con medidas de la muestra, disminuyéndose así la pérdida de información debida al agrupamiento.
6. Una vez determinados los intervalos se debe hacer un **recuento cuidadoso del número de observaciones** que caen dentro de cada intervalo, para construir así la tabla de frecuencias.

En la tabla siguiente se listan los datos medidos por James Short en 1763 sobre la paralaje del Sol en segundos de arco. La paralaje es el ángulo subtendido por la Tierra vista desde el Sol. Se midió observando tránsitos de Venus desde diferentes posiciones y permitió la primera medida de la distancia Tierra-Sol, que es la unidad básica de la escala de distancias en el Sistema Solar (la unidad astronómica).

Datos (en segundos de arco):

8.63	10.16	8.50	8.31	10.80	7.50	8.12
8.42	9.20	8.16	8.36	9.77	7.52	7.96
7.83	8.62	7.54	8.28	9.32	7.96	7.47

1. Recorrido: máximo−mínimo= $10.80 - 7.47 = 3.33$.
2. Número de intervalos: $k = \sqrt{21} = 4.53 \Rightarrow k = 5$. Como se redondea por exceso, la amplitud del intervalo multiplicada por el número de intervalos será mayor que el recorrido y no tendremos problemas en los extremos.
3. Amplitud del intervalo: $3.33/5 = 0.666 \Rightarrow 0.7$.
4. Extremos de los intervalos. Para evitar coincidencias se toma un decimal más. El primer extremo se toma algo menor que el valor mínimo, pero calculándolo de forma que el último extremo sea algo mayor que el valor máximo.

Si tomamos $a_1 = 7.405$ se verifica que es < 7.47 (mínimo), y el último extremo será $7.405 + 5 \times 0.7 = 10.905$ que resulta ser > 10.80 (máximo). Ahora ya podemos calcular los extremos para cada intervalo de clase y las marcas de clase correspondientes.

5. Recuento y construcción de la tabla.

a_i—a_{i+1}	c_i	n_i	f_i	N_i	F_i
7.405 — 8.105	7.755	7	0.333	7	0.333
8.105 — 8.805	8.455	9	0.429	16	0.762
8.805 — 9.505	9.155	2	0.095	18	0.857
9.505 — 10.205	9.855	2	0.095	20	0.952
10.205— 10.905	10.555	1	0.048	21	1.000
Suma		21	1.000		

Representaciones gráficas

- Representaciones gráficas para datos sin agrupar

El diagrama principal para representar datos de variables discretas sin agrupar es el **diagrama de barras**. En este se representan en el eje de abscisas los distintos valores de la variable y sobre cada uno de ellos se levanta una barra de longitud igual a la frecuencia correspondiente. Pueden representarse tanto las frecuencias absolutas *ni* como las relativas *fi*. En la practica se puede graduar simultáneamente el eje de ordenadas tanto en frecuencias absolutas como en relativas en tantos por ciento.

Un diagrama similar es el **polígono de frecuencias**. Este se obtiene uniendo con rectas los extremos superiores de las barras del diagrama anterior. De la misma forma, pueden representarse frecuencias absolutas o relativas, o ambas a la vez.

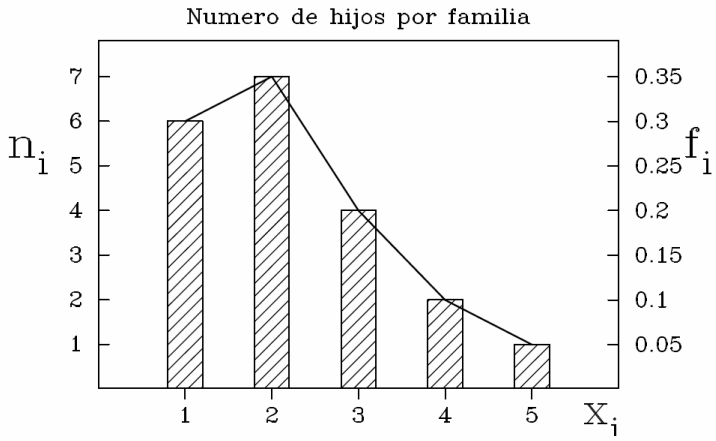

Diagrama de barras y polígono de frecuencias.

