

# Capítulo 2:

## Estadística descriptiva

### Presentación

En este capítulo se introducen los estadísticos y los gráficos más adecuados a cada escala de medida, así como las medidas de posición relativa de las unidades. Todo ello se aplica a la descripción de los participantes en un estudio.

### Objetivos

#### Al terminar este capítulo, un lector que haya realizado los ejercicios:

- Definirá media, mediana, moda, desviación típica, cuartiles y desviación intercuartil.
- Delante de los resultados de un estudio, se preguntará por el grado de dispersión de una variable.
- A partir de la desviación típica y de la media construirá un intervalo aproximado que contenga los valores observados.
- Interpretará la desviación típica como el promedio de las diferencias con la media.
- Interpretará un valor tipificado como la distancia a la media expresada en número de desviaciones típicas.
- Identificará como estadísticamente raro (extremo) un valor que se distancie de la media más de 2 (o 3) desviaciones típicas.
- En distribuciones asimétricas recurrirá a la distancia intercuartil en lugar de a la desviación típica.
- Usará los diagramas de barras y gráficos de sectores para representar variables cualitativas y variables discretas.
- Usará histogramas y diagramas de caja (*box-plot*) para representar variables cuantitativas.
- Observará si los ejes de los gráficos están completamente indicados.
- Deducirá, a partir de un *box-plot*, los valores de los cuartiles.

## Estadístico más adecuado para cada escala de medida

Veamos a continuación cómo la escala de medida puede ayudar a escoger el estadístico con el que se resumirá el conjunto de los datos. Empezaremos con las medidas de posición central, que informan sobre cómo son las observaciones prototípicas.

### Estadísticos de tendencia central

Si las variables están en escala nominal, el parámetro más relevante para caracterizar su distribución es la probabilidad de las categorías más repetidas. En algunas ocasiones, para resumir estas variables, se las representa por su categoría más frecuente, estadístico que se conoce por el nombre de **moda**.

#### Recuerde



*La moda representa a la categoría que más se repite.*

#### Lectura



*Martín et al. (12), al describir los pacientes de su estudio, dice: «Los tumores de estadio II fueron los más frecuentes (55,5%)». Nótese que dan la moda pero que, además, concretan a cuántos casos representa.*

Si las variables están en la escala ordinal, es posible utilizar aquellas medidas que se basan en la posibilidad de ordenar las observaciones. En general, usan las probabilidades acumuladas, que suman las de las categorías anteriores o menores. Así, si se desea situar alrededor de qué valor se encuentran los valores observados, se puede recurrir a la **mediana** o valor del individuo por debajo del cual se encuentra el 50% de las unidades.

Existen otras muchas medidas basadas en el orden de las observaciones. Los percentiles dividen la muestra en 100 partes, los deciles en 10, los quintiles en 5 y los cuartiles en 4. Conviene notar que, para dividir la muestra en cuatro partes, bastan tres cuartiles.

#### Ejemplo 2.1

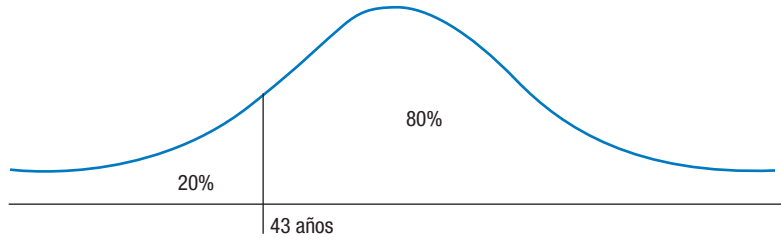


La edad de los pacientes incluidos en un estudio tiene la distribución que se muestra en la figura 2-1. El percentil 20 deja por debajo el 20% de las observaciones, igual que el 2.º decil y el 1.º quintil. Todos ellos toman, por tanto, el mismo valor: 43 años.

#### Ejercicio 2.1



La mediana, ¿a qué percentil corresponde? ¿Y a qué cuartil?



**Figura 2-1** El percentil 20 y el decil 2 son 43 años.

### Ejercicio 2.2



¿A qué percentil corresponde el cuartil 1? ¿Y el cuartil 2?  
¿Y el cuartil 3?

Si las variables están en escala de intervalo, entonces sus valores pueden sumarse, ya que todas ellas se basan en una misma unidad de medida que tiene el mismo significado, independientemente de dónde se haya obtenido: es lo mismo 1 cm aportado por un individuo de 180 cm que 1 cm de otro individuo de 150 cm. Así, para conocer el centro de la distribución puede recurrirse al promedio o **media**: se suman los valores obtenidos en todas las observaciones y se reparten entre el número total de casos.

### Lectura



*El grupo para la Asistencia Médica Integrada Continua de Cádiz (13), al describir sus resultados, sostiene: «La media de pruebas solicitadas por paciente es [...] menor [...] que las del grupo control».*

### Ejemplo 2.2



Estudiemos la media con la ayuda de un ejemplo. Se ha preguntado, a los 5 últimos pacientes que han entrado en la consulta, por el número de parejas que han tenido en los últimos 48 meses. Han contestado que 1, 3, 4, 5 y 7 parejas, respectivamente.

