Parte I

Estadística descriptiva.

Introducción a la Estadística Descriptiva.

Como hemos dicho en la Introducción, la Estadística Descriptiva es la puerta de entrada a la Estadística. En nuestro trabajo o, aún más en general, en nuestra experiencia diaria, las personas nos hemos ido convirtiendo, de forma creciente, en recolectores ávidos de datos. Nuestro hambre de datos se debe a que hemos ido creando cada vez más formas de usarlos, para iluminar nuestra comprensión de algún fenómeno, y para orientar nuestras decisiones.

Pero antes de llegar a ese punto, y poder usar la información para decidir de forma eficaz, tenemos que ser capaces de tomar los datos, que son *información en bruto* y transformarlos en *información estructurada*. En particular, tenemos que desarrollar técnicas para describir, resumir, y representar esos datos. Por un lado, para poder aplicarles métodos avanzados de análisis. En este curso vamos a presentar los más básicos de esos métodos de análisis de datos. Por otro lado, queremos poder *comunicar* a otros la información que contienen esos datos. Por ejemplo, utilizando técnicas gráficas, de visualización.

Todos esos métodos y técnicas, que nos permiten transformar y describir los datos, forman parte de la Estadística Descriptiva. Así que la Estadística Descriptiva se encarga del trabajo directo con los *datos, a los que tenemos acceso*, y con los que podemos hacer operaciones. Una parte del proceso incluye operaciones matemáticas, con su correspondiente dósis de abstracción. Pero, puesto que la Estadística Descriptiva es uno de los puntos de contacto de la Estadística con el mundo real, también encontraremos muchos problemas prácticos. Y en particular, en la era de la informatización, muchos problemas de índole computacional, del tipo "¿cómo consigo que el ordenador haga eso?". No queremos, en cualquier caso, refugiarnos en las matemáticas, obviando esa parte práctica del trabajo. Procesar los datos requiere de nosotros, a menudo, una cierta soltura con las herramientas computacionales, y el dominio de algunos trucos del oficio. En la parte más práctica del curso, los Tutoriales, dedicaremos tiempo a esta tarea.

En esta parte del libro vamos a conocer a algunos actores, protagonistas de la Estadística, que nos acompañarán a lo largo de todo el curso: la media, la varianza, las frecuencias y percentiles, etc. Vamos a tocar, siquiera brevemente, el tema de la visualización y representación gráfica de datos. Hay tanto que decir en ese terreno, que pedimos disculpas al lector por adelantado por lo elemental de las herramientas que vamos a presentar. Entrar con más profundidad en esta materia exigiría un espacio del que no disponemos. Como, por otra parte, nos sucederá más veces a lo largo del curso. No obstante, sí hemos incluido una breve visita a las nociones de precisión y exactitud, y a la vertiente más práctica del trabajo con cifras significativas, porque, en nuestra experiencia, a menudo causa dificultades a los principiantes.

Población y muestra.

También hemos dicho que todas las partes en que se divide la Estadística están interconectadas entre sí. Y no sabríamos cerrar esta introducción a la primera parte del libro, especialmente por ser la primera, sin tratar de tender la vista hacia esas otras partes, que nos esperan más adelante. Así que vamos a extendernos un poco más aquí, para intentar que el lector tenga un poco más de perspectiva.

Como hemos dicho, la Estadística Descriptiva trabaja con datos a los que tenemos acceso. Pero, en muchos casos, esos datos corresponden a una muestra, es decir, a un subconjunto (más o menos pequeño), de una población (más o menos grande), que nos

gustaría estudiar. El problema es que estudiar toda la población puede ser demasiado difícil o indeseable, o directamente imposible. En ese caso surge la pregunta ¿hasta qué punto los datos de la muestra son *representativos* de la población? Es decir, ¿podemos usar los datos de la muestra para *inferir*, o *predecir* las características de la población completa? La Inferencia Estadística, que comenzaremos en la tercera parte del libro, se encarga de dar sentido a estas preguntas, formalizarlas y responderlas. Y es, sin discusión, el auténtico núcleo, el alma de la Estadística.

En la Inferencia clásica, por tanto, trataremos de usar la información que la Estadística Descriptiva extrae de los datos de la muestra para poder hacer predicciones precisas sobre las propiedades de la población. Algunos ejemplos típicos de la clase de predicciones que queremos hacer son las encuestas electorales, el control de calidad empresarial o los ensayos clínicos, que son prototipos de lo que estamos explicando, y que muestran que la Estadística consigue, a menudo, realizar con éxito esa tarea.

¿Por qué funciona la Inferencia? A lo largo del libro tendremos ocasión de profundizar en esta discusión. Pero podemos adelantar una primera respuesta: funciona porque, en muchos casos, cualquier muestra *bien elegida* (y ya daremos más detalles de lo que significa esto), es bastante *representativa* de la población. Dicho de otra manera, si pensamos en el conjunto de todas las posibles muestras bien elegidas que podríamos tomar, la inmensa mayoría de ellas serán coherentes entre sí, y representativas de la población. Un ingrediente clave en este punto, sobre el que volveremos, es el enorme tamaño del conjunto de posibles muestras. Así que, si tomamos una *al azar*, casi con seguridad habremos tomado una muestra representativa. Y puesto que hemos mencionado el *azar*, parece evidente que la manera de hacer que estas frases imprecisas se conviertan en afirmaciones científicas, verificables, es utilizar el lenguaje de la Probabilidad. Por esa razón, necesitamos hablar en ese lenguaje para poder hacer Estadística rigurosa. Y con eso, tenemos trazado el plan de buena parte de este libro y de nuestro curso.

Capítulo 1

Introducción a la estadística descriptiva.

1.1. Tipos de Variables.

A lo largo del curso estudiaremos técnicas para describir y/o analizar características de una población. Los datos que obtengamos los almacenaremos en variables. Podemos pensar en una variable como una especie de "contenedor" en el que guardar los datos. Dependiendo del tipo de característica en la que estemos interesados, usaremos un tipo de variable u otro para almacenar la información a partir de la que empezar a trabajar.

1.1.1. Variables cualitativas y cuantitativas.

A veces se dice que las variables cuantitativas son las variables numéricas, y las cualitativas las no numéricas. La diferencia es, en realidad, un poco más sutil. Una variable es cualitativa nominal cuando sólo se utiliza para establecer categorías, y *no para hacer operaciones con ella*. Es decir, para poner nombres, crear clases o especies dentro de los individuos que estamos estudiando. Por ejemplo, cuando clasificamos a los seres vivos en especies, no estamos *midiendo nada*. Podemos *representar* esas especies mediante números, naturalmente, pero en este caso la utilidad de ese número se acaba en la propia representación, y en la clasificación que los números permiten. Pero no utilizamos las propiedades de los números (las operaciones aritméticas, suma, resta, etc.). Volviendo al ejemplo de las especies, no tiene sentido sumar especies de seres vivos. A menudo llamaremos a estas variables factores, y diremos que los distintos valores que puede tomar un factor son los niveles de ese factor. Por ejemplo, en un estudio sobre cómo afecta al crecimiento de una planta el tipo de riego que se utiliza, podríamos utilizar un factor (variable cualitativa) llamado *riego*, con niveles: *ninguno, escaso, medio, abundante*.

Una variable cuantitativa, por el contrario, tiene un valor numérico, y las operaciones matemáticas que se pueden hacer con ese número son importantes para nosotros. Por ejemplo, podemos medir la presión arterial de un animal y utilizar fórmulas de la mecánica de fluidos para estudiar el flujo sanguíneo.

En la frontera, entre las variables cuantitativas y las cualitativas, se incluyen las cuali-

tativas ordenadas. En este caso existe una ordenación dentro de los valores de la variable.

Ejemplo 1.1.1. *Un ejemplo de este tipo de variables es la gravedad del pronóstico de un enfermo ingresado en un hospital. Como ya hemos dicho, se pueden codificar mediante números de manera que el orden se corresponda con el de los códigos numéricos, como aparece en la Tabla 1.1.*

Pronóstico	*Código*
Leve	1
Moderado	2
Grave	3

Tabla 1.1: Un ejemplo de variable cualitativa ordenada.

Pero no tiene sentido hacer otras operaciones con esos valores: no podemos sumar grave con leve. □

En este caso es *especialmente importante* no usar esos números para operaciones estadísticas que pueden no tener significado (por ejemplo, calcular la media, algo de lo que trataremos en el próximo capítulo).

1.1.2. Variables cuantitativas discretas y continuas.

A su vez, las variables cuantitativas (aquellas con las que las operaciones numéricas tienen sentido) se dividen en discretas y continuas. Puesto que se trata de números, y queremos hacer operaciones con ellos, la clasificación depende de las operaciones matemáticas que vamos a realizar.

Cuando utilizamos los números enteros ($\mathbb{Z}$), que son

$$\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$$

o un subconjunto de ellos como modelo, la variable es discreta. Y entonces con esos números podemos sumar, restar, multiplicar (pero no siempre dividir).

Por el contrario, si usamos los números reales ($\mathbb{R}$), entonces la variable aleatoria es continua. La diferencia entre un tipo de datos y el otro se corresponde en general con la diferencia entre digital y analógico. Es importante entender que la diferencia entre discreto y continuo es, en general, una diferencia que establecemos nosotros al crear un modelo con el que estudiar un fenómeno, y que la elección correcta del tipo de variable es uno (entre otros) de los ingredientes que determinan la utilidad del modelo. Un ejemplo clásico de este tipo de situaciones es el uso de la variable *tiempo*. Cuando alguien nos dice que una reacción química, por ejemplo la combustión en un motor diesel a 1500 rpm, ha transcurrido en 5.6 milisegundos, está normalmente claro que, en el contexto de este problema, nos interesan los valores de la variable tiempo con mucha precisión, y la diferencia entre 5.6 y, por ejemplo, 5.9 milisegundos puede ser fundamental. Además, y aún más importante, en este tipo de situaciones, damos por sentado que la variable tiempo podría tomar *cualquier valor en un cierto intervalo*. Si observáramos esa reacción con aparatos más precisos, a lo mejor podríamos decir que el tiempo de la combustión es de

5.57 milisegundos, y no, por ejemplo, de 5.59 milisegundos. ¡Aunque, por supuesto, ambas cantidades se redondearán a 5.6 milisegundos cuando sólo se usan dos cifras significativas![1] Por el contrario, si decimos que el tratamiento de un paciente en un hospital ha durado tres días, está claro que en el contexto de este problema no queremos decir que el paciente salió por la puerta del hospital exactamente 72 horas (o 259200 segundos) después de haber entrado. El matiz esencial es que *no nos importa la diferencia* entre salir a las 68 o a las 71 horas. Y decidimos usar una unidad de tiempo, el día, que *sólo toma valores enteros, separados por saltos de una unidad.* En este problema hablamos de un día, dos o tres días, pero no diremos que el paciente salió del hospital a los 1.73 días. Eso no significa que no tenga sentido hablar de 1.73 días. ¡Naturalmente que lo tiene! La cuestión es si nos importa, si necesitamos ese nivel de precisión en el contexto del problema que nos ocupa.

Somos conscientes de que esta diferencia entre los tipos de variables y su uso en distintos problemas es una cuestion sutil, que sólo se aclarará progresivamente a medida que el lector vaya teniendo más experiencia con modelos de los diversos tipos: discretos, continuos, y también factoriales. Además este tema toca de cerca varias cuestiones (como la idea de precisión, o el uso de las cifras significativas) sobre los que volveremos más adelante, en la Sección 1.3, y siempre que tengamos ocasión a lo largo del curso.