Representaciones gráficas

- Representaciones gráficas para datos sin agrupar

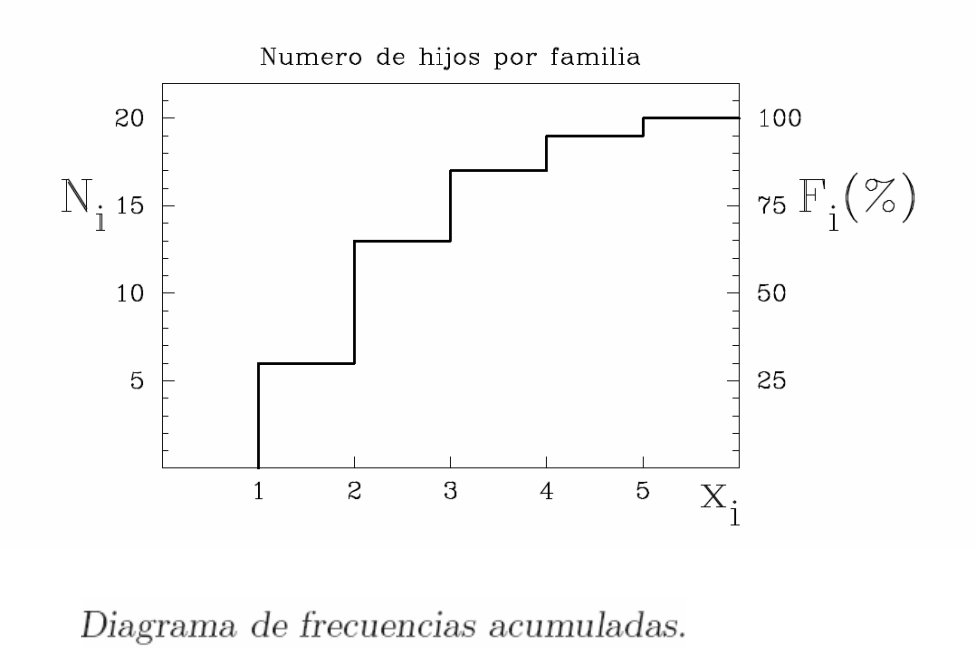

Diagrama de frecuencias acumuladas.

Para representar **las frecuencias, tanto absolutas como relativas, acumuladas** se usa el diagrama de frecuencias acumuladas. Este gráfico, en forma de escalera se construye representando en abscisas los distintos valores de la variable y levantando sobre cada xi una perpendicular cuya longitud será la frecuencia acumulada (*Ni* o *Fi*) de ese valor. Los puntos se unen con tramos horizontales y verticales como se muestra en la figura. Evidentemente la escalera resultante ha de ser siempre ascendente.

Representaciones gráficas

- Representaciones gráficas para variables cualitativas

El **diagrama de rectángulos es similar al diagrama de barras y el histograma para las variables cuantitativas.** Consiste en representar en el eje de abscisas los diferentes caracteres cualitativos y levantar sobre cada uno de ellos un rectángulo (de forma no solapada) cuya altura sea la frecuencia (absoluta o relativa) de dicho carácter.

Un diagrama muy usado es el diagrama de sectores (también llamado diagrama de tarta). En el se representa el valor de cada carácter cualitativo como un sector de un círculo completo, siendo el área de cada sector, o, lo que es lo mismo, el arco subtendido, proporcional a la frecuencia del carácter en cuestión. De forma práctica, cada arco se calcula como 360° multiplicado por la frecuencia relativa. Es además costumbre escribir dentro, o a un lado, de cada sector la frecuencia correspondiente. Este tipo de diagrama proporciona una idea visual muy clara de cuales son los caracteres que más se repiten.

