
Estadística descriptiva

Introducción al análisis de datos

PID_00269798

Àngel J. Gil Estallo



Universitat
Oberta
de Catalunya

Àngel J. Gil Estallo

Doctor en Ciencias Matemáticas por la Universidad de Barcelona desde el año 1996. Profesor titular de escuela universitaria de la Universidad Pompeu Fabra desde 1991. Su actividad docente se centra en temas de matemáticas, estadística e informática en los estudios de Economía de dicha Universidad. Consultor de la Universitat Oberta de Catalunya desde 1998.

La revisión de este recurso de aprendizaje UOC ha sido coordinada por la profesora: Mireia Besalú Mayol (2019)

Cuarta edición: septiembre 2019
© Àngel J. Gil Estallo
Todos los derechos reservados
© de esta edición, FUOC, 2019
Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Ninguna parte de esta publicación, incluido el diseño general y de la cubierta, puede ser copiada, reproducida, almacenada o transmitido de ninguna manera ni por ningún medio, tanto eléctrico como químico, mecánico, óptico, de grabación, de fotocopia, o por otros métodos, sin la autorización previa por escrito de los titulares del copyright.

Índice

Sesión 1

Tipos de datos y su representación gráfica	5
1. Tipos de variables	6
2. Variables cualitativas y variables numéricas discretas que toman un número pequeño de valores diferentes	7
2.1. El diagrama de barras	8
2.2. El diagrama de sectores	9
3. Variables numéricas	10
3.1. El diagrama de puntos	11
3.2. El diagrama de tallo y hojas	12
4. Variables continuas e histograma	13
4.1. Histograma de frecuencias relativas acumuladas	16
4.2. Ejemplos de construcción de histogramas	16
4.3. Interpretación de los histogramas	18
5. Resumen	19
Ejercicios	20

Sesión 2

Medidas de centro y propiedades	24
1. La moda	24
2. La mediana	24
3. La media	26
3.1. La media como valor central	27
3.2. Efecto de los valores alejados	28
3.3. Efecto de las transformaciones lineales	28
4. Comparación media-mediana	30
5. Medidas de centro y datos tabulados	30
6. Resumen	33
Ejercicios	34

Sesión 3

Medidas de dispersión	36
1. Los cuartiles y la mediana	36
1.1. El rango intercuartílico	38
1.2. Los cinco números resumen y el diagrama de caja	38
1.3. Interpretación de los diagramas de caja	40
2. La desviación típica y la media	40
2.1. Propiedades de la varianza y la desviación típica	41
2.2. La regla de Tchebichev	43
2.3. Datos estandarizados	44
3. Usos de la media y la desviación típica o de la mediana y los cinco números resumen	46

4. Varianza y datos tabulados	47
5. Resumen	47
Ejercicios	48

Tipos de datos y su representación gráfica

En esta sesión introduciremos los conceptos básicos de la estadística descriptiva.

A la hora de comenzar a trabajar con estadística, lo primero que hay que tener claro es cuál es la población objeto de nuestro estudio.

La **población** es el conjunto de individuos que se quiere estudiar, en el que los **individuos** pueden ser tanto objetos (yogures, ruedas, etc.) como personas, niños, programadores, etc.

También debemos tener claro si disponemos de datos sobre todo el conjunto de la población o sólo sobre una parte de dicha población, que llamaremos *muestra*.

Una **muestra** es un subconjunto cualquiera de la población.

Muestra y población

Hay que distinguir entre población (la totalidad de los individuos) y muestra (una parte cualquiera de la población). En otro módulo se estudiará cómo extraer muestras de una población.

Ejemplos de poblaciones

Los ficheros almacenados en el disco duro de nuestro ordenador, todos los catalanes que ven alguna vez al día la televisión, los ordenadores que están en este preciso momento en el almacén de una distribuidora determinada, el conjunto de personas en el mundo que disponen de conexión a Internet, etc. son ejemplos de poblaciones.

Representatividad de una muestra

Imaginemos que queremos averiguar cuáles son los programas de la televisión más vistos por los catalanes durante un determinado mes. La población sería el conjunto de todos los catalanes que han visto la televisión alguna vez durante este mes. Como podéis imaginar fácilmente, es imposible acceder a todos los individuos de la población para preguntarles qué programas han visto: lo que se suele hacer es seleccionar una muestra y preguntar sólo a los individuos que la componen. Los resultados que obtendremos de estudiar la muestra serán más fiables cuanto más representativa sea la muestra, es decir, cuanto más reproduzca, a una escala más reducida, la estructura y composición de la población.

Una vez que se ha decidido la población o muestra objeto de estudio, escogemos las variables que conviene estudiar.

Una **variable** es una característica de los individuos objeto de nuestro estudio.

Técnicas de muestreo

¿Os habéis preguntado alguna vez qué programas soléis ver? ¿Tenéis curiosidad por saber cómo es la muestra sobre la que se calculan las audiencias? Las **técnicas de muestreo** estudian varias alternativas para seleccionar muestras a partir de poblaciones.

Ejemplos de variables estadísticas

De los ficheros de nuestro disco duro nos puede interesar el tamaño (en Kbytes), la fecha en que los salvamos por última vez, la aplicación que lo creó, etc. De las audiencias de televisión, las variables serán qué programas se ven, durante cuánto tiempo, cuál gusta más, etc.

Después recogeremos la información para obtener una serie de observaciones del valor de la variable sobre los individuos de la población o de la muestra. Observad que tendremos una observación para cada individuo, aunque, evidentemente, los valores se pueden repetir.

En este módulo os proporcionaremos las pautas básicas de lo que hay que hacer cuando nos encontremos con un conjunto de datos (valores de las variables) correspondientes a una característica determinada de un conjunto de individuos (población o muestra). De momento nos limitaremos a estudiar de forma aislada cada característica de los individuos, es decir, analizaremos cada variable por separado, y por eso diremos que hacemos un análisis univariante.

En esencia, habrá que explorar los datos para llegar a encontrar una descripción de la población (o la muestra) lo más detallada posible. En este sentido, se trata de realizar las funciones siguientes:

- a) Averiguar la distribución de la variable: los valores que toma y cómo los toma (si algunos valores se repiten mucho, si algunos valores son “extraños”, etc.).
- b) Calcular algunos resúmenes numéricos que ayuden a entender cuál es el “centro” de los valores y cómo se distribuyen dichos valores en torno a este “centro”.
- c) Dibujar algunos gráficos que ayuden a visualizar los puntos mencionados anteriormente.

Estas operaciones deben permitirnos elaborar una descripción global de los datos a partir de la cual debemos ser capaces de conseguir lo siguiente:

- Llegar a conclusiones sobre los datos, fundamentadas en los resúmenes numéricos y en los gráficos (como por ejemplo, “tal como se ve en este gráfico...” o “las medidas de centro indican que...”).
- Comparar datos referidos a la misma característica sobre dos conjuntos diferentes de individuos (como por ejemplo, “en el gráfico correspondiente al primer colectivo se ve..., mientras que en el del segundo colectivo se ve...” o bien “los resúmenes numéricos del primer colectivo muestran como..., mientras que en el segundo colectivo vemos que...”).
- Plantearnos preguntas más complejas (como por ejemplo, sobre relaciones entre variables –regresión– o sobre si hay diferencias entre dos colectivos).
- Ver, en una muestra, qué preguntas nos podemos hacer con respecto a las características globales de la población –inferencia.

La importancia de los gráficos

Un buen gráfico siempre es de gran ayuda: ¡representad siempre los datos!

Veréis la inferencia estadística y la regresión en otros módulos.



1. Tipos de variables

Cuando estudiamos una población o muestra determinada, seleccionamos unas cuantas variables relevantes. Estas variables pueden ser de diferentes tipos:

1) **Variables cualitativas:** son aquellas que no se expresan numéricamente, sino como categorías o características de los individuos. A veces reciben el nombre de **variables categóricas**.

2) **Variables cuantitativas:** son las que se expresan de forma numérica. De entre estas últimas podemos distinguir los tipos siguientes:

a) **Variables cuantitativas discretas:** sólo toman valores enteros. Generalmente provienen de contar unidades de una clase determinada.

b) **Variables cuantitativas continuas:** pueden tomar cualquier valor en un intervalo. Acostumbran a ser el resultado de medir algún fenómeno.

En la práctica muchas variables cualitativas se codifican (se asigna un número a cada categoría) para facilitar su estudio.

Ejemplo de codificación de variables cualitativas

La variable “Tipo de ordenador que utiliza normalmente para conectarse a Internet” tiene un número finito de posibilidades diferentes. Si se asigna un número entero a cada tipo de ordenador posible (por ejemplo, 1 = PC, 2 = Mac, 3 = Otros), obtenemos una codificación numérica que describe la variable que queremos estudiar.

De la misma forma, una variable discreta que toma un número finito de valores se puede tratar como una variable cualitativa, en la que todos los individuos en los que la variable toma un valor determinado forman una categoría.

Ejemplo de categorización de una variable discreta

La variable “Número de hijos de una pareja” es cuantitativa discreta (y además toma pocos valores, quizá quince o dieciséis valores diferentes como máximo). Por otro lado, todas las parejas en las que la variable vale 1 constituyen la categoría de las parejas con hijo único, todas las parejas cuya variable vale 2 constituyen la categoría de las parejas con dos hijos, etc.

2. Variables cualitativas y variables numéricas discretas que toman un número pequeño de valores diferentes

Supongamos que tenemos que estudiar una variable categórica que puede tomar k valores posibles, o bien una variable cuantitativa discreta que puede tomar k valores diferentes, donde k es un número relativamente pequeño. Estos dos tipos de variables admiten un tratamiento similar, en el que hay que comenzar por averiguar las características siguientes:

1) El número total de individuos de los que se disponen datos; designaremos este número con N .

2) La **frecuencia absoluta** de cada valor de la variable: es decir, el número de individuos para los cuales la variable toma este valor. Designaremos con n_i la frecuencia absoluta del valor i .

3) La **frecuencia relativa** de cada valor de la variable: es decir, la proporción de individuos en los que la variable toma este valor. Designaremos con f_i la frecuencia relativa del valor i , y la acostumbraremos a dar en porcentaje.

Variables cualitativas y cuantitativas

El color de los ojos, la profesión de una persona y el tipo de ordenador que utiliza son variables cualitativas.

El número de hijos de una persona (contamos hijos), los puntos obtenidos por un equipo en un partido de básquet (contamos puntos), etc., son variables cuantitativas (numéricas) discretas.

El peso de una persona, la cotización del euro con respecto al dólar, el tiempo de acceso a una base de datos, etc., son variables numéricas continuas.

Frecuencias absoluta y relativa de una categoría

La frecuencia absoluta de una categoría es el número de individuos que pertenecen a la categoría.

La frecuencia relativa de una categoría es la proporción de individuos que pertenecen a ella.

Con esto tendremos la **distribución de frecuencias de la variable**, que es el conjunto de los valores que adopta la variable y la frecuencia con que los adopta.

Utilidad de la frecuencia relativa

La frecuencia relativa nos será muy útil si debemos comparar la misma variable en dos poblaciones diferentes que tienen diferente número de individuos.

Dado que la variable debe tomar alguno de los k valores posibles, es evidente que $n_1 + n_2 + \dots + n_k = N$. Además, f_i se obtiene de dividir el número de individuos en los que la variable toma el valor x_i y por el total de individuos, es decir:


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

Es muy fácil ver que $f_1 + f_2 + \dots + f_k = 1$, es decir, la suma de las frecuencias relativas de todos los valores es siempre igual a 1.

En caso de que tenga sentido ordenar los valores de la variable, para cada valor x_j se pueden definir la **frecuencia absoluta acumulada**, que es la suma de las frecuencias de los valores menores o iguales que x_j , y la **frecuencia relativa acumulada**, que es la suma de las frecuencias relativas de los valores menores o iguales que x_j . Más formalmente, si tenemos N observaciones de una variable que toma k valores diferentes $x_1, x_2, x_3, \dots, x_k$, de manera que $x_1 < x_2 < x_3 < \dots < x_k$, definimos:

- La **frecuencia absoluta acumulada** del valor x_j como $N_j = n_1 + n_2 + n_3 + \dots + n_j$.
- La **frecuencia relativa acumulada** del valor x_j como $F_j = f_1 + f_2 + f_3 + \dots + f_j$.

De estas definiciones, resulta evidente que $N_k = n_1 + n_2 + n_3 + \dots + n_k = N$ y que $F_k = f_1 + f_2 + f_3 + \dots + f_k = 1$.

Normalmente, las frecuencias se representan en forma de tabla. Lo veremos en seguida en el ejemplo de los tipos de ordenadores. Sin embargo, antes trataremos de representar gráficamente la distribución de la variable que se estudia; os sugerimos dos formas muy útiles y muy simples. 

2.1. El diagrama de barras

Para hacer una representación en forma de diagrama de barras, se dispone un eje horizontal sobre el que se sitúan tantas barras como categorías o valores toma la variable, separadas por pequeños espacios, y con un título claro en la base de cada barra que indique a qué categoría nos referimos.

Procedimiento para dibujar un diagrama de barras.

En el eje vertical del diagrama de barras marcaremos una escala que permita leer bien la altura de cada barra. Esta altura será o bien la frecuencia absoluta o bien la frecuencia relativa de la categoría correspondiente a la barra. En el eje vertical hay que indicar qué frecuencia se representa. En el eje horizontal también marcaremos, si es preciso, una escala con las unidades correspondientes.

2.2. El diagrama de sectores

Un diagrama de sectores consiste en un círculo que se reparte en diferentes sectores (o porciones) de manera que las superficies de los sectores sean proporcionales a las frecuencias de cada una de las categorías.

Procedimiento para dibujar un diagrama de sectores.

Si tenemos que hacerlo a mano, para obtener el ángulo que corresponde a cada sector tenemos que multiplicar los 360° de la circunferencia por la frecuencia relativa de cada clase; después hay que utilizar el transportador de ángulos.

Ejemplo del tipo de ordenadores

Hemos hecho una encuesta entre estudiantes de la UOC en la que se pregunta el tipo de ordenador que utilizan habitualmente en casa para conectarse al campus virtual. A continuación damos las respuestas abreviadas:

PC, PC, MAC, PCP, PCP, PCP, MAC, O, O, PC
PC, PC, PC, MACP, O, MAC, O, PCP, PCP, PC

donde:

- 1 = PC = "PC compatible de sobremesa"
- 2 = PCP = "PC portátil"
- 3 = MAC = "MAC"
- 4 = MACP = "MAC portátil"
- 5 = O = "Otros"

Con estos datos podemos calcular fácilmente la distribución de frecuencias de la variable. Primero contamos las respuestas (20), con lo que $N = 20$. A partir de aquí confeccionamos una tabla en la que se indiquen las frecuencias absolutas y relativas de cada categoría y con una hilera adicional para las sumas totales, para comprobar que no nos hemos equivocado.

	Frecuencia	Frecuencia relativa
Tipo de ordenador	n_i	f_i
1 = PC	$n_1 = 7$	$f_1 = 7/20 = 0,35 = 35\%$
2 = PCP	$n_2 = 5$	$f_2 = 5/20 = 0,25 = 25\%$
3 = MAC	$n_3 = 3$	$f_3 = 3/20 = 0,15 = 15\%$
4 = MACP	$n_4 = 1$	$f_4 = 1/20 = 0,05 = 5\%$
5 = Otros	$n_5 = 4$	$f_5 = 4/20 = 0,2 = 20\%$
Totales	$N = 20$	100%

Tabla de distribución de frecuencias

Esta tabla muestra la distribución de frecuencias de la variable.