1.1.3. Notación para las variables. Tablas de frecuencia. Datos agrupados.

En cualquier caso, vamos a tener siempre una lista o vector x de valores (datos, observaciones, medidas) de una variable, que representaremos con símbolos como

$$x_1, x_2, \ldots, x_n \text{ o también } x = (x_1, x_2, \ldots, x_n)$$

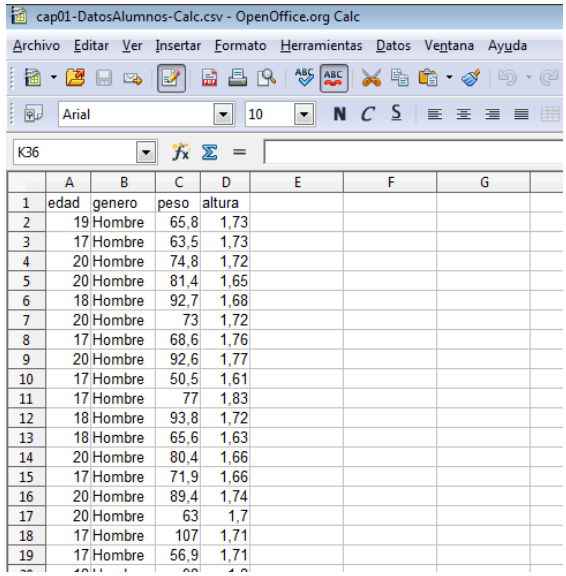
El número n se utiliza habitualmente en Estadística para referirse al número total de valores de los que se dispone. Por ejemplo, en el fichero cap01-DatosAlumnos.csv (hay una versión adecuada para usarla en la hoja de cálculo Calc, usando comas para los decimales: cap01-DatosAlumnos-Calc.csv) tenemos una tabla con datos de los 100 alumnos de una clase ficticia. No te preocupes de los detalles técnicos del fichero, por el momento. En los primeros tutoriales del curso explicaremos cómo usar este fichero con el ordenador. Para cada alumno tenemos, en una fila de la tabla, un valor de cada una de las variables *género, peso , altura , edad*. En la Figura 1.1 se muestra una parte de los datos que contiene este fichero, abierto con la hoja de cálculo Calc.

Vamos a utilizar estos datos para ilustrar algunas de las ideas que iremos viendo.

Una observación: si utilizamos $p_1, p_2, \ldots, p_{100}$ para referirnos, por ejemplo, a los datos de peso de esa tabla, entonces p_1 es el dato en la segunda fila, p_2 el dato en la tercera, y p_{35} el dato de la fila 36. Porque, como veremos, puede ser cómodo y conveniente conservar los nombres de las variables en la primera fila de la tabla de datos. Además, en estos casos puede ser una buena idea introducir una columna adicional con el índice i que corresponde a p_i (i es el número de la observación).

Un mismo valor de la variable puede aparecer repetido varias veces en la serie de observaciones. En el fichero de alumnos del que estamos hablando, la variable *edad* toma estos

[1] Hablaremos con detalle sobre cifras significativas en la Sección 1.3



	A	B	C	D	E	F	G
1	edad	genero	peso	altura			
2	19	Hombre	65,8	1,73			
3	17	Hombre	63,5	1,73			
4	20	Hombre	74,8	1,72			
5	20	Hombre	81,4	1,65			
6	18	Hombre	92,7	1,68			
7	20	Hombre	73	1,72			
8	17	Hombre	68,6	1,76			
9	20	Hombre	92,6	1,77			
10	17	Hombre	50,5	1,61			
11	17	Hombre	77	1,83			
12	18	Hombre	93,8	1,72			
13	18	Hombre	65,6	1,63			
14	20	Hombre	80,4	1,66			
15	17	Hombre	71,9	1,66			
16	20	Hombre	89,4	1,74			
17	20	Hombre	63	1,7			
18	17	Hombre	107	1,71			
19	17	Hombre	56,9	1,71			

Figura 1.1: El contenido del fichero `cap01-DatosAlumnos.csv`, en Calc.

cuatro valores distintos:

$$17, \qquad 18, \qquad 19, \qquad 20$$

Pero, naturalmente, cada uno de esos valores aparece repetido unas cuantas veces; no en vano ¡hay 100 alumnos! Este número de repeticiones de un valor es lo que llamamos la frecuencia de ese valor. Por ejemplo, el valor 20 aparece repetido 23 veces, lo que significa obviamente que hay 23 alumnos de 20 años de edad en esa clase. ¿Cómo hemos sabido esto? Desde luego, no los hemos contado "a mano". Una de las primeras cosas que haremos en los tutoriales del curso es aprender a obtener la frecuencia en un caso como este.

El número de repeticiones de un valor, del que hemos hablado en el anterior párrafo, se llama frecuencia absoluta, para distinguirlo de la frecuencia relativa, que se obtiene dividiendo la frecuencia absoluta por n (el total de observaciones). La frecuencia relativa es un *tanto por uno*, y se convierte fácilmente en un porcentaje, multiplicándola por 100. Volveremos sobre este tema en el Capítulo 2 (ver la página 27).

Cuando tratamos con variables cualitativas o discretas, muchas veces, en lugar del valor de cada observación la información que tenemos es la de las frecuencias de cada uno de los posibles valores distintos de esas variables. Esto es lo que se conoce como una tabla de frecuencias. Por ejemplo, la Tabla 1.2 (pág.9) es la tabla de frecuencia de la variable edad en este ejemplo

¿Qué sucede en este ejemplo con la variable peso? ¿Podemos calcular una tabla de frecuencias? Sí, en principio, podemos. Pero hay demasiados valores distintos, y la información presentada así no es útil. De hecho, como el peso *es una variable (cuantitativa) continua*, si nos dan los pesos de los alumnos en kilos, con, por ejemplo, dos cifras decimales, algo como 56.41kg, *es muy posible que no haya dos alumnos con el mismo valor de la variable*

edad	frecuencia
17	17
18	37
19	23
20	23

Tabla 1.2: Tabla de frecuencia. variable edad en el ejemplo de una clase ficticia.

peso. Por otra parte, si los pesos de varios alumnos se diferencian en unos pocos cientos de gramos, seguramente preferiremos representarlos por un valor común (el mismo para todos los alumnos de pesos parecidos). En el caso de variables continuas, lo habitual es *dividir el recorrido de posibles valores de esa variable continua en intervalos*, que también se llaman clases. Y además se elige a un valor particular, llamado la marca de clase, como representante de todos los valores que pertenecen a ese intervalo. Si el intervalo es $(a, b]$ (es decir, los valores x que cumplen $a < x \leq b$), lo habitual es tomar como marca de clase el punto medio de ese intervalo; es decir, el valor:

$$\frac{a+b}{2}$$

Por cierto, tomamos los intervalos de la forma $(a, b]$ para evitar dudas o ambigüedades sobre a qué intervalo pertenecen los extremos.

Una tabla de frecuencia por intervalos muestra, para estas variables, cuantos de los valores observados caen dentro de cada uno de los intervalos. En el ejemplo que estamos utilizando, podemos dividir arbitrariamente los valores del peso en intervalos de 10 kilos, desde 40 hasta 110, y obtenemos la tabla de frecuencias (se muestra en disposición horizontal, dividida en dos filas):

Peso (kg) entre	(40,50]	(50,60]	(60,70]	(70,80]
Número de alumnos	1	20	21	29
Peso (kg) entre	(80,90]	(90,100]	(100,110]	
Número de alumnos	20	7	2	

Tabla 1.3: Tabla de frecuencia, variable peso agrupada en intervalos.

Algunos comentarios adicionales sobre esta tabla:

1. El proceso para obtener estas tablas de frecuencias por intervalos es algo más complicado. De nuevo nos remitimos a los tutoriales, en este caso al Tutorial01, en el que veremos en detalle cómo se hace esto en una hoja de cálculo. Además, este proceso está relacionado con la distinción entre valores cuantitativas discretas y continuas (ver pág. 6). Ya dijimos que esa diferencia era una cuestión sutil, que iría quedando más clara con la experiencia.

2. Los intervalos, insistimos, se han elegido de manera arbitraria en este ejemplo. Invitamos al lector a pensar cómo cambiaría la información de la tabla de frecuencias si eligiéramos un número distinto de intervalos, o si, por ejemplo, los intervalos no fueran todos de la misma longitud.

Cuando los valores de una variable continua se presentan en forma de tabla de frecuencias por intervalos hablaremos de datos agrupados. En cualquier caso, conviene recordar que una tabla de frecuencias es una forma de resumir la información, y que al pasar del conjunto de datos inicial a las tablas de frecuencias de Peso y Género generalmente se pierde información.

1.2. Tablas y representación gráfica de datos.

Una vez organizados y resumidos los datos en tablas queremos extraer la información que contienen. En primera instancia es recomendable hacer una exploración visual, para lo que resulta extremadamente útil trasladar el contenido de las tablas a gráficas. Vamos a ver, en este apartado, algunos de los tipos básicos (y clásicos) de diagramas que se pueden utilizar para visualizar las tablas de frecuencia. Pero no queremos dejar de decir que el tema de la visualización de datos es muy amplio, que es un campo donde la actividad es ahora mismo febril, y que a lo largo de los próximos capítulos iremos viendo otros ejemplos de representación gráfica de la información.

1.2.1. Diagramas de sectores y barras.

Los diagramas de sectores y barras se utilizan cuando queremos mostrar frecuencias (o porcentajes, recuentos, etcétera). Se pueden utilizar para ilustrar las frecuencias de variables tanto cualitativas como cuantitativas. A continuación vamos a describir un ejemplo de cada uno de estos tipos de diagrama, usando en ambos casos los datos del fichero cap01-DiagramaBarrasSectores.csv . Este fichero contiene 1500 números enteros aleatorios, del 1 al 6. La tabla de frecuencias es esta:

Valor	1	2	3	4	5	6
Frecuencia	72	201	423	512	222	70

Los diagramas de sectores circulares, como el de la Figura 1.2, son útiles para mostrar proporciones, pero sólo cuando los valores son bastante distintos entre sí. Porque, pese a su popularidad, en muchas ocasiones pueden resultar confusos o poco precisos. Por ejemplo, en esa figura ¿qué frecuencia es mayor, la del grupo 2 o la del grupo 5?

Los diagramas de barras o columnas tienen, en general, más precisión que los de sectores. En la parte (a) de la Figura 1.3 se muestra el mismo conjunto de valores que antes vimos en el diagrama de sectores. Y ahora es evidente que, aunque son muy parecidas, la frecuencia del valor 2 es menor que la del valor 5. Además, los diagramas de barras se pueden utilizar para mostrar varios conjuntos de datos simultáneamente, facilitando la comparación entre ellos, como en la parte (b) de la Figura 1.3.

En los tutoriales aprenderemos a dibujar este tipo de gráficos.

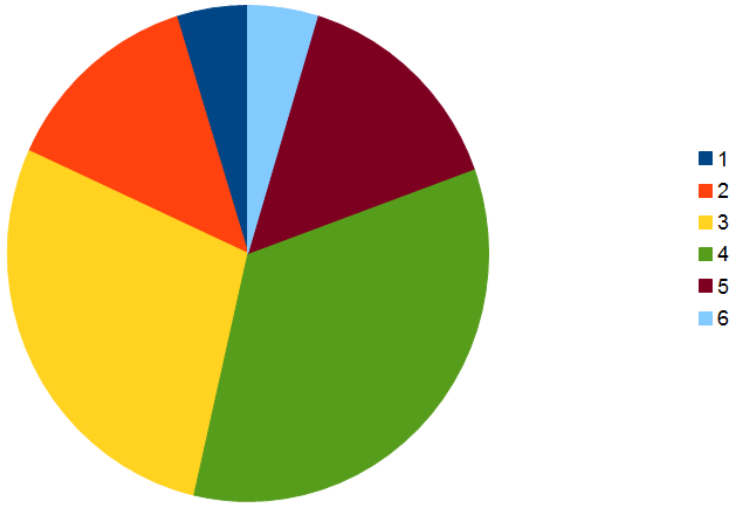


Figura 1.2: Diagrama de sectores circulares, dibujado con Calc.

1.2.2. Histogramas.

Un histograma es un tipo especial de diagrama de barras que se utiliza para variables cuantitativas agrupadas en intervalos (clases) (recuerda la discusión que precedía a la Tabla 1.3, pág. 9). Puedes ver un ejemplo en la Figura 1.5. Las dos propiedades básicas que caracterizan a un histograma son:

1. Las *bases de cada una de las barras se corresponden con los intervalos* en los que hemos dividido el recorrido de los valores de la variable continua.