Las notas de una asignatura de Físicas (en la UCM) del curso académico 95/96 se distribuyeron de acuerdo a la siguiente tabla para los alumnos presentados en junio:

Nota	n_i	f_i	N_i	F_i	α_i
Suspenso (SS)	110	0.46	110	0.46	165.6
Aprobado (AP)	90	0.38	200	0.84	136.8
Notable (NT)	23	0.10	223	0.94	36.0
Sobresaliente (SB)	12	0.05	235	0.99	18.0
Matrícula de Honor (MH)	2	0.01	237	1.00	3.6

Representaciones gráficas

- Representaciones gráficas para variables cualitativas

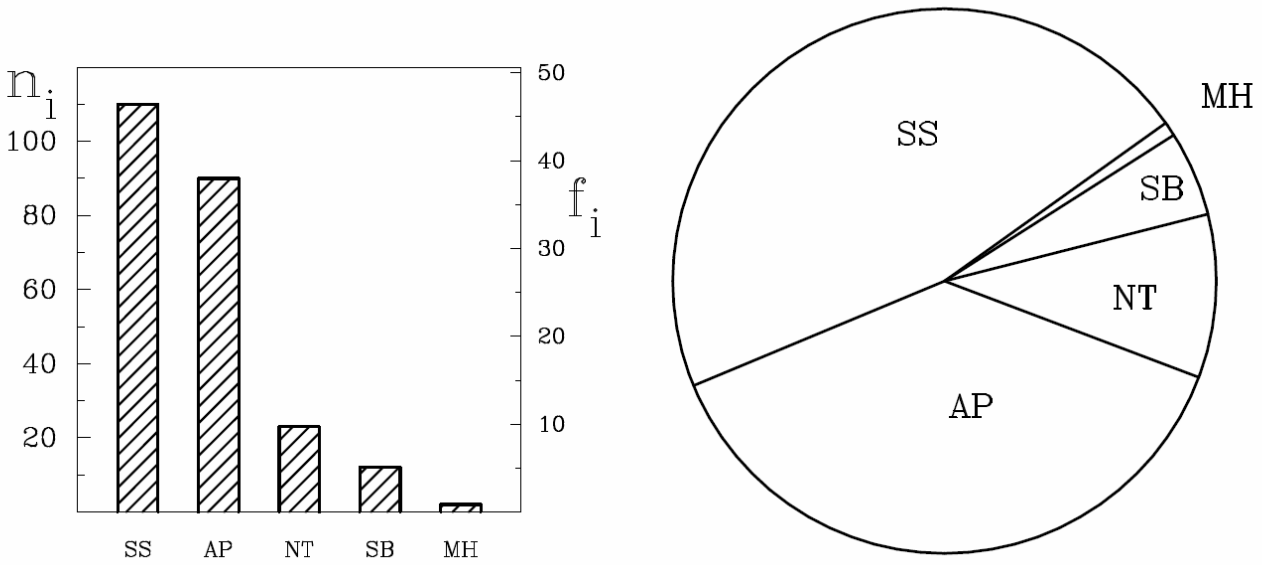

Diagrama de rectángulos (izquierda) y de sectores (derecha) para las notas del ejemplo I–7. Las frecuencias relativas están dadas en tanto por ciento. Los ángulos de cada sector circular se determinan como $\alpha_i = f_i \times 360$ *(grados).*

Medidas características de una distribución

Medidas de centralización

Media aritmética

Supongamos que tenemos una muestra de tamaño N, donde la variable estadística x toma los valores $x_1, x_2, \ldots, x_N$. Se define la **media aritmética** $\overline{x}$, o simplemente **media**, de la muestra como

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}.$$

Es decir, la media se calcula sencillamente sumando los distintos valores de x y dividiendo por el número de datos. En el caso de que los diferentes valores de x aparezcan *repetidos*, tomando entonces los valores $x_1, x_2, \ldots, x_k$, con frecuencias absolutas $n_1, n_2, \ldots, n_k$, la media se determina como

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

pudiéndose expresar también en función de las frecuencias relativas mediante

$$\overline{x} = \sum_{i=1}^{k} x_i f_i.$$

Medidas características de una distribución

Medidas de centralización

Media aritmética

Calcularemos la media aritmética para los datos del ejemplo I–5.