Observad que la suma de la columna n_i es:

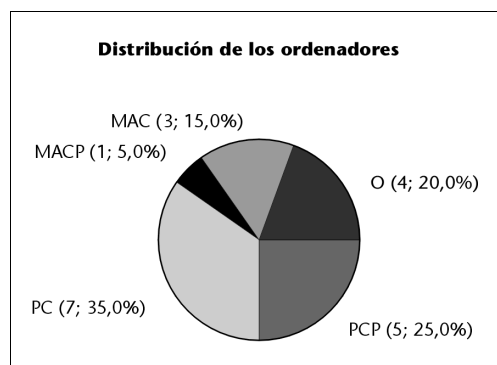
$$7 + 5 + 3 + 1 + 4 = 20 = N$$

y que la suma de las frecuencias relativas es:

$$100\% = 1$$

Si haciendo las sumas no se obtienen estos valores, seguro que hay algún error (salvo problemas de redondeo).

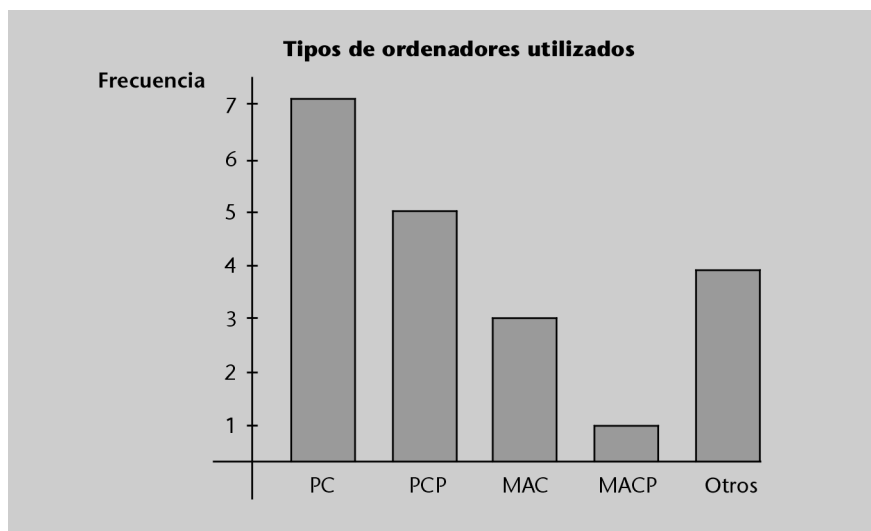
Representamos en primer lugar un diagrama de sectores, en el que vemos que el "pastel" correspondiente a la población está repartido en cinco categorías y que las porciones más grandes corresponden a ordenadores personales (PC) y personales portátiles (PCP), que en total incluyen al 60% de los usuarios.



Cómo dibujar las secciones circulares

¿Recordáis el transportador de ángulos? La mayoría de las veces haremos el gráfico con ayuda de algún programa informático, así que no tendremos que preocuparnos por él.

A continuación dibujamos el diagrama de barras con la frecuencia absoluta de cada clase. Observamos que las barras más altas corresponden a los PC, bien sean de sobremesa, bien sean portátiles. Con este gráfico obtenemos una visión global del reparto de individuos por categorías.



Ejemplo de las notas de estadística

En el semestre anterior las notas de la asignatura de *Estadística* de un grupo de alumnos fueron las siguientes:

4, 5, 5, 5, 6, 7, 8, 4, 9, 9, 10, 4, 7, 7, 7, 7, 8, 6, 7, 8

A continuación calcularemos la distribución de frecuencias de la variable y añadiremos las frecuencias acumuladas:

	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
Nota	n_i	f_i	N_i	F_i
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	3	0,15 = 15%	3	0,15 = 15%
5	3	0,15 = 15%	6	0,3 = 30%
6	2	0,1 = 10%	8	0,4 = 40%
7	6	0,3 = 30%	14	0,7 = 70%
8	3	0,15 = 15%	17	0,85 = 85%
9	2	0,1 = 10%	19	0,95 = 95%
10	1	0,05 = 5%	20	1 = 100%
Totales	20			

A partir de la tabla se pueden deducir los hechos siguientes:

- Un 15% de los alumnos ha sacado un 5 (frecuencia relativa del valor 5).
- 17 alumnos han aprobado (suma de frecuencias absolutas de los valores 5 a 10).
- El 85% de los alumnos ha aprobado (suma de frecuencias relativas de los valores 5 a 10).

3. Variables numéricas

Disponemos ahora de un conjunto de datos correspondientes a una variable numérica. Aparte de las técnicas introducidas para el caso de las variables discretas

con pocos valores diferentes y especialmente en caso de que tengamos una variable discreta con muchos valores diferentes o bien una variable continua, lo primero que hay que averiguar es:

- Cuántos individuos aparecen en el estudio (N).
- El **máximo** (*máx*) y el **mínimo** (*mín*) de los valores que toma la variable.
- El **rango de la variable**, es decir, la diferencia entre el valor máximo y el mínimo.

A continuación veremos diferentes gráficos que se pueden utilizar para representar este tipo de variables.

3.1. El diagrama de puntos

El diagrama de puntos consta de un único eje horizontal con una escala fijada en la que los individuos se representan por puntos dibujados encima del valor que les corresponde. En caso de valores repetidos, situamos un punto sobre el otro.

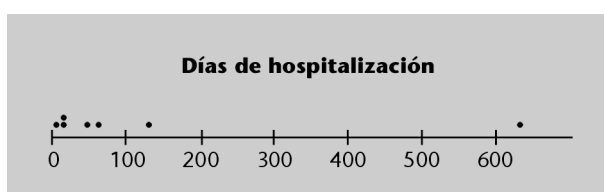
Procedimiento para dibujar un diagrama de puntos.

Ejemplo de los días de hospitalización

Se ha elaborado un estudio sobre los días de hospitalización de un grupo de pacientes sometidos a un mismo tratamiento, estudio del que se han obtenido los datos siguientes: 15, 15, 3, 46, 623, 126, 64 días.

Así pues, tenemos $N = 7$ datos, que van desde *mín* = 3 hasta *máx* = 623.

Dibujamos un eje que contenga estos valores y representamos a cada paciente con un punto un poco por encima de los días que ha estado hospitalizado, de forma que obtenemos el gráfico siguiente, en el que podemos ver que hay una concentración de los datos entre los valores 0 y 70 y un punto muy alejado de los otros (el 623):



Diremos que un dato alejado de los otros o que está fuera del patrón previsible en el gráfico es un **dato atípico** o **dato extremo**.

Nomenclatura

Utilizaremos de forma indistinta los términos *anómala*, *extrema*, *insólita*, *atípica* o su expresión inglesa: *outlier*.

Si nos encontramos un dato con estas características, de entrada hay que comprobar que no se trate de un error en la introducción de los datos. Una vez comprobado que no lo es, conviene averiguar la causa de este comportamiento anómalo.

3.2. El diagrama de tallo y hojas

Para hacer un diagrama de tallo y hojas, hay que llevar a cabo las operaciones siguientes:

- 1) Partimos cada una de las observaciones de la variable en dos partes: la primera contiene todos los dígitos menos el de más a la derecha (ésta será el **tallo**), y la segunda contiene el último dígito (ésta será la **hoja**).
- 2) Situamos los tallos uno debajo de otro, ordenados de manera creciente.
- 3) Colocamos al lado de cada tallo las hojas que le correspondan, también en orden creciente.

Para separar el tallo de las hojas, se puede dibujar una línea vertical. Como veremos a continuación, este diagrama tiene un aspecto similar a un diagrama de barras colocado en posición vertical. En este diagrama se encuentran representadas todas las observaciones de la variable y, por tanto, debemos escribir tantas hojas repetidas como sea necesario. Si el gráfico nos parece poco descriptivo (porque tiene pocos niveles, por ejemplo), podemos optar por desdoblar algún nivel.

Ejemplo del jugador de ajedrez

Consideremos el número de minutos que un jugador de ajedrez de gran nivel ha necesitado para ganar al programa de ordenador *Deep Yellow* en quince partidas consecutivas:

54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22

El último dígito es la hoja, de manera que, en este caso, en el tallo habrá representadas las decenas:

2		25
3		45
4		1166679
5		449
6		0

En este gráfico podemos ver que los datos se concentran en el “nivel” de los 40 minutos y no hay datos extremos; parece un jugador muy regular.

Si consideramos que el gráfico es poco descriptivo, podemos desdoblar cada decena en dos niveles: el primero contendrá las hojas del 0 al 4 y el segundo, las hojas del 5 al 9, de manera que resultará un diagrama como éste:

2		2
2		5
3		4
3		5
4		11
4		66679
5		44
5		9
6		0

Aquí se aprecia mejor que la acumulación de datos se da en la zona que va de 45 a 49.

Procedimiento para dibujar un diagrama de tallo y hojas.

Nomenclatura

En inglés el diagrama de tallo y hojas se llama *stem and leaf diagram*.

4. Variables continuas e histograma

En caso de que nuestra variable numérica tenga muchos valores diferentes o que haya muchos individuos (más de cien, por ejemplo), el diagrama de tallo y hojas puede ser demasiado cargado y difícil de interpretar. También podemos encontrar con tablas de distribución de frecuencias muy “aburridas”, en las que todas las frecuencias sean iguales a 1 (éste sería el caso si todos los valores de la variable fuesen diferentes). Para simplificar estas situaciones, se suelen agrupar los datos y representar no las frecuencias de cada valor, sino las de las diferentes agrupaciones.

Estas consideraciones conducen a la necesidad de definir lo que se entiende por **distribución de frecuencias** de una variable numérica continua o bien discreta con muchos valores diferentes.

Para obtener una expresión de la **distribución de frecuencias**, haremos lo siguiente:

Procedimiento para calcular la distribución de frecuencias en el caso continuo.

- 1) Agrupamos las observaciones en intervalos (generalmente todos con el mismo ancho) llamados **clases**. Los intervalos deben ser adyacentes (¡no vale dejar agujeros!) y deben cubrir, como mínimo, todo el rango, desde el mínimo hasta el máximo.
- 2) Calculamos el punto medio de cada intervalo, llamado **marca de clase**. La marca de clase será la “representante” de todas las observaciones que caen en el intervalo.
- 3) Una vez que tenemos las clases, calculamos la **frecuencia absoluta** de cada una (el número de observaciones que caen en el intervalo) y su **frecuencia relativa** (la proporción de observaciones que caen en el intervalo).
- 4) Si es preciso, también podemos calcular la **frecuencia absoluta acumulada** de la clase (el número de observaciones que caen en el intervalo más las que caen en intervalos anteriores) y la **frecuencia relativa acumulada** de la clase (suma de las frecuencias relativas de la clase más la de todas las clases anteriores).

En este caso, una manera muy eficiente de representar la distribución de frecuencias es por medio de los llamados *histogramas*.

Un **histograma** es más que un diagrama de barras: de hecho, es un diagrama de rectángulos. Cada rectángulo del diagrama representará una de las clases en las que tenemos distribuidos los valores de la variable, de manera que la proporción sobre el área total del área de cada rectángulo es precisamente la fre-

cuencia relativa de la clase que representa. En el caso de que la suma del área de todos los rectángulos sea 1, el área de cada rectángulo será igual a la frecuencia relativa de la clase que representa, y diremos que hemos construido un **histograma de densidad**.

Para construir un **histograma de densidad**, debemos dar los pasos siguientes:

Procedimiento para dibujar un histograma de densidad.

- 1) Seleccionamos una escala adecuada en el eje de las x .
- 2) Marcamos en el eje x todas las clases en las que se distribuye la variable.
- 3) Seleccionamos una escala adecuada en el eje de las y .
- 4) Para cada clase representamos un rectángulo que tiene como base la clase misma y como altura, la frecuencia relativa de la clase dividida por el ancho de ésta, que, evidentemente, es la longitud de la clase.

En este caso es fácil ver que la suma de las áreas de los rectángulos es 1.

Procedimiento para dibujar un histograma de frecuencias relativas.

En el caso bastante habitual de que todas las clases tengan el mismo ancho, podemos confeccionar otro tipo de histograma que consiste simplemente en poner la frecuencia relativa de la clase como altura de cada rectángulo. De esta manera la suma de todas las áreas de los rectángulos es precisamente el ancho de las clases, y si dividimos el área de uno de los rectángulos por el área total, obtenemos precisamente la frecuencia relativa de la clase. Entonces decimos que se trata de un **histograma de frecuencias relativas**.

Normalmente, y siempre que sea posible, construiremos histogramas de frecuencias relativas, ya que son más fáciles de interpretar y, de hecho, tienen el mismo aspecto visual que un histograma de densidad (si las clases presentan el mismo ancho). En todo caso, siempre hay que dejar claro qué tipo de histograma se construye e indicar si en el eje de las y se representan frecuencias relativas o densidad.

Recomendaciones

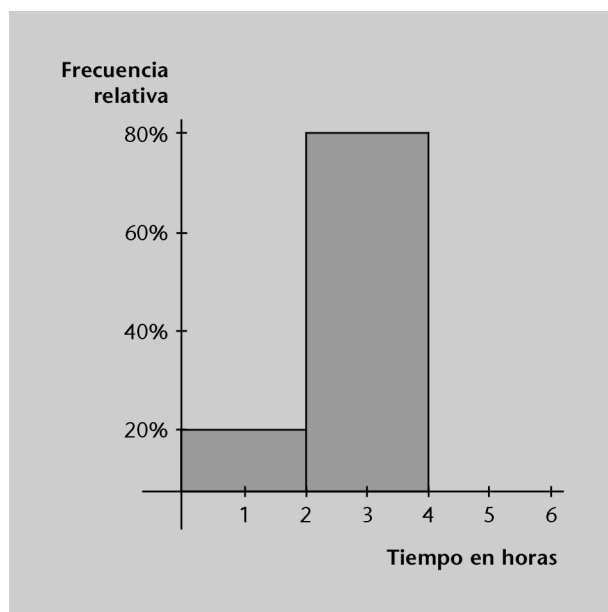
Especificad siempre las unidades en las que medís la escala de cada eje. Indicad también los títulos que convenga.
Si las clases tienen anchos diferentes, es aconsejable trabajar con histogramas de densidad.

Ejemplo de elaboración de un diagrama de frecuencias relativas

Se ha llevado a cabo un estudio sobre el número de horas diarias que los estudiantes de la UOC ven la televisión y se ha constatado que el 20% la ve menos de 2 horas y que el 80% restante la ve 2 horas o más, pero menos de 4 horas. Representamos los datos en forma de tabla:

Clases	f_i
[0,2)	20%
[2,4)	80%

Si dibujamos el histograma correspondiente (por ejemplo, de frecuencias relativas, ya que el ancho de todas las clases es igual a 2) obtenemos el gráfico siguiente:



En este gráfico la suma de las áreas de todos los rectángulos es:

$$2 \times 20\% + 2 \times 80\% = 2.$$

El área del primer rectángulo es $2 \times 20\% = 0,4$, que es precisamente el 20% de 2 (el área total), y el área del segundo rectángulo es $2 \times 80\% = 1,6$, que es precisamente el 80% del área total (que es 2).