Dado que la suma total de parejas es 20, el promedio «que le corresponde a cada uno» es de 4 parejas:

$$\begin{array}{r} 1 \\ + 3 \\ + 4 \\ + 5 \\ + 7 \\ \hline \end{array}$$

$$\text{suma} = \sum_{i=1,5} X_i = 20$$

$$\text{De donde el promedio o media es: } \sum_{i=1,5} X_i / n = 20/5 = 4$$

**Nota técnica**



$\sum_{i=1,5} X_i$  representa la suma de los valores que la variable  $X$  toma en los individuos 1 a 5. Simboliza el «sumatorio desde  $i = 1$  hasta  $i = 5$  de  $X$  sub  $i$ ».

Con un promedio de 4 parejas por paciente, un investigador descuidado, que se olvidara de la riqueza de la variabilidad y de la existencia de diferencias entre las unidades, podría decir que *cada uno* de estos 5 pacientes ha tenido 4 parejas en los últimos 48 meses. ¡Qué sorpresa para el de 1 pareja! Y qué forma de decir mentiras. La tabla 2-1 muestra cuánto valen estas mentiras. Nótese que su suma es igual a 0.

Dicen ellos	Se les asigna	Mentira resultante
1	4	+3
3	4	+1
4	4	0
5	4	-1
7	4	-3
Suma	20	0

**Tabla 2-1** Mentira o error resultante si se interpreta que cada paciente tiene exactamente el valor de la media

**Estadísticos de dispersión**

La media representa el centro de la distribución, pero ¿hasta qué punto representa a cada individuo? Sería ingenuo creer que todas las observaciones se sitúan en la media. Además, esta simplicidad implicaría perder toda la información contenida en su diversidad. Por ello, la siguiente medida de interés consiste en estudiar cuál es la distancia que suelen tener las observaciones respecto a ese centro que representa la media.

**Definición**



La **desviación típica** o desviación estándar (DE) representa el alejamiento prototípico con el centro.

**Ejemplo 2.2 (Cont.)**



Si a cada uno de ellos se le dice que ha tenido 4 parejas, las mentiras respectivas serán +3, +1, 0, -1 y -3. Ahora bien, como el investigador descuidado es, además, terco, insiste en que su cálculo es acertado ya que la suma de las mentiras da 0 y, por tanto, su mentira promedio es también 0. La media, como centro de gravedad de la variable, tiene

Ejemplo 2.2 (Cont.)



esta propiedad: se compensan los desvíos positivos con los negativos. Para poder valorar la «mentira promedio», se elevan estas distancias al cuadrado antes de sumarlas:

Dicen ellos	Se les asigna	Mentira resultante	Mentira <sup>2</sup>
1	4	+3	9
3	4	+1	1
4	4	0	0
5	4	-1	1
7	4	-3	9
Suma	20	0	20

Tabla 2-2 Cuadrado de la mentira si se interpretara que cada paciente tiene el valor medio

Ahora, la suma de las mentiras cuadradas es 20 parejas<sup>2</sup>. Si las parejas que han tenido entre todos se reparten «equitativamente» en los 5 casos, se observa una «mentira<sup>2</sup> promedio» de 4 parejas<sup>2</sup>, cálculo conocido por el nombre de **varianza**. Para eliminar ese engorroso «cuadrado», se hace la raíz cuadrada, de donde se obtiene que la mentira prototípica es de 2 parejas. Este valor representa, pues, la distancia o desvío (con la media) típico de todas las observaciones. Por esta razón recibe el nombre de desviación típica.

Ejemplo 2.3



Uso de la media y de la desviación típica. Cien niños tratados han tenido fiebre durante una media de 3 días. La desviación típica (o estándar) ha sido de 1 día. Se están describiendo los resultados obtenidos en la muestra: el centro se ha situado en 3 días y los niños se alejaban de este centro, en promedio, 1 día (se entiende que se alejaban por arriba y por abajo).

Para interpretar si la desviación típica es grande o pequeña es útil el siguiente truco. Como promedio de todas las distancias, quiere decir que habrá distancias mayores y distancias menores, que se equilibrarán mutuamente. Así, para «compensar» un valor que coincida exactamente con la media, es decir, que tenga un desvío igual a 0, se necesita otro valor que tenga un desvío que sea el doble de la desviación típica.

Ejemplo 2.4



Si la media de la fiebre era de 3 días y la desviación típica de 1 día, puede interpretarse que, para tener un desvío promedio de 1 día, los casos se distanciarán aproximadamente entre 0 y 2 días de la media. Así, la distancia máxima con

**Ejemplo 2.4 (Cont.)**

la media será, en este cálculo aproximado, de 2 días. Por tanto, en general, los niños han tenido fiebre entre 1 y 5 días. (Nota: éste es un cálculo aproximado que más adelante se afinará teniendo en cuenta la forma de la distribución.)

**Ejemplo 2.5**

Soriano et al. (14). «La edad media (DE) de los 11 pacientes con infección de prótesis total de cadera era de 69 (10) años [...]. Se interpreta que el centro de la distribución está en 69 años. Pero esto no significa que todos los pacientes tengan 69 años, sino que están a su alrededor. La distancia o desviación típica que mantienen con el centro vale 10. Esta cifra representa el alejamiento “típico” de 69. Así, algunos casos estarían más cerca y otros más lejos. De manera aproximada, puede decirse que un caso que está justo en el centro (y tiene una distancia de 0) se compensa con un caso que tiene una distancia que dobla la desviación típica (20 años). Así, en esta primera aproximación, cabe imaginar que estos 11 pacientes tienen edades comprendidas entre los 49 y los 89 años.»