x_i	n_i	f_i	$x_i \times n_i$	$x_i \times f_i$
1	6	0.30	6	0.30
2	7	0.35	14	0.70
3	4	0.20	12	0.60
4	2	0.10	8	0.40
5	1	0.05	5	0.25
Total	20	1.00	45	2.25

$$\overline{x} = \frac{\sum_1^5 x_i\, n_i}{N} = \frac{45}{20} = 2.25,$$

o también usando las frecuencias relativas mediante la ecuación

$$\overline{x} = \sum_1^5 x_i\ f_i = 2.25.$$

Medidas características de una distribución

Medidas de centralización

Media aritmética

Calcularemos la media aritmética para los datos del ejemplo I–5.

x_i	n_i	f_i	$x_i \times n_i$	$x_i \times f_i$
1	6	0.30	6	0.30
2	7	0.35	14	0.70
3	4	0.20	12	0.60
4	2	0.10	8	0.40
5	1	0.05	5	0.25
Total	20	1.00	45	2.25

$$\overline{x} = \frac{\sum_1^5 x_i\, n_i}{N} = \frac{45}{20} = 2.25,$$

o también usando las frecuencias relativas mediante la ecuación

$$\overline{x} = \sum_1^5 x_i\ f_i = 2.25.$$

Medidas características de una distribución

Medidas de centralización

Mediana

Una medida de centralización importante es la **mediana** M_e. Se define ésta como una medida central tal que, con los datos ordenados de menor a mayor, el 50 % de los datos son inferiores a su valor y el 50 % de los datos tienen valores superiores. Es decir, la mediana divide en dos partes iguales la distribución de frecuencias o, gráficamente, divide el histograma en dos partes de áreas iguales. Vamos a distinguir diversos casos para su cálculo:

1. Supongamos en primer lugar que los diferentes valores de la variable no aparecen, en general, repetidos. En este caso, y suponiendo que tenemos los datos ordenados, la mediana será el valor central, si N is impar, o la media aritmética de los dos valores centrales, si N es par. Por ejemplo, si $x = 1, 4, 6, 7, 9$, la mediana sería 6. Por otro lado, si $x = 1, 4, 6, 7$ la mediana es $M_e = (4 + 6)/2 = 5$.

Medidas características de una distribución

Medidas de centralización

Mediana

2. En el caso de que tengamos una variable discreta con valores repetidos sobre la cual hemos elaborado una tabla de frecuencias se calcula en primer lugar el número de observaciones N dividido entre 2. Podemos distinguir entonces dos casos. El primero de ellos es cuando dicho valor $N/2$ coincide con la frecuencia absoluta acumulada N_j de un valor x_j de la variable (o, lo que es lo mismo, cuando la frecuencia relativa acumulada $F_j = 0.5$). En este caso la mediana se ha de situar entre este valor de la variable y el siguiente ya que de esta forma dividirá la distribución de frecuencias en 2. Es decir, se calcula como la media aritmética de dicho valor de la variable y su superior

$$M_e = \frac{x_j + x_{j+1}}{2}$$

Si $N/2$ no coincidiese con ningún valor de la columna de frecuencias acumuladas (como suele ocurrir) la mediana sería el primer valor de x_j con frecuencia absoluta acumulada N_j mayor que $N/2$, ya que el valor central de la distribución correspondería a una de las medidas englobadas en ese x_j.

Medidas características de una distribución

Medidas de centralización

Mediana

Ejemplo I–5

(Continuación.)

Usando los datos del número de hijos del ejemplo I–5, tenemos

x_i	N_i
1	6
2	13
3	17
4	19
5	20

1–1–1–1–1–1–2–2–2–2–2–2–2–2–3–3–3–3–4–4–5

$$N/2 = 10$$

La mediana será el primer valor de x_i con frecuencia absoluta acumulada $N_i > 10$, es decir