Ejemplo de elaboración de un diagrama de densidad

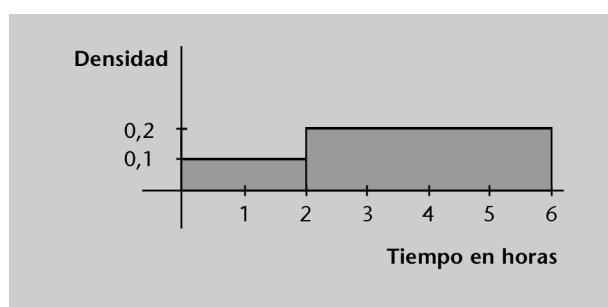
Otro estudio sobre el número de horas diarias que los estudiantes de la UOC ven la televisión ha constatado que el 20% de los estudiantes la ve menos de 2 horas y que el 80% restante la ve 2 horas o más, pero menos de 6. Cuando representamos los datos en una tabla, obtenemos lo siguiente:

Clases	f_i
[0,2)	20%
[2,6)	80%

En este caso tenemos que dibujar un histograma de densidad, en el que la altura de cada rectángulo viene dada por los datos siguientes:

Clases	f_i	Altura del rectángulo
[0,2)	20%	$20\% / 2 = 0,1$
[2,6)	80%	$80\% / 4 = 0,2$

El histograma resultante es el que vemos a continuación (este histograma y el anterior están dibujados con la misma escala en el eje de las x para facilitar su comparación):



Interpretación de los histogramas de densidad

El diagrama de densidad del ejemplo nos informa del hecho de que a cada una de las horas de la clase [0,2) le corresponde un $0,1 = 10\%$ de la población, mientras que a cada una de las horas que van de la 2 a la 6 le corresponde un $0,2 = 20\%$ de la población.

Ahora la superficie total es:

$$2 \times 0,1 + 4 \times 0,2 = 1.$$

El área del primer rectángulo es $2 \times 0,1 = 0,2 = 20\%$, que es precisamente la frecuencia relativa de la primera clase, y el área del segundo rectángulo es $4 \times 20\% = 0,8 = 80\%$, que es la frecuencia relativa de la segunda clase.

Los histogramas de densidad dan una idea de cómo se distribuyen las observaciones dentro de cada clase.

4.1. Histograma de frecuencias relativas acumuladas

Para representar de manera gráfica las frecuencias relativas acumuladas, se suele utilizar un histograma. En el histograma de frecuencias relativas acumuladas la base del rectángulo es la clase y la altura, la frecuencia relativa acumulada de la clase. En este histograma la altura de los rectángulos crece hasta llegar a la altura de la última clase, que es evidentemente 1.

La función de distribución acumulada

Al hablar de probabilidad y de variables aleatorias, utilizaremos la función de distribución acumulada. En determinados casos la representación de esta función es un histograma de frecuencias relativas acumuladas.

4.2. Ejemplos de construcción de histogramas

Como podéis observar, la descripción de cómo se calcula la distribución de frecuencias para estas variables y de cómo se hace un histograma son muy generales y dependen de los intervalos que seleccionemos. No existe una norma universal que explique cuántas ni qué clases debemos considerar; cada vez tendremos que decidirlo según el tipo y la forma de los datos, sin perder de vista el objetivo final.

La finalidad de la construcción de histogramas es obtener una representación gráfica que resuma la distribución de los datos de una manera entendedora, fácil de asimilar y que muestre aspectos relevantes de los datos.

Muchas veces tendremos que hacer varias pruebas hasta conseguir la representación que mejor ilustre la forma de la distribución o que argumente mejor nuestras opiniones. Si tenemos que utilizar un histograma, es conveniente tener en cuenta los aspectos siguientes:

- Siempre que sea posible, se cogerán clases del mismo ancho.
- El número de clases debe ser aproximadamente igual a \sqrt{N} , y mejor si está entre 7 y 15 (N es, como siempre, el número de observaciones).
- Para determinar dónde debe comenzar y acabar cada clase, podemos guiarnos por las consideraciones siguientes:
 - Debe quedar muy claro a qué clase deben pertenecer los puntos de cambio de intervalo; generalmente utilizaremos intervalos de la forma $[x, y)$, de manera que contienen el extremo izquierdo, pero no el derecho.

Histogramas por ordenador

La mayoría de los programas de ordenador permiten escoger el número y la forma de los intervalos: podemos hacer pruebas hasta encontrar una "buena" representación.

La importancia del número de clases

Si el número de clases es demasiado pequeño, tendremos histogramas con pocas barras y bastante altas; si hay demasiadas clases, parecerá un conjunto de palos de alturas parecidas y con agujeros.

- Cuando los datos toman valores enteros como, por ejemplo, 11, 13, 15, etc., es aconsejable considerar clases de manera que su punto medio coincida con los valores enteros; en este caso definiríamos las clases $[10,5-11,5)$, $[11,5-12,5)$, etc.

Los resultados de un test de estadística

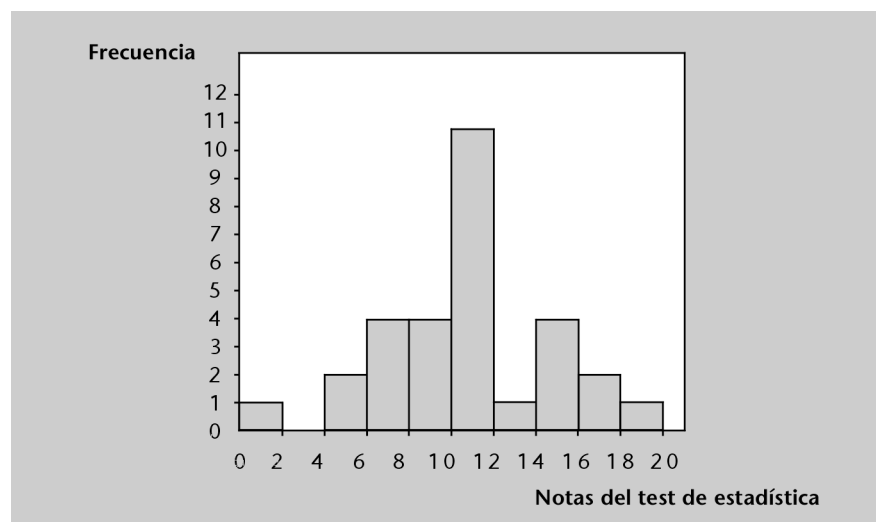
A continuación mostramos las notas de un test de estadística pasado a treinta estudiantes de un curso de la UOC durante el semestre de otoño. La puntuación posible del test va de 0 a 20 puntos.

11,5	7	5,25	19	8,5	11	10	6	15,5	11,75
9,75	15	16	16	7,5	8	11,5	0	10	11
13	10,25	10	8	14	6	14,5	10	5,6	10

Primero calculamos el valor máximo ($máx = 19$), mínimo ($mín = 0$) y el rango ($máx - mín = 19$). En el paso siguiente se agrupan los datos en intervalos de manera que cubran todo el rango. En este caso podríamos agrupar los datos en intervalos de longitud 1, comenzando por el 0 y acabando por el 20. Puesto que salen demasiadas clases, optamos por hacer clases de longitud 2, de manera que obtenemos una tabla como ésta, en la que también se han calculado las frecuencias acumuladas.

		Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
Intervalo	Marca de clase	n_i	f_i	N_i	F_i
$[0,2)$	$1 = (0 + 2) / 2$	1	$1/30 = 0,03$	1	0,03
$[2,4)$	$3 = (2 + 4) / 2$	0	$0 = 0$	1	0,03
$[4,6)$	5	2	$2/30 = 0,07$	3	0,10
$[6,8)$	7	4	$4/30 = 0,13$	7	0,23
$[8,10)$	9	4	$4/30 = 0,13$	11	0,37
$[10,12)$	11	11	$11/30 = 0,37$	22	0,73
$[12,14)$	13	1	$1/30 = 0,03$	23	0,77
$[14,16)$	15	4	$4/30 = 0,13$	27	0,90
$[16,18)$	17	2	$2/30 = 0,07$	29	0,97
$[18,20)$	$19 = (18 + 20) / 2$	1	$1/30 = 0,03$	30	1,00
Totales		30	1		

Con estos datos podemos dibujar el histograma siguiente:



Cálculo del valor medio

Para calcular el punto medio de un intervalo, se suman los extremos y se divide el resultado por 2.

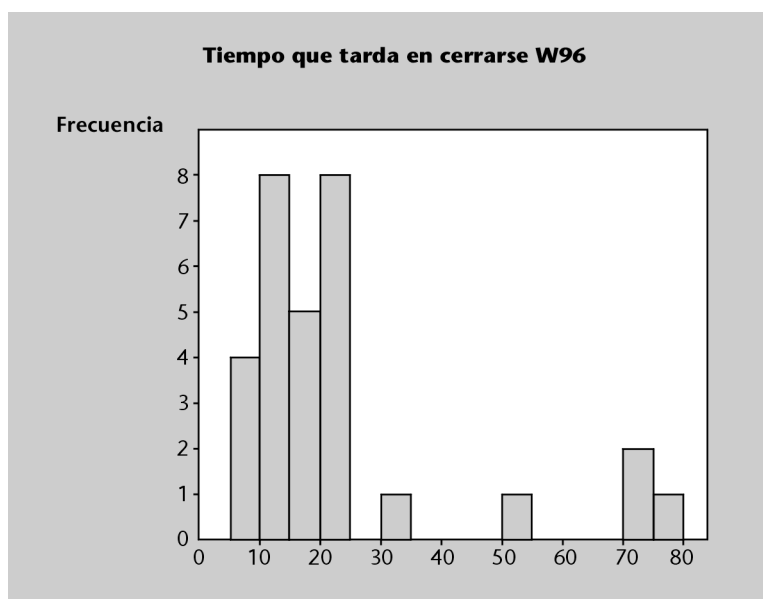
Donde se aprecia una cima destacada que corresponde a la clase $[10,12)$ y una cierta simetría alrededor de ésta. También se observa una clase vacía –la $[2,4)$ tiene frecuencia 0.

Tiempo que tarda en cerrarse el ordenador

En este segundo ejemplo de histograma utilizaremos los datos siguientes, que corresponden al tiempo que ha tardado en cerrarse nuestro ordenador (equipado con Windows 96) desde que hemos dado la instrucción adecuada, las últimas treinta veces que lo hemos utilizado:

16	32	10	24	23	12	15	21	16	10
24	20	21	71	12	50	76	15	19	8
8	10	12	23	9	71	13	14	20	6

Con un programa estándar hemos obtenido el histograma siguiente:



Como podéis ver, aquí las clases son $[0,5)$, $[5,10)$... hasta $[75,80)$. Hay dos picos, correspondientes a los intervalos $[10,15)$ y $[20,25)$, aunque el comportamiento es bastante irregular y nada simétrico. También hay datos alejados o extremos en torno a los 70 segundos y parece que haya tres agrupaciones en los datos: de 0 a 30, de 50 a 60 y de 70 a 80 segundos. Cada una de estas agrupaciones podría estar relacionada con las circunstancias en las que lo apagamos y con los programas que se han ejecutado.

4.3. Interpretación de los histogramas

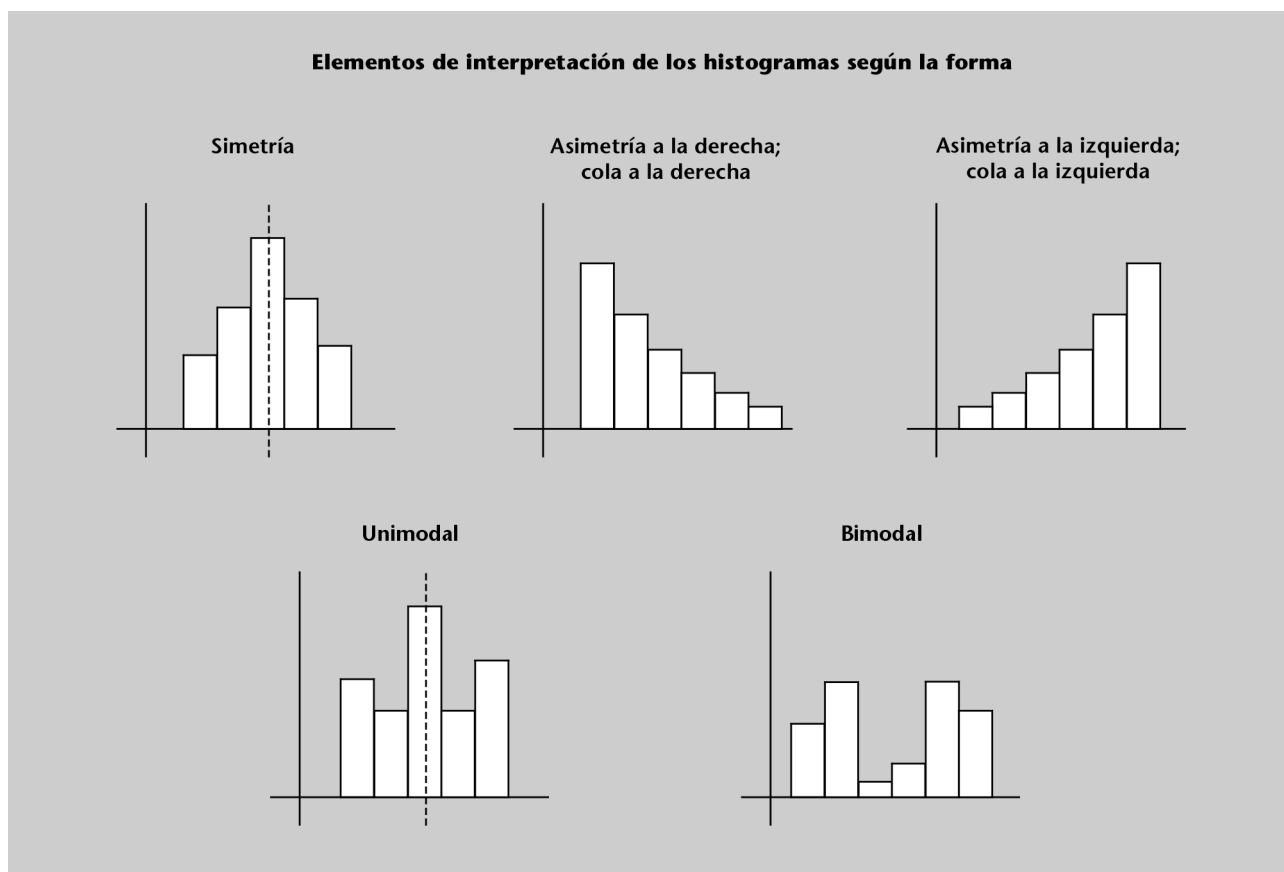
En la interpretación de los histogramas es conveniente destacar los aspectos siguientes:

- a) La **simetría**: el histograma será simétrico si es posible trazar un eje vertical de manera que la parte que hay a la derecha del eje sea aproximadamente igual a la imagen reflejada en un espejo de la parte izquierda.
- b) Los **picos**: corresponden a intervalos en los que se tienden a concentrar los valores de la variable (son intervalos con frecuencias grandes). Si sólo hay un pico destacado, diremos que es un **caso unimodal**; si tenemos dos picos de altura similar, será un **caso bimodal**.
- c) Las **colas**: si no hay simetría, es posible que uno de los lados del histograma se extienda mucho más lejos que el otro. En este caso diremos que hay una

cola en el lado más extenso (llamado **caso de cola larga**) o que hay asimetría hacia este mismo lado.

d) Hay que detectar los **datos extremos**, las **clases vacías** y si estas clases vacías separan a la población en grupos.

A continuación representamos de forma gráfica algunos de estos casos e indicamos el eje de simetría en los que hay:



5. Resumen

En esta sesión hemos introducido el vocabulario básico de la estadística (población/muestra, individuo, variable, frecuencia, distribución de la variable, etc.). Según el tipo de variable considerada, se han presentado varias formas de representar su distribución (diagrama de barras y de sectores para variables cualitativas y numéricas discretas con pocos valores y diagramas de puntos, de tallo y hojas e histograma para las cuantitativas).

Ejercicios

1. Considerad los cuatro programas más vistos durante una semana de una cadena de TV determinada un día concreto. El número de espectadores (en miles) que ven cada uno de los programas es el que se muestra en la tabla siguiente:

Programa	Número de espectadores
"Viva la estadística"	875
"El hermano"	925
"Matanza sangrienta"	742
"Informativo del día"	682

a) Representad gráficamente los datos de la manera que creáis más oportuna.

b) Si la semana siguiente el número de espectadores de estos mismos programas es el que refleja la tabla que hay a continuación, representad los datos de manera que se pueda comparar el número de espectadores en las dos semanas.