**Comentario**

*Un cálculo mental aproximado de la desviación típica, en una variable con distribución simétrica, consiste en dividir entre 2 la distancia entre el valor más alto (o el más bajo) y la media.*

**Ejercicio 2.3**

El personal de cierto hospital camina a una velocidad media de 3 km/h, siendo los extremos de velocidad 2 y 4 km/h aproximadamente. ¿Qué valor cree que puede tener la desviación típica?

**Recuerde**

*La varianza es el promedio de las distancias con la media elevadas al cuadrado. La desviación típica es su raíz cuadrada y valora el promedio de las distancias con la media: representa la distancia típica o esperada de una observación con la media.*

**Recuerde (Cont.)**

La desviación típica muestral se representa por  $S$ . En Medicina Clínica se representa por  $DE$  (desviación estándar) y en las revistas inglesas por  $SD$  (standard deviation).

**Ejercicio 2.4**

Los 21 pacientes tenían una edad media ( $DE$ ) de 82 (8) años. Interprete la media y la desviación típica. ¿Entre qué márgenes aproximados cabe esperar que fluctúe la edad de estos pacientes?

La desviación típica es el estadístico por excelencia para valorar las dispersiones, pero requiere que exista escala de intervalo.

**Comentario**

Se ha visto que existe escala de intervalo cuando hay unidad de medida. Es decir, cuando un aumento de una unidad siempre significa lo mismo. Esta situación es verosímil cuando la variable es simétrica. Si, por ejemplo, se estudia la variable salario, ¿significa lo mismo un aumento mensual de 100 € para quien gana 500 € que para quien gana 5.000 €? Posiblemente tampoco significa lo mismo un aumento de las GOT de 10 a 40 que de 110 a 140.

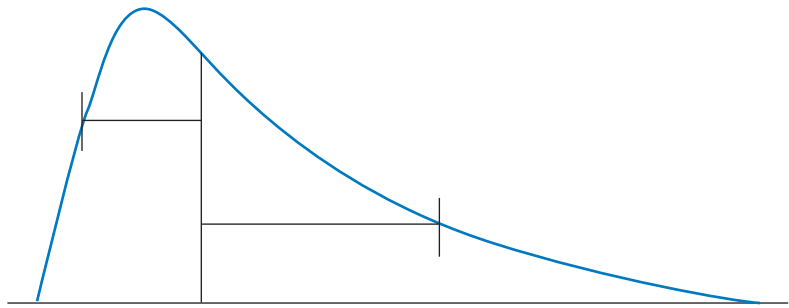
**Comentario**

Las variables salario y GOT tienen una marcada asimetría, con una cola muy larga en el extremo superior (fig. 2-2). En esta situación, la desviación típica pierde sentido, ya que no puede interpretarse de la misma forma en ambas colas de la distribución.

Por ejemplo, una persona que gane poco, se distanciará del salario promedio menos que una persona que gane mucho. Por tanto, un mismo estadístico, la desviación típica, no puede representar bien desvíos que son diferentes a ambos lados de la media.

**Recuerde**

Para poder interpretar, con la misma desviación típica, distancias por encima como por debajo de la media, se requiere que la distribución sea simétrica.



**Figura 2-2** Si la distribución es asimétrica, la desviación típica no puede representar simultáneamente los desvíos superiores e inferiores.

Si las variables son muy asimétricas puede recurrirse también a los cuartiles. Asimismo, para valorar la dispersión en la escala ordinal es muy útil la diferencia entre el primer y el tercer cuartil, conocida como **distancia intercuartil**.

La tabla 2-3 muestra los estadísticos de tendencia central y de dispersión que pueden aplicarse en las diferentes escalas de medida, así como las propiedades mínimas que requiere cada estadístico. Por ejemplo, la media sólo puede ser utilizada en escala de intervalo, pero la moda puede ser empleada en cualquier escala.

Escala	Propiedades	Tendencia central	Dispersión
Nominal	Equivalencia	Moda	
Ordinal	Orden	Mediana	Distancia intercuartil
Intervalo	Unidad	Media	Desv. típica = $\sqrt{\text{Varianza}}$

**Tabla 2-3** Estadísticos apropiados según la escala de medida

### Ejercicio 2.5



a) Suponga que se ha medido la presión arterial sistólica a 5 pacientes: 115, 117, 124, 135 y 142 mmHg. Sin hacer el cálculo, diga qué valor aproximado le parece correcto para la media:

- i) 115 mmHg
- ii) 125 mmHg
- iii) 135 mmHg

b) Suponga ahora que el resultado observado en los 5 pacientes ha sido 100, 125, 130, 135 y 160 mmHg, con una media de 130 mmHg. Sin hacer el cálculo, diga qué valor aproximado le parece correcto para la desviación típica:

- i) 5 mmHg
- ii) 20 mmHg
- iii) 35 mmHg

**Ejercicio de Navegación**

Entre en la página que se indica al final del párrafo, dentro del apartado Statistics → Statoscope. Este *applet* calcula los estadísticos de interés. Existen dos opciones: introducir manualmente un conjunto de datos o simular un conjunto de datos con una determinada media y desviación estándar. Introduzca manualmente algunos datos e intente adivinar los valores de la media y de la desviación típica.