2. El *área de cada barra es proporcional a la frecuencia correspondiente a ese intervalo.*

Una consecuencia de estas propiedades es que las columnas de un histograma no tienen porque tener la misma anchura, como se ve en la Figura 1.5.

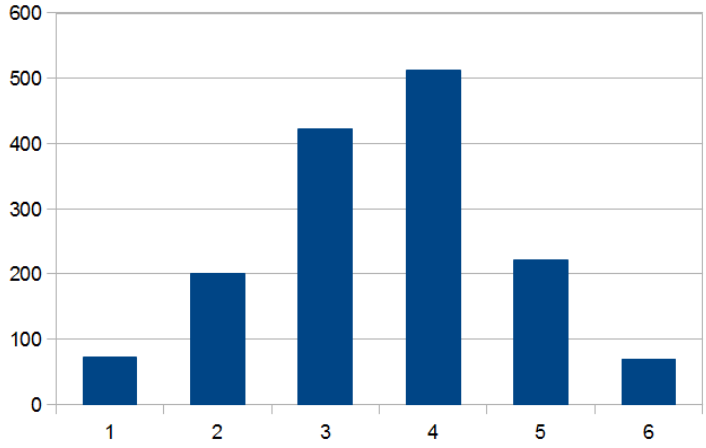
Dos observaciones adicionales: en primer lugar, puesto que los intervalos deben cubrir todo el recorrido de la variable, en un histograma no hay espacio entre las barras. Y, como práctica recomendable, para que la visualización sea efectiva, no es conveniente utilizar un histograma con más de 10 o 12 intervalos, ni con menos de cinco o seis.

En el caso de variables cuantitativas discretas, normalmente los intervalos se extienden a valores intermedios (que la variable no puede alcanzar) para que no quede espacio entre las barras del histograma.

Los pasos para obtener el histograma, en el caso en el que todos los intervalos son de la misma longitud, son estos:

1. Si no nos los dan hechos, debemos empezar por determinar los intervalos. Para ello podemos localizar el valor máximo y el mínimo de los valores, restarlos y obtenemos el *recorrido* de la variable (daremos más detalles en el Capítulo 2).

(a)



(b)

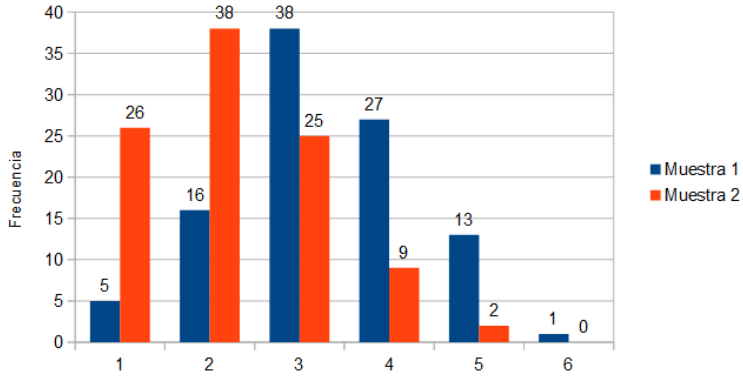


Figura 1.3: Diagrama de barras para (a) un conjunto de datos, (b) dos conjuntos de datos.

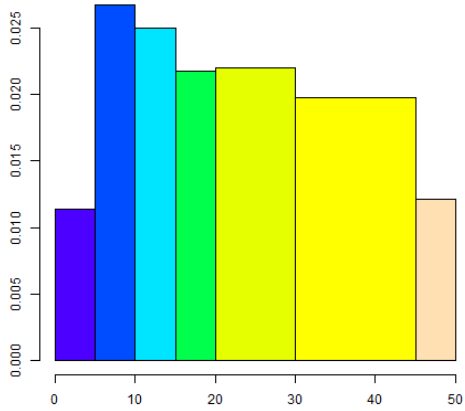


Figura 1.4: Histograma.

2. Dividimos ese recorrido entre el número de intervalos deseados, para obtener la longitud de cada uno de los intervalos. Construimos los intervalos y la tabla de frecuencias correspondiente.

3. Calculamos la altura de cada barra, teniendo en cuenta que área = base · altura, y que el área (¡no la altura!) es proporcional a la frecuencia. Por lo tanto podemos usar:

$$\text{altura} = \frac{\text{frecuencia}}{\text{base}} = \frac{\text{frecuencia del intervalo}}{\text{longitud del intervalo}}$$

   para calcular la altura de cada una de las barras.

Quizá la mejor manera de entender la propiedad más importante (y más útil) de un histograma sea viendo un *falso histograma*, un histograma mal hecho.

Ejemplo 1.2.1. *En la Tabla 1.4 se muestra la tabla de frecuencia de un conjunto de datos, agrupados por intervalos (clases). Observa que la longitud del último intervalo, el intervalo* $(8, 12]$*, es el doble de las longitudes de los restantes intervalos, que son todos de longitud* 2*.*

Clase	[0,2]	(2,4]	(4,6]	(6,8]	(8,12]
Frecuencia	1320	3231	1282	900	1105

Tabla 1.4: Datos para el Ejemplo 1.2.1

En la parte (a) de la Figura 1.5 se muestra un falso histograma, en el que la altura de las columnas se corresponde con esas frecuencias. Para un observador que no disponga de

la Tabla 1.4 (e incluso si dispone de ella, en muchos casos), la sensación que transmite ese gráfico es que el número de casos que corresponden al intervalo $(8, 12]$ *es mucho mayor que los del intervalo* $(6, 8]$*. Resulta poco claro, en esta representación gráfica, el hecho relevante de que esa frecuencia mayor se corresponde con un intervalo el doble de ancho. El sistema perceptivo humano tiende a dar más importancia a las figuras con mayor área, especialmente si sus alturas son parecidas.*

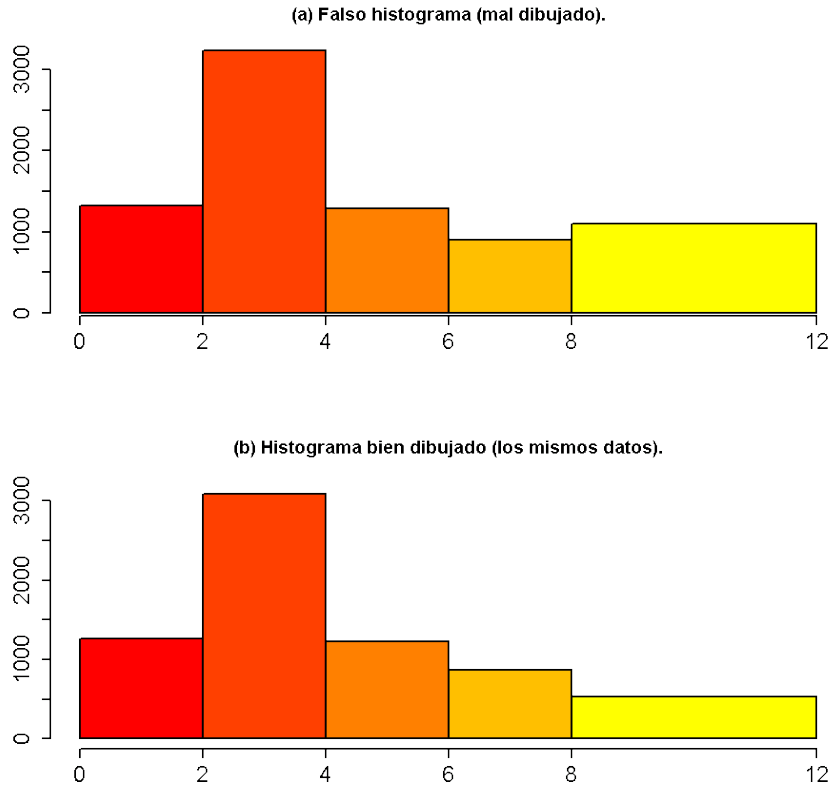


Figura 1.5: Representación de los datos del Ejemplo 1.2.1, con un (a) falso histograma (con *la altura* proporcional a la frecuencia), y (b) el histograma correcto para esos mismos datos (con *el área* proporcional a la frecuencia).

En la parte (b) de esa Figura, por otra parte, aparece el histograma correctamente dibujado. Como puede apreciarse, el hecho de hacer que sea el área de la columna lo que se corresponda con la frecuencia, ayuda a captar visualmente la importancia relativa del intervalo $(8, 12]$. De esta manera queda de manifiesto que la anchura de ese intervalo es distinta de las otras, sin sobrevalorar la frecuencia que le corresponde. □

1.3. Precisión y exactitud. Cifras significativas.

Vamos a aprovechar la sección final de este capítulo para introducir algunas herramientas de lenguaje, y procedimientos de trabajo con datos numéricos que usaremos a lo largo de todo el libro. Hemos repetido varias veces en este capítulo que la diferencia entre variables cuantitativas discretas y continuas es bastante sutil. En particular, en el caso de datos agrupados en clases (ver el apartado 1.1.3 y especialmente la discusión de la pág. 9), surge la pregunta de cómo definir el límite entre dos clases. Aunque en los tutoriales veremos cómo hacer esto en la práctica, podemos preparar el terreno. Esta cuestión, a su vez, está estrechamente ligada a la cuestión de las unidades de medida que se utilizan, y de la precisión con la que obtenemos esas medidas. Volviendo al ejemplo de los alumnos de una clase, es muy extraño pensar que alguien nos va a decir que uno de esos alumnos pesa 65.2365789 kilogramos. ¿De verdad nos creemos que tiene sentido expresar así el peso de una persona, cuando la pérdida de un sólo cabello[2] cambiaría esa cifra en una escala mucho mayor que la supuesta "precisión" de la medida? Naturalmente que no. Por esa razón, al hablar del peso de una persona lo más *práctico* es trabajar en kilos, a lo sumo en cientos o decenas de gramos. Al hacer esto, sucede algo interesante: si usamos los kilos como unidad de medida, sin preocuparnos de diferencias más finas, diremos que un alumno pesa 57 kilos y otro 58 kilos, pero no diremos nunca que pesa 55'5 o 55'32 kilos. Es decir, que al trabajar de esa manera, estaremos usando el peso *como si fuera una variable discreta*, que cambia a saltos, de kilo en kilo. El lector estará pensando ¡pero el peso ES continuo! Y lo que queremos es invitarle a descubrir que el peso no es ni continuo ni discreto. En distintos problemas usamos distintos modelos, y matemáticas distintas, para trabajar con las medidas de peso. Y la decisión sobre cuál es el modelo más adecuado depende muchas veces de la precisión y exactitud con las que deseamos trabajar.

Aprovechemos la ocasión para establecer una distinción entre las nociones de precisión y exactitud. Aunque a menudo se usan indistintamente en el lenguaje cotidiano[3], estas dos nociones tienen significados técnicos distintos. No queremos entrar en una discusión demasiado técnica, así que vamos a recurrir, para ilustrar la diferencia entre ambas nociones a la imagen, que se usa a menudo de una diana a la que estamos tratando de acertar. La idea se ilustra en la Figura 1.6 (pág. 16). Como puede verse, la idea de exactitud se relaciona con la distancia al objetivo (con el tamaño del error que se comete) y con el hecho de que esos disparos estén *centrados* en el blanco. En cambio, la idea de precisión tiene que ver con el hecho de que los disparos estén más o menos agrupados o *dispersos* entre sí.

A lo largo del curso, y muy especialmente en el próximo capítulo, vamos a tener sobradas ocasiones de volver sobre estas dos ideas. Pero ya que vamos a trabajar muy a menudo con valores numéricos, vamos a hablar del concepto de cifras significativas, que está muy relacionado con la idea de precisión de las medidas.

Todos los números que proceden de mediciones tienen una precisión limitada, ligada a menudo al propio aparato o proceso de medición. Por ejemplo, y para que no suene muy abstracto, si medimos una longitud con una regla típica, la precisión de la medida sólo llega al milímetro, porque esas son las divisiones de la escala en nuestra regla. De la misma forma un termómetro doméstico no suele afinar más allá de las décimas de grado, la balanza de

[2] Una persona tiene aproximadamente cien mil pelos en la cabeza, cada uno de unos miligramos de peso.