$$M_e = x_2 = 2.$$

Modificando la tabla de datos para estar en el otro caso mencionado

x_i	N_i
1	6
2	10
3	15
4	17
5	20

1–1–1–1–1–1–2–2–2–2–3–3–3–3–3–4–4–5–5–5

$$N/2 = 10 = N_2,$$

entonces

$$M_e = \frac{x_2 + x_{2+1}}{2} = \frac{2+3}{2} = 2.5.$$

Medidas características de una distribución

Medidas de centralización

Mediana

3. Supongamos ahora que tenemos una muestra de una variable continua cuyos valores están agrupados en intervalos de clase. En este caso pueden ocurrir dos situaciones. En primer lugar, si $N/2$ coincide con la frecuencia absoluta acumulada N_j de un intervalo (a_j, a_{j+1}) (con marca de clase c_j), la mediana será sencillamente el extremo superior a_{j+1} de ese intervalo. En el caso general de que ninguna frecuencia absoluta acumulada coincida con $N/2$ será necesario interpolar en el polígono de frecuencias acumuladas (Fig. 3.1). Supongamos que el valor $N/2$ se encuentra entre las frecuencias N_{j-1} y N_j, correspondientes a los intervalos (a_{j-1}, a_j) y (a_j, a_{j+1}) respectivamente, la mediana se situará en algún lugar del intervalo superior (a_j, a_{j+1}). Para calcular el valor exacto se interpola según se observa en la Figura 3.1

$$\frac{a_{j+1} - a_j}{N_j - N_{j-1}} = \frac{M_e - a_j}{N/2 - N_{j-1}}$$

$$\Rightarrow M_e = a_j + \frac{N/2 - N_{j-1}}{N_j - N_{j-1}}(a_{j+1} - a_j) = a_j + \frac{N/2 - N_{j-1}}{n_j}(a_{j+1} - a_j).$$

Medidas características de una distribución

Medidas de centralización

Mediana

Ejemplo I–6

(Continuación.)

Volviendo de nuevo a las medidas agrupadas del ejemplo I–6, podemos calcular la mediana recordando el agrupamiento en intervalos que realizamos en su momento.

a_i—a_{i+1}	n_i	N_i
7.405—8.105	7	7
8.105—8.805	9	16
8.805—9.505	2	18
9.505—10.205	2	20
10.205—10.905	1	21

$$N/2 = 10.5 \neq N_i$$

$$(N_1 = 7) < (N/2 = 10.5) < (N_2 = 16)$$

La mediana se situará entonces en el intervalo 8.105—8.805,

$$8.105 < M_e < 8.805.$$

$$M_e = a_j + \frac{N/2 - N_{j-1}}{n_j}(a_{j+1} - a_j) = a_2 + \frac{10.5 - N_1}{n_2}(a_3 - a_2) =$$

$$= 8.105 + \frac{10.5 - 7}{9}(8.805 - 8.105) = 8.105 + 0.388 \times 0.7 = 8.38.$$

Compárese este resultado con $\overline{x} = 8.52$.

Medidas características de una distribución

Medidas de centralización

Mediana y Media

En comparación con la media aritmética la mediana, como medida de centralización, tiene propiedades muy distintas, presentando sus ventajas e inconvenientes. Por un lado, la mayor ventaja de la media es que se utiliza toda la información de la distribución de frecuencias (todos los valores particulares de la variable), en contraste con la mediana, que solo utiliza el orden en que se distribuyen los valores. Podría pues considerarse, desde este punto de vista, que la media aritmética es una medida más fiable del valor central de los datos. Sin embargo, como hemos visto anteriormente, la media es muy poco robusta, en el sentido de que es muy sensible a valores extremos de la variable y, por lo tanto, a posibles errores en las medidas. La mediana, por otro lado, es una medida robusta, siendo muy insensible a valores que se desvíen mucho. Por ejemplo, supongamos que la variable x toma los valores $x = 2, 4, 5, 7, 8$, la media y la mediana serían en este caso muy parecidas ($\overline{x} = 5.2$, $M_e = 5$). Pero si sustituimos el último valor 8 por 30, la nueva media se ve muy afectada ($\overline{x} = 9.6$), no siendo en absoluto una medida de la tendencia central, mientras que el valor de la mediana no cambia ($M_e = 5$). Podríamos poner como contraejemplo el caso de las longitudes de barras (en cm) inicialmente idénticas calentadas a temperaturas desconocidas en distintos recipientes: 1.80/1.82/1.85/1.90/2.00, cuya media y mediana son $\overline{x} = 1.874$ y $M_e = 1.85$. Si la temperatura de uno de esos recipientes varía, y la longitud mayor aumenta de 2.00 a 2.20 cm, la mediana no varía, pero la media pasa a $\overline{x} = 1.914$ y nos informa del cambio.