<http://www.stat.duke.edu/sites/java.html>

**Comentario**

*En las sociedades industriales predominaba el paradigma de la uniformidad, hasta el punto de que las diferencias con el patrón estándar, con la media, recibían el nombre de desvíos. En la sociedad de la información se abre paso el paradigma biológico de la diversidad, y las diferencias empiezan a ser consideradas un valor positivo y los ordenadores intentan imitar las redes neuronales para acercarse a la inteligencia natural.*

**Historieta**

*Demos pues la bienvenida a la diversidad y olvidemos las connotaciones negativas del término desviación. Un término de connotaciones menos negativas, especialmente en el ejemplo de las parejas, podría ser «diversión típica». Seguiremos buscando...*

**Nota técnica**

El cálculo de la varianza presentado ha dividido la suma de las distancias cuadradas por el número de observaciones, pero puede verse que una de las observaciones no estaba aportando ninguna distancia. Si tuviéramos sólo una observación se podría estimar la media, pero no la dispersión. El hecho de estimar la media y la desviación típica en la misma muestra implica gastar una pieza de información, lo que se denomina «perder un grado de libertad». El estadístico más habitual para el cálculo de la varianza divide entre «n-1» (número de casos menos uno) en lugar de entre «n». Los libros de estadística matemática explican sus ventajas.

**Definición**

$$\text{Varianza muestral } S^2 = \frac{\sum (x_i - \bar{x})^2}{(n-1)}$$

$$\text{Desviación típica muestral } S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n-1)}}$$

$$\text{Fórmulas abreviadas } S^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{\sum x_i - \bar{x}^2 \cdot n}{n-1}$$

donde  $x_i$  representa el valor de la observación  $i$ -ésima y  $\bar{x}$  la media muestral.

**Ejercicio 2.6**

- a) Suponga ahora que el resultado observado en los 5 pacientes ha sido 100, 110, 120, 130 y 140 mmHg. Calcule la media, la varianza y la desviación típica.
- b) Suponga que se ha medido la presión arterial sistólica al mismo paciente 5 veces en la última visita, habiendo observado 125, 128, 130, 132 y 135 mmHg. Calcule la media, la varianza y la desviación típica.
- CONSEJO: hágalo con una hoja de cálculo.

**Comentario**

*La desviación típica del segundo enunciado es muy inferior, ya que sólo incluye las oscilaciones debidas a las fluctuaciones intracaso, que pueden ser debidas a cambios en el individuo o a errores en el proceso de medida. En el primer caso, aparecen las fluctuaciones a causa de las diferencias entre individuos. Nótese que la variabilidad entre casos es mayor que la variabilidad intracasos. Si esta última fuera mayor tendríamos una baja repetibilidad de los valores, lo que indicaría que la determinación es poco fiable y cuestionaría la utilidad del proceso de medida.*

**Medidas de posición relativa de los individuos**

La existencia de diferencias representa información. El hecho de que seamos diferentes nos permite distinguarnos. Para ello, puede resultar muy útil conocer cuál es la posición de una unidad respecto a otras unidades de su entorno.

**Ejemplo 2.6**

Vamos a visitar a un amigo al que hemos conocido en un chat de internet. Pongamos que vive en un poblado de África y que, para identificarlo, nos dice que él mide 170 cm. A medida que nos acercamos a su poblado nos entran dudas sobre si seremos capaces de reconocerlo. ¿Cuál debe ser la altura típica de su poblado? Podría ser que fueran muy altos. O todo lo contrario. Saber la media de la altura puede ser una gran ayuda. Pongamos que en su poblado dicha media sea de 150 cm. Por tanto, consideraremos como «altos» a todos los que midan más de 150 cm y «bajos» a los que midan menos. Ahora ya sabemos que tenemos que mirar hacia los altos, ya que nuestro conocido tiene una distancia positiva de 20 cm con la media del poblado. Ahora bien, podría ser que en dicho poblado existiera una gran dispersión y nuestro conocido pasara desapercibido dentro de los altos. O podría ser que todos los habitantes estuvieran muy cerca de la media y nuestro co-

### Ejemplo 2.6 (Cont.)



nocido enseguida resaltara. Ahora queremos saber cuánto vale la desviación típica. Si fuera de 20 cm, nuestro conocido sería alto, pero sin destacar entre los altos: sería un «alto típico». En cambio, si la desviación típica fuera de 2 cm, sabemos que la altura de nuestro conocido resaltará mucho entre las de sus vecinos.

### Definición



El procedimiento estadístico de tipificar o estandarizar el valor de una variable consiste en restarle la media y dividirlo por la desviación típica.

$$z = \text{desvío tipificado} = \frac{\text{valor observado} - \text{media}}{\text{desviación típica}}$$

Valores de  $z$  alrededor de 1 o  $-1$  representan distancias típicas al valor central. Valores cercanos a 0 representan valores muy próximos al centro de la distribución. Y valores de  $z$  mayores que 2 (o menores que  $-2$ ) representan individuos que se están alejando más del doble de lo que se aleja el individuo típico.

### Ejemplo 2.7



Si la desviación típica del poblado de nuestro amigo africano es de 20 cm, el desvío tipificado de nuestro amigo vale 1:

$$z_1 = \frac{170 - 150}{20} = 1$$

En cambio, si la desviación típica del poblado fuera 2 cm, el desvío tipificado de nuestro amigo sería 10:

$$z_2 = \frac{170 - 150}{2} = 10$$

### Comentario



Regla «a ojo de buen cubero». Hemos visto que si la desviación típica representa la distancia promedio, quiere decir que por cada caso que coincida con la media, que no se aleje nada, habrá un caso que se aleje 2 desviaciones típicas.