[3] El Diccionario de la Real Academia Española (ver enlace [3]) nos parece especialmente poco atinado en estas dos entradas...

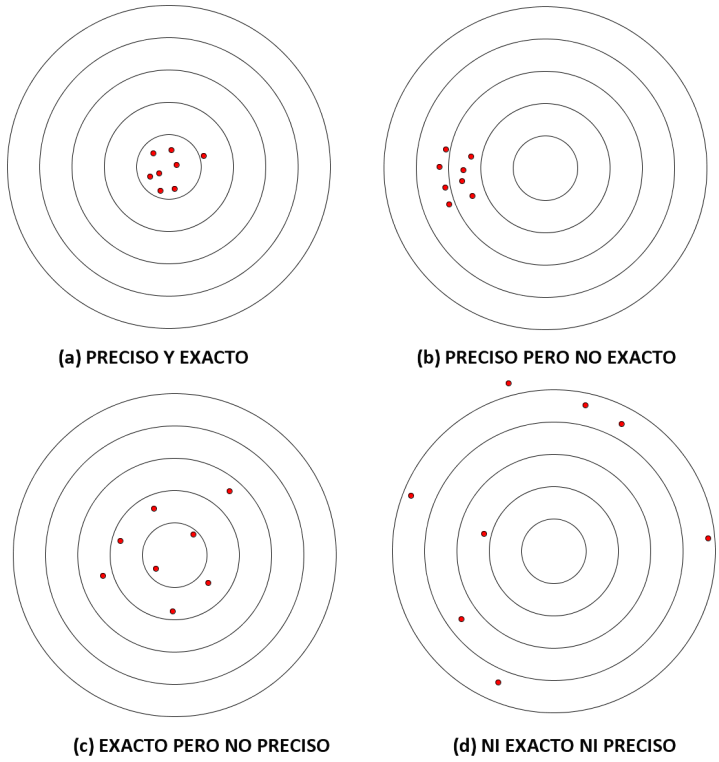


Figura 1.6: Precisión y exactitud.

cocina distingue normalmente, a lo sumo, gramos, etcétera.

Por esa razón, si hemos medido con la regla una longitud de 5cm, o sea 50mm, y lo hemos hecho teniendo cuidado de precisar hasta el milímetro, sabemos que en realidad sólo hemos sido capaces de asegurar que el valor de la longitud está entre

$$50 - 1 = 49, \qquad \text{y} \qquad 50 + 1 = 51 \text{ mm}.$$

Hasta aquí las cosas son relativamente fáciles. El problema viene, habitualmente, cuando se hacen operaciones con los resultados de las medidas. Por ejemplo, si dividimos esa longitud en tres trozos iguales, ¿cuánto medirán esos tres trozos? Si tecleamos en una calculadora 50/3 podemos terminar respondiendo algo como que esos trozos miden:

$$16.66666667 \text{ mm}.$$

Así, mediante el procedimiento mágico de aporrear las teclas de una calculadora, resulta que una medida que sólo conocíamos con una precisión de un milímetro se ha convertido en un resultado preciso casi hasta la escala atómica. Evidentemente esta no es la forma correcta de trabajar. Es necesario algún proceso de redondeo para obtener un resultado preciso.

Uno de los objetivos secundarios de este curso es proporcionar al lector una formación básica en el manejo de los números como instrumentos de comunicación científica. Vamos a

empezar, en este apartado, por familiarizarnos con la noción de cifras significativas, y poco a poco, en sucesivas visitas a este tema, iremos aprendiendo cómo se manejan correctamente situaciones como la que hemos descrito, en la que hacemos operaciones con números aproximados. Trataremos, también en esto, de darle un enfoque siempre eminentemente práctico a lo que hagamos.

Así que, en lugar de empezar tratando de definir qué son las cifras significativas, comencemos con algunos ejemplos, para ilustrar la forma de proceder. La idea intuitiva, en cualquier caso, es buscar el número con una cierta cantidad de cifras más cercano al número de partida. Precisar esta idea requiere tener en cuenta varios detalles técnicos intrincados, pero como ilustran los siguientes ejemplos el resultado es un procedimiento mecánico muy sencillo de aplicar.

Ejemplo 1.3.1. *Supongamos que nos dan el número*

$$1.623698$$

y nos piden que lo redondeemos a cuatro cifras significativas. Se trata, por tanto, de aprender a redondear un número dado, en notación decimal, a una cierta cantidad de cifras significativas (cuatro, en este ejemplo). El procedimiento es este:

1. *empezando desde la primera cifra del número (la situada más a la izquierda), buscamos la primera cifra que no sea un cero. En el ejemplo esa cifra es* 1*, la primera del número por la izquierda.*

   $$\begin{array}{cccccccc} \mathbf{1} & . & 6 & 2 & 3 & 6 & 9 & 8 \\ \uparrow & & & & & & & \end{array}$$

   Para este paso no importa la posición del punto decimal. La única duda que se puede plantear es si hay ceros a la izquierda, y ese caso lo veremos enseguida, más abajo.

2. *Como queremos cuatro cifras significativas, empezamos a contar desde esa primera cifra (inclusive) hacia la derecha, hasta llegar a cuatro cifras.*

   $$\begin{array}{cccccccc} 1 & . & 6 & 2 & 3 & 6 & 9 & 8 \\ \uparrow & & \uparrow & \uparrow & \uparrow & & & \\ 1^o & & 2^o & 3^o & 4^o & & & \end{array}$$

3. *Ahora miramos la siguiente cifra, en este caso la quinta (que es un seis). Y aplicamos esta regla de decisión: si la quinta cifra es mayor o igual que 5, sumamos 1 a la cuarta cifra, con acarreo si es necesario (veremos esto en el siguiente ejemplo). En el ejemplo,*

   $$\begin{array}{cccccccc} 1 & . & 6 & 2 & 3 & \mathbf{6} & 9 & 8 \\ & & & & & \uparrow & & \\ & & & & & 5^o & & \end{array}$$

   Como la quinta cifra es 6, y por lo tanto mayor o igual a 5, sumamos 1 a la última cifra de 1.623 (las cuatro primeras cifras no nulas del número original) y obtenemos:

   $$1.62\mathbf{4}.$$

   Este es el valor de 1.623698 *redondeado a cuatro cifras significativas.*

Veamos ahora un ejemplo más complicado, en el que entran en juego reglas adicionales de redondeo. De nuevo nos dan un número, en este caso

$$0.00337995246$$

y vamos a redondearlo, ahora a cinco cifras significativas. Aplicamos el mismo esquema:

1. *Empezando desde la primera cifra del número (la situada más a la izquierda), buscamos la primera cifra que no sea un cero. En el ejemplo esa cifra es* 3*, en realidad la cuarta cifra del número por la izquierda (la tercera después del punto decimal).*

   $$0\ .\ 0\ 0\ \mathbf{3}\ 3\ 7\ 9\ 9\ 5\ 2\ 4\ 6$$
   $$\uparrow$$

   Los ceros a la izquierda no se tienen en cuenta para el total de cifras significativas.

2. *Como queremos cinco cifras significativas, empezamos a contar desde el 3 que hemos localizado en el paso anterior, y hacia la derecha, hasta llegar a cinco cifras.*

   $$0\ .\ 0\ 0\ \mathbf{3}\ \mathbf{3}\ \mathbf{7}\ \mathbf{9}\ \mathbf{9}\ 5\ 2\ 4\ 6$$
   $$\uparrow\ \uparrow\ \uparrow\ \uparrow\ \uparrow$$
   $$1^o\ 2^o\ 3^o\ 4^o\ 5^o$$

3. *Miramos la siguiente cifra, que en este caso es un cinco.*

   $$0\ .\ 0\ 0\ 3\ 3\ 7\ 9\ 9\ \mathbf{5}\ 2\ 4\ 6$$
   $$\uparrow$$

   Como esa cifra es mayor o igual a 5, sumamos 1 a la última cifra de 0.0033799 *(la parte precedente del número original) y obtenemos:*

   $$0.0033800.$$

   Fíjate en que hemos hecho la suma con acarreo (dos acarreos, porque había dos nueves al final). Y que, al hacer esto, conservamos los ceros que aparecen a la derecha. Es importante hacer esto, porque esos ceros sí que son cifras significativas (a diferencia de los ceros de la izquierda, que no cuentan). Así que el número, redondeado a cinco cifras significativas es 0.0033800*.*

   Un último ejemplo. Hasta ahora, en los dos ejemplos que hemos visto, el proceso de redondeo ocurría a la derecha del punto decimal. Pero si nos piden que redondeemos el número 324755 *a tres cifras significativas, acabaremos con el número* 325000*. Los ceros a la derecha son, en este caso, imprescindibles. Este último ejemplo pretende ayudar a clarificar un hecho básico: el proceso de redondeo a cifras significativas,* nunca *afecta a la posición de la coma decimal en el número.*

□

Naturalmente, esta receta no agota la discusión, ni mucho menos. Para empezar, no hemos dicho nada sobre la forma de *operar* con números aproximados. Si tengo dos números con cuatro cifras significativas y los multiplico, ¿cuántas cifras significativas tiene el producto? ¿Y qué sucede si calculo la raíz cuadrada de un número con tres cifras significativas? Veamos un ejemplo sencillo, para que el lector comprenda de que se trata:

Ejemplo 1.3.2. *Tenemos los dos números*

$$\begin{cases} a = 10000 \\ b = 2.1 \end{cases}$$

y suponemos que los dos tienen dos cifras significativas, que se han redondeado usando el procedimiento que hemos descrito. En el caso de a, y con las reglas de redondeo que hemos dado esto significa que sólo podemos asegurar que a cumple:

$$10000 - 499 < a < 10000 + 499$$

Y en particular, al calcular la suma $a + b$ no tiene ningún sentido decir que es

$$a + b = 10002.1$$

porque esto parece estar diciendo que conocemos $a + b$ con mucha más precisión de la que conocemos el propio número a. Lo razonable en este caso es decir que

$$a + b \approx a$$

donde el símbolo $\approx$ se lee aproximadamente*, e indica el efecto del redondeo. Al sumar, el número b "ha desaparecido". En cambio, si multiplicamos, está claro que debe suceder algo como*

$$a \cdot b \approx 21000.$$

Y aún pueden suceder cosas peores. Imagínate que tenemos los números

$$\begin{cases} c = 43.12 \\ d = 43.11 \end{cases}$$

ambos con cuatro cifras significativas, y los restamos. ¿Cuántas cifras significativas tiene el resultado? Como puede verse, es necesario algo más de reflexión para operar acertadamente con números aproximados. □

Este tipo de preguntas tienen su respuesta detallada en una parte de las Matemáticas llamada Análisis (o Cálculo) Numérico. En general, cada operación con números aproximados supone una pérdida de precisión. Pero aquí no queremos extendernos, y vamos a dejar sin respuesta esas preguntas. Por el momento, nos basta con que el lector comprenda este procedimiento de redondeo a un número de cifras significativas dado. En la práctica, dado que usaremos el ordenador para la mayor parte de las operaciones, vamos a asumir que, en casi todos los casos, la precisión con la que trabaja la máquina es suficiente para compensar la pérdida de precisión asociada a las operaciones que hacemos. A la larga, ese punto de vista se revela como una ingenuidad, pero de momento no necesitamos más.

Capítulo 2

Valores centrales y dispersión.

Ahora que ya sabemos resumir la información de los datos en tablas y presentarlos gráficamente, vamos a dar un paso más. Vamos a sintetizar esa información en un número, que llamaremos "valor central". Una de las ideas centrales de este capítulo es que ese valor central tiene que ser un buen representante del conjunto de datos que estamos usando. También veremos que, como no podía ser de otra manera, la elección del representante adecuado depende de la tarea para la que lo vayamos a utilizar.
Pero además, una vez elegido un representante de un conjunto de datos, querremos saber cómo de representativo es ese valor central, respecto del conjunto de datos que describe. Eso nos llevará a hablar de la idea de dispersión. La dispersión es, precisamente, una medida de la calidad del valor central, como representante de un conjunto de datos. Es una noción directamente emparentada con la idea de precisión, de la que hablamos en el capítulo anterior (ver Figura 1.6 en la pág. 16).

