ANÁLISIS DESCRIPTIVO

ÍNDICE

1.0. Conceptos generales

La observación de la sociedad y la naturaleza, el intento de dar una explicación a los hechos que en ellas se producen y, en consecuencia, la búsqueda de soluciones que den respuesta a las necesidades existentes, es un elemento crucial en el camino para producir conocimiento, y así progresar.

De forma que, cuando se intenta dar una explicación a un determinado fenómeno observado en una población (o en una parte de ella), es necesario disponer de una herramienta rigurosa que permita extraer conclusiones sobre dicha población y eliminar, en gran medida, la posible subjetividad presente en el observador. Esta herramienta es la estadística.

La estadística es la rama de las matemáticas aplicadas que permite interpretar información caracterizada por una condición esencial: la variabilidad de los datos. De este modo facilita el estudio de una característica que puede expresarse numéricamente, bien porque es medible por naturaleza, o porque de alguna manera puede ser formulada numéricamente. Dado que la estadística se basa en la interpretación de la información, debe prestarse mucha atención en garantizar la calidad de los datos recogidos, sea a través de esfuerzos puntuales o de bases poblacionales, tales como, por ejemplo, los registros de mortalidad y de natalidad. Las aportaciones de John Graunt (1620-1674) y Willian Petty (1623-1687) en esta materia, a mediados del siglo XVII, fueron de gran relevancia, ya que establecieron las bases para los sistemas de recolección y organización de la información que se utilizan en la actualidad [1].

Uno de los primeros estudios que empleó métodos estadísticos fue debido a Pierre Charles Alexandre Louis (1787-1872), publicado en 1835. En este trabajo, Pierre Louis empleó lo que él denominó *La Méthode Numérique* (el método numérico) para valorar la eficacia de la sangría como tratamiento de la neumonía; para ello comparó la evolución de los pacientes que habían sido sangrados en los primeros días de la enfermedad (de 1 a 4 días) con los que habían recibido dicho tratamiento ya avanzada la enfermedad (de 5 a 9 días). Como resultado de este estudio observó que la duración de la enfermedad se reducía una media de tres días en el grupo de personas sangradas a principio de la enfermedad en comparación con el otro grupo, pero en el primer grupo había un mayor porcentaje de defunciones con respecto al que había sido sangrado a partir de los 5 días de enfermedad. Pierre Louis concluyó que la sangría era inadecuada para el tratamiento de la neumonía y la recomendaba en situaciones muy específicas [2][3].

Dentro de la estadística existen dos ramas bien diferenciadas: la *estadística descriptiva* y la *estadística inferencial*. La estadística descriptiva es la parte de la disciplina que se encarga de ordenar, resumir y analizar un conjunto de datos mediante una serie de técnicas y métodos, donde los resultados proporcionados no pretenden ir más allá del propio conjunto de datos. Se podría decir que es el recurso que nos permite conocer de manera descriptiva cómo es la realidad bajo investigación y ha sido caracterizada como "el arte de perder información" [4], debido a que una vez aplicada obtenemos básicamente medidas de resumen y a partir de ellas no se podría recuperar la información original. En rigor, cabe subrayarlo, es frecuente que sea exactamente eso lo que se desea: desembarazarnos de datos no esenciales, no característicos, para quedarnos con pocos elementos que permitan hacer una caracterización "a grandes rasgos" de los datos procesados. La estadística inferencial estudia las técnicas mediante las cuales pueden extraerse conclusiones sobre una población a partir de los resultados obtenidos en una muestra. Debe tenerse en cuenta que antes de realizar cualquier estudio más o menos complejo es necesario describir los datos por medio de las técnicas empleadas en el análisis descriptivo, lo que también permite detectar posibles errores, como por ejemplo de grabación.

Un concepto importante en la estadística es el de *variable aleatoria* que, de manera informal, puede definirse como cualquier característica que se pueda medir o clasificar (e.g. el peso de un bebé al nacer, el número de trabajadores de una empresa o el estado civil de una persona).

Según los valores que tomen las variables aleatorias se clasifican en cualitativas o cuantitativas.

Las variables cualitativas, categóricas o atributos son aquellas que no se pueden asociar de forma natural a un número, por lo que no es apropiado realizar operaciones algebraicas con ellas. Sin embargo, no es infrecuente que se asignen códigos numéricos a los valores de una variable cualitativa cuando se vuelcan en una base de datos, aunque la magnitud de tales números no es relevante. Tal maniobra de codificación puede contribuir a minimizar los errores de tecleo, facilitar el manejo de los datos y favorecer en general la manipulación de la información. Las variables categóricas, a su vez, se dividen en *nominales* y *ordinales* en función de la escala de medida. La escala nominal es aquella que permite distinguir categorías, definiendo si una es igual o distinta de otra, pero sin establecer un orden entre ellas; es el caso del grupo sanguíneo de una persona, la raza o su país de residencia. Una variable nominal con sólo dos categorías se llama *dicotómica*, y concierne generalmente a la presencia o no de una determinada característica (e.g., fuma-no fuma). La escala ordinal permite, además de distinguir categorías, establecer un orden entre ellas, aunque sin entrañar diferencias métricas entre las categorías. Ejemplos de este tipo son: la intensidad de dolor (ausente, leve, moderado y fuerte) o los grados militares (soldado, sargento, teniente, etc). Este último ejemplo permite ver con claridad un rasgo caracterizador de las variables ordinales: si todos los sujetos pasaran a estar en la siguiente categoría contigua, las relaciones de subordinación se mantendrían sin cambios.

Las variables *cuantitativas* o *numéricas* son aquellas que adoptan valores numéricos. Según los valores que pueden tomar, se clasifican a su vez en *discretas* y *continuas*. Si un conjunto de observaciones numéricas, cuando se dibujan en una escala numérica, pueden situarse sólo en ciertos puntos aislados y no en los puntos intermedios, entonces se dice que es un conjunto de datos discretos; estas variables suelen tomar valores enteros como consecuencia de la acción de contar; un ejemplo es el número de cigarrillos fumados al día. Si un conjunto de observaciones puede caer, teóricamente, en cualquier lugar de un intervalo de una escala numérica, entonces se dice que es un conjunto de datos continuos, tal y como ocurre con la estatura de una persona. Una manera de distinguir las variables continuas de las restantes consiste en lo siguiente: si x_1 y x_2 son dos valores posibles para la variable, entonces cualquier valor real intermedio que se ubique entre dichos valores, también es posible. Una persona puede tener 4 hijos y otra puede tener 5; pero nadie puede tener 4,3 hijos por ejemplo (se trata de una variable discreta).

Es posible transformar las variables cuantitativas en cualitativas mediante un proceso de categorización; es decir, creando categorías a partir de los valores que toma la variable. Por ejemplo, la edad de una persona en años se podría categorizar en los siguientes grupos de edad: menores de 20 años, de 20 a 39, de 40 a 59 y 60 años o más; de esta forma, en este ejemplo, se pasa de una variable cuantitativa discreta a una cualitativa ordinal.

Las variables medidas en un conjunto de individuos se pueden describir mediante tablas que resumen sus valores, bien empleando técnicas gráficas, bien calculando medidas numéricas de resumen.

Las opciones incluidas en Epidat 4 para realizar un análisis descriptivo de un conjunto de datos son las siguientes:

- Tablas:
  - Tablas de frecuencias
  - Tablas de contingencia
- Medidas numéricas de resumen:
  - Estadísticos descriptivos
  - Coeficiente de correlación
- Gráficos:
  - Gráfico de barras
  - Gráfico de sectores
  - Gráfico de líneas
  - Gráfico de dispersión
  - Histograma
  - Diagrama de cajas
  - Intervalos de confianza

Todas las opciones de este módulo, exceptuando el gráfico de líneas y el de intervalos de confianza, comparten las siguientes características:

- La entrada de datos se realiza única y exclusivamente de forma automática, a través de un asistente para la obtención de datos, que permite abrir un archivo e identificar las variables necesarias para el análisis que se desee realizar.
- Es posible establecer filtros en los datos, definiendo condiciones lógicas a partir de las variables del archivo, de modo que se puede circunscribir el examen a un subconjunto de los datos.
- Los resultados se pueden *segmentar* en función de las categorías de una variable cualitativa (vale decir, obtener indicadores descriptivos para cada una de las subpoblaciones definidas por dichas categorías).