### Ejemplo 2.7 (Cont.)



El desvío tipificado de nuestro amigo de 1 en el poblado de desviación típica de 20 indica que es un alto típico. En cambio, el desvío de 10 (correspondiente al hipotético poblado con una desviación de 2 cm) indica que nuestro amigo tiene una altura atípica, extraordinariamente alta. Desde un punto de vista estadístico, se trata de un caso «raro» o **extremo**.

**Recuerde**

*Un caso que se aleje más de 2 desviaciones típicas está **fuera de la banda (outlier)** y puede considerarse como extremo en una primera aproximación.*

**Ejemplo 2.8**

Los «errores» en la duplicación del ADN introducen ciertas variaciones que se traducen en individuos de diferentes características. La evolución de las especies se produce porque el entorno selecciona a las unidades mejor adaptadas. La selección natural precisa, por tanto, de la existencia de variabilidad.

**Ejercicio 2.7**

En cierta población, el colesterol total tiene una media de 200 mg/dl y una desviación típica de 50 mg/dl. Un paciente con colesterol de 175, ¿qué desvío tipificado le corresponde? ¿Cómo interpreta este valor? ¿Y para un paciente con 350 mg/dl?

**Ejemplo 2.9**

Serían ejemplos de observaciones extremas, un individuo que midiera más de 210 cm (criterio univariante) y otro de 180 cm que pesara 55 kg (criterio bivariante).

Conviene distinguir entre situaciones imposibles (p. ej., 300 cm) o situaciones raras pero posibles (p. ej., 227 cm). Un *outlier* alerta sobre posibles errores de transcripción, o posibles contaminaciones de la muestra, pero no es ninguna prueba definitiva de dato erróneo, por lo que se deben consultar y revisar estas anomalías. No se aconseja eliminar un caso por criterios de «rareza» estadística.

Digamos, para terminar, que la variabilidad no tiene por qué ser necesariamente molesta. Al contrario, puede ser fuente de información y de mejora.

**Ejemplo 2.10**

Ciertas rutinas de programación generan, al azar, muchas posibles soluciones de un problema. Luego se mejoran, se seleccionan y se vuelve a añadir ruido para reiniciar este pequeño ciclo.

**Ejercicio 2.8**

La variable RFS tiene una media de 400 y una desviación típica de 150. Defina criterios para detectar datos «sospechosos» en las semanas 0, 6, 12 y 24 del estudio. ¿Qué hará con estos casos?

**Ejercicio 2.9**

Si consulta al investigador que generó los datos, ¿cuándo le parece más oportuno?

## Descripción de los participantes en un estudio

Para que el lector pueda apreciar hasta qué punto los resultados de un estudio pueden ser aplicados en su propio entorno, los autores de artículos científicos deben describir las condiciones en las que han sido recogidos los datos y las características de sus unidades estadísticas, sean pacientes, voluntarios sanos, determinaciones analíticas o muestras de tejido. Las recomendaciones CONSORT para informe de ensayos clínicos, en su ítem 15 dicen:

**Lectura**

*The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration (9). «La información basal se presenta de manera eficiente en una tabla (tabla 2-4). En lo que se refiere a las variables continuas, tal como el peso corporal o la presión arterial, se debería indicar la variabilidad de los datos junto con los valores medios. Las variables continuas pueden ser resumidas en cada grupo mediante los valores correspondientes a la media y a la desviación estándar. En los casos en los que los datos continuos presentan una distribución asimétrica, un planteamiento preferible puede ser el de presentar los datos correspondientes a la mediana y al rango de percentiles (quizá, los percentiles 25 y 75). El error estándar y el intervalo de confianza no son apropiados para describir la variabilidad debido a que ambos son parámetros estadísticos de tipo inferencial más que descriptivo. Las variables constituidas por un número pequeño de categorías ordenadas (como los estadios I a IV de la enfermedad) no deben ser consideradas como variables continuas; en estos casos, es necesario presentar los números y las proporciones correspondientes a cada categoría.»*

**Ejercicio 2.10**

¿Cómo representaría los resultados de éstas variables?

- Glicemia en ayuno en personas sanas
- Transaminasas en enfermos
- Grado de cardiopatía (nivel I a IV) según NYA
- Presión arterial sistólica

Características	Grupo vitaminas (n = 141)	Grupo placebo (n = 142)
Edad media $\pm$ SD, y	28,9 $\pm$ 6,4	29,8 $\pm$ 5,6
Fumadores, n (%)	22 (15,6)	14 (9,9)
Índice de masa corpora media $\pm$ SD, kg/m <sup>2</sup>	25,3 $\pm$ 6,0	295,6 $\pm$ 5,6
Tensión arterial media $\pm$ SD, mmHg		
Sistólica	112 $\pm$ 15	110 $\pm$ 12
Diastólica	67 $\pm$ 11	68 $\pm$ 10
Paridad, n (%)		
0	90 (65)	87 (61)
1	39 (28)	42 (30)
2	9 (6)	8 (6)
>2	2 (1)	5 (4)
Enfermedad concomitante, n (%)		
Hipertensión idiopática	10 (7)	7 (5)
Lupus o síndrome antifosfolipídico	4 (3)	1 (1)
Diabetes	2 (1)	3 (2)

**Tabla 2-4** Ejemplo de tabla con las características iniciales, clínicas y demográficas, de los grupos en comparación (9)