2.1. La media aritmética.

Vamos a aprovechar este concepto, que suponemos ya conocido del lector, para introducir parte de la notación abstracta, típica de las Matemáticas, que utilizaremos a lo largo del curso. Esta sección puede servir de "chequeo preliminar" para el lector. Si tienes muchas dificultades con la notación en este punto inicial del curso, es probable que necesites reforzar tus habilidades matemáticas para poder seguir adelante. En los tutoriales 1 y 2 aprenderemos, entre otras cosas, a realizar estos cálculos en el ordenador.

2.1.1. Definición de la media aritmética.

La idea de media aritmética apenas necesita presentación. Dados n valores de una variable cuantitativa, sean $x_1, x_2, \ldots, x_n$, su media aritmética (en inglés, *arithmetic mean* o *average*) es:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (2.1)$$

Algunos comentarios sobre la notación. El símbolo $\bar{x}$ refleja la notación establecida en Estadística: la media de un vector de datos se representa con una barra sobre el nombre de ese vector. Y el símbolo $\sum_{i=1}^{n} x_i$, que suponemos que el lector ya conoce, es un sumatorio, y representa en forma abreviada, la frase "suma todos estos valores x_i donde i es un número que va desde 1 hasta n".

Insistimos en esto: la **media aritmética sólo tiene sentido para variables cuantitativas** (discretas o continuas). Aunque una variable cualitativa se represente numéricamente, la media aritmética de esos números seguramente sea una cantidad sin ningún significado estadístico.

La media aritmética es "la media" por excelencia. Pero hay otros conceptos de media que juegan un papel importante en algunos temas: la media geométrica, la media armónica, etc. Pero no las vamos a necesitar en este curso, así que no entraremos en más detalles.

Ejemplo 2.1.1. *Dado el conjunto de valores (son $n = 12$ valores)*

$$9,\ 6,\ 19,\ 10,\ 17,\ 3,\ 28,\ 19,\ 3,\ 5,\ 19,\ 2,$$

su media aritmética es:

$$\bar{x} = \frac{9+6+19+10+17+3+28+19+3+5+19+2}{12} =$$

$$= \frac{140}{12} \approx 11.67,$$

(cuatro cifras significativas). Proponemos al lector como ejercicio que piense si el número $\bar{x} = 11.67$ se puede considerar, en este caso, un buen representante de este conjunto de datos. □

El siguiente ejemplo sirve para presentar una característica de la media aritmética que debemos tener siempre presente:

Ejemplo 2.1.2. *Ahora consideramos el mismo conjunto de valores, al que añadimos el número* 150 *(en la última posición, aunque su posición es irrelevante para lo que sigue):*

$$9,\ 6,\ 19,\ 10,\ 17,\ 3,\ 28,\ 19,\ 3,\ 5,\ 19,\ 2,\ 150$$

La media aritmética ahora es:

$$\bar{x} = \frac{9+6+19+10+17+3+28+19+3+5+19+2+150}{13} =$$

$$= \frac{290}{13} \approx 22.31,$$

(con cuatro cifras significativas). ¿Sigue siendo posible, en este caso, considerar a la media aritmética $\bar{x} = 22.31$ como un buen representante de los datos? Por ejemplo, si elegimos al azar uno cualquiera de esos números, ¿es de esperar que se parezca a la media? □

Volveremos sobre la pregunta que plantean estos ejemplos en la Sección 2.2 (pág. 25). Pero antes vamos a pensar un poco más sobre la forma de calcular la media aritmética, si los datos vienen descritos mediante una tabla de frecuencias.

2.1.2. La media aritmética a partir de una tabla de frecuencias.

Supongamos que tenemos una tabla de frecuencias de unos valores, correspondientes a una variable cuantitativa. Es decir, una tabla como esta :

Valor	**Frecuencia**
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_k	f_k

y queremos calcular la media aritmética a partir de esta tabla.

Aquí los valores *distintos* de la variable[1] son $x_1, \ldots, x_k$ y sus frecuencias absolutas respectivas son $f_1, f_2, \ldots, f_k$. Está claro entonces que:

$$f_1 + f_2 + \cdots + f_k = (\text{núm. de observ. de } x_1) + \cdots + (\text{núm. de observ. del valor } x_k) =$$

$$= (\text{suma del número de observaciones de todos los valores distintos}) = n$$

Recordemos que para calcular la media tenemos que sumar el valor de todas (las n observaciones). Y como el valor x_i se ha observado f_i veces, su contribución a la suma es

$$x_i \cdot f_i = x_i + x_i + \cdots + x_i \quad (\text{sumamos } f_i \text{ veces})$$

Teniendo en cuenta la contribución de cada uno de los k valores distintos, vemos que para calcular la media debemos hacer:

$$\bar{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \cdots + x_k \cdot f_k}{f_1 + f_2 + \cdots + f_k} = \frac{\displaystyle\sum_{i=1}^{k} x_i \cdot f_i}{\displaystyle\sum_{i=1}^{k} f_i}.$$

Ejemplo 2.1.3. *En una instalación deportiva el precio de la entrada para adultos es de 10€ y de 4€ para menores. Hoy han visitado esa instalación* 230 *adultos y* 45 *menores. ¿Cuál es el ingreso medio por visitante que recibe esa instalación?*
Tenemos dos posibles valores de la variable x =“precio de la entrada”, *que son* $x_1 = 10$ *y* $x_2 = 4$. *Además sabemos las frecuencias correspondientes:* $f_1 = 230$ *y* $f_2 = 45$. *Por lo tanto:*

$$\bar{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2}{f_1 + f_2} = \frac{10 \cdot 230 + 4 \cdot 45}{230 + 45} = 9.02$$

El ingreso medio es de 9.02€ por visitante. □

[1] Acuérdate de que tenemos n observaciones de la variable, pero puede haber valores repetidos. Aquí estamos usando el número de valores distintos, sin repeticiones, y ese número es k.

2.1.3. Media aritmética con datos agrupados.

Si lo que queremos es calcular la media aritmética a partir de la tabla de frecuencias agrupadas por intervalos de una variable cuantitativa (ver el final de la Sección 1.1.3), las cosas son (sólo un poco) más complicadas. En este caso vamos a tener una tabla de frecuencias por intervalos (recuerda que los intervalos a veces se llaman también clases) como esta:

Intervalo	**Frecuencia**
$[a_1, b_1)$	f_1
$[a_2, b_2)$	f_2
$\vdots$	$\vdots$
$[a_k, b_k)$	f_k

Comparando esta tabla con el caso anterior está claro que lo que nos falta son los valores $x_1, \ldots, x_k$ y, en su lugar, tenemos los intervalos $[a_1, b_1), \ldots, [a_k, b_k)$. Lo que hacemos en estos casos es *fabricar* unos valores x_i a partir de los intervalos. Se toma como valor x_i el punto medio del intervalo $[a_i, b_i)$; es decir:

Marcas de clase

$$x_i = \frac{a_i + b_i}{2}, \quad \text{para } i = 1, \ldots, n. \tag{2.2}$$

Estos valores x_i se denominan marcas de clase (o marcas de intervalo). Una vez calculadas las marcas de clase, podemos usar la misma fórmula que en el caso anterior.

Ejemplo 2.1.4. *La Tabla 2.1.4 muestra la tabla de frecuencias de un conjunto de* 100 *datos agrupado por clases. En la última columna se muestran, además, las correspondientes marcas de clase.*

Clase	Frecuencia	Marca de clase
[0,4)	3	2
[4,8)	27	6
[8,12)	32	10
[12,16)	25	14
[16,20)	7	18
[20,24)	2	22
[24,28]	4	26

Tabla 2.1: Tabla de valores agrupados por clases del Ejemplo 2.1.4

A partir de la Tabla 2.1.4 es fácil calcular la media aritmética usando la Ecuación 2.2:

$$\bar{x} = \frac{3 \cdot 2 + 27 \cdot 6 + 32 \cdot 10 + 25 \cdot 14 + 7 \cdot 18 + 2 \cdot 22 + 4 \cdot 26}{100} = \frac{1112}{100} = 11.12$$

El fichero cap02-EjemploMediaAritmetica-ValoresAgrupadosClases.csv *contiene los* 100 *datos originales, sin agrupar por clases. Con los métodos que aprenderemos en los tutoriales es posible comprobar que la media aritmética de esos datos, calculada directamente, es, con seis cifras significativas, igual a* 11.1158*. Así que, por un lado vemos que la media calculada a partir de los datos agrupados no coincide con la media real. Pero, por otro lado, en ejemplos como este, el error que se comete al agrupar es relativamente pequeño.* □

2.2. Mediana, cuartiles, percentiles y moda.

Aunque la media aritmética es el valor central por excelencia, no siempre es la que mejor refleja el conjunto de datos al que representa. La razón es, como hemos comprobado en el Ejemplo 2.1.2 (pág. 22), que la media es muy sensible a la presencia de valores mucho más grandes (o mucho más pequeños, tanto da) que la mayoría de los valores. Un nuevo ejemplo puede ayudar a reafirmar esta idea:

Ejemplo 2.2.1. *Examinemos esta afirmación con un ejemplo muy sencillo. Los conjuntos de datos*

$$\{1, 2, 3, 4, 35\} \qquad y \qquad \{7, 8, 9, 10, 11\}$$

tienen la misma media, que vale 9*. Sin embargo, en el primer caso, el de la izquierda, casi todos los valores son menores o iguales que* 4*, y el hecho de que aparezca un dato anormalmente alto, el* 35*, aleja la media del grueso de los datos. No ocurre así con la segunda serie de datos. Si jugamos con los números, pueden darse muchas situaciones diferentes.* □

Este ejemplo busca ponernos en guardia y motivar los conceptos que vamos a ver continuación.

2.2.1. Mediana.

Como en el caso de la media aritmética, vamos a suponer que tenemos n observaciones de una variable cuantitativa

$$x_1, x_2, \ldots, x_n.$$

y suponemos que los datos no están agrupados en una tabla de frecuencia. Más abajo veremos el caso de datos agrupados.

Como los x_i son números, vamos a suponer que los hemos ordenado de menor a mayor:

$$x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n.$$

Entonces, la mediana (en inglés, *median*) de ese conjunto de datos es el *valor central* de esa serie ordenada. Es decir:

> **Caso impar:** si tenemos una cantidad impar de datos, sólo hay un valor central, y ese es la mediana. Por ejemplo, para siete datos:
>
> $$\underbrace{x_1 \leq x_2 \leq x_3}_{\text{mitad izda.}} \leq \underset{\substack{\uparrow \\ \text{mediana}}}{\mathbf{X}_4} \leq \underbrace{x_5 \leq x_6 \leq x_7}_{\text{mitad dcha.}}$$

Caso par: por contra, si el número de datos es par, entonces tomamos el valor máximo de la mitad izquierda, y el valor mínimo de la mitad derecha y hacemos la media. Por ejemplo, para seis datos:

$$\underbrace{x_1 \leq x_2 \leq x_3}_{\text{mitad izda.}} \leq \underset{\substack{\uparrow \\ \text{mediana}}}{\frac{x_3 + x_4}{2}} \leq \underbrace{x_4 \leq x_5 \leq x_6}_{\text{mitad dcha.}}$$

En el caso de un número impar de datos la mediana siempre coincide con uno de los datos originales. Pero en el caso de un número par de datos la mediana pueden darse los dos casos.