Programa	Número de espectadores
"Viva la estadística"	100
"El hermano"	200
"Matanza sangrienta"	100
"Informativo del día"	682

2. Una empresa utiliza como procesador de textos el famoso programa Macrohard Phrase. Se ha pedido a los usuarios de este programa que anoten cuántas veces se les ha "colgado" el ordenador durante un mes mientras trabajaban con éste. Se han obtenido los resultados siguientes:

47	23	28	0	11	12	35	36	40	30	37	14
----	----	----	---	----	----	----	----	----	----	----	----

Representad gráficamente los datos de la manera que creáis más adecuada.

3. Dado que el comportamiento del programa Macrohard parece harto preocupante, hemos decidido recoger datos sobre el número de veces que se ha "colgado" en un determinado mes en cincuenta ordenadores diferentes:

Ejemplo: el histograma de Macrohard.

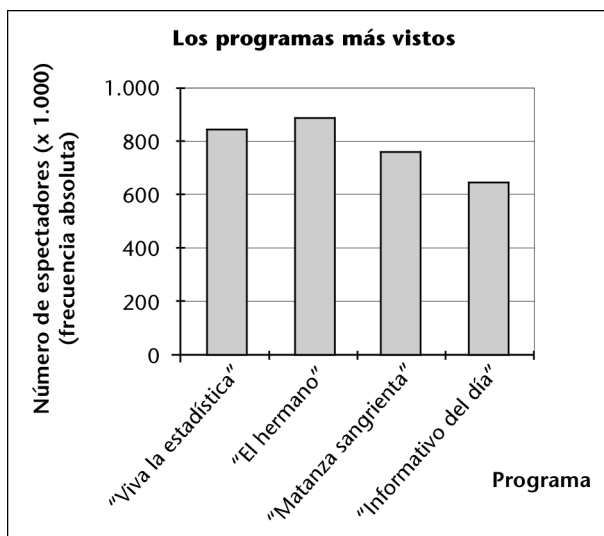
0	9	12	14	19
2	10	12	14	20
4	10	12	14	20
5	10	12	15	21
6	11	12	15	22
6	11	12	17	29
6	11	12	17	29
7	11	13	18	32
8	11	13	18	39
9	12	14	19	39

Encontrad la distribución de frecuencias de esta variable, representadla en forma de histograma y comentad sus características. Representad el histograma de frecuencias relativas acumuladas.

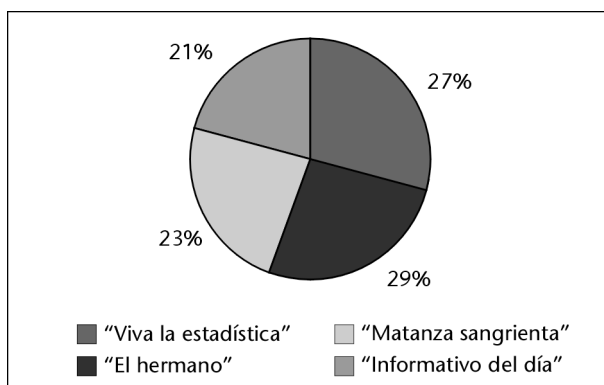
Solucionario

1.

a) Podemos hacer un gráfico de barras como éste (por ejemplo, con Excel):



También podemos representar un diagrama de sectores con las frecuencias relativas:



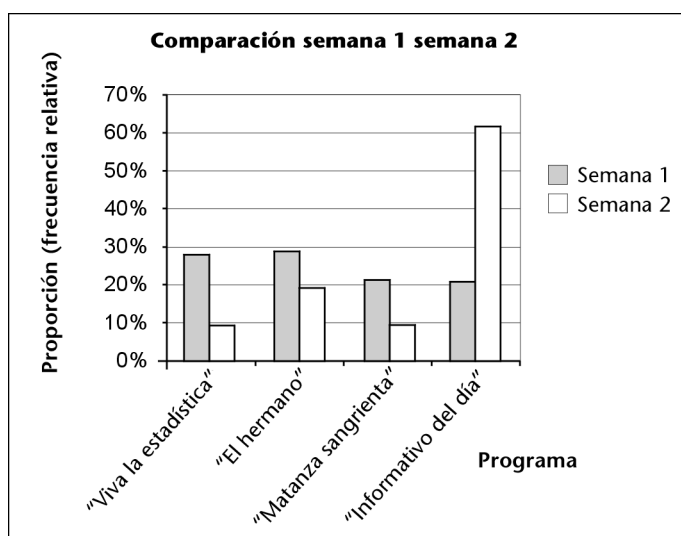
en el que podemos ver que las audiencias están muy igualadas.

b) Para poder comparar las dos semanas, es mejor calcular primero las frecuencias relativas de cada programa en cada semana:

Programa	Frecuencia absoluta		Frecuencia relativa	
	Semana 1	Semana 2	Semana 1	Semana 2
Viva la estadística	875	100	27,14%	9,24%

Programa	Frecuencia absoluta		Frecuencia relativa	
	Semana 1	Semana 2	Semana 1	Semana 2
El hermano	925	200	28,69%	18,48%
Matanza sangrienta	742	100	23,01%	9,24%
Informativo del día	682	682	21,15%	63,03%
Total	3.224	1.082	100% = 1	1

A partir de estos datos podemos crear el gráfico que se muestra a continuación, en el que para cada programa se representan dos barras, una para cada semana:



En el gráfico se ve claramente que el "Informativo del día" ha aumentado en gran medida su cuota de pantalla, de modo que supera ampliamente los otros programas, aunque el número de espectadores se ha mantenido constante.

2. Dado que tenemos pocos datos, podemos optar por hacer un diagrama de tallo y hojas:

0	0
1	124
2	38
3	0567
4	07

Como podemos ver, se trata de un diagrama bastante simétrico, distribuido en torno a los valores entre 20 y 30, con cierta tendencia a desplazarse hacia los valores grandes (30 en adelante). Realmente, el programa no parece demasiado fiable.

3. El valor mínimo que toma la variable es 0, el máximo es 39 y hay 50 observaciones. Puesto que el rango es aproximadamente 40, podemos hacer 10 clases de ancho 4, comenzando en 0 y acabando en 40. Probamos esta opción. De entrada, tabulamos los datos para obtener las frecuencias de cada clase:

Terminología

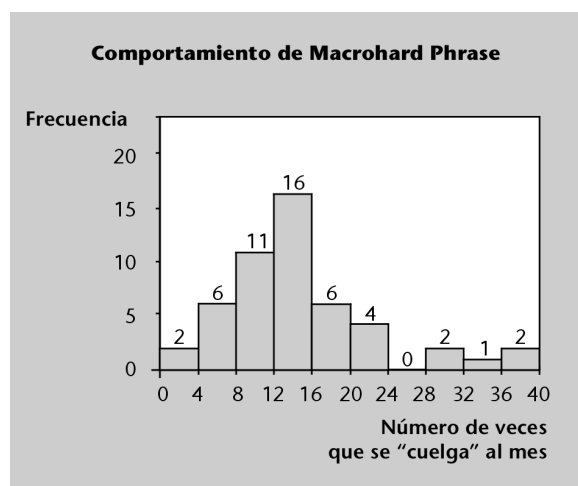
En el caso específico de las audiencias de televisión, la frecuencia relativa se denomina *cuota de pantalla*.

Comenzar en 0

Es razonable comenzar en 0, ya que no puede ser que un ordenador se "cuelgue" menos de cero veces.

		Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
Intervalo	Marca de clase	n_i	f_i	N_i	F_i
[0,4)	2	2	4%	2	0,04
[4,8)	6	6	12%	8	0,16
[8,12)	10	11	22%	19	0,38
[12,16)	14	16	32%	35	0,7
[16,20)	18	6	12%	41	0,82
[20,24)	22	4	8%	45	0,9
[24,28)	26	0	0%	45	0,9
[28,32)	30	2	4%	47	0,94
[32,36)	34	1	2%	48	0,96
[36,40)	38	2	4%	50	1
Totales		50	1		

Y después dibujamos el histograma, en el que, por comodidad, hemos escrito las frecuencias de cada clase.



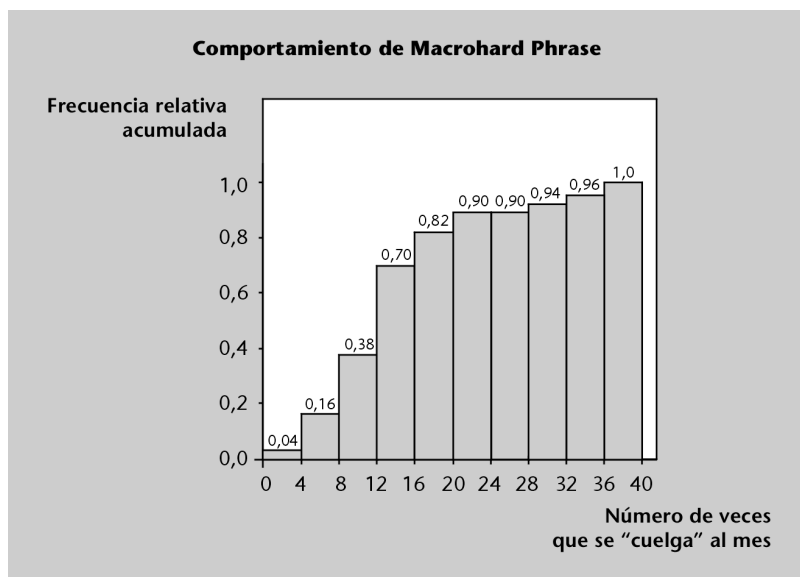
Caso particular

En caso de que alguna observación fuese exactamente 40, podríamos optar por considerar el último intervalo de la forma [36,40].

Interpretación

Dos ordenadores se han "colgado" entre 0 y 4 veces, 6 entre 4 y 8 veces, etc.

El histograma de frecuencias relativas acumuladas resulta de la manera siguiente:



Interpretación

El 70% de los ordenadores se ha "colgado" menos de 16 veces; el 90%, menos de 28 veces, etc.

Medidas de centro y propiedades

En esta sesión aprenderemos a asignar, calcular e interpretar diferentes indicadores numéricos que nos ayudan a describir el lugar “central” de los datos. Evidentemente, existen varias formas de describir el “centro”; nosotros nos concentraremos en tres medidas: la moda, la mediana y la media, y analizaremos la información que nos aportan y la manera como se pueden utilizar para comparar los valores de la variable en muestras diferentes.

1. La moda

La moda es el valor más repetido, es decir, el valor que tiene una mayor frecuencia.

El término *moda*

Una cosa es moda cuando es muy frecuente, es decir, ¡cuando se lleva mucho!

Si aparecen diferentes valores con la máxima frecuencia, tendremos varias modas y la distribución será bimodal o trimodal.

Ejemplos de cálculo de la moda

En el ejemplo de los tipos de ordenadores, la moda es PC. En el ejemplo de los días de hospitalización, la moda es 15. En el ejemplo del test de estadística, la moda es 10.

Si los datos son: 2, 2, 3, 4, 4, las modas son 2 y 4, ya que ambos valores tienen la mayor frecuencia: 2.

Observad el ejemplo del tipo de ordenadores en el apartado 2.2. y el de los días de hospitalización en el 3.1. de la sesión “Tipos de datos y su representación gráfica”.

2. La mediana

La **mediana** es un valor numérico que ocupa el lugar central una vez ordenadas las observaciones de más pequeñas a más grandes.

Atención

Vigilad: para calcular la mediana, primero hay que ordenar los datos.

La mediana ocupa justamente el valor que se sitúa en el medio de todas las observaciones. De este modo, al menos la mitad de las observaciones deben ser menores o iguales que la mediana y al menos la mitad tienen que ser mayores o iguales que la mediana.

La mediana como valor central

A veces hay diferentes valores que satisfacen la definición de mediana; en seguida veréis cuál es la práctica habitual en estos casos y daremos una regla para encontrar la mediana según el número de individuos.

Precisiones en la definición de mediana

Quizá os preguntéis por qué en la definición de mediana utilizamos la expresión *al menos*. A continuación vemos un caso. Considerad los datos siguientes:

1	12	13	15	15	15	17	18	19
---	----	----	----	----	----	----	----	----

Aquí la mediana es 15 (el 15 que está destacado), pero hay seis observaciones con valor menor o igual que 15 y seis con valor mayor o igual que 15, ¡y sólo tenemos nueve observaciones!

Con un par de ejemplos lo veremos del todo claro:

a) Cálculo de la mediana cuando el número de datos es impar

La mediana de las nueve observaciones –12, 13, 11, 16, 17, 19, 18, 15, 14– es 15, ya que si las ordenamos, nos quedan de esta manera:

11	12	13	14	15	16	17	18	19
----	----	----	----	----	----	----	----	----

Donde se ve claramente que el 15 se encuentra en el lugar central y cinco observaciones son menores o iguales que 15 y cinco, mayores o iguales.

b) Cálculo de la mediana cuando el número de datos es par

En el caso de las ocho observaciones siguientes:

11	12	13	14	15	16	17	18
----	----	----	----	----	----	----	----

seguramente dudaríamos entre los valores 14 ó 15, ya que los dos, sin ser el “centro”, están muy cerca del mismo (hay cuatro observaciones menores o iguales que 14 y cinco mayores o iguales, mientras que cinco menores o iguales que 15 y cuatro mayores o iguales). En este caso, que se da cuando el número de datos es par, la práctica habitual es salomónica: se opta por tomar la mediana como el valor medio entre los dos valores centrales. De esta manera, la mediana es $(14 + 15) / 2 = 14,5$, que deja cuatro observaciones por debajo y cuatro por encima (la mitad por debajo y la mitad por encima).

Dado que en el cálculo de la mediana el factor importante es la posición que ocupan los datos una vez ordenados, a continuación precisamos una regla que nos será muy útil en el caso de tener un número de datos muy elevado. Para obtener la mediana de un conjunto de N datos, debemos dar los pasos siguientes:

Procedimiento para calcular la mediana.

- 1) Ordenamos las observaciones de menor a mayor.
- 2) Si N es impar, la mediana es la observación que ocupa el lugar $(N + 1) / 2$.
- 3) Si N es par, calculamos el valor $(N + 1) / 2$, y la mediana será el valor medio de las dos observaciones que ocupan los lugares más cercanos a $(N + 1) / 2$.

De este procedimiento, hay que destacar los dos aspectos siguientes: 

- a) $(N + 1) / 2$ es “la dirección” de la mediana, no su valor.
- b) Si N es par, $(N + 1) / 2$ no es un número entero.

Error habitual

Es un error frecuente decir que la mediana de un conjunto de N datos es $(N + 1) / 2$. Recordad que $(N + 1) / 2$ sólo nos ayuda a encontrar la posición que ocupa la mediana; para conocer su valor, debemos observar los datos de que disponemos.

Ejemplo de cálculo de la mediana

Supongamos que nos proporcionan un conjunto de datos en forma de diagrama de tallo y hojas:

8	15
9	22
10	3456789
11	0
12	
13	09

Y nos piden calcular su mediana. En primer lugar escribimos los valores de la variable ordenados (juntamos cada hoja con el tallo correspondiente):

81, 85, 92, 92, 103, 104, 105, 106, 107, 108, 109, 110, 130, 139

Dado que hay catorce observaciones, $(N + 1) / 2 = 15/2 = 7,5$ y, por tanto, la mediana se obtiene como el valor medio de las observaciones que ocupan los lugares 7 y 8; es decir, la mediana es:

$$(105 + 106) / 2 = 105,5$$

3. La media

La **media** de las observaciones de una variable x (denotada por \bar{x}) se obtiene dividiendo la suma de todas las observaciones por el número de individuos; es, pues, el valor medio de todas las observaciones de la variable.