Hay dos opciones para las cuales el funcionamiento difiere del resto (gráfico de líneas y gráfico de intervalos de confianza). En estos casos, Epidat 4 no opera con la información de la base de datos para hacer el gráfico, sino que representa los valores introducidos por el usuario, los cuales deben tener una estructura determinada, como se verá más adelante. Por este motivo, también es posible cargar los datos de forma manual, además de importarlos a partir de un archivo.

Ejemplo

En el año 2005 se implantó en Galicia un *Sistema de Información sobre Conductas de Riesgo* (SICRI) que realiza encuestas telefónicas anuales en la población general adulta mediante un sistema CATI (Computer Asisted Telephone Interview). La encuesta de 2010 estaba dirigida a la población de 16 años y más residente en Galicia, e incluyó n=7.845 personas seleccionadas por muestreo aleatorio estratificado a partir del registro poblacional de Tarxeta Sanitaria. El cuestionario incluyó, además de preguntas sociodemográficas (sexo, edad, estado civil, nivel de estudios, situación laboral), bloques sobre estado de salud, consumo de tabaco y medidas antropométricas, entre otros. Para ilustrar los métodos incluidos en el módulo de análisis descriptivo de Epidat 4 se utilizará una submuestra de 2.000 personas de la encuesta SICRI-2010 y un subconjunto de variables. Los datos se encuentran en el archivo SICRI-2010.xls, que contiene las siguientes variables:

- ID: Nº de identificación.
- SEXO: 1-Hombre, 2-Mujer.
- EDAD: Edad en años en el momento de la encuesta.
- GEDAD: Grupo de edad: 1- 16 a 24, 2- 25 a 44, 3- 45 a 64, 4- 65 años y más.
- ESTUDIOS: Máximo nivel de estudios completados: 1-Sin estudios, 2-Nivel básico, 3-Nivel medio, 4-Nivel superior.
- E_CIVIL: Estado civil: 1-Casado/vive en pareja, 2-Soltero, 3-Separado, 4-Viudo
- ESALUD: Estado de salud autopercibida: 1-Muy bueno, 2-Bueno, 3-Regular, 4-Malo, 5-Muy malo.
- TABACO: Relación con el tabaco: 1-Fumador, 2-Exfumador, 3-Nunca fumador.
- PESO: Peso en Kg.
- TALLA: Talla en cm.
- IMC: Índice de masa corporal en Kg./m^2.
- IMC_CAT: Categorías de IMC: 1-Bajo peso (IMC<18,5), 2-Peso normal (18,5≤IMC<25), 3-Sobrepeso (25≤IMC<30), 4-Obesidad (IMC≥30).

1.1. Tablas de frecuencias

Las tablas de frecuencias resumen los valores que toma una variable en forma de frecuencias, porcentajes y porcentajes acumulados; estos últimos se calculan como la suma acumulada de porcentajes y tienen sentido cuando los valores de la variable tienen una ordenación.

Epidat 4 da la posibilidad de incluir o excluir de la tabla los valores ausentes ("missing") y ordenar la tabla por frecuencias, de forma ascendente o descendente. El hecho de ordenar la tabla puede suponer que, en aquellos casos donde tengan sentido los porcentajes acumulados, la correcta interpretación de tales acumulados resulte imposible. Veámoslo con un ejemplo:

En la encuesta SICRI-2010, destinada a conocer la prevalencia de diferentes factores de riesgo relacionados con la salud, se preguntó a los encuestados por su peso y talla, para así calcular el índice de masa corporal (IMC=peso/talla^2, con la talla medida en metros). La tabla de frecuencias correspondiente a esta variable, categorizada según los criterios de la Sociedad Española para el Estudio de la Obesidad [5], fue:

Valor	Frecuencia	Porcentaje	Porcentaje acumulado
Bajo peso (IMC<18,5)	20	1,13	1,13
Peso normal (18,5<=IMC<25)	810	45,61	46,73
Sobrepeso (25<=IMC<30)	692	38,96	85,70
Obesidad (IMC>=30)	254	14,30	100,00
Total	1.776	100,00	

Con estos datos puede decirse que el 46,73% de los encuestados no tiene problemas de exceso de peso. Cabe señalar que el número de observaciones de la tabla (1.776) es inferior al número de registros del archivo utilizado para obtenerla (2000); la diferencia se debe a los valores ausentes de la variable IMC.

La tabla resultante de ordenar las frecuencias de forma ascendente es:

Valor	Frecuencia	Porcentaje	Porcentaje acumulado
Bajo peso (IMC<18,5)	20	1,13	1,13
Obesidad (IMC>=30)	254	14,30	15,43
Sobrepeso (25<=IMC<30)	692	38,96	54,39
Peso normal (18,5<=IMC<25)	810	45,61	100,00
Total	1.776	100,00	

Al ordenar la tabla, ya no se dispone del porcentaje de encuestados sin problemas de exceso de peso, dado que la ordenación alteró el orden jerárquico en los valores de esta variable.

Las tablas de frecuencias permiten ver cómo se distribuyen los valores de una variable, y también son útiles en un análisis exploratorio para detectar errores o para orientar al investigador a la hora de definir puntos de corte que establezcan categorías.

Para realizar una tabla de frecuencias en Epidat 4 hay que importar los datos individuales a partir de un archivo en formato de Excel (*.xls, *.xlsx) o de OpenOffice (*.ods) por medio del asistente de datos. Deben identificarse una o varias variables para resumir (categóricas o numéricas) y, opcionalmente, una variable categórica para segmentar los resultados. Epidat 4 no presenta tablas con más de 200 filas.

Ejemplo

Para describir el perfil demográfico de los 2.000 encuestados en el SICRI-2010 hay que conocer la distribución por sexo y grupos de edad. En Epidat 4 las dos tablas de frecuencias se pueden hacer simultáneamente identificando SEXO y GEDAD como variables para resumir.

Resultados con Epidat 4:

Entrada automática:

Archivo de trabajo: C:\Archivos de programa\Epidat 4\Ejemplos\Análisis descriptivo\SICRI-2010.xls
Tabla: Datos
Variables:
Resumir: SEXO, GEDAD

Datos:

Valores ausentes:	Excluir
Ordenar la tabla por frecuencias:	No
Filtro:	No

Resultados:

Frecuencias para la variable SEXO:

VALOR	Frecuencia	Porcentaje	Porcentaje acumulado
Hombres	955	47,75	47,75
Mujeres	1.045	52,25	100,00
TOTAL	2.000	100,00	

Frecuencias para la variable GEDAD:

VALOR	Frecuencia	Porcentaje	Porcentaje acumulado
16-24	229	11,45	11,45
25-44	702	35,10	46,55
45-64	580	29,00	75,55
65 y más	489	24,45	100,00
TOTAL	2.000	100,00	

Los datos indican que en la muestra hay aproximadamente la misma proporción de mujeres que de hombres, con una ligera diferencia a favor de las mujeres, y que casi la mitad de los encuestados (47%) tienen menos de 45 años.

Si la variable SEXO se utiliza para segmentar los resultados se obtiene la distribución de la muestra por grupos de edad separadamente para hombres y mujeres.

Resultados con Epidat 4:

Resultados para SEXO=1:

Frecuencias para la variable GEDAD:

VALOR	Frecuencia	Porcentaje	Porcentaje acumulado
16-24	117	12,25	12,25
25-44	351	36,75	49,01
45-64	285	29,84	78,85
65 y más	202	21,15	100,00
TOTAL	955	100,00	

Resultados para SEXO=2:

Frecuencias para la variable GEDAD:

VALOR	Frecuencia	Porcentaje	Porcentaje acumulado
16-24	112	10,72	10,72
25-44	351	33,59	44,31
45-64	295	28,23	72,54
65 y más	287	27,46	100,00
TOTAL	1.045	100,00	

A la vista de estas tablas puede decirse que la proporción de menores de 45 es mayor en hombres (SEXO=1) que en mujeres (SEXO=2).