Ejemplo 2.2.2. *Por ejemplo, si tenemos estos seis datos ordenados:*

$$2 \leq 5 \leq 6 \leq 7 \leq 11 \leq 15,$$

Entonces la mediana es 6.5

$$2 \leq 5 \leq 6 \leq \mathbf{6.5} \leq 7 \leq 11 \leq 15,$$

que no aparecía en el conjunto original (fíjate en que, como pasaba con la media aritmética, aunque todos los datos originales sean enteros, la mediana puede no serlo). Mientras que si tenemos estos seis datos, con los dos datos centrales iguales:

$$2 \leq 5 \leq 6 \leq 6 \leq 11 \leq 15,$$

Entonces la mediana es 6*,*

$$2 \leq 5 \leq 6 \leq \mathbf{6} \leq 6 \leq 11 \leq 15,$$

que ya estaba (repetido) entre los datos originales. □

¿Qué *ventajas* aporta la mediana frente a la media aritmética? Fundamentalmente, la mediana se comporta mejor cuando el conjunto de datos contiene datos atípicos (en inglés, *outliers*). Es decir, datos cuyo valor se aleja *mucho* de la media. Todavía no podemos precisar esto porque para hacerlo necesitamos un poco más de vocabulario que vamos a ver enseguida. Pero la idea intuitiva es que si tenemos un conjunto de datos, e introducimos un dato adicional que se aleja mucho de la media aritmética inicial, entonces en el nuevo conjunto de datos podemos tener una media aritmética bastante distinta de la inicial. En cambio la mediana sufre modificaciones mucho menores frente a esos datos atípicos. Podemos hacernos una primera impresión con un par de ejemplos, basados en conjuntos de datos que ya hemos examinado antes.

Ejemplo 2.2.3. *En el Ejemplo 2.1.2 (pág. 22) hemos visto que la media aritmética del conjunto de datos:*

$$9,\ 6,\ 19,\ 10,\ 17,\ 3,\ 28,\ 19,\ 3,\ 5,\ 19,\ 2,\ 150$$

es

$$\bar{x} = \frac{290}{13} \approx 22.31.$$

Y, al comparar este resultado con el del Ejemplo 2.1.1, hemos concluido que la presencia del valor 150 *(que es atípico, como veremos), tenía un efecto muy grande en la media aritmética, hasta el punto de hacerla poco útil como representante del conjunto de datos. Para calcular la mediana, empezamos por ordenar los datos de menor a mayor:*

$$2,\ 3,\ 3,\ 5,\ 6,\ 9,\ 10,\ 17,\ 19,\ 19,\ 19,\ 28,\ 150.$$

Puesto que son 13 *números, la mediana es el valor que ocupa la séptima posición; es decir, la mediana vale* 10*. Y como se ve, es mucho más representativa de la mayoría de los números de este conjunto.*

Además, veamos lo que sucede si eliminamos el valor 150*, para volver al conjunto de datos del Ejemplo 2.1.1 y, después de eliminarlo, volvemos a calcular la mediana. Los datos restantes, ordenados, son estos* 12 *números:*

$$2,\ 3,\ 3,\ 5,\ 6,\ 9,\ 10,\ 17,\ 19,\ 19,\ 19,\ 28.$$

Y ahora la mediana será la media entre los números de la sexta y séptima posiciones. Por lo tanto la mediana es 9.5*. Como puede verse, el cambio en la mediana, debido a la presencia de* 150*, es muy pequeño, comparado con el que sufre la media aritmética. Y, de hecho, si sustituimos* 150 *por un valor aún más exagerado, como* 2000*, veremos que la mediana cambia exactamente igual.*

Como pone de manifiesto este ejemplo, *la mediana no atiende a tamaños, sino a posiciones.* Eso la hace muy adecuada para representar un conjunto de valores del que sospechamos que puede contener valores con tamaños muy alejados de los de la mayoría.

Y entonces, ¿por qué no se usa siempre la mediana en lugar de la media aritmética? La respuesta es que la Estadística basada en la mediana utiliza unas matemáticas bastante más complicadas que la que se basa en la media aritmética. En años recientes, a medida que el ordenador ha ido convirtiéndose en una herramienta más y más potente, la importancia de los métodos basados en la mediana ha ido aumentado en paralelo. Pero los métodos que usan la media aritmética, que dominaron la Estadística clásica, siguen siendo los más comunes.

Mediana y tablas de frecuencias relativas y acumuladas.

Puede darse el caso de que queramos calcular la mediana a partir de una tabla de frecuencias. Empecemos suponiendo que se trata de valores no agrupados. Para obtener la mediana vamos a tener que dar un pequeño rodeo, e introducir un par de conceptos nuevos. Concretando, vamos a utilizar las nociones de frecuencia relativa y frecuencia acumulada.

Si tenemos una tabla de datos $x_1, \ldots, x_k$ (estos son los valores distintos), con frecuencias $f_1, \ldots, f_k$, de manera que

$$f_1 + \cdots + f_k = n$$

es el número total de datos, entonces las frecuencias relativas se definen mediante:

$$f'_1 = \frac{f_1}{n}, f'_2 = \frac{f_2}{n}, \ldots, f'_k = \frac{f_k}{n}.$$

Por lo tanto las frecuencias relativas son un "tanto por uno", y se convierten fácilmente en porcentajes multiplicando por 100. Veamos un ejemplo.

Ejemplo 2.2.4. *La Tabla 2.2 muestra, en las dos primeras columnas, la tabla de frecuencias absolutas de un conjunto de valores (del* 1 *al* 6*). En la última columna aparecen las frecuencias relativas. En este ejemplo las cosas son especialmente fáciles, porque la suma de las frecuencias absolutas es* 100*. Así que cada frecuencia relativa se limita a traducir en un tanto por uno el porcentaje del total de datos que representa cada valor. Así, por ejemplo, vemos que el* 31 % *de los valores son iguales a* 4*.*

Valor x_i	Frecuencia absoluta f_i	Frecuencia relativa f_i'.
1	2	0.02
2	25	0.25
3	31	0.31
4	31	0.31
5	8	0.08
6	3	0.03
Suma	100	1

Tabla 2.2: Tabla de frecuencias relativas del Ejemplo 2.2.4

Para que sirva de comparación, en la Tabla 2.3 tienes otra tabla de frecuencias absolutas y relativas (redondeadas, estas últimas, a dos cifras significativas). En este caso, el número de datos (la suma de frecuencias absolutas) es 84*. Así que para obtener las frecuencias relativas hay que usar la fórmula:*

$$f_i' = \frac{f_i}{n}.$$

Con esto, por ejemplo,

$$f_3 = \frac{24}{84} \approx 0.29$$

(con dos cifras significativas). Este resultado nos informa de que el valor 3 *aparece en aproximadamente el* 29 % *de los datos.*

Valor x_i	Frecuencia absoluta f_i	Frecuencia relativa f_i' (aprox).
1	20	0.24
2	29	0.35
3	24	0.29
4	9	0.11
5	2	0.02
Sum	84	1

Tabla 2.3: Otra tabla de frecuencias relativas para el Ejemplo 2.2.4. Frecuencias relativas redondeadas a dos cifras significativas.

□

Las frecuencias relativas, como ilustran esos ejemplos, sirven, por tanto, para responder fácilmente a preguntas como "¿qué porcentaje de los datos tiene el valor x_2?". Además, es

importante darse cuenta de que la suma de todas las frecuencias relativas siempre es 1:

$$f'_1 + \cdots + f'_k = \frac{f_1 + \cdots + f_k}{n} = \frac{n}{n} = 1.$$

Conviene observar que, puesto que son simplemente un recuento, las frecuencias relativas se pueden usar con cualquier tipo de variable (cualitativa o cuantitativa).

¿Qué son las frecuencias acumuladas (en inglés, *cumulative frequencies*)? Este tipo de frecuencias sólo son útiles para variables cuantitativas, que además vamos a suponer ordenadas, de forma que los valores (distintos) del conjunto de datos cumplen:

$$x_1 < x_2 < \ldots < x_k.$$

En tal caso, las frecuencias acumuladas se definen así:

$$f''_1 = f_1, \quad f''_2 = f_1 + f_2, \quad f''_3 = f_1 + f_2 + f_3, \text{ etc., hasta } f''_k = f_1 + f_2 + \cdots + f_k.$$

Es decir, cada frecuencia absoluta es la suma de todas las frecuencias (ordinarias) precedentes. Veamos, de nuevo, un par de ejemplos.

Ejemplo 2.2.5. *La Tabla 2.4, que usa el mismo conjunto de datos que en la Tabla 2.2 del Ejemplo 2.2.4, muestra, en la última columna, la tabla de frecuencias acumuladas de ese conjunto de valores.*

Valor x_i	Frecuencia absoluta f_i	Frecuencia acumulada f''_i.
1	2	2
2	25	27=2+25
3	31	58=27+31=2+25+31
4	31	89=58+31=2+25+31+31
5	8	97=89+8=2+25+31+31+8
6	3	100=97+3=2+25+31+31+8+3
Suma	100	**373**
		↑
		¡¡Esta suma es inútil!!

Tabla 2.4: Tabla de frecuencias acumuladas del Ejemplo 2.2.5

Junto a cada frecuencia acumulada f'_i se muestra cómo se ha obtenido, sumando todos los valores precedentes de la tabla. O, de forma alternativa, y más eficiente, sumando la frecuencia absoluta f_i con la frecuencia acumulada de la fila anterior f'_{i-1}. Como se ve, la última frecuencia acumulada coincide con n, el número total de datos, que es la suma de las frecuencias absolutas (y que en este ejemplo resulta ser 100*, pero que, desde luego, puede ser cualquier valor). Hemos incluido, destacada, la suma de las frecuencias absolutas, pero sólo para dejar claro que esa suma carece de sentido.* Acumular ya es sumar*, así que no tiene sentido volver a sumar lo que ya hemos sumado.*

Para el segundo conjunto de datos del Ejemplo 2.2.4, los de la Tabla 2.3, se obtienen las frecuencias acumuladas de la Tabla 2.5.

Valor x_i	Frecuencia absoluta f_i	Frecuencia acumulada f_i''.
1	20	20
2	29	49=20+29
3	24	73=49+24
4	9	82=73+9
5	2	84=82+2
Suma	84	

Tabla 2.5: Tabla de frecuencias acumuladas para los datos de la Tabla 2.3

Esta vez sólo hemos calculado las frecuencias relativas por el método más eficiente, y no hemos incluido la suma de las frecuencias absolutas, porque, como ya hemos dicho, carece de sentido.

□

Las frecuencias acumuladas sirven para contestar preguntas como, por ejemplo, "¿cuántos, de los datos, son menores o iguales a x_3?". La respuesta a esa pregunta sería f_3''. Para que esto funcione, está claro que los datos tienen que estar ordenados. La última de estas frecuencias acumuladas siempre es igual a n, el número total de datos:

$$f_1'' + \cdots + f_k'' = n.$$

Además, estas frecuencias acumuladas satisfacen otra propiedad, de recursividad, que hemos usado en el Ejemplo 2.2.5 para calcularlas, y que nos resultará muy útil a la hora de calcularlas. Se tiene que:

$$f_1'' = f_1, \quad f_2'' = f_2 + f_1'', \quad f_3'' = f_3 + f_2'', \ldots, f_k'' = f_k + f_{k-1}''.$$

Es decir, cada frecuencia acumulada se obtiene sumando la correspondiente frecuencia absoluta con la frecuencia acumulada precedente.

Para volver al cálculo de la mediana, y otras medidas de posición como los percentiles, tenemos que combinar ambas ideas, definiendo las que se conocen como frecuencias relativas acumuladas (en inglés, *relative cumulative frequencies*), o de forma equivalente, las frecuencias acumuladas relativas (porque es indiferente acumular primero y dividir después por el total, o empezar calculando las frecuencias relativas, y después acumularlas).