Cálculo de la media

Para calcular la media:

- 1) sumamos todas las observaciones;
- 2) dividimos el resultado por el número de individuos.

Con algo de notación podemos escribir esta definición formalmente. Si x_1, x_2, \dots, x_N son las observaciones de la variable que consideramos, entonces la media se define de la manera siguiente:

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

Ejemplo de cálculo de la media

Supongamos que los sueldos mensuales (en miles de euros) de los compañeros de un equipo de baloncesto son 1, 4, 2, 4, 1, 6, 12. En este caso la media es:

$$(1 + 4 + 2 + 4 + 1 + 6 + 12) / 7 = 4,285 \text{ miles de euros}$$

Observad que hemos repartido la suma total de los sueldos entre los siete compañeros, de manera que si todos tuviesen que cobrar lo mismo –y la cantidad total se mantuviese fija– tocarían 4,285 miles de euros a cada uno, lo que no significa que efectivamente los cobren. Sin embargo, observad que sólo dos personas cobran más de 4 miles de euros al mes.

La media saldrá a menudo a lo largo de la asignatura y, por tanto, es conveniente tener sus propiedades muy claras y, por qué no, también sus limitaciones. ■

La media es un reparto de la suma total de los valores considerados entre todas las observaciones. Matemáticamente, si multiplicamos la media por el número total de individuos, obtenemos la suma total de las observaciones de la variable. En efecto, si en la definición de la media se aísla la suma de los valores de las observaciones, obtenemos:

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

Una propiedad importante

Si multiplicamos la media por el número de individuos, obtenemos la suma de todas las observaciones de la variable.

3.1. La media como valor central

Para ver en qué sentido la media es el “centro” de los valores, deberemos hacer algunos cálculos basados en el estudio de las desviaciones con respecto a la media:

Las diferencias entre los valores de las observaciones y su media ($x_i - \bar{x}$) se denominan **desviaciones con respecto a la media**.

Si se calculan las operaciones siguientes, es fácil apreciar que la suma de todas las desviaciones da 0:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_N - \bar{x}) = (x_1 + x_2 + \dots + x_N) - N\bar{x} = 0$$

Las desviaciones a lado y lado de la media se “equilibran”.

donde la última igualdad se deduce de la relación entre la media y la suma de los valores de la variable.

Por tanto, podemos concluir que la suma de las desviaciones con respecto a la media es 0.

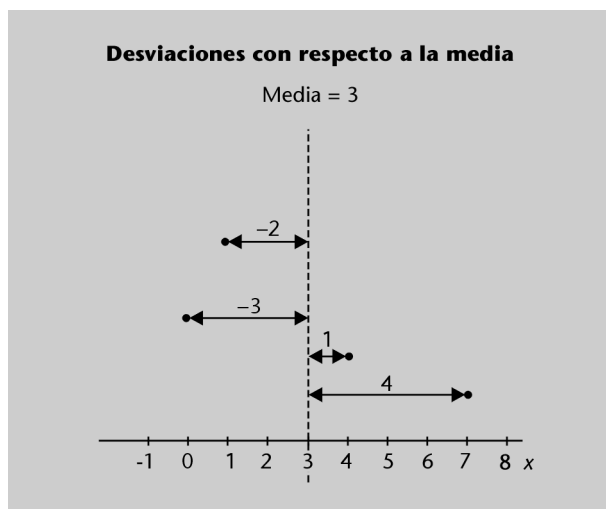
A veces se dice que la media representa el centro de gravedad de la distribución.

Cálculo de las desviaciones con respecto a la media

Consideremos los valores 0, 1, 4 y 7, que tienen de media $\bar{x} = 3$. Si calculamos sus desviaciones:

x_i	$(x_i - \bar{x}) = (x_i - 3)$
0	-3
1	-2
4	1
7	4
Total	-3 - 2 + 1 + 4 = 0

Al hacer la suma de las desviaciones, las positivas se compensan con las negativas y el resultado es 0. Gráficamente, el significado de las desviaciones con respecto a la media se puede visualizar de la manera siguiente:



3.2. Efecto de los valores alejados

Los valores alejados afectan en gran medida a la media. En ocasiones se dice que la media es poco resistente a los valores extremos, o bien que es poco robusta. Además, si hay pocos datos, el efecto es todavía más importante.

La media tiene tendencia a desplazarse hacia los datos alejados.

Ejemplos del efecto de los valores extremos en la media

a) Supongamos que un jugador de baloncesto ha marcado 1 punto en cada uno de los cinco partidos que ha disputado; evidentemente, la media de puntos de este jugador es 1. Imaginemos que juega otro partido y marca 50 puntos. Con este nuevo dato, la media del jugador ha pasado de 1 a 9,16. ¿Decir que la media de puntos es 9,16 describe lo suficiente al jugador? En este caso más bien no, ya que sólo una vez ha marcado más de un punto.

b) Una empresa fabrica biberones. Hacemos un estudio sobre su capacidad y obtenemos los datos (en mililitros) siguientes:

270, 270, 271, 271, 271 y 272

Con estos datos, la media de la capacidad es 270,83 ml. Si consideramos ahora los datos siguientes:

270, 270, 271, 271, 271, 272 y 390

obtenemos que la nueva media es 287,85. Como podéis ver, el valor añadido (que es un dato atípico) afecta en gran medida al valor de la media. Estos ejemplos nos muestran dos cosas: que los datos extremos afectan mucho al valor de la media y que si hay pocos datos, este efecto es todavía más importante.

3.3. Efecto de las transformaciones lineales

A continuación examinaremos cómo afectan a la media algunas transformaciones en los datos iniciales. Estudiaremos tres casos, los tres basados en una situación inicial que consta de N observaciones de una variable con valores x_1, x_2, \dots, x_N , cuya media es \bar{x} .

1) Si todas las observaciones tienen el mismo valor, es decir, si $x_1 = x_2 = \dots = x_N = v$, entonces la media tiene el mismo valor $x_1 = v$, tal como se demuestra a continuación:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\overbrace{v + v + \dots + v}^N}{N} = \frac{Nv}{N} = v$$

2) Si sumamos una misma cantidad K a todas las observaciones, la media de los nuevos valores (que son $y_1 = x_1 + K$, $y_2 = x_2 + K$, ..., $y_N = x_N + K$), es igual a $\bar{x} + K$, tal como se puede apreciar si se calcula la media de las y :

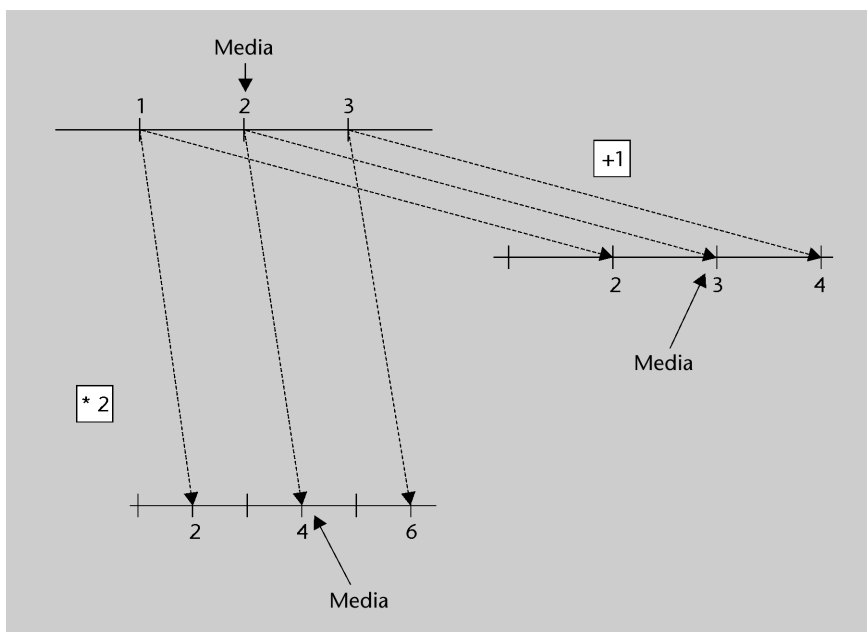
$$\begin{aligned}\bar{y} &= \frac{y_1 + y_2 + \dots + y_N}{N} = \frac{(x_1 + K) + (x_2 + K) + \dots + (x_N + K)}{N} \\ &= \frac{(x_1 + x_2 + \dots + x_N) + NK}{N} = \frac{(x_1 + x_2 + \dots + x_N)}{N} + \frac{NK}{N} = \bar{x} + K\end{aligned}$$

3) Si multiplicamos todas las observaciones por una misma cantidad K , la media de los nuevos valores (que ahora son $y_1 = Kx_1$, $y_2 = Kx_2$, ..., $y_N = Kx_N$) resulta ser igual a $K\bar{x}$, tal como se ve cuando hacemos:

$$\begin{aligned}\bar{y} &= \frac{y_1 + y_2 + \dots + y_N}{N} = \frac{Kx_1 + Kx_2 + \dots + Kx_N}{N} \\ &= \frac{K(x_1 + x_2 + \dots + x_N)}{N} = K\bar{x}\end{aligned}$$

Ejemplos de efecto de algunas transformaciones en la media

El sueldo de tres compañeros es de 1, 2 y 3 miles de euros respectivamente; por tanto, la media de los sueldos es $\bar{x} = 2$. Si el sueldo de cada uno de los compañeros aumenta en 1.000 euros, la nueva media será $\bar{y} = \bar{x} + 1 = 3$. Si nuestros amigos multiplican por 2 sus sueldos, la nueva media será $\bar{y} = 2\bar{x} = 2 \cdot 2 = 4$ miles de euros. Podemos ver las transformaciones de los sueldos en el gráfico siguiente, en el que se aprecia cómo se desplaza la media de la misma forma que lo hacen las observaciones: se suma 1 en el primer caso y se multiplica por 2 en el segundo:



En resumen

- 1) Si todas las observaciones son iguales, la media coincide con el valor de las observaciones.
- 2) Si sumamos un mismo número a todas las observaciones, la media aumenta en este mismo número.
- 3) Si multiplicamos todas las observaciones por un mismo número, la media se multiplica por este número.

4. Comparación media-mediana

De las dos medidas de centro, la media y la mediana, ¿cuál es la mejor? Desgraciadamente, en el mundo de la estadística los números son los que son, y no podemos decir que sean ni buenos ni malos. Unas veces va mejor la media y otras la mediana; en otros casos depende de lo que se quiera hacer con los datos.

Comparación media-mediana

Supongamos que la media de los días de estancia de los pacientes con un tratamiento concreto en el hospital es 113,1, y la mediana es 31. Si un paciente con este tratamiento nos pregunta cuánto tiempo es previsible que esté ingresado, ¿qué deberemos responder? Si respondemos la media, lo asustaremos demasiado, ya que de los ocho pacientes sólo dos han estado ingresados más de 113 días. En este caso podríamos optar por informarle del valor de la mediana y de su sentido: la mitad de estos pacientes están ingresados 31 días o menos. En cambio, si sabemos el coste diario de hospitalizar a este tipo de paciente y queremos calcular el coste total, parece que será más útil la media, ya que si multiplicamos la media de días de estancia por el coste diario y por el número de pacientes del tratamiento, obtendremos el coste total que comporta este tipo de paciente.

Otro aspecto que hay que destacar es que si la distribución de los datos presenta una forma simétrica, la media y la mediana coinciden y, si bien la media tiene tendencia a desplazarse hacia la cola larga de la distribución o los datos alejados, la mediana no se ve tan afectada.

Ejemplo de efecto de los datos alejados sobre la media y la mediana

Si disponemos de los datos siguientes, correspondientes a la lectura de los tres termómetros que tenemos en casa (uno al lado del otro): 20,9, 21 y 21,1, resulta que tanto la media como la mediana son 21. Si otro día los datos son 20,9, 21 y 36,1, la nueva media es 26, mientras que la mediana se mantiene en 21. (En este caso seguro que ha pasado algo raro: ¡o no hemos leído bien uno de los termómetros o uno de éstos se ha estropeado!).

5. Medidas de centro y datos tabulados

Muchas veces accedemos a información que se presenta en forma de tabla de frecuencias (absolutas o relativas) de los valores de una variable. En este caso podemos calcular fácilmente la moda, la mediana y la media de las observaciones resumidas en la tabla; lo explicaremos con varios ejemplos.

Ejemplo de los virus atacantes (I)

Consideremos una variable x , cuyas observaciones corresponden al número de virus que han atacado cada uno de los 31 ordenadores de nuestra empresa durante el año 2000. Después de agrupar los datos, obtenemos la tabla siguiente:

x_i	n_i
0	10
5	15
9	6

En otras ocasiones la información nos llega en forma de tabla de frecuencias relativas. En este caso no conocemos los valores de todas las observaciones de la variable, sino que sólo sabemos el porcentaje de veces en que se ha observa-

do cada valor. Entonces, a partir de la fórmula $\bar{x} = \frac{\sum_{i=1}^k n_i x_i}{N}$ de la media, deducimos:

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

Ejemplo de la página de Guillermo Puertas

Es conocido que una de las peores páginas personales de Internet es la de Guillermo Puertas, propietario de una conocida empresa de *hardware*. Se ha hecho un estudio del tiempo (en segundos) que tarda un internauta en salir de la página a partir del momento en que aparece la foto de Puertas, y se han obtenido los resultados siguientes:

Clase	m_i	f_i
[0,1)	0,5	80%
[1,5)	3	10%
[5,15)	10	10%

Según estos datos, el 80% de los internautas tarda menos de 1 segundo en salir de la página web; el 10%, entre 1 y 5 segundos, y el 10% restante, entre 5 y 15 segundos (el extremo superior no se considera dentro de los intervalos). En este caso la media del tiempo que se tarda en salir es:

$$\bar{x} = 80\% \times 0,5 + 10\% \times 3 + 10\% \times 10 = 1,7 \text{ segundos}$$

Todavía podemos considerar otros casos en los que la información nos llega en forma de tabla de frecuencias relativas o absolutas de las clases en las que hemos distribuido los valores que toma una variable determinada; en este caso procederemos como en los dos anteriores, pero tomaremos como representante de cada clase la marca correspondiente, es decir:

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

donde m_i es la marca correspondiente a cada clase y k es el número de clases en las que hemos distribuido la variable.

6. Resumen

En esta sesión se han presentado tres medidas estadísticas que permiten describir el centro de la distribución de una variable según diferentes puntos de vista. La moda, que es el valor más frecuente; la mediana, que es el valor central una vez que los datos están ordenados, y la media, que hace que la suma de las desviaciones con respecto a este valor sea 0, se han definido y estudiado. Se explican las propiedades de la media y se comparan las características de la mediana y de la media. También se han tratado los casos en los que se dispone de datos en forma de tabla de frecuencias (absolutas o relativas).

Ejercicios

1. Los sueldos en € de los analistas contratados por una empresa informática son:

1.150	1.200	1.300	1.450	1.000	1.350	1.900	1.230	1.450	1.010	1.500	1.450
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

Calculad el sueldo medio y la mediana de los sueldos.

¿Cuánto aumentan la media y la mediana de los sueldos en las situaciones siguientes?

- Si se decide un aumento lineal (la misma cantidad para todos).
- Si se decide un aumento de sueldo de un 12% para todos los trabajadores.
- Si se aumenta en 100 euros el sueldo del trabajador que cobra más, si se aumenta en 100 euros el sueldo del trabajador que cobra menos y si se aumenta en 100 euros el sueldo de un trabajador cualquiera.

2. Un estudio de la universidad ha recopilado el tiempo en minutos que sus estudiantes dedican semanalmente a los *chats*, y se ha obtenido la tabla siguiente:

x_i	f_i
[0,10)	1%
[10,30)	15%
[30,200)	44%
[200,400)	40%

Calculad la media del tiempo que los estudiantes dedican a chatear y representad gráficamente la distribución de la variable.