### Comentario



*Observe que esta guía dice que el error estándar y los intervalos de confianza (todavía no estudiados) no sirven para describir las condiciones iniciales de los casos.*

### Ejemplo 2.11



Bobes et al. (15). «Descripción de la muestra: Las características basales de los 168 sujetos incluidos en el estudio (52 pacientes estables, 116 pacientes inestables) se describen en la tabla 2-5. Los pacientes fueron en su mayoría mujeres (el 85 y 82%, respectivamente), con una media (DE) de edad de 47 (12) y 45 (13) años, respectivamente, y nivel de estudios primario. En ambos grupos, la mayoría de pacientes estaba en situación laboral activa (el 35 y el 47%), si bien también fue importante el porcentaje de amas de casa incluidas (el 29 y el 35%). El diagnóstico mayoritario fue el trastorno depresivo mayor de episodio único (el 31 y el 20% en pacientes estables e inestables, respectivamente) o recidivante (el 33 y 42%,

Variables	Pacientes estables (n = 52)	Pacientes inestables (n = 116)
Edad (años), media (DE)	47,5 (12,1)	45,2 (12,8)
Sexo		
Varones	8 (15,4)	21 (18,3)
Mujeres	44 (84,6)	94 (81,7)
Nivel de educación		
Sin estudios	3 (5,9)	8 (7,0)
Estudios primarios	33 (64,7)	72 (62,6)
Estudios secundarios	9 (17,6)	19 (16,5)
Estudios universitarios	6 (11,8)	16 (13,9)
Situación laboral		
Trabaja fuera de casa	18 (34,6)	53 (47,3)
Parado	2 (3,8)	7 (6,3)
Jubilado	2 (3,8)	3 (2,7)
Incapacidad laboral o invalidez permanente	13 (25,0)	9 (8,0)
Ama de casa	15 (28,8)	39 (34,8)
Estudiante	2 (3,8)	1 (0,9)
Diagnóstico (código DSM-IV)		
Trastorno depresivo mayor, episodio único (296,2)	16 (30,8)	23 (19,8)
Trastorno depresivo mayor, recidivante (296,3)	17 (32,7)	49 (42,2)
Trastorno distímico (300,4)	12 (23,1)	20 (17,2)
Trastorno adaptativo con depresión (309,0)	7 (13,5)	24 (20,7)
Tiempo de evolución del trastorno		
0-3 meses	7 (14,3)	35 (30,7)
4 meses-1 año	14 (28,6)	32 (28,1)
> 1 año	28 (57,1)	47 (41,2)
Gravedad del trastorno		
Un poco enfermo	2 (3,8)	7 (6,0)
Levemente enfermo	22 (42,3)	84 (72,4)
Moderadamente enfermo	22 (42,3)	24 (20,7)
Gravemente enfermo	6 (11,5)	1 (0,9)
Entre los casos más graves de la enfermedad		

**Tabla 2-5** Características sociodemográficas y clínicas de los pacientes en estudio

### Ejemplo 2.11 (Cont.)



respectivamente). En el grupo de pacientes estables, el siguiente diagnóstico en importancia fue el de trastorno distímico (23%), mientras que para el grupo de pacientes inestables fue el de trastorno adaptativo con depresión (21%). En ambos grupos, la mayoría de los pacientes presentó una duración del trastorno superior a un año (el 57 y el 41%, respectivamente), y la gravedad del trastorno, en función de la impresión clínica global (ICG), fue moderada o grave en más de la mitad de los casos (el 54% en el grupo de pacientes estables y el 93% en el grupo de pacientes inestables).»

## Comentario



*Se trata de una descripción de los casos observados: cómo eran y cómo evolucionan. Queda pendiente por aclarar qué información (y cuánta) aportan estos pacientes sobre cómo cabe esperar que evolucionen otros casos futuros de las mismas características.*

## Gráficos según escala de medida

Veamos a continuación cómo el tipo de variable y la escala de medida pueden ayudar a escoger el gráfico con el que se resumirá el conjunto de los datos. En los capítulos sucesivos se irán presentando los gráficos más adecuados para cada tipo de análisis.

## Lectura



*González et al. (16). «Cuando las relaciones entre variables son complejas, los procesos temporales juegan un papel primordial y el componente aleatorio enmascara los procesos en estudio, entonces la representación gráfica deviene una herramienta imprescindible. La biomedicina, en su sentido más amplio, desde las actividades de investigación hasta las de asistencia o de gestión, es un ámbito con estas características y donde los gráficos, bien utilizados, permiten una aproximación nueva y enriquecedora a la información disponible.»*

## Variables discretas

Los dos gráficos más apropiados para la representación de este tipo de variables son el **gráfico de sectores** y los **diagramas de barras**.

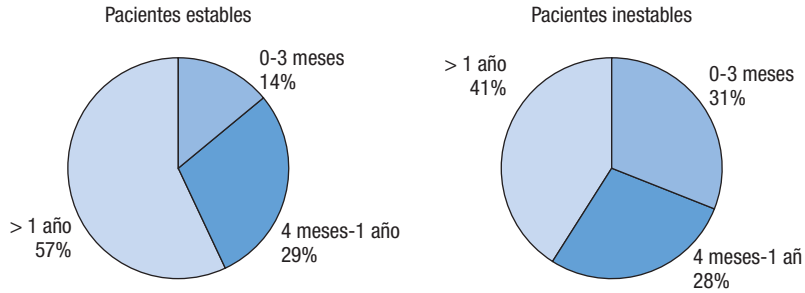
## Gráfico de sectores

Consiste en un círculo segmentado en sectores de tamaño proporcional a la frecuencia de cada uno de los valores de la variable. Este gráfico es apropiado cuando los valores de la variable, preferiblemente pocos, son excluyentes entre ellos.