Se definen así (mostramos varias expresiones equivalentes):

$$\begin{cases} f_1''' = \dfrac{f_1}{n} = \dfrac{f_1''}{n} = f_1' \\ f_2''' = \dfrac{f_1 + f_2}{n} = \dfrac{f_1''}{n} = f_1' + f_2' \\ f_2''' = \dfrac{f_1 + f_2 + f_3}{n} = \dfrac{f_3''}{n} = f_1' + f_2' + f_3' \\ \quad \vdots \\ f_n''' = \dfrac{f_1 + f_2 + \cdots + f_n}{n} = \dfrac{f_n''}{n} = f_1' + f_2' + \cdots + f_n' \end{cases} \tag{2.3}$$

Veamos un ejemplo:

Ejemplo 2.2.6. *Para el segundo conjunto de datos del Ejemplo 2.2.4, los de las Tablas 2.3 y 2.5, se obtienen estas frecuencias relativas acumuladas de la Tabla 2.6.*

Valor x_i	Frec. absoluta f_i	Frec. relativa f_i'.	F. acumulada relativa f_i'''.
1	20	0.24	0.24
2	29	0.35	$0.59 \approx 0.24 + 0.35$
3	24	0.29	$0.87 \approx 0.58 + 0.29$
4	9	0.11	$0.98 \approx 0.87 + 0.11$
5	2	0.02	$1 \approx 0.98 + 0.02$
Suma	84	1	

Tabla 2.6: Tabla de frecuencias acumuladas relativas (o relativas acumuladas) para los datos de la Tabla 2.3

□

Las frecuencias relativas acumuladas son, en definitiva, los *tantos por uno acumulados*. Y por lo tanto sirven para contestar una pregunta que es la combinación de las dos que hemos visto: "¿qué porcentaje de valores es menor o igual que x_k?" Ahora debería estar clara la relación con la mediana. Si localizamos, en la tabla de frecuencias relativas acumuladas, el primer valor para el que la frecuencia relativa acumulada es mayor o igual que 1/2, habremos localizado la mediana de los datos.

Ejemplo 2.2.7. (Continuación del Ejemplo 2.2.6) *Un vistazo a la Tabla 2.2.6 nos muestra que el menor valor para el que la frecuencia relativa acumulada es mayor o igual a* 1/2 *es el valor* 2*. Por lo tanto, la mediana de ese conjunto de datos es* 2*.* □

2.2.2. La mediana en el caso de datos cuantitativos agrupados en intervalos.

¿Y si lo que necesitamos es calcular la mediana a partir de la tabla de frecuencias de una variable cuantitativa agrupada en intervalos? En este caso, el método que se utiliza para definir la mediana es algo más complicado. Nos viene bien, para entender lo que sucede, la idea de histograma. Con ayuda de la noción de histograma podemos definir así la mediana: es el valor de la variable (por lo tanto es el punto del eje horizontal) que divide el histograma en dos mitades con el mismo área. Existen fórmulas para calcular la mediana en estos casos (usando un método matemático que se conoce como interpolación) pero aquí no nos vamos a entretener en los detalles técnicos. Preferimos insistir en que el cálculo de la mediana, en este caso, es más complicado de lo que, ingenuamente, podría esperarse. Tenemos por un lado una idea informal de lo que debe ser la mediana: un valor que divide a los datos en dos mitades "del mismo tamaño". El problema es que la forma de medir el "tamaño" de las dos mitades, en la práctica, es mediante el área que representan en el histograma. Y, para empezar, el propio histograma depende de la forma en la que hemos agrupado los datos, así que como se ve hay mucho margen de maniobra en esa "definición".

Vamos a ver, a continuación, algunas otras situaciones parecidas: tenemos una noción informal, intuitiva, de lo que significa cierto valor, pero cuando llega el momento de calcularlo, descubriremos que los detalles del cálculo son complicados.

2.2.3. Cuartiles y percentiles.

Hemos visto que la mediana es, intuitivamente, el valor que deja a la mitad de los datos a cada lado. Esta idea se puede generalizar fácilmente, mientras nos movamos en el terreno de la intuición: el valor que deja al primer cuarto de los datos a su izquierda es el primer cuartil de ese conjunto de datos. Dicho de otra forma: la mediana divide a los datos en dos mitades, la mitad izquierda y la mitad derecha. Pues entonces el primer cuartil es la mediana de la mitad izquierda. Y de la misma forma el tercer cuartil es la mediana de la mitad derecha. Y, por tanto, es el valor que deja a su derecha al último cuarto de los datos. Por si el lector se lo está preguntando, sí, la mediana se puede considerar como el segundo cuartil (aunue pocas veces la llamaremos así, claro), y de hecho la mayor parte de los programas estadísticos de ordenador permiten calcular un segundo cuartil, que coincide siempre con la mediana. Veremos varios ejemplos de este tipo de cálculos en los tutoriales.

Otra forma de ver esto es que los cuartiles (incluyendo la mediana entre ellos) son los valores que señalan la posición del 25 %, el 50 % y el 75 % de los datos. Por esa razón se denomina a estos valores como medidas de posición.

Llegados a este punto, es fácil generalizar aún más la idea de los cuartiles, que ya son una generalización de la idea de mediana. Como hemos dicho, el primer cuartil deja a su izquierda el 25 % de los datos. Si pensamos en el valor que deja a su izquierda el 10 % de los datos, estaremos pensando en un percentil, concretamente en el percentil 10. Los percentiles se suelen dar en porcentajes, pero también en tantos por uno, es decir en números comprendidos entre 0 y 1.

El cálculo de los cuartiles y percentiles, en casos prácticos, plantea los mismos problemas que el de la mediana. Hay muchas formas posibles de medir el "tamaño" de las partes en que un percentil divide a los datos, más allá del mero hecho de contarlos. Como el propio nombre indica, queremos un valor que nos de una medida posicional. Es bueno, para entender que hay varias posibilidades, pensar en el ejemplo de una balanza clásica, con dos platos que han de equilibrase. Y pensemos en los datos como si fueran unas monedas que colocamos en esos platos. Podríamos pensar que el equilibrio se alcanza cuando los dos platos tienen el mismo número de monedas. Y esa sería una noción de equilibrio que se obtendría *simplemente contando*. Pero al pensar así, damos por sentado que todas las monedas son iguales. ¿Y si todas las monedas del plato izquierdo son más grandes que las del derecho? Entonces la balanza no estará en equilibrio, aunque los números sean iguales. Y hay otras posibilidades: supongamos que los dos brazos de la balanza no son de la misma longitud. Entonces aunque las monedas sean iguales, y haya el mismo número en ambos platos, seguiremos sin alcanzar el equilibrio... Todos estos ejemplos pretenden transmitir la idea de que, cuando descendemos a los detalles, las medidas de posición se tienen que definir con una idea clara de lo que se espera de ellas. No hay una definición universal, sino distintos métodos para problemas distintos. En el programa R, por ejemplo, se pueden encontrar hasta nueve métodos distintos de cálculo. El artículo [HF96], contiene mucha información, bastante técnica, sobre este problema. Nosotros, por el momento, nos vamos a conformar con la idea intuitiva de lo que significan, y en los tutoriales veremos cómo calcularlos con el ordenador.

2.2.4. Moda.

La media aritmética y la mediana se utilizan exclusivamente para variables cuantitativas. La moda en cambio puede utilizarse además con variables de tipo cualitativo (y es,

de los que vamos a ver, el único tipo de valor promedio que puede usarse con variables cualitativas). La moda de una serie de valores agrupados en una tabla de frecuencias es el valor con la frecuencia más alta.

Puesto que puede haber dos o más valores que tengan la misma frecuencia, hay conjuntos de datos que tienen más de una moda. Hablaremos en este caso de conjuntos de datos unimodales, bimodales, etcétera. Por ejemplo, en la Figura 2.1 se muestra el histograma de un conjunto de datos bimodal, con dos cumbres de aproximadamente la misma altura, El

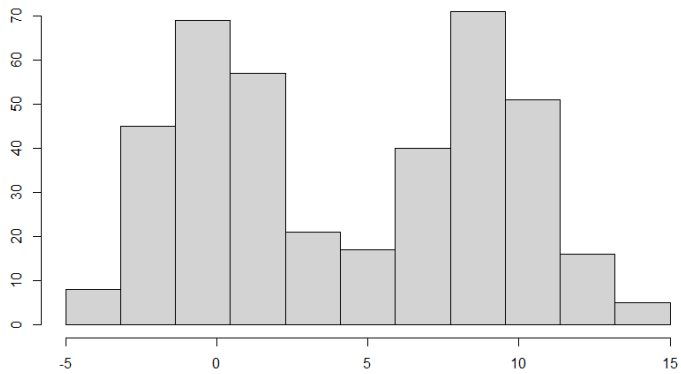


Figura 2.1: Un conjunto de datos bimodal.

cálculo de la moda (o modas) es inmediato, a partir de las tablas de frecuencias, y en los tutoriales comentaremos brevemente cómo realizarlo.

2.3. Medidas de dispersión.

Hasta ahora hemos estado calculando *valores centrales*, que nos sirvieran como buenos representantes de una colección de datos. Sin embargo, es fácil entender que hay muchas colecciones de datos, muy distintas entre sí, que pueden tener la misma media aritmética o la misma mediana, etcétera. El mismo representante puede corresponder a colecciones de datos con *formas* muy diferentes.

Por lo tanto, no sólo necesitamos un valor representativo, además necesitamos una forma de medir *la calidad de ese representante.* ¿Cómo podemos hacer esto? La idea que vamos a utilizar es la de dispersión. Una colección de números es poco dispersa cuando los datos están muy concentrados alrededor de la media. Dicho de otra manera, si los datos son poco dispersos, entonces se parecen bastante a la media (o al representante que estemos usando). En una colección de datos poco dispersos, la *distancia típica* de uno de los datos al valor central es pequeña.

Esa es la idea intuituiva, y como se ve está muy relacionada con el concepto de *precisión* del que hablamos en la Sección 1.3 (ver la Figura 1.6, página 16). Pero ahora tenemos que

concretar mucho más si queremos definir un valor de la dispersión que se pueda calcular. ¿cómo podemos medir eso? En esta sección vamos a introducir varios métodos de medir la dispersión de una colección de datos.

2.3.1. Recorrido (o rango) y recorrido intercuartílico.

La idea más elemental de dispersión es la de recorrido, que ya hemos encontrado al pensar en las representaciones gráficas. El recorrido es simplemente la diferencia entre el máximo y el mínimo de los valores. Es una manera rápida, pero excesivamente simple, de analizar la dispersión de los datos, porque depende exclusivamente de dos valores (el máximo y el mínimo), que pueden ser muy poco representativos. No obstante, es esencial, como primer paso en el estudio de una colección de datos, empezar por calcular el recorrido, porque nos ayuda a *enmarcar* nuestro trabajo, y evitar errores posteriores.

Un comentario sobre la terminología. El recorrido se denomina a veces, rango. Por razones que quedarán más claras en el Apéndice A (donde usaremos *rango* para otra noción distinta), nosotros preferimos el término *recorrido* para este concepto. La confusión se debe a la traducción como *rango* de las dos palabras inglesas *range*, que nosotros traducimos como *recorrido*, y *rank*, que traducimos como *rango*.

Si queremos ir un paso más allá, para empezar a entender la forma de los datos, podemos usar las medidas de posición. En concreto, la mediana y los cuartiles se pueden utilizar para medir la dispersión de los datos, calculando el recorrido intercuartílico (en inglés, *interquartile range*, IQR) , que se define como la diferencia entre el tercer y el primer cuartil.

IQR, recorrido intercuartílico.

El recorrido intercuartílico es:

$$IQR = (\text{tercer cuartil}) - (\text{primer cuartil})$$

Ejemplo 2.3.1. *Para el conjunto de datos del Ejemplo 2.1.2, que eran estos:*

$$9,\ 6,\ 19,\ 10,\ 17,\ 3,\ 28,\ 19,\ 3,\ 5,\ 19,\ 2,\ 150$$

el programa de ordenador (R, en este ejemplo) nos dice que el primer cuartil es 5*, y que el tercer cuartil es* 19*. Por lo tanto,*

$$IQR = 19 - 5 = 14.$$

□

Los datos que son mucho menores que el primer cuartil o mucho mayores que el tercer cuartil se consideran valores atípicos (en inglés, *outlier*). ¿Cómo de lejos tienen que estar de los cuartiles para considerarlos *raros o excepcionales*? La forma habitual de proceder es considerar que un valor mayor que el tercer cuartil, y cuya diferencia con ese cuartil es mayor que 1.5 veces el recorrido intercuartílico es un valor atípico. De la misma forma, también es un valor atípico aquel valor menor que el tercer cuartil, cuya diferencia con ese cuartil es mayor que 1.5·IQR. Ya hemos discutido que existen muchas formas distintas de definir los cuartiles, así que el recorrido intercuartílico depende, naturalmente, del método que se use para calcular los cuartiles. Nosotros siempre lo calcularemos usando el ordenador (con R, la hoja de cálculo o algún otro programa), y nos conformaremos con los valores por defecto que producen esos programas.