Solucionario

1.

a) La media es 1.332,5 euros. Los datos ordenados y sus respectivas posiciones una vez ordenados son los siguientes:

1	2	3	4	5	6	7	8	9	10	11	12
1.000	1.010	1.150	1.200	1.230	1.300	1.350	1.450	1.450	1.450	1.500	1.900

Puesto que hay 12 observaciones, $N = 12$ y $(N + 1) / 2 = 6,5$. Por tanto, la mediana es $(1.300 + 1.350) / 2 = 1.325$ euros, ligeramente inferior a la media.

a) En general, si el sueldo aumenta en K euros, la media y la mediana también aumentan en K euros.

b) Aumentar el sueldo en un 12% equivale a multiplicar cada sueldo por 1,12. Por tanto, la media y la mediana también se multiplicarán por 1,12, es decir, aumentarán en un 12%.

c) La media aumenta en $100/12$ €, ya que si x_1, x_2, \dots, x_n son los sueldos de los trabajadores, la media después de subir 100 euros a cualquier trabajador será:

$$\frac{x_1 + x_2 + \dots + x_{12} + 100}{12} = \frac{x_1 + x_2 + \dots + x_{12}}{12} + \frac{100}{12} = \bar{x} + \frac{100}{12}$$

donde \bar{x} es la media de los sueldos originales. Si aumentamos el sueldo al trabajador que cobra más o que cobra menos, la mediana no cambia, ya que las observaciones que permiten calcularla no se ven afectadas. La mediana sólo cambia en estos tres casos:

- Si cambiamos el sueldo de 1.230 a 1.330, la nueva mediana será $(1.330 + 1.350)/2$
- Si cambiamos el sueldo de 1.300 a 1.400, la nueva mediana será $(1.350 + 1.400)/2$
- Si cambiamos el sueldo de 1.350 a 1.450, la nueva mediana será $(1.300 + 1.450)/2$

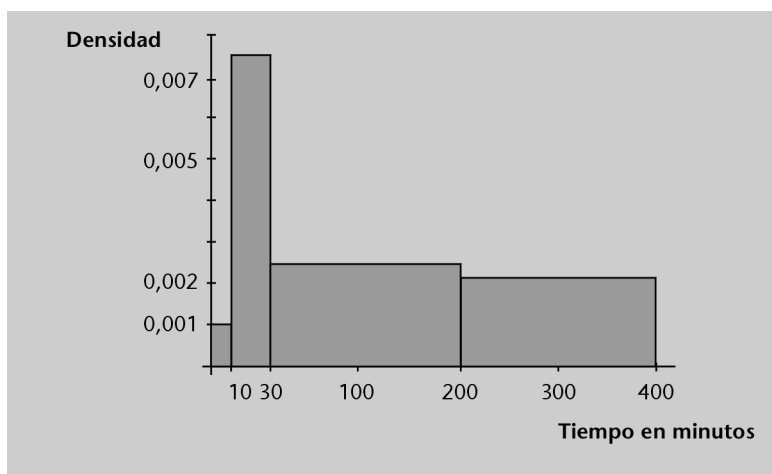
2. Añadimos una columna a la tabla:

x_i	f_i	Altura del rectángulo
[0,10)	1%	$1\% / 10 = 0,001$
[10,30)	15%	$15\% / 20 = 0,0075$
[30,200)	44%	$44\% / 170 = 0,0025$
[200,400)	40%	$40\% / 200 = 0,002$

A partir de las marcas de clase, la media será la siguiente:

$$\bar{x} = 1\% \times 5 + 15\% \times 20 + 44\% \times 115 + 40\% \times 300 = 173,65 \text{ minutos}$$

Dado que las clases no tienen el mismo ancho, representaremos gráficamente la distribución de la variable con un histograma de densidad que elaboraremos a partir de los datos de la tabla.



Medidas de dispersión

En esta tercera sesión aprenderemos a asignar, calcular e interpretar diferentes indicadores numéricos que nos ayudan a describir cómo se distribuyen los datos en torno a su lugar “central”. Evidentemente, la manera de distribuir los datos depende de qué entendamos por “centro”. Por eso estudiaremos por separado la dispersión con respecto a la mediana y con respecto a la media. Comencemos con un ejemplo introductorio:

Necesidad de las medidas de dispersión

Con la media no es suficiente: hay que buscar algún indicador que nos ayude a averiguar cómo se distribuyen los datos en torno a este valor, que nos hable de las posibles fluctuaciones del valor de la variable. Lo mismo nos pasará con la mediana.

Motivación para la introducción de las medidas de dispersión

Las cotizaciones de las acciones de una importante compañía de telecomunicaciones al cierre y durante tres días consecutivos han sido de 9, 10 y 11 euros. Los mismos días las cotizaciones de las acciones de otra compañía de telecomunicaciones han sido de 1, 10 y 19 euros. A simple vista se ve que las medias de las cotizaciones han sido iguales para las dos compañías, 10 euros, aunque el comportamiento de las cotizaciones es bastante diferente: las de la segunda han experimentado unos cambios muy grandes en su cotización, ya que han tenido una mayor variabilidad o más dispersión.

1. Los cuartiles y la mediana

Comenzaremos con el estudio del caso en el que la mediana es un buen indicador del centro de los datos e introduciremos herramientas para describir cómo se distribuyen los datos en cada uno de los dos grupos en los que la mediana divide a la población. La idea de *cuartil* proviene del interés en dividir los datos ordenados en cuatro grupos con aproximadamente el mismo número de individuos para poder ver “el espacio que ocupa” cada grupo en relación con los demás. Por tanto, definimos los elementos siguientes:

a) **Primer cuartil (Q1)**: es aquel valor numérico tal que al menos el 25% de las observaciones son menores o iguales que aquél, y al menos el 75%, mayores o iguales.

b) **Segundo cuartil (Q2)**: es la mediana.

c) **Tercer cuartil (Q3)**: es aquel valor numérico tal que al menos el 75% de las observaciones son menores o iguales que aquél, y al menos el 25%, mayores o iguales.

A continuación presentamos un procedimiento de cálculo de los cuartiles que aprovecha las operaciones efectuadas para calcular la mediana:

1) Ordenamos las observaciones en sentido ascendente y calculamos $(N+1)/2$ y la mediana.

Procedimiento de cálculo para calcular los cuartiles Q1 y Q3.

2) Distribuimos las observaciones en dos grupos iguales:

- el grupo de las que ocupan una posición estrictamente menor que $(N + 1) / 2$;
- el grupo de las que ocupan una posición estrictamente mayor que $(N + 1) / 2$.

3) Una vez hecho esto:

- El primer cuartil (Q1) es la mediana del grupo de los datos que ocupan una posición estrictamente menor que $(N + 1) / 2$.
- El tercer cuartil (Q3) es la mediana de los datos que ocupan una posición estrictamente mayor que $(N + 1) / 2$.
- El segundo cuartil (Q2) es la mediana.

Ejemplos de cálculo de cuartiles

a) Consideremos las observaciones siguientes: 17, 16, 12, 13, 15, 14. Aplicaremos el procedimiento que acabamos de describir para calcular los cuartiles:

1) Primero ordenamos los datos:

12	13	14	15	16	17
----	----	----	----	----	----

También tenemos que $N = 6$ y $(N + 1) / 2 = 3,5$. Por tanto, la mediana es $Q2 = 14,5$.

2) Distribuimos los datos en dos grupos:

Ocupan una posición estrictamente menor que 3,5	12	13	14			
Ocupan una posición estrictamente mayor que 3,5				15	16	17

3) Calculamos Q1 como la mediana del primer grupo y Q3 como la mediana del segundo; por tanto, $Q1 = 13$ y $Q3 = 16$. Vemos que hay dos observaciones menores o iguales que 13, y 5 mayores o iguales; puesto que el 25% de 6 es 1,5 y el 75% de 6 es 4,5, al menos el 25% de los datos son menores o iguales que Q1 y al menos el 75% son mayores o iguales.

b) Consideremos las observaciones de cierta variable:

12	13	14	15	16	17	200
----	----	----	----	----	----	-----

En este caso tenemos $N = 7$ y $(N + 1) / 2 = 4$. La mediana es $Q2 = 15$. Como en el ejemplo anterior, separamos los datos en dos grupos:

Ocupan una posición estrictamente menor que 4	12	13	14			
Ocupan una posición estrictamente mayor que 4				16	17	200

por tanto, $Q1 = 13$ y $Q3 = 17$.

c) Consideremos las observaciones de cierta variable:

12	13	14	15	15	15	200
----	----	----	----	----	----	-----

Ahora, $N = 7$ y $(N + 1) / 2 = 4$. La mediana es $Q2 = 15$. Si separamos los datos en los dos grupos de antes:

Ocupan una posición estrictamente menor que 4	12	13	14			
Ocupan una posición estrictamente mayor que 4				15	15	200

por tanto, $Q1 = 13$ y $Q3 = 15$. En este caso $Q2 = Q3 = 15$.

Las definiciones de mediana y cuartiles

La mediana y los cuartiles no tienen una definición única. En diferentes libros aparecen diferentes métodos de cálculo y los programas de ordenador utilizan diferentes algoritmos, que dan lugar a resultados diferentes. En general, si hay muchos datos y pocos "agujeros" entre éstos, los resultados serán muy similares. En cualquier caso, es conveniente informarse de cómo se calculan los valores mencionados en cada caso.

1.1. El rango intercuartílico

El rango intercuartílico es la diferencia entre el tercer y el primer cuartil, es decir:

$$\text{Rango intercuartílico} = Q_3 - Q_1$$

Dado que entre Q_1 y Q_3 se distribuyen aproximadamente el 50% de las observaciones centrales de la variable, el rango intercuartílico es una medida de la dispersión de este colectivo. Así, un rango intercuartílico pequeño significa que los datos centrales están muy apretados, mientras que un rango intercuartílico grande indica una fuerte dispersión.

1.2. Los cinco números resumen y el diagrama de caja

Los cinco números resumen de la distribución de una variable son: el mínimo, Q_1 , la mediana, Q_3 y el máximo.

Existe una manera muy eficiente de representar los cuartiles y los cuatro grupos en los que los cuartiles dividen las observaciones: es el diagrama de caja o *box-plot*.

Procedimiento para dibujar un diagrama de caja.

Un **diagrama de caja** se construye de la manera siguiente:

1) Se marca un eje (horizontal o vertical) con la escala adecuada de la variable y se marcan los cinco números resumen de la distribución.

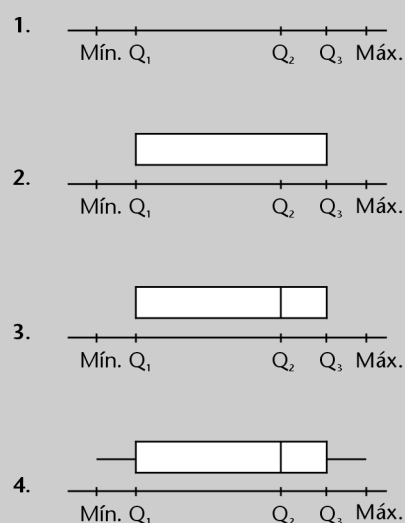
2) Se dibuja un rectángulo (por encima o al lado derecho del eje) que tenga un lado en el valor correspondiente a Q_1 y otro en el valor Q_3 .

3) Dentro del rectángulo se dibuja una línea de lado a lado en el valor que corresponde a la mediana.

4) Se dibujan dos brazos que van desde el punto medio de los lados construidos a Q_1 y Q_3 hasta los valores mínimo y máximo, respectivamente.

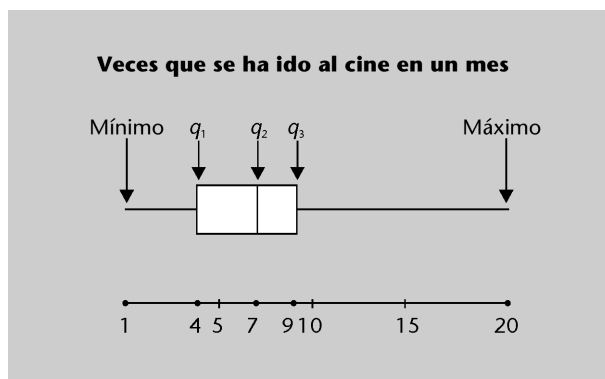
Como siempre, es más difícil describir su aspecto con palabras que dibujar uno.

Procedimiento para dibujar un diagrama de caja



Ejemplo de elaboración de un diagrama de caja: veces que se va al cine

Preguntamos a algunas personas cuántas veces han ido al cine en los últimos treinta días. Después de procesar los datos, obtenemos los siguientes cinco números resumen: mínimo = 1, $Q_1 = 4$, mediana = 7, $Q_3 = 9$, máximo = 20. El diagrama de caja correspondiente es el siguiente:

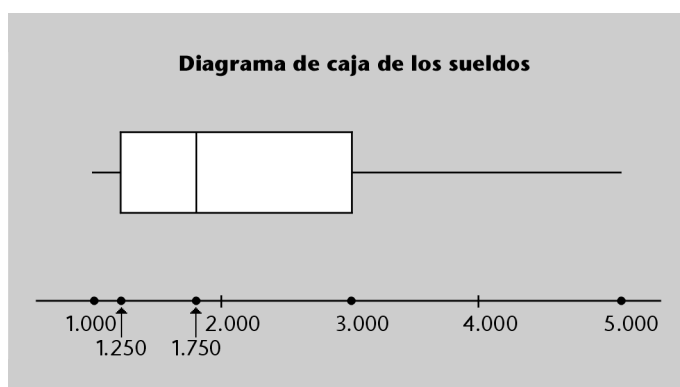


Observamos que la población que estudiamos se divide en cuatro grupos con aproximadamente el mismo número de personas. El primer grupo ha ido al cine el último mes entre 1 y 4 veces, el segundo entre 4 y 7, el tercero entre 7 y 9 y el último grupo, entre 9 y 20. El gráfico muestra que no existe simetría con respecto a la mediana, que hay una cola hacia la derecha (el brazo de Q_3 al máximo es más largo que el del mínimo a Q_1), que los datos están más apretados en el intervalo que va de Q_2 a Q_3 (recordad que en los cuatro grupos hay aproximadamente el mismo número de individuos y, por tanto, aproximadamente el 25% de los encuestados fue entre 7 y 9 veces al cine) y que los datos están más dispersos en el intervalo que va de Q_3 al máximo (lo que significa que la cuarta parte de la población que va más al cine –entre 9 y 20 veces– lo hace con frecuencias bastante diferentes). Con respecto al 50% de la población con valores más centrales, podemos ver que el rango intercuartílico es $Q_3 - Q_1 = 5$ y que, por tanto, este colectivo distribuye el número de veces que ha ido al cine entre 6 posibilidades, de 4 a 9 veces.

Es un error frecuente pensar que los diagramas de caja son todos iguales y que determinan cuatro regiones iguales: el diagrama de caja distribuye las observaciones en cuatro grupos que contienen aproximadamente el mismo número de observaciones, pero no todos los grupos ocupan el mismo “territorio”.

Ejemplo de distribución de las observaciones en un diagrama de caja: los sueldos

Un estudio muestra que los cinco números resumen correspondientes a la distribución de los sueldos mensuales de una empresa determinada son: mínimo = 1.000, $Q_1 = 1.250$, mediana = 1.750, $Q_3 = 3.000$, máximo = 5.000. Dibujamos el diagrama de caja correspondiente y obtenemos:



En este diagrama se ve que, a medida que los sueldos crecen, también son más diferentes entre sí: mientras que la cuarta parte de los trabajadores que cobran menos tienen sueldos muy similares (entre 1.000 y 1.250), la cuarta parte de los trabajadores que cobran más tienen sueldos muy diferentes (van de 3.000 a 5.000). El rango intercuartílico en este caso es $3.000 - 1.250 = 1.750$; esto significa que en el intervalo del 50% de los sueldos más centrales hay una oscilación de 1.750 euros entre el más bajo y el más alto.