## Ejemplo 2.11 (Cont.)



Bobes et al. (15). Si se representa la variable tiempo de evolución del trastorno para cada grupo de pacientes se obtiene la figura 2-3.



**Figura 2-3** Tiempo de evolución de pacientes, estables e inestables.

## Diagrama de barras

Este tipo de gráfico se emplea para variables nominales, ordinales y cuantitativas discretas. Consiste en un eje de coordenadas en el que se colocan los distintos valores de la variable en el eje horizontal, con un rectángulo cada uno de ellos de altura proporcional a la frecuencia del valor. En el eje vertical se presenta la escala que va desde 0 hasta, como mínimo, la frecuencia del valor modal.

### Ejemplo 2.12

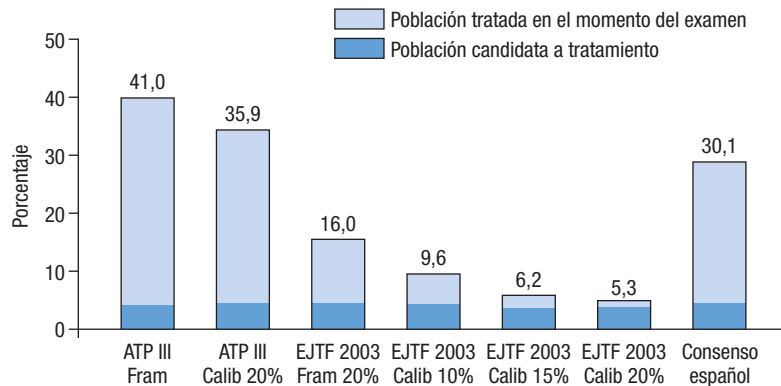


Ramos et al. (17). Hay que destacar las notables diferencias en la proporción poblacional de tratados entre las distintas recomendaciones de tratamiento de la hipercolesterolemia que se observa en la figura 2-4.

### Comentario



*Para que el gráfico proporcione una correcta impresión visual el eje de ordenadas debe empezar en 0. De no ser así, debe resaltarse para alertar al lector.*



**Figura 2-4** Tratamiento de la hipercolesterolemia.

## Variables continuas

En el caso de las variables continuas, existen multitud de gráficos, entre los que presentamos el **histograma** y el *box-plot*.

### Histograma

El histograma es una extensión del diagrama de barras que dibuja los rectángulos unidos entre sí, indicando de este modo que existe continuidad en los valores de las variables. Un histograma es, por tanto, un gráfico de variable continua dividida en intervalos de los que se eleva un rectángulo con área proporcional a su frecuencia.

#### Comentario



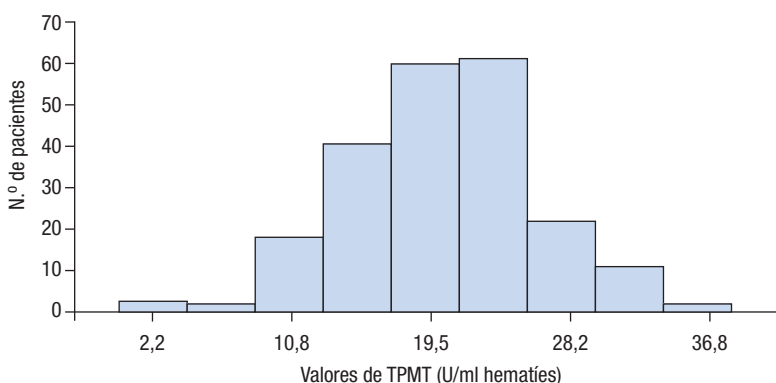
*Obsérvese que lo que es proporcional es el área, no la altura, lo que permite intervalos de diferente amplitud.*

Habitualmente, los intervalos son de igual amplitud.

#### Ejemplo 2.13



Figura 2-5. Distribución de los valores de tiopurina metiltransferasa (TPMT) en los pacientes con hepatitis autoinmune. Gisbert et al. (18).



**Figura 2-5** Distribución de los valores de tiopurina metiltransferasa (TPMT) en los pacientes con hepatitis autoinmune. Gisbert et al. (18)

#### Ejercicio de Navegación



Entre en la página que se indica al final del párrafo, dentro del apartado Distributions → Histograms. En este *applet* se encuentran representados en un histograma datos de las erupciones de un géiser. Se permite cambiar la amplitud de los intervalos, ¿puede cambiar la interpretación de los resultados según la amplitud del intervalo escogida?

<http://www.stat.duke.edu/sites/java.html>

A partir de un histograma pueden construirse otros tipos de gráficos. Por ejemplo, los gráficos de línea consisten en unir los puntos medios de todos los intervalos contiguos mediante una recta, construyendo así un polígono de frecuencias.

### Box-plot o diagrama de caja

En este gráfico (fig. 2-6) se representan los cuartiles. Un diagrama de caja o *box-plot* consta de un rectángulo cuya longitud es igual a la amplitud intercuartil, con una recta en su interior que representa la mediana; siendo los cuartiles 1 y 3, los límites inferior y superior de la caja. Por fuera de ésta, se dibujan dos rectas que, sin que tengan una longitud superior a una vez y media el rango intercuartil, llegan hasta el valor mínimo o máximo de la distribución.