1.2. Tablas de contingencia

Mediante las tablas de contingencia se clasifica un conjunto de observaciones en función de los valores de dos variables cualitativas que definen, respectivamente, las filas y columnas de la tabla. Una tabla M×N es la que tiene M filas y N columnas, y las celdas pueden representar frecuencias, porcentajes de fila, porcentajes de columna o porcentajes del total de observaciones.

La utilidad de este tipo de tablas es que permiten evaluar la relación entre dos variables y, además, proporcionan la información necesaria para contrastar si hay asociación entre ambas mediante la prueba ji-cuadrado de Pearson.

Para realizar una tabla de contingencia en Epidat 4 hay que importar los datos individuales a partir de un archivo en formato de Excel (*.xls, *.xlsx) o de OpenOffice (*.ods) por medio del asistente de datos. Deben identificarse una variable para las filas y otra para las columnas, ambas categóricas y, opcionalmente, otra variable categórica para segmentar los resultados. Estas tablas en Epidat 4 están limitadas a 200 filas y 10 columnas.

Las tablas de contingencia que calcula Epidat 4 pueden incluir, simultáneamente, frecuencias absolutas, porcentajes de fila, columna o total, según las opciones elegidas por el usuario. Además, el programa da la posibilidad de mostrar o no los valores ausentes de las variables como una fila o columna más de la tabla.

Ejemplo

La distribución de la muestra de 2.000 individuos de la encuesta SICRI-2010 por grupos de edad para cada sexo puede obtenerse haciendo una tabla de contingencia, y se obtienen los mismos resultados que en el ejemplo anterior.

Resultados con Epidat 4:

Entrada automática:

Archivo de trabajo: C:\Archivos de programa\Epidat 4\Ejemplos\Análisis descriptivo\SICRI-2010.xls
Tabla: Datos
Variables:
Definir filas: GEDAD
Definir columnas: SEXO

Datos:

Valores ausentes:	Excluir
Celdas de la tabla:	Porcentajes de columna
Filtro:	No

Resultados:

VALORES	Hombres	Mujeres	TOTAL
16-24	12,25	10,72	11,45
25-44	36,75	33,59	35,10
45-64	29,84	28,23	29,00
65 y más	21,15	27,46	24,45
TOTAL	100,00	100,00	100,00

1.3. Estadísticos descriptivos

Para describir la distribución de valores de una variable cuantitativa se suele recurrir a determinadas medidas numéricas de resumen que permitan resaltar las características más destacables de dicha variable: el número de observaciones, medidas de tendencia central, medidas de dispersión, percentiles y medidas de forma.

En este submódulo de Epidat 4, cuando la variable que se resume tiene valores ausentes, los cálculos prescinden de ellas. Por eso, al resumir simultáneamente varias variables de una misma base de datos, los resultados de cada una pueden basarse en un número diferente de observaciones.

Para calcular estadísticos descriptivos en Epidat 4 hay que importar los datos individuales a partir de un archivo en formato de Excel (*.xls, *.xlsx) o de OpenOffice (*.ods) por medio del asistente de datos. Debe identificarse al menos una variable cuantitativa para resumir y, opcionalmente, una variable categórica para segmentar los resultados.

1.3.1. Medidas de tendencia central

Las medidas de tendencia central indican en torno a qué valor parecen agruparse los datos. Epidat 4 da la posibilidad de calcular la media, la mediana, la moda y la media geométrica de un conjunto de observaciones.

Media

La media, también llamada media aritmética o promedio, es una de las medidas de tendencia central más conocida y utilizada. Su cálculo se realiza sumando todas las observaciones (x_1, x_2, …, x_n) y dividiendo la suma entre el número total de sumandos (n), es decir:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Por tanto, en este sencillo cálculo intervienen todas las observaciones y se obtiene un valor único. Sin embargo, la media debe utilizarse con precaución cuando los datos siguen una distribución muy asimétrica (con valores extremos, muy alejados de la media, colocados a un lado de la distribución), ya que es muy sensible cuando la serie incluye tales valores. Estos valores "tiran" de la media hacia ellos, por lo que su interpretación podría producir una falsa ilusión de que la media refleja un valor "característico" o "típico" de la serie. . En estos casos es recomendable utilizar la mediana como medida de tendencia central o, al menos, presentar ambas medidas.

El siguiente ejemplo ilustra elocuentemente esta situación: en una muestra de 237 pacientes ingresados con síndrome coronario agudo la estancia media en la unidad coronaria fue de 4,4 días. La tabla de frecuencias de la variable, obtenida con Epidat 4, es la siguiente:

Valor	Frecuencia	Porcentaje	Porcentaje acumulado
0	2	0,84	0,84
1	23	9,70	10,55
2	68	28,69	39,24
3	53	22,36	61,60
4	37	15,61	77,22
5	19	8,02	85,23
6	11	4,64	89,87
7	10	4,22	94,09
8	7	2,95	97,05
10	1	0,42	97,47
11	1	0,42	97,89
12	1	0,42	98,31
13	1	0,42	98,73
15	1	0,42	99,16
35	1	0,42	99,58
185	1	0,42	100,00
Total	237	100,00	

Puede observarse que para el 90% de los pacientes la duración de la estancia no superó una semana, en tanto que la estancia de uno de ellos fue muy superior a la del resto (185 días). Si

se recalcula la media eliminando este paciente, el resultado se reduce a 3,6 días, lo que supone una diferencia considerable. El interés de calcular la media sin ese valor extremo reside justamente en que 3,6 representa mejor que 4,4 al valor en torno al cual se ubican los datos.

Una generalización de la media aritmética es la *media ponderada*, que se basa en la idea de que las observaciones no tengan igual peso o importancia, y se calcula de la siguiente manera:

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

donde (w_1, w_2,…,w_n) son los *pesos* correspondientes a las observaciones (x_1, x_2,…,x_n).

Epidat 4 no contempla el cómputo de esta generalización; únicamente realiza el cálculo de la media aritmética es decir, cuando todos los pesos de las observaciones toman el mismo valor.

Mediana

La mediana es el valor de la variable que tiene la propiedad de dividir a la distribución en dos partes iguales, de tal manera que deja por debajo al 50% de las observaciones y por encima al otro 50%, una vez ordenados los datos en función de su magnitud.

A diferencia de lo que ocurre con la media, la mediana no es tan sensible a valores extremos, ya que está basada en la posición que ocupan las observaciones y no en su magnitud. Si el número de observaciones es impar, la mediana es el valor que ocupa la posición central, es decir, el que está en el lugar (n+1)/2 de los datos ordenados de menor a mayor. Con un número par de resultados, la mediana se calcula como la media aritmética de los dos valores situados en el centro, que son los que ocupan las posiciones n/2 y (n/2)+1.

Siguiendo con el ejemplo de la estancia en la unidad coronaria, la duración mediana calculada con los datos de los 237 pacientes es de 3 días, la misma que se obtiene si se elimina el paciente que permaneció 185 días ingresado en esa unidad.

Moda

La moda es el valor que se presenta más frecuentemente en un conjunto de observaciones. Este valor puede no ser único, de forma que cuando sólo existe una moda se dice que la distribución de los datos es unimodal, cuando existen dos modas se dice que es bimodal, y así sucesivamente. Esta característica le resta eficacia como medida de tendencia central por lo que no resulta útil en la práctica.

Un ejemplo en el que no tendrían sentido la media ni la mediana y en el que sería adecuada la moda es el siguiente: un profesor de estadística propone a sus 30 alumnos que resuelvan un ejercicio consistente en calcular la varianza de un conjunto de datos, y anota en la pizarra el resultado obtenido por cada uno de ellos. Muy probablemente, la moda de esos 30 valores coincide con el resultado correcto del ejercicio.