Medidas características de una distribución

Repaso percentiles

Cálculo de los percentiles

En primer lugar buscamos la clase donde se encuentra $\frac{k \cdot N}{100}$, $k = 1,2,..99$, en la tabla de las frecuencias acumuladas.

$$P_k = L_i + \frac{\frac{k \cdot N}{100} - F_{i-1}}{f_i} \cdot a_i \qquad k = 1,2,..99$$

$\mathbf{L_i}$ es el límite inferior de la clase donde se encuentra el percentil.

N es la suma de las frecuencias absolutas.

$\mathbf{F_{i-1}}$ es la **frecuencia acumulada** anterior a la clase del percentil.

$\mathbf{a_i}$ es la amplitud de la clase.

Medidas características de una distribución

Repaso percentiles

Cálculo de los percentiles

En primer lugar buscamos la clase donde se encuentra $\frac{k \cdot N}{100}$, $k = 1,2,..99$, en la tabla de las frecuencias acumuladas.

$$P_k = L_i + \frac{\frac{k \cdot N}{100} - F_{i-1}}{f_i} \cdot a_i \qquad k = 1,2,..99$$

$\mathbf{L_i}$ es el límite inferior de la clase donde se encuentra el percentil.

N es la suma de las frecuencias absolutas.

$\mathbf{F_{i-1}}$ es la **frecuencia acumulada** anterior a la clase del percentil.

$\mathbf{a_i}$ es la amplitud de la clase.

Medidas características de una distribución

Medidas de centralización

Moda

Se define la **moda** M_o de una muestra como aquel valor de la variable que tiene una frecuencia máxima. En otras palabras, es el valor que más se repite. Hay que indicar que puede suceder que la moda no sea única, es decir que aparezcan varios máximos en la distribución de frecuencias. En ese caso diremos que tenemos una distribución bimodal, trimodal, etc. Evidentemente, en el caso de una variable discreta que no toma valores repetidos, la moda no tiene sentido. Cuando sí existen valores repetidos su cálculo es directo ya que puede leerse directamente de la tabla de distribución de frecuencias.

Medidas características de una distribución

Medidas de centralización

Moda

Consideremos de nuevo el caso del número de hijos por familia.

x_i	n_i	f_i	N_i	F_i
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00

El valor que más se repite es 2 hijos, que ocurre en siete familias de la muestra ($n_i = 7$). La moda es por tanto $M_o = 2$ y en este caso coincide con la mediana.

Medidas características de una distribución

Medidas de centralización

Moda

Cálculo de la moda para datos agrupados

1º Todos los intervalos tienen la misma amplitud.

$$Mo = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \cdot a_i$$

L_i es el límite inferior de la clase modal.

f_i es la frecuencia absoluta de la clase modal.

f_{i--1} es la frecuencia absoluta inmediatamente inferior a la clase modal.

f_{i-+1} es la frecuencia absoluta inmediatamente posterior a la clase modal.

a_i es la amplitud de la clase.

Medidas características de una distribución

Medidas de centralización

Moda

También se utiliza otra **fórmula** de la **moda** que da un **valor aproximado** de ésta:

$$Mo = L_i + \frac{f_{i+1}}{f_{i-1} + f_{i+1}} \cdot a_i$$

Ejemplo:

Calcular la **moda** de una distribución estadística que viene dada por la siguiente tabla:

	f_i
[60, 63)	5
[63, 66)	18
[66, 69)	42
[69, 72)	27
[72, 75)	8
	100

$$Mo = 66 + \frac{(42-18)}{(42-18)+(42-27)} \cdot 3 = 67.846$$

Medidas características de una distribución

Medidas de centralización

Moda

En primer lugar tenemos que hallar las alturas.