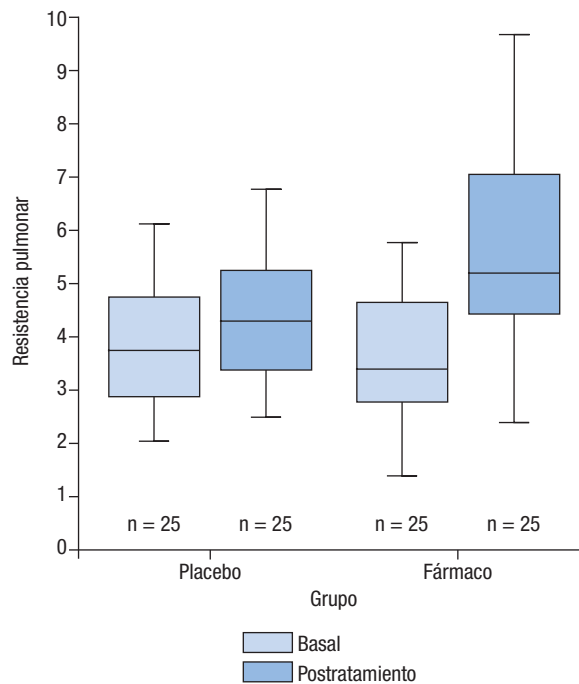
Este gráfico es muy útil, entre otros aspectos, para detectar rápidamente los valores extremos o atípicos (*outliers*), que el *box-plot* define como los individuos que se ubican por encima o por debajo de un rango y medio intercuartil, es decir, fuera de las dos rectas complementarias a la caja.

#### Ejercicio de Navegación



Entre nuevamente en la página que se indica al final del párrafo, dentro del apartado Statistics → Statoscope. Vuelva a simular conjuntos de datos y observe el *box-plot* correspondiente a cada uno de ellos.

<http://www.stat.duke.edu/sites/java.html>



**Figura 2-6** Ejemplo de *box-plot* o diagrama de caja. González et al. (16).

## Soluciones a los ejercicios

2.1 La mediana se corresponde con el percentil 50 y el cuartil 2.

2.2 El cuartil 1 se corresponde con el percentil 25; el cuartil 2, con el 50 y el cuartil 3, con el 75.

2.3 Si podemos aceptar que alguien que camina muy despacio va a 2 km/h y alguien muy rápido a 4 km/h, cabe esperar una desviación típica próxima al valor 0,5 km/h, dado que la mitad de  $4 - 2 = 2$  es 1, es 0,5.

2.4 El doble de la desviación típica es 16, que restado y sumado de 82, da 66 y 98. Se trata de una población anciana (82 años) pero que cubre un amplio margen, ya que fluctúa entre 66 y 98.

2.5 a) ii) De hecho, el valor exacto es 126,6 mmHg.

b) ii) De hecho, el valor exacto es 21,5 mmHg.

2.6 a) Media = 120 mmHg; varianza  $S^2 = \frac{\sum (x_i - \bar{x})^2}{(n - 1)} = 1000 / 4 = 250 \text{ mmHg}^2$ ,

luego la desviación típica es  $S = \sqrt{250 \text{ mmHg}^2} \approx 15,8 \text{ mmHg}$ .

b) Media = 130 mmHg; varianza  $S^2 = 58 / 4 \approx 14,5 \text{ mmHg}^2$ , luego la desviación típica es

$S = \sqrt{14,5} \approx 3,8 \text{ mmHg}$ .

2.7 Al paciente con un valor de 175 mg/dl le corresponde un desvío típico de  $-0,5$ , lo que indica que está ligeramente por debajo, ya que es negativo pero no alcanza la distancia habitual ( $= 1$ ) que guardan los valores bajos con la media. El paciente con un valor de 350 mg/dl tiene un desvío típico de  $+3$ , lo que indica que está muy por encima, ya que su distancia es 3 veces mayor que la distancia típica de todos los que están por encima. Estadísticamente, se trataría de un caso extremo.

2.8 Con esta media y esta desviación típica, la regla de «buen cubero» (aproximada) dice que los casos deberían estar comprendidos entre:

Valores = media  $\pm$  2 desviación típica =  $400 \pm 2 \cdot 150 \approx 400 \pm 300 = [100, 700]$

Así, los valores que fueran inferiores a 100 o superiores a 700 serían sospechosos de acuerdo con este criterio univariante. Con un criterio bivariante, podría establecerse como sospechoso a un paciente que sufriera variaciones de su CD4 superiores al, por ejemplo, 50%.

Estos casos deberían ser contrastados con mucho cuidado, de acuerdo con su historia clínica, a la búsqueda de posibles errores de transcripción. Si no se encuentran errores, el valor debe darse por bueno.

2.9 Por supuesto, lo más próximo al momento en el que se generó el dato. De lo contrario, puede llegar a ser imposible verificarlo.

2.10 a) Media y desviación típica, ya que por experiencia previa cabe esperar una distribución simétrica.

b) Mediana y cuartiles 1 y 3 (o percentiles 25 y 75, que son lo mismo), ya que no parece simétrica.

c) Frecuencias y porcentajes de cada nivel I–IV.

d) Media y desviación típica, ya que parece simétrica.

Y recuerde que hay que informar, siempre, del número total,  $n$ , de casos.