Cuando los datos tienen una distribución aproximadamente simétrica y unimodal, es decir, se distribuyen de forma similar a ambos lados de la media, entonces la media, la mediana y la moda coinciden o tienen valores muy próximos. Cuando los datos no son simétricos,

generalmente la mediana es la medida de tendencia central más adecuada para describirlos, debido a la limitación que presenta la media de verse muy influida por valores extremos.

Media geométrica

La media geométrica es un tipo de media poco usual, pero más adecuada que la media aritmética para describir crecimiento proporcional. Se define como la raíz n-ésima del producto de n observaciones; es decir, es el valor que multiplicado por si mismo tantas veces como datos haya, resulta igual al producto de todos ellos. Formalmente, la fórmula es:

$$x_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

donde $\prod_{i=1}^{n} x_i$ representa el producto de todos los valores de la serie.

Es fácil comprobar que la media geométrica puede calcularse también como la exponencial de la media aritmética del logaritmo neperiano de los valores de la variable:

$$x_g = \exp\left\{\frac{\sum_{i=1}^{n} \mathrm{Ln}(x_i)}{n}\right\}$$

La media geométrica es menos sensible a valores atípicos que la media, puesto que la transformación logarítmica "contrae" los datos; pero también resulta más difícil de interpretar. Un ejemplo que puede ilustrar bien su uso es el siguiente: el número de casos de cierta enfermedad en una población se ha reducido un 87% en los últimos años, pasando de 1.509 en el año 2000 a 203 en 2009. La siguiente tabla recoge los casos anuales y los porcentajes de cada año con respecto al año previo:

Año	Casos	% respecto al año previo
2000	1509	-
2001	1360	90,1
2002	1303	95,8
2003	1255	96,3
2004	1055	84,1
2005	985	93,4
2006	851	86,4
2007	736	86,5
2008	636	86,4
2009	203	31,4

La media geométrica de los nueve porcentajes es 80, y este valor caracteriza el descenso anual del número de casos durante el período, pues calculando sucesivamente el 80% empezando en los casos del año 2000 se obtiene finalmente el valor de 2009:

$$1.509 \times 0{,}8 = 1.207 \text{ (2001)}$$

$$1.225 \times 0{,}8 = 966 \text{ (2002)}$$

$$\ldots$$

$$253 \times 0{,}8 = 203 \text{ (2009)}$$

Por tanto, puede decirse que el porcentaje de descenso anual de casos en el período 2000-2009 es del 20%. Sin embargo, la media aritmética de los porcentajes (83,4) no tiene esta propiedad; si se aplica el 83,4% sucesivamente al número de casos desde el año 2000, se obtiene un valor de 295 casos para el año 2009.

Para calcular la media geométrica es necesario que todos los valores sean mayores que cero, ya que el logaritmo de cero o de un número negativo no existe. Epidat 4 no muestra resultados cuando la variable toma algún valor negativo, pero sí cuando existe algún valor cero y el resto son positivos; en este caso elimina los valores nulos y realiza el cálculo de la media geométrica con el resto de observaciones.

1.3.2. Medidas de dispersión

Grupos diferentes de observaciones pueden tener la misma media, mediana o moda, incluso tratándose de series muy diferentes en cuanto a la dispersión entre las observaciones individuales que las componen; por lo tanto, son necesarias algunas medidas descriptivas de esta variación, que complementen a las medidas de tendencia central. Estas medidas, llamadas *de dispersión*, hacen referencia a cómo quedan agrupados los datos alrededor de una medida de centralización. Epidat 4 da la posibilidad de calcular las siguientes: desviación típica, varianza, coeficiente de variación, mínimo, máximo, recorrido y recorrido intercuartílico.

Varianza

La varianza, denotada por s^2, es una medida de dispersión que cuantifica el grado de variabilidad de los datos en torno a la media. Se calcula como la media aritmética del cuadrado de las distancias entre cada observación y la media de todas ellas y, por tanto, es un valor positivo o nulo; este último caso se da cuando todas las observaciones son iguales entre sí y, en consecuencia, no hay variabilidad. La razón de elevar al cuadrado las distancias de cada observación a la media es convertirlas en positivas, ya que la media aritmética tiene la propiedad de estar en el "medio" de los datos; es decir, unas distancias son positivas y otras negativas, y la suma de todas ellas es igual a cero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

La varianza permite hacerse una idea del grado de dispersión de una variable, de forma que a mayor valor de la varianza, mayor dispersión de los datos. Sin embargo, más allá de esta interpretación general, no es sencillo valorar el significado de su magnitud, ya que está expresada en las unidades empleadas para las observaciones (que son las de la media, claro está) pero al cuadrado. Si, por ejemplo, tuviéramos las estaturas medidas en metros para 100 sujetos, la media vendría dada también en metros, pero la varianza es una magnitud en metros al cuadrado.

Si en la fórmula de la varianza se sustituye el número de observaciones *n* por *n-1*, se obtiene la cuasivarianza. Esta medida tiene mejores propiedades que la varianza como estimador de

la varianza poblacional; por esta razón su uso está muy extendido, sobre todo en el mundo de la bioestadística, hasta el punto de que es habitual ver definida la varianza como la cuasivarianza. Por este motivo, Epidat 4 calcula la varianza usando la fórmula de la cuasivarianza y, en adelante, el término varianza siempre hará referencia a la cuasivarianza.

Desviación típica

La desviación típica o desviación estándar (s) se calcula como la raíz cuadrada positiva de la varianza y, por tanto, está expresada en las mismas unidades que la media. Esta medida da idea de la dispersión de los datos con respecto a su media aritmética; así, al comparar dos grupos de datos, el grupo con menor variabilidad exhibe menor desviación estándar . Debe tenerse en cuenta que la desviación estándar tiene unidades de medida, las mismas que la media, por lo que carece de sentido comparar las desviaciones de dos variables que no estén relacionadas o que estén expresadas en distintas unidades.

La pareja de valores conformada por la media y la desviación típica de un conjunto de datos, permite en muchas ocasiones caracterizar su distribución de valores. Si la distribución es aproximadamente simétrica y unimodal, puede decirse que aproximadamente el 95% de los valores se encuentran en el intervalo $(\bar{x} \pm 2s)$. Para otras situaciones, la desigualdad de Chebychev [6] permite afirmar que para cualquier número k≥1, por lo menos el $[1-(1/k)^2]$ de las observaciones están dentro de k desviaciones estándar de su media; por ejemplo, si k=2, el intervalo $(\bar{x} \pm 2s)$ contiene al menos el 75% de los datos. Esta regla es menos específica que la anterior, pero es independiente de la forma de los datos.

Coeficiente de variación

El coeficiente de variación (CV) es una medida de variabilidad relativa que relaciona la desviación estándar de un conjunto de observaciones con su media, ya que, por ejemplo, una desviación estándar de 10 cm no significa lo mismo en un conjunto de datos con media 10 que si la media es 1.000; en el primer caso, la variabilidad es el 100% de la media mientras que en el segundo es solo el 1%. El CV se calcula como el cociente entre la desviación estándar y la media, que están expresados en las mismas unidades, de modo que el resultado es un coeficiente adimensional. En la práctica es habitual presentarlo multiplicado por 100, aunque Epidat 4 no lo muestra de esa manera.

El CV es una herramienta útil para comparar la dispersión de variables que tienen distintos valores medios, o que emplean distintas unidades, lo que impide una comparación directa de sus desviaciones típicas ya que, normalmente, la variabilidad aumenta con la media. Por ejemplo, el peso medio al nacer de los niños nacidos en Galicia durante el año 2005 fue de 3.219 gr., con una desviación estándar de 533 gr.; en una muestra de niños gallegos de 12 años seleccionados en 2005 para participar en un estudio de salud bucodental, el peso medio fue de 47 Kg. con una desviación estándar de 10,1 Kg. Para comparar la variabilidad del peso en las dos poblaciones es obvio que se podrían pasar todos los valores a las mismas unidades, gr. o Kg., pero las medias son muy distintas, por lo que es más adecuado utilizar el coeficiente de variación, que es del 17% en el caso de los recién nacidos y del 22% para los niños de 12 años.