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

La clase modal es la que tiene mayor altura.

$$Mo = L_i + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} \cdot a_i$$

La **fórmula** de la **moda aproximada** cuando existen distintas amplitudes es:

$$Mo = L_i + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot a_i$$

Medidas características de una distribución

Medidas de centralización

Moda

Ejemplo:

En la siguiente tabla se muestra las calificaciones (suspenso, aprobado, notable y sobresaliente) obtenidas por un grupo de 50 alumnos. **Calcular la moda.**

	f_i	h_i
[0, 5)	15	3
[5, 7)	20	10
[7, 9)	12	6
[9, 10)	3	3
	50	

$$Mo = 5 + \frac{10-3}{(10-3)+(10-6)} \cdot 2 = 6.27$$

$$Mo = 5 + \frac{6}{3+6} \cdot 2 = 6.33$$

Medidas características de una distribución

Medidas de centralización

Moda

En el caso de variables continuas agrupadas en intervalos de clase existirá un intervalo en el que la frecuencia sea máxima, llamado intervalo modal. Es posible asociar la moda a un valor determinado de la variable dentro de dicho intervalo modal. Para ello supongamos que sea (a_j, a_{j+1}) el intervalo con frecuencia máxima n_j. Si n_{j-1} y n_{j+1} son las frecuencias de los intervalos anterior y posterior al modal, definimos $\delta_1 = n_j - n_{j-1}$ y $\delta_2 = n_j - n_{j+1}$ (ver el histograma de la Figura 3.2). En este caso, el valor exacto de la moda se puede calcular como

$$M_o = a_j + \frac{\delta_1}{\delta_1 + \delta_2}(a_{j+1} - a_j)$$

(ver demostración en el libro de Quesada). Es decir, la moda estará más próxima a a_j cuanto menor sea la diferencia de frecuencias con el intervalo anterior, y al revés. Si, por ejemplo, $n_{j-1} = n_j$ $(\delta_1 = 0)$, la moda será efectivamente a_j. Por el contrario si $n_{j+1} = n_j$ $(\delta_2 = 0)$ la moda será a_{j+1}, estando situada entre dos intervalos.

Medidas características de una distribución

Medidas de centralización

Moda

Ejemplo I–6

(Continuación.)

Para el caso de las medidas de la paralaje solar (ejemplo I–6), se estudia el intervalo con frecuencia máxima (intervalo modal) que en este caso es $(a_j, a_{j+1}) = (8.105\ ,\ 8.805)$,

a_i—a_{i+1}	c_i	n_i
7.405—8.105	7.755	7
8.105—8.805	8.455	9 ←
8.805—9.505	9.155	2
9.505—10.205	9.855	2
10.205—10.905	10.555	1

$$j = 2; \quad n_{j-1} = 7; \quad n_j = 9; \quad n_{j+1} = 2$$

$$\delta_1 = n_j - n_{j-1} = 9 - 7 = 2$$

$$\delta_2 = n_j - n_{j+1} = 9 - 2 = 7$$

$$M_o = a_j + \frac{\delta_1}{\delta_1 + \delta_2}(a_{j+1} - a_j) = 8.105 + \frac{2}{2+7}(8.805 - 8.105) = 8.26.$$

Medidas características de una distribución

Medidas de dispersión

Recorrido

Una evaluación rápida de la dispersión de los datos se puede realizar calculando el **recorrido** (también llamado rango), o diferencia entre el valor máximo y mínimo que toma la variable estadística. Con el fin de eliminar la excesiva influencia de los valores extremos en el recorrido, se define el **recorrido intercuartílico** como la diferencia entre el trecer y primer cuartil

$$R_I = Q_{3/4} - Q_{1/4}.$$

Está claro que este recorrido nos dará entonces el rango que ocupan el 50 % central de los datos. En ocasiones se utiliza el **recorrido semiintercuartílico**, o mitad del recorrido intercuartílico

$$R_{SI} = \frac{Q_{3/4} - Q_{1/4}}{2}.$$

Medidas características de una distribución

Medidas de dispersión

Varianza y desviación típica

Sin lugar a dudas la medida más usada para estimar la dispersión de los datos es la desviación típica. Esta es especialmente aconsejable cuando se usa la media aritmética como medida de tendencia central. Al igual que la desviación media, está basada en un valor promedio de las desviaciones respecto a la media. En este caso, en vez de tomar valores absolutos de las desviaciones, para evitar así que se compensen desviaciones positivas y negativas, se usan los cuadrados de las desviaciones. Esto hace además que los datos con desviaciones grandes influyan mucho en el resultado final. Se define entonces la **varianza** de una muestra con datos repetidos como