Ejemplo 2.3.2. *Como habíamos anunciado, vamos a ver que, para el conjunto de datos del Ejemplo 2.1.2, el valor* 150 *es un valor atípico. En el Ejemplo 2.3.1 hemos visto que el tercer cuartil de esos valores era* 19*, y que el recorrido intercuartílico era* 14*. Así que un valor será atípico si es mayor que*

$$\textit{(tercer cuartil)} + 1.5 \cdot IQR = 19 + 1.5 \cdot 14 = 19 + 21 = 40.$$

Desde luego, queda claro que 150 *es un valor atípico, en ese conjunto.* □

La mediana, los cuartiles y el recorrido intercuartílico se utilizan para dibujar los diagramas llamados de caja y bigotes (en inglés, *boxplot*), como el que se muestra en la Figura 2.2. En estos diagramas se dibuja una caja cuyos extremos son el primer y tercer cuartiles. Dentro de esa caja se dibuja el valor de la mediana. Los valores atípicos se suelen mostrar como puntos individuales (fuera de la caja, claro), y finalmente se dibujan segmentos que unen la caja con los datos más alejados que no son atípicos. Hasta hace muy poco, las hojas de cálculo no ofrecían la posibilidad de dibujar diagramas de cajas, y de hecho, nosotros recomendamos utilizar programas especializados para dibujarlos. Aprenderemos a hacerlo en el Tutorial02, donde también veremos como calcular el recorrido intercuartílico.

2.3.2. Varianza y desviación típica.

El recorrido intercuartílico se expresa en términos de cuartiles (o percentiles), y por lo tanto tiene más que ver con la mediana que con la media aritmética. Sin embargo, uno de los objetivos más importantes (si no el más importante) de la Estadística es hacer inferencias desde una muestra a la población. Y cuando se trate de hacer inferencias, vamos a utilizar en primer lugar la media aritmética como valor central o representativo de los datos. Por eso estas medidas de dispersión relacionadas con la mediana, y no con la media, no son las mejores para hacer inferencia. Necesitamos una medida de dispersión relacionada con la media aritmética.

Varianza poblacional y cuasivarianza muestral.

Tenemos, como siempre, un conjunto de n datos,

$$x_1, x_2, \ldots, x_n$$

que corresponden a n valores de una variable cuantitativa. La primera idea que se nos puede ocurrir es medir la diferencia entre cada uno de esos valores y la media (la *desviación individual* de cada uno de los valores):

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x},$$

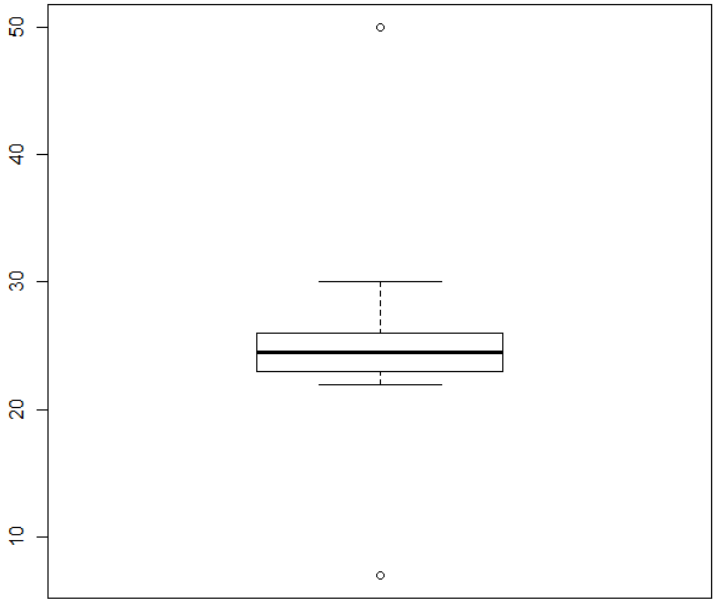
Y para tener en cuenta la contribución de todos los valores podríamos pensar en hacer la media de estas desviaciones individuales:

$$\frac{(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x})}{n}.$$

El problema es que esta suma siempre vale cero. Vamos a fijarnos en el numerador (y recuerda la definición de media aritmética):

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) = (x_1 + x_2 + \cdots + x_n) - n \cdot \bar{x} = 0. \qquad (2.4)$$

(a)



(b)

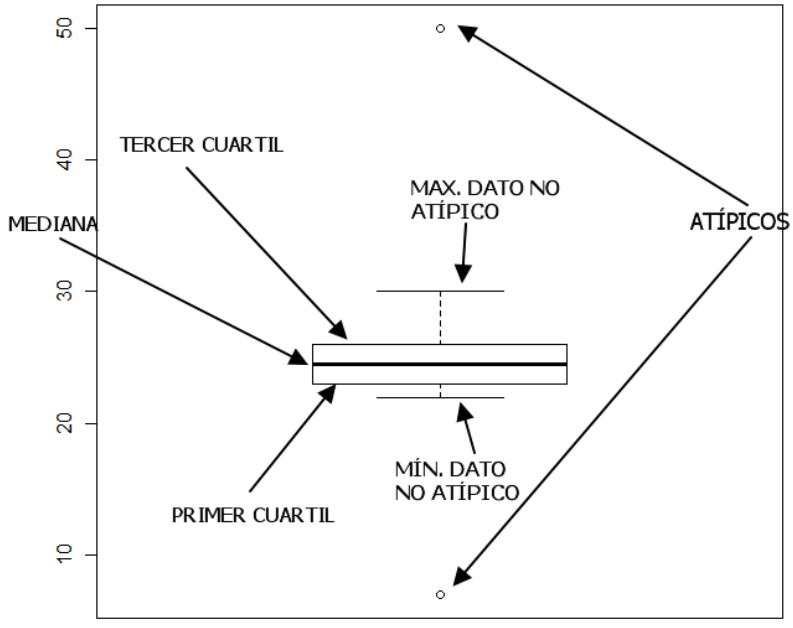


Figura 2.2: Un boxplot (a) y su estructura (b).

Está claro que tenemos que hacer algo más complicado, para evitar que el signo de unas desviaciones se compense con el de otras. A partir de aquí se nos abren dos posibilidades, usando dos operaciones matemáticas que eliminan el efecto de los signos. Podemos usar el valor absoluto de las desviaciones individuales:

$$\frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n},$$

o podemos elevarlas al cuadrado:

$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}.$$

Las razones para elegir entre una u otra alternativa son técnicas: vamos a usar la que mejor se comporte para hacer inferencias. Y, cuando se hacen inferencias sobre la media, la mejor opción resulta ser la que utiliza los cuadrados. En otros tipos de inferencia, no obstante, se usa la definición con el valor absoluto.

La varianza (poblacional) (o desviación cuadrática media) (en inglés, *variance*) del conjunto de datos $x_1, x_2, \ldots, x_n$ es:

Varianza (poblacional)

$$Var(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}. \quad (2.5)$$

En muchos libros, incluso sin hablar de la varianza, se define una cantidad relacionada, llamada varianza muestral o cuasivarianza muestral, que es el nombre que nosotros vamos a usar, mediante la fórmula

Cuasivarianza muestral

$$s^2(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \quad (2.6)$$

Como puede verse, la única diferencia es que en el denominador de la fórmula aparece $n-1$ en lugar de n. En particular, si n es muy grande, ambas cantidades son prácticamente iguales, aunque la cuasivarianza siempre es ligeramente mayor.

El concepto de cuasivarianza muestral será importante cuando hablemos de inferencia, y entonces entenderemos el papel que juega la cuasivarianza muestral, y su relación con la varianza (poblacional) tal como la hemos definido. Lo que sí es muy importante, usando software o calculadoras, es que sepamos si el número que se obtiene es la varianza o la cuasivarianza muestral.

Ejemplo 2.3.3. *Para el conjunto de valores*

9, 6, 19, 10, 17, 3, 28, 19, 3, 5, 19, 2,

del Ejemplo 2.1.1 (pág. 22), que ya hemos usado en varios ejemplos, su media aritmética es:

$$\bar{x} = \frac{140}{12} \approx 11.67.$$

Así que la varianza (poblacional) es:

$$Var(x) = \frac{\left(9 - \frac{140}{12}\right)^2 + \left(6 - \frac{140}{12}\right)^2 + \cdots + \left(19 - \frac{140}{12}\right)^2 + \left(2 - \frac{140}{12}\right)^2}{12} =$$

$$= \frac{\frac{2360}{3}}{12} = \frac{2360}{36} \approx 65.56$$

con cuatro cifras significativas. La cuasivarianza muestral se obtiene dividiendo por 11 *en lugar de* 12*, y es:*

$$s^2 = \frac{\left(9 - \frac{140}{12}\right)^2 + \left(6 - \frac{140}{12}\right)^2 + \cdots + \left(19 - \frac{140}{12}\right)^2 + \left(2 - \frac{140}{12}\right)^2}{11} =$$

$$= \frac{\frac{2360}{3}}{11} = \frac{2360}{33} \approx 71.52,$$

también con cuatro cifras significativas.

Dejamos como ejercicio para el lector comprobar que, para los datos del Ejemplo 2.1.2, que incluyen el valor atípico 150*, la varianza poblacional y la cuasivarianza muestral son (con cuatro cifras significativas)*

$$Var(x) \approx 1419, \quad s^2 \approx 1538.$$

Como puede verse, con la presencia del valor atípico la dispersión del conjunto ha aumentado mucho. □

Varianza a partir de una tabla de frecuencias.

Cuando lo que tenemos son datos descritos mediante una tabla de frecuencias, debemos proceder así:

1. La Ecuación 2.5 se sustituye por:

Varianza (poblacional) a partir de una tabla de frecuencias

$$Var(x) = \frac{\sum_{i=1}^{k} \mathbf{f_i} \cdot (x_i - \bar{x})^2}{\sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{f_i}}.$$

   donde, ahora, $x_1, \ldots, x_k$ son los valores *distintos* de la variable, y $f_1, \ldots, f_k$ son las correspondientes frecuencias.

2. En el caso de datos agrupados por intervalos, los valores x_i que utilizaremos serán las marcas de clase.

En los tutoriales tendremos ocasión sobrada de practicar este tipo de operaciones.

Desviación típica.

La varianza, como medida de dispersión, tiene un grave inconveniente: puesto que hemos elevado al cuadrado, las unidades en las que se expresa son el cuadrado de las unidades originales en las que se medía la variable x. Y nos gustaría que una medida de dispersión nos diera una idea de, por ejemplo, cuantos metros se alejan de la media los valores de una variable medida en metros. Dar la dispersión en metros cuadrados es, cuando menos, extraño. Por esa razón, entre otras, vamos a necesitar una nueva definición.

La desviación típica es la raíz cuadrada de la varianza:

Desviación típica (poblacional)

$$DT(x) = \sqrt{Var(x)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}.$$

Y, si es a partir de una tabla de frecuencias, entonces:

$$DT(x) = \sqrt{\frac{\sum_{i=1}^{k} \mathbf{f_i} \cdot (x_i - \bar{x})^2}{\sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{f_i}}}.$$

También existe una cuasidesviación típica muestral s, que es la raíz cuadrada de la cuasivarianza muestral, y con la que nos volveremos a encontrar muchas veces en el resto del curso.

El cálculo de la desviación típica tiene las mismas características que el de la varianza. Y, de nuevo, es muy importante, usando software o calculadoras, que sepamos si el número que se obtiene es la desviación típica o la cuasidesviación típica muestral.

Ejemplo 2.3.4. *Para los datos del Ejemmplo 2.3.3, y tomando raíces cuadradas, se obtiene una desviación típica poblacional aproximadamente igual a* 8.097 *y una cuasidesviación típica muestral aproximadamente igual a* 8.457. □