Para utilizar el CV, se recomienda que la variable tome solo valores positivos.

Recorrido

El recorrido (R) mide la amplitud de las observaciones y se calcula como la diferencia entre los valores máximo y mínimo. El hecho de que este coeficiente utilice sólo dos valores de las observaciones disponibles hace que sea una medida ineficiente, muy sensible a valores extremos. Por este motivo, resulta más conveniente utilizar la varianza y la desviación típica para medir la dispersión.

Recorrido intercuartílico

El recorrido intercuartílico (RI) se calcula como la diferencia entre el tercer y el primer cuartil y se corresponde con el recorrido de los datos que ocupan el 50% central de las observaciones.

1.3.3. Cuantiles

Los cuantiles son valores que dividen un conjunto de datos en grupos de igual tamaño. Para obtener N grupos es necesario definir N-1 cuantiles, que reciben distintos nombres en función del valor de N: percentiles (N=100), deciles (N=10), quintiles (N=5), cuartiles (N=4) o mediana (N=2) [7].

Los percentiles son útiles en el análisis exploratorio de datos y en el análisis descriptivo porque permiten valorar la dispersión, la simetría y la distribución de los datos, sobre todo de forma visual mediante los diagramas de caja que se describirán más adelante. También suelen utilizarse para categorizar variables continuas como, por ejemplo, el nivel de colesterol, de forma que se clasifica a los individuos en grupos de igual tamaño. Esto facilita la presentación de los datos en forma de tablas o gráficos, aunque supone una pérdida de información, que será mayor cuanto más grandes sean los grupos.

Una aplicación muy extendida de los percentiles se realiza en pediatría, para valorar el crecimiento de los niños. Las curvas de crecimiento desarrolladas a partir de estudios longitudinales, como por ejemplo las de la Fundación Orbegozo [8], proporcionan una estimación de los percentiles de peso y talla para cada edad y sexo, y esos valores se usan en las revisiones infantiles como referencia de un adecuado crecimiento.

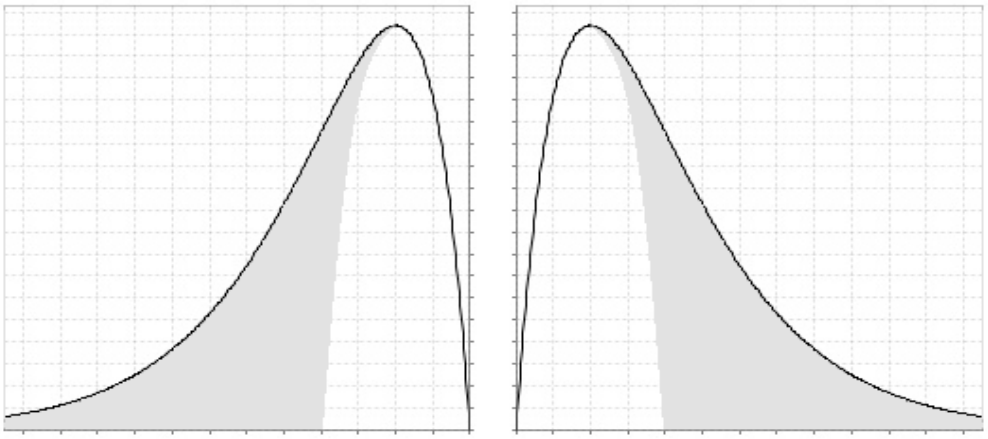
1.3.4. Medidas de forma

Las distribuciones pueden diferir entre ellas en términos de su valor central y en cómo se agrupan los valores individuales alrededor de esa medida; pero también existen distribuciones de frecuencias con la misma media y desviación típica que difieren en su forma. Para caracterizar el perfil de una distribución de valores existen dos coeficientes, llamados genéricamente *medidas de forma*, útiles para describir la forma de una distribución: los coeficientes de asimetría y de curtosis, propuestos por Ronald Fisher [9].

Asimetría

El *coeficiente de asimetría* cuantifica en qué medida las observaciones de un conjunto de datos se distribuyen simétricamente alrededor de la media. Su interpretación, que solo tiene sentido cuando la distribución es unimodal, es la siguiente: si la variable es simétrica entonces el coeficiente de asimetría toma el valor cero; cuando la distribución de valores presenta una cola hacia la izquierda, el coeficiente toma un valor negativo (asimetría

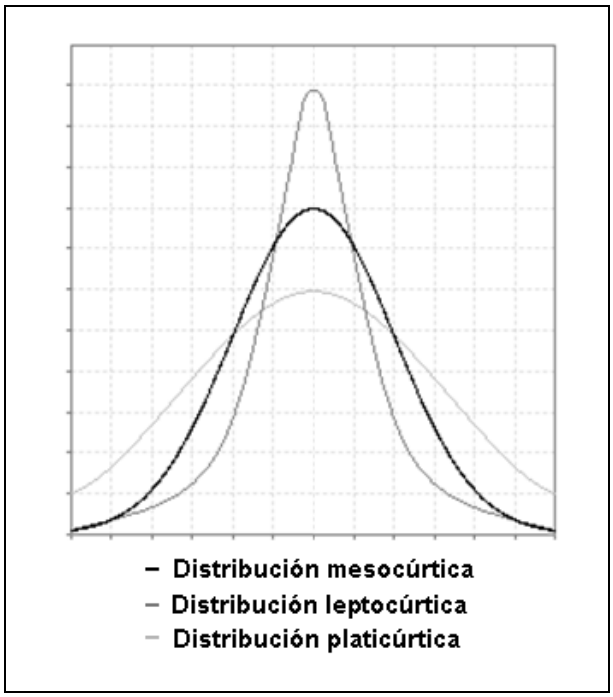
negativa) y si la cola es hacia la derecha el valor del coeficiente es positivo (asimetría positiva).



Asimetría negativa **Asimetría positiva**

Curtosis

El *coeficiente de curtosis* (coeficiente de apuntamiento o exceso de curtosis) mide el grado de apuntamiento de una distribución con respecto a la distribución normal con la misma media y varianza. La interpretación de este coeficiente tiene sentido siempre que la distribución sea unimodal y esencialmente simétrica, de forma que, si la distribución presenta el mismo perfil que la normal con la misma media y varianza, entonces el coeficiente de curtosis toma el valor cero (distribución mesocúrtica); cuando la distribución es más apuntada que la normal correspondiente, el valor del coeficiente es positivo (distribución leptocúrtica) y, por último, si la distribución es más "aplastada" se tiene un valor del coeficiente negativo (distribución platicúrtica).



Ejemplo

Para caracterizar la distribución del índice de masa corporal en la muestra de adultos jóvenes del SICRI-2010, se calculan estadísticos descriptivos de esta variable para hombres y mujeres por separado en el grupo de 25 a 44 años. En Epidat, trabajando con el archivo SICRI-2010.xls, hay que seleccionar la variable IMC para resumir, segmentar por SEXO y definir un filtro con la condición GEDAD=2.