Asimetría en la economía

Siempre que se estudian sueldos, gastos, ingresos, etc., nos encontramos con situaciones asimétricas hacia la derecha. Esto se debe a los valores extremos.

1.3. Interpretación de los diagramas de caja

Los diagramas de caja dan mucha información sobre la simetría de los datos y sobre cómo se distribuyen las observaciones en cuatro grupos que contienen aproximadamente el mismo número de observaciones. También permiten visualizar el rango intercuartílico. Entre los “defectos” que presentan hay que destacar que, tal como hemos descrito aquí, el aspecto de estos diagramas depende bastante de los valores máximo y mínimo (por tanto, de dos valores individuales) y hay que tener presente esta información cuando investiguemos brazos que van del mínimo a Q1 y de Q3 al máximo.

Los diagramas de caja resultan muy útiles para comparar la distribución de una variable en diferentes poblaciones o muestras.

Consultad la utilidad de los diagramas de cajas para comparar muestras diferentes en el ejercicio 1 de este apartado.

2. La desviación típica y la media

Una manera de saber cómo se distribuyen los datos en torno a la media podría ser calcular el valor medio de las desviaciones con respecto a la propia media. Desgraciadamente, este valor medio es 0, porque la suma de las desviaciones es 0 (los valores positivos se cancelan con los negativos). Para evitar estas cancelaciones, podemos tomar el valor de las desviaciones al cuadrado (que son todas positivas). Esta idea motiva la definición de varianza.

La **varianza** (también llamada **varianza poblacional**) y que se denota por s^2 se calcula como la suma del cuadrado de las desviaciones con respecto a la media dividido por N , donde N es el número de las observaciones, es decir:

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{N} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

Varianza muestral y varianza poblacional

Existe “otra” varianza, llamada *varianza muestral*, que se obtiene como la poblacional pero dividiendo por $N - 1$. La varianza muestral tiene una gran importancia cuando se estudia la inferencia.

Es fácil comprobar que la varianza se puede calcular también a partir de la fórmula siguiente:

$$s_x^2 = \frac{\sum_{i=1}^N x_i^2}{N} - (\bar{x})^2$$

Escribimos s_x^2 si queremos destacar que calculamos la varianza de la variable x .

La **desviación típica** (denotada por s o s_x) se define como la raíz cuadrada positiva de la varianza, es decir:

$$s = \sqrt{s^2}$$

Ejemplo de cálculo de la desviación típica

Consideremos los valores 7, 4, 6, 5, 5. Para calcular su varianza, organizamos los cálculos en una tabla como ésta:

	x_i	$x_i - \bar{x} = x_i - 5,4$	$(x_i - \bar{x})^2 = (x_i - 5,4)^2$
	7	$7 - 5,4 = 1,6$	$(1,6)^2 = 2,56$
	4	$4 - 5,4 = -1,4$	$(-1,4)^2 = 1,96$
	6	$6 - 5,4 = 0,6$	$(0,6)^2 = 0,36$
	5	$5 - 5,4 = -0,4$	$(-0,4)^2 = 0,16$
	5	$5 - 5,4 = -0,4$	$(-0,4)^2 = 0,16$
Suma de la columna	27	0	$2,56 + 1,96 + 0,36 + 0,16 + 0,16 = 5,2$

Donde la media es:

$$\bar{x} = \frac{7 + 4 + 6 + 5 + 5}{5} = \frac{27}{5} = 5,4$$

la varianza es:

$$s_x^2 = \frac{5,2}{5} = 1,04$$

y la desviación típica es:

$$s = \sqrt{s^2} = \sqrt{1,04} = 1,02$$

Desviación típica muestral y poblacional

Evidentemente, también disponemos de dos desviaciones típicas: la muestral y la poblacional. Muchas calculadoras tienen una tecla para cada una: σ_n para la poblacional y σ_{n-1} para la muestral.

Cálculo de la varianza

Para calcular la varianza, hay que dar los pasos siguientes:

- 1) calcular la media;
- 2) calcular las desviaciones con respecto a la media;
- 3) calcular el cuadrado de las desviaciones con respecto a la media;
- 4) sumar el cuadrado de las desviaciones y dividir por el número de individuos.

Cálculo de la desviación típica

Para calcular la desviación típica, hacemos la raíz cuadrada positiva de la varianza.

2.1. Propiedades de la varianza y la desviación típica

Dado que en la definición de la varianza y la desviación típica interviene la media, y ésta es sensible a los datos extremos, la varianza y la desviación típica también se ven afectadas por los valores extremos.

Las transformaciones lineales en los datos iniciales afectan a la varianza y a la desviación típica de la manera siguiente:

- 1) Si todas las observaciones tienen el mismo valor, es decir, si $x_1 = x_2 = \dots = x_n$, la varianza y la desviación típica valen 0:

$$\begin{aligned} s^2 &= \frac{(x_1 - \bar{x}) + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{N} = \\ &= \frac{(x_1 - x_1)^2 + (x_1 - x_1)^2 + \dots + (x_1 - x_1)^2}{N} = 0 \end{aligned}$$

ya que todas las observaciones son iguales e iguales a su media.

La varianza de los valores 3, 3, 3, 3 es 0 y su media 3.

Observación

Fijaos en que la varianza siempre es positiva.

2) Si sumamos una misma cantidad K a todas las observaciones, la varianza y la desviación típica no cambian, tal como se ve si se calcula la varianza de los valores nuevos (que son $y_1 = x_1 + K$, $y_2 = x_2 + K$, ..., $y_N = x_N + K$), y utilizando que la media de los nuevos valores sea $\bar{y} = \bar{x} + K$:

$$\begin{aligned} s_y^2 &= \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \dots + (y_N - \bar{y})^2}{N} = \\ &= \frac{((x_1 + K) - (\bar{x} + K))^2 + ((x_2 + K) - (\bar{x} + K))^2 + \dots + ((x_N + K) - (\bar{x} + K))^2}{N} = \\ &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{N} = s_x^2 \end{aligned}$$

Observad de qué forma K "desaparece".

3) Si multiplicamos todas las observaciones por una misma cantidad K , la varianza de los valores nuevos (que ahora son $y_1 = Kx_1$, $y_2 = Kx_2$, ..., $y_N = Kx_N$) se obtiene multiplicando la varianza de los datos originales por K^2 ; la desviación típica de los nuevos valores se obtiene multiplicando la desviación típica de los datos originales por $|K|$. Recordemos que en este caso la media de los nuevos valores es $\bar{y} = K\bar{x}$ y que, por tanto:

$$\begin{aligned} s_y^2 &= \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \dots + (y_N - \bar{y})^2}{N} = \\ &= \frac{(Kx_1 - K\bar{x})^2 + (Kx_2 - K\bar{x})^2 + \dots + (Kx_N - K\bar{x})^2}{N} = \\ &= \frac{K^2(x_1 - \bar{x})^2 + K^2(x_2 - \bar{x})^2 + \dots + K^2(x_N - \bar{x})^2}{N} = \\ &= K^2 s_x^2 \end{aligned}$$

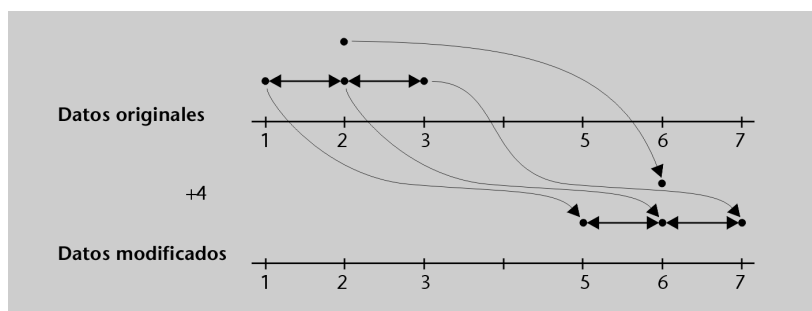
$|K|$ es el valor absoluto de K .

Finalmente: $s_y = \sqrt{s_y^2} = \sqrt{K^2 s_x^2} = |K| s_x$.

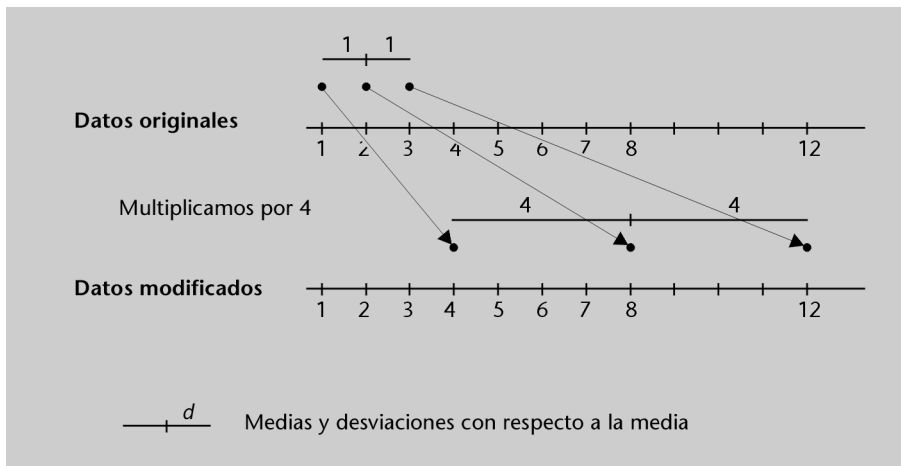
El valor absoluto $|K|$ hace que s_y sea positiva, aunque K sea negativa.

Propiedades de la varianza y la desviación típica

Si partimos de los valores 1, 2, 3 que tienen media 2 y varianza $2/3$, y sumamos 4 a todos, obtenemos los valores 5, 6 y 7, que tienen media $6 = 2 + 4$ y varianza $2/3$, igual que la de los datos iniciales. En el gráfico observamos que la media se desplaza, mientras que las desviaciones (y, por tanto, la varianza y la desviación típica) se mantienen.



Si partimos de los valores 1, 2 y 3, que tienen media 2 y varianza $2/3$, y los multiplicamos todos por 4, obtenemos los valores 4, 8 y 12, que tienen una media de $8 = 4 \cdot 2$, una varianza de $10,66 = 4^2 \cdot 2/3$ y una desviación típica de $|4| \cdot \sqrt{2/3} = 3,26$.



2.2. La regla de Tchebichev

La regla de Tchebichev permite entender mejor el papel que tiene la desviación típica como medida de dispersión de los datos en torno a la media. Comenzaremos por enunciar la regla, después daremos algunos casos particulares de la misma y estudiaremos su uso, pero no lo justificaremos aquí.

La regla de Tchebichev afirma que, dado cualquier conjunto de datos x_1, x_2, \dots, x_n con media \bar{x} y desviación típica s_x , si m es un número cualquiera, entonces la proporción de datos que pertenecen al intervalo $(\bar{x} - ms_x, \bar{x} + ms_x)$ es, como mínimo:

$$1 - \frac{1}{m^2}$$

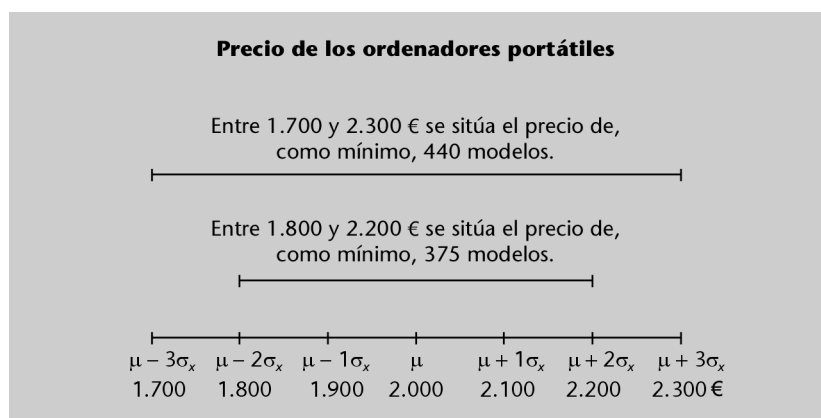
La tabla siguiente registra algunos casos concretos:

m	Intervalo	Porcentaje mínimo de observaciones en el intervalo previsto por la regla de Tchebichev
1	$(\bar{x} - 1s_x, \bar{x} + 1s_x)$	$1 - \frac{1}{1^2} = 0$
2	$(\bar{x} - 2s_x, \bar{x} + 2s_x)$	$1 - \frac{1}{2^2} = 0,75 = 75\%$
3	$(\bar{x} - 3s_x, \bar{x} + 3s_x)$	$1 - \frac{1}{3^2} = 0,88 = 88\%$
4	$(\bar{x} - 4s_x, \bar{x} + 4s_x)$	$1 - \frac{1}{4^2} = 0,93 = 93\%$
...
m	$(\bar{x} - ms_x, \bar{x} + ms_x)$	$1 - \frac{1}{m^2}$

Por ejemplo, podemos estar seguros de que, sean cuales sean nuestros datos, en el intervalo $(\bar{x} - 4s_x, \bar{x} + 4s_x)$ siempre encontraremos más del 93% de los datos, y así con otros valores de m (excepto $m = 1$ sobre el cual, curiosamente, la regla no dice nada).

Ejemplo de los ordenadores portátiles

Recogemos información sobre el precio de todos los modelos de ordenador portátil que se pueden encontrar en el mercado. Supongamos que en total tenemos 500 modelos diferentes, que la media de los precios es 2.000 euros y la desviación típica es 100 euros. En este caso podemos estar seguros de que el precio de, como mínimo, 375 (= 75% de 500) portátiles está en el intervalo $(2.000 - 2 \cdot 100, 2.000 + 2 \cdot 100) = (1.800, 2.200)$ y de que el precio de como mínimo 440 ordenadores portátiles (88% de 500) está en el intervalo $(2.000 - 3 \cdot 100, 2.000 + 3 \cdot 100) = (1.700, 2.300)$. Gráficamente:



Como veremos en los ejercicios, la regla de Tchebichev es muy “pesimista”: muchas veces encontraremos más datos de los que predice; esto se debe a que sirve para todos los conjuntos de datos, sean cuales sean y, por tanto, debe ser poco restrictiva.

2.3. Datos estandarizados

Otro concepto que ayuda a interpretar las relaciones entre la media y la desviación típica es el de datos estandarizados. Dado un conjunto de N observaciones de una variable con valores x_1, x_2, \dots, x_N , cuya media es \bar{x} con desviación típica s_x , el valor estandarizado de x se define de la manera siguiente:

$$z = \frac{x - \text{Media de las } x}{\text{Desviación típica de las } x} = \frac{x - \bar{x}}{s_x}$$

Habitualmente consideraremos un conjunto de observaciones con la media y la desviación típica correspondientes y estandarizaremos todas las observaciones, es decir, transformaremos sus valores originales en los valores estandarizados correspondientes:

$$x_i \xrightarrow{\text{Estandarizar}} z_i = \frac{x_i - \bar{x}}{s_x}$$

Dado un conjunto de observaciones cualquiera, los **valores estandarizados** correspondientes tienen media 0 y desviación típica 1, ya que si los datos estandarizados son z_1, z_2, \dots, z_N , se verifican las propiedades siguientes:

$$\begin{aligned} \text{a) } \bar{z} &= \frac{z_1 + z_2 + \dots + z_N}{N} = \frac{\frac{x_1 - \bar{x}}{s_x} + \dots + \frac{x_N - \bar{x}}{s_x}}{N} = \\ &= \frac{x_1 + \dots + x_N - N\bar{x}}{Ns_x} = \frac{0}{Ns_x} = 0 \end{aligned}$$

$$\begin{aligned} \text{b) } s_z^2 &= \frac{(z_1 - \bar{z})^2 + \dots + (z_N - \bar{z})^2}{N} = \frac{(z_1 - 0)^2 + \dots + (z_N - 0)^2}{N} = \\ &= \frac{\left(\frac{x_1 - \bar{x}}{s_x}\right)^2 + \dots + \left(\frac{x_N - \bar{x}}{s_x}\right)^2}{N} = \frac{(x_1 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{s_x^2 N} = \\ &= \frac{1}{s_x^2} \frac{(x_1 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{N} = \frac{s_x^2}{s_x^2} = 1 \end{aligned}$$

c) Otra propiedad interesante es que el valor estandarizado correspondiente a una observación determinada nos indica cuántas desviaciones típicas dista de la media. Esto se ve muy fácilmente si se aísla la x en la expresión del valor estandarizado:

$$z = \frac{x - \bar{x}}{s_x} \xrightarrow{\text{implica}} x = \bar{x} + zs_x$$

de donde se obtiene que el valor de una observación determinada es igual a la media más el producto de su valor estandarizado por la desviación típica.