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

Evidentemente la varianza no tiene las mismas unidades que los datos de la muestra. Para conseguir las mismas unidades se define la **desviación típica** (algunas veces llamada desviación estándar) como la raíz cuadrada de la varianza

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

En el caso de que los datos no se repitan, estas definiciones se simplifican a

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1} \quad ; \quad s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}.$$

Medidas características de una distribución

Medidas de dispersión

Varianza y desviación típica

En el caso de una variable discreta

x_i	n_i	$x_i \times n_i$	$x_i^2 \times n_i$
1	6	6	6
2	7	14	28
3	4	12	36
4	2	8	32
5	1	5	25
Total	20	45	127

$$s^2 = \frac{\sum_1^5 x_i^2\, n_i - \frac{1}{20}(\sum_1^5 x_i\, n_i)^2}{20-1}$$

$$s^2 = \frac{127 - \frac{1}{20}45^2}{19} = 1.355$$

$$s = \sqrt{1.355} = 1.16$$

Medidas características de una distribución

Medidas de dispersión

Varianza y desviación típica

En el caso de datos agrupados en intervalos de clase

c_i	n_i	$c_i \times n_i$	$c_i^2 \times n_i$
7.755	7	54.285	420.980
8.455	9	76.095	643.383
9.155	2	18.310	167.628
9.855	2	19.710	194.242
10.555	1	10.555	111.408
Total	21	178.955	1537.641

$$s^2 = \frac{\sum_1^5 c_i^2\, n_i - \frac{1}{20}(\sum_1^5 c_i\, n_i)^2}{21-1}$$

$$s^2 = \frac{1537.641 - \frac{1}{21}178.955^2}{20} = 0.632$$

$$s = \sqrt{0.632} = 0.795$$

(sin agrupar en intervalos se obtiene $s = 0.900$)

Medidas características de una distribución

Medidas de dispersión

Desviación típica

En el caso de datos agrupados en intervalos de clase

c_i	n_i	$c_i \times n_i$	$c_i^2 \times n_i$
7.755	7	54.285	420.980
8.455	9	76.095	643.383
9.155	2	18.310	167.628
9.855	2	19.710	194.242
10.555	1	10.555	111.408
Total	21	178.955	1537.641

$$s^2 = \frac{\sum_1^5 c_i^2\, n_i - \frac{1}{20}\left(\sum_1^5 c_i\, n_i\right)^2}{21-1}$$

$$s^2 = \frac{1537.641 - \frac{1}{21}178.955^2}{20} = 0.632$$

$$s = \sqrt{0.632} = 0.795$$

(sin agrupar en intervalos se obtiene $s = 0.900$)

Medidas características de una distribución

Medidas de dispersión

Coeficiente de variación

Un problema que plantean las medidas de dispersión vistas es que vienen expresadas en las unidades en que se ha medido la variable. Es decir, son medidas absolutas y con el único dato de su valor no es posible decir si tenemos una dispersión importante o no. Para solucionar esto, se definen unas medidas de dispersión relativas, independientes de la unidades usadas. Estas dispersiones relativas van a permitir además comparar la dispersión entre diferentes muestras (con unidades diferentes). Entre estas medidas hay que destacar el **coeficiente de variación de Pearson**, definido como el cociente entre la desviación típica y la media aritmética

$$CV = \frac{s}{|\overline{x}|}.$$

Nótese que este coeficiente no se puede calcular cuando $\overline{x} = 0$. Normalmente CV se expresa en porcentaje, multiplicando su valor por 100. Evidentemente, cuanto mayor sea CV, mayor dispersión tendrán los datos.

$$CV = s/|\overline{x}| = 1.16/2.25 = 0.516 \qquad 52\,\%.$$