Resultados con Epidat 4:

Entrada automática:

Archivo de trabajo: C:\Archivos de programa\Epidat 4\Ejemplos\Análisis descriptivo\SICRI-2010.xls
Tabla: Datos
Variables:
Resumir: IMC
Segmentar resultados : SEXO

Datos:

Filtro: GEDAD = 2

Resultados para SEXO=1:

	IMC
n	343
Media	26,048
Mediana	25,469
Desviación típica	3,532
Varianza	12,478
Mínimo	18,588
Máximo	39,792
Cuartiles	
P25	23,766
P50	25,469
P75	27,682

Resultados para SEXO=2:

	IMC
n	336
Media	23,934
Mediana	23,029
Desviación típica	4,138
Varianza	17,122
Mínimo	16,33
Máximo	46,312
Cuartiles	
P25	21,297
P50	23,029
P75	25,766

Los resultados obtenidos indican que el promedio de IMC en los hombres de la muestra supera ligeramente al punto de corte que indica sobrepeso (25 Kg./m^2), mientras que en las mujeres, la media está por debajo de dicho valor. En ambos casos, la media y la mediana tienen valores próximos, lo que es indicativo de que la variable tiene una distribución simétrica. En cuanto a la variabilidad, ocurre lo contrario que con la media, es mayor en las mujeres, como indican los valores de varianza y desviación típica, así como el rango de variación de la variable; el IMC máximo en esta muestra de mujeres es de 46,3, valor que está en el rango de obesidad de tipo III (mórbida); en los hombres el máximo es próximo a 40, y se clasifica como obesidad de tipo II [5]. Los cuartiles se pueden interpretar del modo siguiente: el 50% de los hombres tienen un IMC igual o superior a 25,5, es decir, más de la mitad tienen sobrepeso u obesidad (IMC ≥ 25 Kg./m^2); en las mujeres, en cambio, el percentil 75 es 25,8, por lo que algo más de la cuarta parte tienen sobrepeso u obesidad. No se calcularon la asimetría y la curtosis porque no son necesarios para este análisis.

1.4. Correlación

En términos generales, se dice que dos características o variables están correlacionadas si al cambiar una de ellas tiende a cambiar la otra, en el mismo sentido o en sentido opuesto; por ejemplo, en general el peso aumenta con la talla, por lo que hay una correlación positiva entre estas dos variables. El concepto estadístico de correlación fue introducido en 1888 por Sir Francis Galton y de sus trabajos, y las contribuciones de Edgeworth y Pearson, surgió el llamado coeficiente de correlación de Pearson, que cuantifica el grado de relación lineal entre dos variables cuantitativas así como la dirección, positiva o negativa, de dicha relación [10].

El coeficiente de correlación de Pearson es adimensional, es decir, no depende de las unidades de medida de las variables, y toma valores entre -1 y 1, donde el signo indica el sentido de la relación. Por otra parte, a medida que aumenta el valor absoluto del coeficiente aumenta el grado de relación lineal entre las variables. Un valor de -1 o de 1 indica una relación lineal perfecta entre las dos variables, en el primer caso negativa y en el segundo positiva; de modo que al representar los datos en un diagrama de dispersión, se disponen formando una línea recta decreciente o creciente, respectivamente. Si el coeficiente de correlación fuera 0, entonces las variables no estarían relacionadas linealmente, aunque no se puede descartar que exista otro tipo de relación entre ellas distinta de la lineal; sin embargo, independencia implica incorrelación, es decir, cuando dos variables son independientes, el coeficiente de correlación toma el valor cero.

Ejemplo

El índice de masa corporal es un indicador que se utiliza para caracterizar la obesidad, y que se calcula como el cociente entre el peso (en Kg.) y la talla (en metros) al cuadrado. Los datos de la encuesta SICRI-2010 indican que el IMC está muy correlacionado con el peso, pero no con la talla. Para calcular estos coeficientes de correlación con Epidat 4 hay que seleccionar las variables PESO, TALLA e IMC para resumir y, dado que la antropometría de hombres y mujeres es diferente, podemos segmentar los resultados por SEXO.

Resultados con Epidat 4:

Entrada automática:

Archivo de trabajo: C:\Archivos de programa\Epidat 4\Ejemplos\Análisis descriptivo\SICRI-2010.xls
Tabla: Datos
Variables:
Analizar: IMC, PESO, TALLA
Segmentar resultados : SEXO

Resultados para SEXO=1:

Coeficiente de correlación de Pearson:

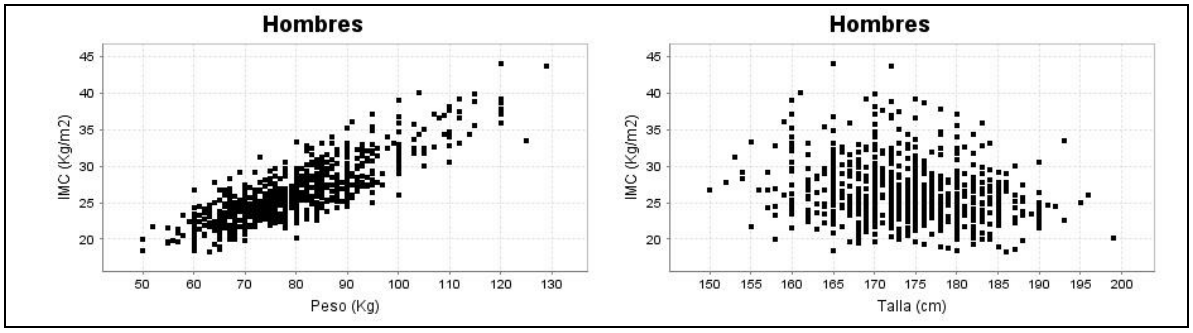
	IMC	PESO	TALLA
IMC	1		
PESO	0,832	1	
TALLA	-0,226	0,347	1

Resultados para SEXO=2:

Coeficiente de correlación de Pearson:

	IMC	PESO	TALLA
IMC	1		
PESO	0,872	1	
TALLA	-0,255	0,244	1

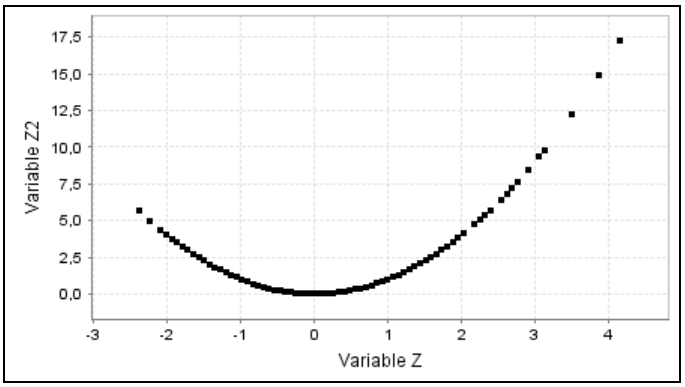
Los diagramas de dispersión entre PESO-IMC y TALLA-IMC en hombres son coherentes con los coeficientes de correlación obtenidos:



A la hora de interpretar el coeficiente de correlación deben tenerse en cuenta las siguientes recomendaciones:

- *Correlación* significa relación lineal [10]. Dos variables pueden estar fuertemente relacionadas de forma no lineal y tener un coeficiente de correlación bajo. Por esta razón, es recomendable representar gráficamente los datos mediante un diagrama de dispersión antes de calcular el coeficiente de correlación. Por ejemplo, llamemos Z a la variable peso estandarizada (se obtiene al restarle su media y dividir por su desviación estándar) y Z2 a dicha variable al cuadrado. Con los datos del SICRI-2010

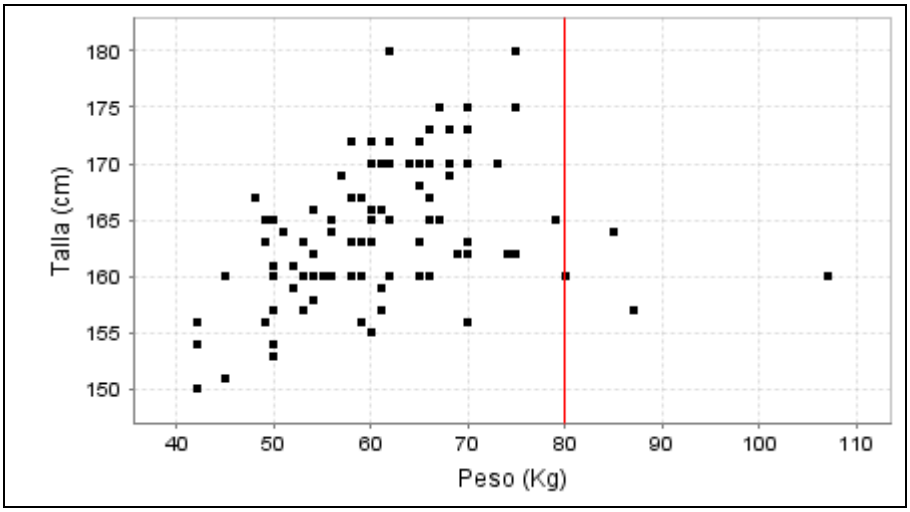
se obtiene que la correlación entre Z y Z2 es baja (r=0,346) y, sin embargo, hay una clara relación entre las dos variables, como se aprecia en el diagrama de dispersión:



- *Correlación* no implica causalidad [11][12]. Puede ocurrir que dos variables estén muy correlacionadas (muchas veces debido a que las dos están causalmente relacionadas con una tercera variable), pero que no haya relación causal entre ellas. También puede darse la situación de que dos variables sin ninguna relación entre ellas, como por ejemplo la tasa de mortalidad infantil y la prevalencia de caries en escolares, calculadas para cada año de un determinado período, presenten una tendencia decreciente durante ese lapso por lo que, probablemente, estarán positivamente correlacionadas.