Si la distribución de los datos es lo bastante simétrica, podemos utilizar la media y la desviación típica para describirla de forma rápida. En este caso el uso de los valores estandarizados nos permite entender muy rápido dónde se encuentra un dato en relación con su media y comparar muy fácilmente datos que pertenecen a conjuntos de observaciones diferentes.

Ejemplo de estandarización de variables estadísticas

Supongamos un conjunto de datos con media 15 y desviación típica 3. Si una observación tiene un valor estandarizado de 3, entonces:

$$z = \frac{x - \bar{x}}{s_x} = 3 \longrightarrow x = \bar{x} + 3s_x = 15 + 3 \cdot 3 = 24$$

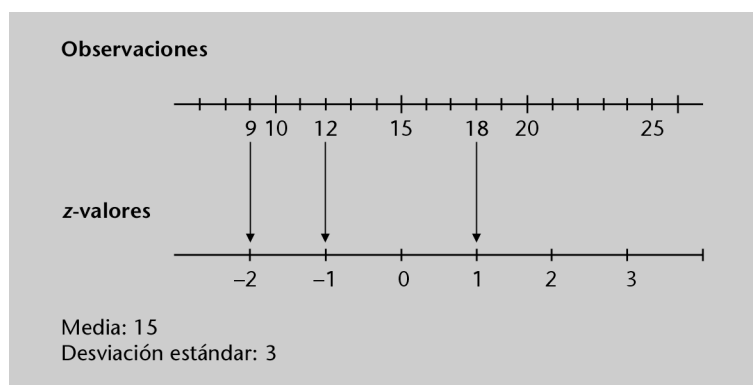
Si el valor de una observación es $x = 9$, entonces su valor estandarizado es -2 , que indica que está a -2 desviaciones típicas de la media de la distribución. Observad estas situaciones en el gráfico siguiente:

Nomenclatura

Los valores estandarizados también se denominan *z-valores*.

Recordad que la suma de los valores de las observaciones es el producto de la media por el número de individuos.

El concepto *estandarización* es clave en el estudio de las variables aleatorias y del modelo normal.



Las notas de álgebra y estadística

Las notas de los exámenes de álgebra y estadística de cinco alumnos han sido las que se muestran a continuación:

Álgebra	1	2	6	4	5
Estadística	5	7	6	8	9

Supongamos que un alumno ha sacado un 6 en las dos materias; ¿en cuál ha sacado mejor nota? Evidentemente, un 6 es un 6, pero ahora nos planteamos cómo está situado este 6 en relación con el resto de las notas en cada materia. Si denominamos x_i a las notas de álgebra e y_i a las notas de estadística, la media de las notas de álgebra es $\bar{x} = 3,6$ y la desviación típica es $s_x = 1,8547$, mientras que en el caso de estadística la media es $\bar{y} = 7$ y la desviación típica es $s_y = 1,4142$. Así pues, en relación con las notas de álgebra, un 6 tiene un valor estandarizado de $(6 - 3,6) / 1,8547 = 1,29$, mientras que en relación con la estadística el mismo 6 tiene un valor estandarizado de $(6 - 7) / 1,4142 = -0,7071$ (un valor estandarizado negativo indica que estamos a la izquierda de la media). Por tanto, sacar un 6 en álgebra equivale a estar 1,29 desviaciones típicas por encima de la media de la nota de álgebra, mientras que sacar un 6 en estadística equivale a situarse 0,7071 desviaciones típicas por debajo de la media (el valor estandarizado es negativo). De aquí podemos concluir que en relación con estos dos conjuntos de notas destaca más un 6 en álgebra que un 6 en estadística.

3. Usos de la media y la desviación típica o de la mediana y los cinco números resumen

Hasta ahora hemos visto diferentes maneras de describir con números y gráficos los diferentes valores que puede tomar una variable (la distribución de la variable); ahora podemos preguntarnos cuándo conviene utilizar cada una de estas maneras. Como podéis suponer, no podemos dar reglas fijas. Dado que es muy difícil argumentar el uso de los resúmenes numéricos, veremos qué opina sobre el tema un importante autor:

“Una distribución asimétrica con unas pocas observaciones en la cola larga de la distribución tendrá una desviación típica grande. En tal caso, la desviación estándar no proporciona una información demasiado útil. Como en una distribución muy asimétrica la dispersión de las colas es muy distinta, es imposible describir bien la dispersión con un solo número. Los cinco números-resumen, con los dos cuartiles y los dos valores extremos, proporcionan una información mejor. Es preferible utilizar los cinco números-resumen en lugar de la media y la desviación típica para describir una distribución asimétrica. Utiliza la media y la desviación típica sólo para distribuciones razonablemente simétricas. [...] Recuerda que la mejor visión global de una distribución la da un gráfico. Las medidas numéricas de centro y de dispersión reflejan características concretas de una distribución,

pero no describen completamente su forma. Los resúmenes numéricos no detectan, por ejemplo, la presencia de múltiples picos ni de espacios vacíos. [...] REPRESENTA SIEMPRE TUS DATOS GRÁFICAMENTE.”

D.S. Moore. *Estadística aplicada básica*.

4. Varianza y datos tabulados

De la misma manera que se ha hecho con la media, hay que adaptar las fórmulas de cálculo de la varianza al caso en que los datos estén en forma de tabla, ya sea de frecuencias relativas o absolutas. Así, utilizaremos las fórmulas siguientes:

$$s^2 = \frac{\sum_{i=1}^k n_i (x_i - \bar{x})^2}{N} = \frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{N} \text{ o bien } s^2 = \frac{\sum_{i=1}^k n_i (m_i - \bar{x})^2}{N} = \frac{\sum_{i=1}^k f_i (m_i - \bar{x})^2}{N}$$

en caso de que los datos estén agrupados en intervalos con marca de clase m_i .

Ejemplo de cálculo de la varianza a partir de datos tabulados

En el ejemplo de los virus atacantes (I), en el que se calcula el número de virus que han atacado los diferentes ordenadores de nuestra empresa durante el año 2000, la media vale 4,16. Para calcular la varianza, nos puede ayudar la tabla siguiente:

x_i	n_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$n_i \cdot (x_i - \bar{x})^2$
0	10	-4,16	17,3056	173,056
5	15	0,84	0,7056	10,584
9	6	4,84	23,4256	140,5536
Totales	31			324,1936

Y, por tanto, la varianza es $324,1936 / 31 = 10,45$.

5. Resumen

En esta sesión se han presentado varias medidas que permiten estudiar la distribución de los datos en torno a una medida de centro.

Si describimos el centro con la mediana, podemos complementar la información que aporta con los cuartiles y el rango intercuartílico; en este caso resulta especialmente interesante el diagrama de caja.

La varianza y la desviación típica ayudan a entender cómo se distribuyen los datos en torno a la media; por eso hay que estudiar las propiedades de la varianza y la desviación típica, entre las cuales destaca la regla de Tchebichev.

Finalmente, se define el concepto de **estandarización de datos**, que da el número de desviaciones típicas que un valor dista de la media.

Ejercicios

1. Se ha medido el tiempo en segundos que tarda en arrancar la última versión del programa Macrohard Phrase en los ordenadores de nuestra empresa según el sistema operativo con el que funcionan. Los resultados han sido los siguientes:

- En los ordenadores equipados con Doors95: 27, 25, 50, 33, 25, 86, 28, 31, 34, 36, 37, 44, 20, 59 y 85 segundos.
- En los ordenadores equipados con Doors98: 33, 7, 25, 14, 5, 31, 19, 10, 29 y 18 segundos.

Calculad los cinco números resumen y la media de la distribución correspondiente al tiempo que el programa tarda en arrancar y dibujad algunos gráficos que os parezcan relevantes para comparar el tiempo que el programa tarda en arrancar según el sistema operativo. A partir de estos gráficos, comparad el comportamiento del programa según el sistema operativo y explicad si creéis que hay diferencia entre utilizar Doors95 y Doors98.

2.

a) Confeccionad una lista de 10 números tales que mínimo = 2, máximo = 20, primer cuartil = 5, mediana = 10 y tercer cuartil = 19.

b) Igual que antes, pero con la media = 11.

c) En las condiciones del punto a, ¿la media podría ser igual a 21? (Recordad que la suma de las desviaciones con respecto a la media ha de ser 0).

3. Los siguientes datos corresponden al número de veces que el programa Minihard Word se “colgó” durante un mes en cada uno de los 50 ordenadores de nuestra empresa. Comprobad que se satisface la regla de Tchebichev para $m = 1$, $m = 2$ y $m = 3$.

0	9	12	14	19
2	10	12	14	20
4	10	12	14	20
5	10	12	15	21
6	11	12	15	22
6	11	12	17	29
6	11	12	17	29
7	11	13	18	32
8	11	13	18	39
9	12	14	19	39

Solucionario

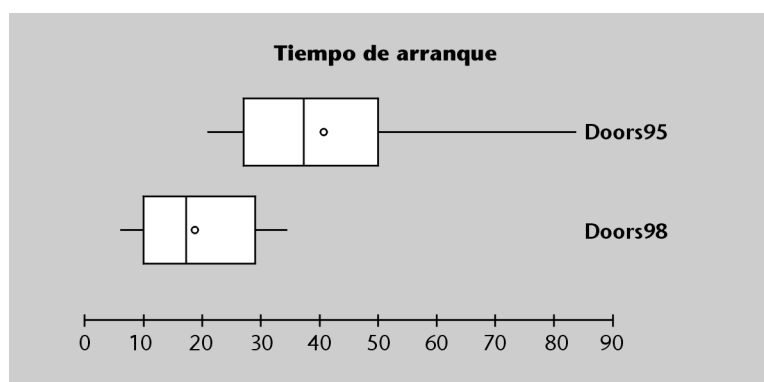
1. Los datos ordenados son:

- Doors95: 20, 25, 25, 27, 28, 31, 33, 34, 36, 37, 44, 50, 59, 85, 86
- Doors98: 5, 7, 10, 14, 18, 19, 25, 29, 31, 33

Esta tabla registra los cinco números resumen y la media:

Sistema	Media	Mínimo	Q1	Q2	Q3	Máximo
Doors95	41,33	20	27	34	50	86
Doors98	19,1	5	10	18,5	29	33

Si representamos simultáneamente y con la misma escala los diagramas de caja correspondientes a las dos distribuciones, obtenemos el gráfico siguiente, en el que también se han marcado con círculos las medias correspondientes:



A partir de los resúmenes numéricos y de los gráficos podemos concluir que Doors98 tarda mucho menos en arrancar que Doors95; hay que destacar los aspectos siguientes:

- Tanto la media como la mediana de Doors98 son aproximadamente la mitad que las de Doors95.
- La mediana de Doors98 es menor que el mínimo de Doors95; esto significa que la mitad de las veces Doors98 tarda menos en arrancar que la vez que más rápido ha arrancado Doors95.
- El máximo de Doors98 es menor que la mediana (es decir, la mitad de las veces que arrancamos Doors95 tarda más que la vez que más ha tardado con Doors98).
- El diagrama de caja de Doors98 es mucho más simétrico y no tiene ninguna cola hacia valores alejados o mayores de los normales. Esta simetría su-

giere que Doors98 tiene un comportamiento más regular y más estable que Doors95.

2.

a) Supongamos que tenemos una lista ordenada de 10 números. La mediana ocupará la posición $10 + 1/2 = 5,5$ y, por tanto, se calculará como valor medio de la 5.^a y 6.^a observaciones. En estas circunstancias el primer cuartil debe ser la 3.^a observación y el tercer cuartil, la 8.^a, tal como se ve en esta tabla, en la que también escribimos el mínimo y el máximo (M1 y M2 indican las observaciones que se necesitan para calcular la mediana).

Posición	1	2	3	4	5	6	7	8	9	10
	Mínimo		Q1		M1	M2		Q3		Máximo
	2		5					19		20

Observamos que en esta tabla el máximo, el mínimo y el primer y tercer cuartiles están fijados. Podemos acabar de rellenar la tabla de diferentes maneras, siempre que las observaciones aparezcan ordenadas y que la mediana tenga el valor 10; por ejemplo:

Posición	1	2	3	4	5	6	7	8	9	10
	Mínimo		Q1		M1	M2		Q3		Máximo
	2	3	5	9	10	10	18	19	19	20
	2	2	5	6	8	12	13	19	19,5	20
	2	3	5	5	5	15	18	19	20	20

b) Si la media ha de ser 11, debemos tener cuidado en que la suma de los valores sea igual a $10 \times 11 = 110$. Dado que las posiciones 1, 3, 8 y 10 tienen el valor fijado y la suma de los valores 5.^o y 6.^o debe ser 20, la suma de los valores que aparecen en las posiciones 2, 4, 7 y 9 tiene que ser $110 - 20 - 2 - 5 - 19 - 20 = 44$. Así, por ejemplo, podemos hacer:

Posición	1	2	3	4	5	6	7	8	9	10
	Mínimo		Q1					Q3		Máximo
	2	<u>3</u>	5	<u>5</u>	8	12	<u>17</u>	19	<u>19</u>	20

c) La media no puede superar la observación mayor. Una posible razón es que si lo fuese, todas las desviaciones con respecto a la media serían estrictamente negativas y, por tanto, la suma de las desviaciones no podría ser 0.

3. Si calculamos la media y la desviación típica, obtenemos $\bar{x} = 14,28$; $\sigma_x = 8,11$; con lo que los intervalos interesantes, las proporciones de observa-

ciones en cada intervalo y el mínimo predicho por la regla de Tchebichev se pueden ver en la tabla siguiente:

m	Intervalo	Observaciones en el intervalo	Porcentaje de observaciones en el intervalo	Porcentaje mínimo previsto por la regla de Tchebichev
1	$(\bar{x} - 1 s_x, \bar{x} + 1 s_x) = (6, 18, 22, 38)$	38	$38/50 = 76\%$	0%
2	$(\bar{x} - 2 s_x, \bar{x} + 2 s_x) = (-1, 92, 30, 48)$	47	$47/50 = 94\%$	75%
3	$(\bar{x} - 3 s_x, \bar{x} + 3 s_x) = (-10, 03, 38, 56)$	48	$48/50 = 96\%$	88%

Primero se calculan los intervalos sustituyendo los valores de m , \bar{x} y s_x .

El porcentaje de observaciones...

... en el intervalo se calcula contando cuántas observaciones caen dentro del intervalo.

Vemos que en todos los casos tenemos más observaciones en los intervalos señalados que las previstas por la regla. Recordad que la regla indica cuáles son los porcentajes más bajos que podemos encontrar en cada intervalo.