- El coeficiente de correlación depende del rango de variación de las variables implicadas [10]. Por ejemplo, la edad y la estatura están muy correlacionadas en los niños, y de forma positiva, mientras que en los adultos la correlación es baja y negativa (-0,254 en mayores de 25 años, según los datos del SICRI-2010).

- El coeficiente de correlación se ve muy influido por la presencia de valores extremos [12]. Por ejemplo, la correlación entre el peso y la talla en mujeres de 16 a 24 años del SICRI-2010 vale 0,38. Si se eliminan los valores de peso superiores a 80 Kg., que para este rango de edad pueden considerarse muy elevados (véase figura debajo), la correlación aumenta a 0,59. Cuando ocurre una situación como ésta, puede ser adecuado aplicar a las observaciones una transformación, como la logarítmica, que cambie la escala y minimice el efecto de los valores atípicos.



- El tamaño de la muestra debe tenerse en cuenta a la hora de interpretar el coeficiente de correlación, que calculado con pocas observaciones está afectado por una elevada variabilidad [10].
- El coeficiente de correlación no debe utilizarse para valorar el grado de acuerdo entre dos mediciones realizadas de forma repetida a los mismos individuos. En ese caso, es más adecuado el coeficiente de correlación intraclase [13] o el método gráfico de Bland y Altman [14]), ambos incluidos en el módulo de Concordancia y consistencia de Epidat 4, y que están descritos con detalle en la ayuda de dicho módulo.

En un contexto descriptivo, como el que nos ocupa en este módulo, el coeficiente de correlación lineal de Pearson puede calcularse con cualquier par de variables. Sin embargo, para hacer inferencia sobre este coeficiente, mediante un intervalo de confianza o una prueba de significación, es necesario –al menos teóricamente- que los datos sigan una distribución normal. Si los datos son ordinales o su distribución se aleja de la normal, se recomienda utilizar el *coeficiente de correlación de Spearman*, que se obtiene aplicando la fórmula del coeficiente de Pearson a los rangos de las observaciones [12], y que tiene la misma interpretación que este último. Por tanto, el coeficiente de Spearman no tiene en cuenta el valor de las observaciones, sino el orden que ocupan, lo que hace que también sea más adecuado que el coeficiente de correlación de Pearson cuando las series contienen valores extremos.

Por otra parte, puede ocurrir que dos variables tengan una correlación baja a pesar de estar asociadas mediante una relación no lineal. Si la relación es monótona, entonces, los rangos de las observaciones ordenadas pueden tener una relación lineal [10], en cuyo caso será más adecuado calcular el coeficiente de correlación de Spearman.

Para calcular coeficientes de correlación, de Pearson o de Spearman, en Epidat 4 hay que importar los datos individuales a partir de un archivo en formato de Excel (*.xls, *.xlsx) o de OpenOffice (*.ods) por medio del asistente de datos. Deben identificarse al menos dos variables cuantitativas para resumir y, opcionalmente, una variable categórica para segmentar los resultados. El programa presenta los coeficientes calculados para cada par de variables en forma de matriz con unos en la diagonal y, por ser una matriz simétrica, solo se muestran los valores por debajo de la diagonal. Esta matriz se acompaña, de forma opcional, de otra similar en la que se muestra el tamaño efectivo de muestra utilizado para calcular cada coeficiente de correlación, pues las observaciones ausentes no se tienen en cuenta, y pueden variar de unas variables a otras. El número de variables a resumir en este módulo de Epidat 4 está limitado a 20.

1.5. Gráficos

Las representaciones gráficas proporcionan, respecto a las tablas, otra manera de describir un conjunto de datos, de forma que, quizás de un simple vistazo es posible captar sus características más destacables.

A la hora de elaborar un gráfico, el primer paso es decidir qué información desea presentarse, y si el gráfico es la mejor herramienta para ello. Se recomienda utilizar gráficos solo para mostrar información que no pueda ser resumida fácilmente de otro modo, ni con texto ni mediante una tabla. A continuación, habrá que identificar las principales características que condicionarán la construcción del gráfico (por ejemplo, el tipo de variables: cualitativas o cuantitativas) y elegir el formato adecuado. El resultado debe ser un

gráfico autoexplicativo, que contenga toda la información suficiente para poder interpretarlo [15].

Tal como sugiere Molinero [16], un gráfico debe comunicar ideas complejas con precisión, claridad y eficiencia, de tal manera que:

- Induzca a pensar en el contenido más que en la apariencia.
- No distorsione la información proporcionada por los datos.
- Favorezca la comparación entre grupos, si éste es su objetivo.

La calidad de un gráfico radica en su capacidad de presentar datos complejos con sencillez. Sin embargo, la disponibilidad de herramientas informáticas para la elaboración de gráficos favorece la proliferación de representaciones con sofisticados efectos (como los diseños tridimensionales) que no solo no aportan valor al gráfico, sino que resultan confusos y a menudo engañosos. Otras recomendaciones a la hora de elaborar un gráfico son: evitar la duplicidad de información (por ejemplo, no presentar los mismos datos en tabla y en gráfico) y que no haya discrepancias con el texto del trabajo. En el artículo de González-Alastrué [15] se ilustran estas ideas con un detallado ejemplo.

Para profundizar más en el tema se recomienda el libro "The visual display of quantitative information" publicado por primera vez en 1983 [17] y que tiene una segunda edición de 2009 [18].

Epidat 4 incluye, en su módulo de Análisis descriptivo, varias opciones para realizar gráficos, que responden a las necesidades más frecuentes en el análisis exploratorio de datos:

1.5.1. Gráfico de barras

1.5.2. Gráfico de sectores

1.5.3. Gráfico de líneas

1.5.4. Gráfico de dispersión

1.5.5. Histograma

1.5.6. Diagrama de cajas

1.5.7. Gráfico de intervalos de confianza

Un elemento destacable de este submódulo de Epidat es el editor de gráficos, que permite personalizar en gran medida los gráficos realizados con el programa, así como guardarlos con formato imagen (*.jpg o *.png). El editor de gráficos tiene una serie de elementos comunes a todos los gráficos de Epidat como son, por ejemplo, las opciones generales (título, formato de texto, color, borde y tamaño). Además, hay otras opciones que permiten modificar características de los ejes o de los elementos que se representan, y que dependen del tipo de gráfico elegido. No se describirán con detalle las propiedades del editor, porque su manejo es sencillo e intuitivo, y algunas se comentarán en cada gráfico particular. Sin embargo, tres puntos merecen ser destacados:

- Cuando se realizan simultáneamente varios gráficos (por ejemplo, al segmentar por una variable cualitativa), es posible modificar todos los gráficos a la vez activando la opción "*Aplicar a todos los gráficos*" de la pestaña "*Opciones generales*". Todos los cambios que se realicen mientras esté marcada esta opción se aplicarán a todos los gráficos.

- Los gráficos generados pueden guardarse en formato imagen (*.jpg o *.png) desde el propio editor.
- Una vez que el gráfico se presenta en la ventana de resultados, es posible volver a abrirlo con el editor haciendo doble click en él o a través de la opción "*Editar gráfico*" (botón derecho del ratón).

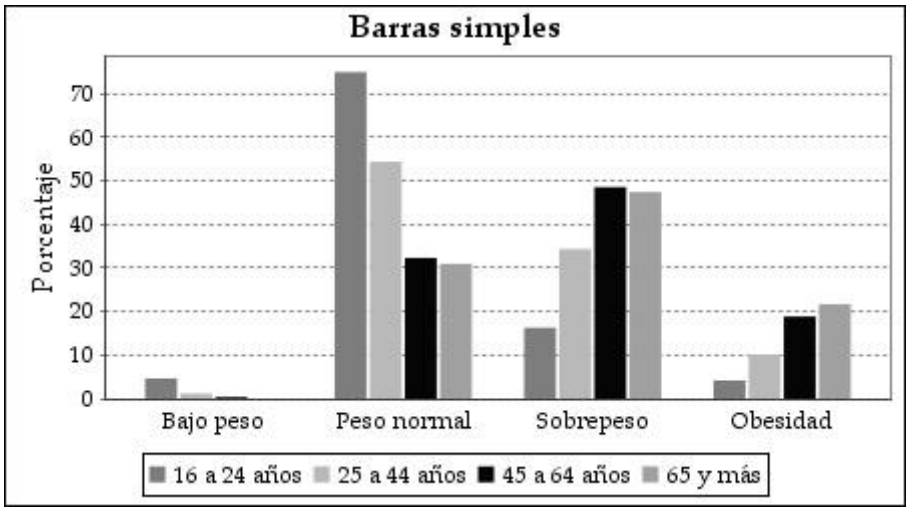
1.5.1. Gráfico de barras

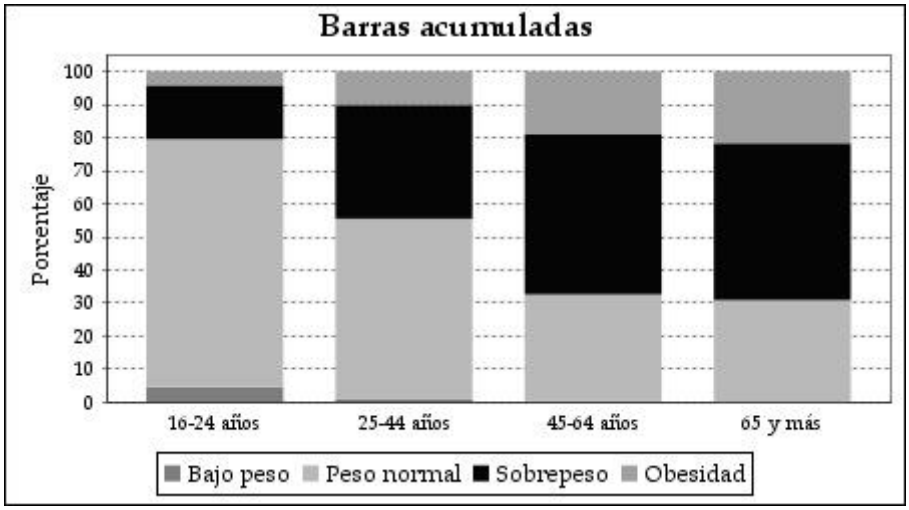
El *gráfico de barras*, también llamado diagrama de barras, es la representación más habitual para describir la distribución de frecuencias de una variable cualitativa. Este recurso representa en el eje de abscisas (eje X) cada una de las categorías de la variable y en el eje de ordenadas (eje Y) las frecuencias o porcentajes de cada categoría, en forma de rectángulos con la misma base. También puede utilizarse para describir variables discretas que tomen pocos valores como, por ejemplo, el número de hijos.

Para realizar un gráfico de barras en Epidat 4 hay que importar los datos individuales a partir de un archivo en formato de Excel (*.xls, *.xlsx) o de OpenOffice (*.ods) por medio del asistente de datos. Debe identificarse al menos una variable cualitativa para resumir y, opcionalmente, una variable categórica para segmentar los resultados y/o una variable categórica para definir grupos en el gráfico. El programa da la posibilidad de elegir la orientación de las barras (horizontal o vertical) y seleccionar el tipo de barras (simples o acumuladas).

Ejemplo

Para representar cómo se distribuyen las categorías del IMC en la muestra del SICRI-2010 por grupos de edad, utilizando el archivo SICRI-2010.xls incluido en Epidat 4, se puede hacer un gráfico de barras simples con las variables IMC_CAT para resumir y GEDAD para definir grupos; otra posibilidad es hacer un gráfico de barras acumuladas con las mismas variables. El resultado es el siguiente:





A la vista de los resultados, podemos decir que el gráfico de barras acumuladas representa mejor la distribución de la variable de interés en cada grupo de edad. Puede observarse, por ejemplo, cómo disminuye claramente el porcentaje de sujetos con peso normal a medida que aumenta la edad, así como que esta reducción ocurre a costa de un aumento en el sobrepeso y la obesidad; este porcentaje (sobrepeso y obesidad conjuntamente) pasa del 20% en el grupo más joven al 70% en los mayores de 65 años.

El gráfico de barras simples sería más claro si solo se comparasen dos grupos; por ejemplo, si se hicieran los mismos gráficos sustituyendo el grupo de edad por el sexo.

1.5.2. Gráfico de sectores

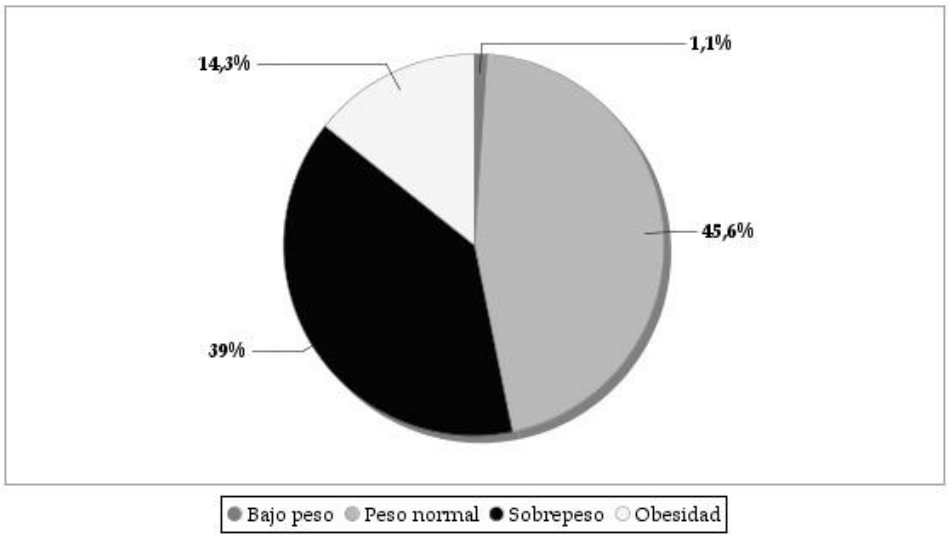
El gráfico de sectores, también llamado diagrama de sectores, gráfico "de pastel", o gráfico circular, representa la frecuencia de cada una de las categorías de una variable cualitativa a través de sectores de un círculo cuyas áreas son proporcionales a las frecuencias. También puede utilizarse para describir variables discretas que tomen pocos valores como, por ejemplo, el número de hijos.

Este gráfico se recomienda cuando el número de categorías de la variable es reducido (quizás no más de 6), pues en caso contrario puede resultar muy difícil distinguir las secciones más pequeñas.

Para realizar un gráfico de sectores en Epidat 4 hay que importar los datos individuales a partir de un archivo en formato de Excel (*.xls, *.xlsx) o de OpenOffice (*.ods) por medio del asistente de datos. Debe identificarse una variable cualitativa, o discreta con pocos valores, para resumir y, opcionalmente, una variable categórica para segmentar los resultados. El programa da la posibilidad de mostrar en el gráfico las frecuencias o los porcentajes.

Ejemplo

Siguiendo con el ejemplo del diagrama de barras, se puede representar la distribución de las categorías del índice de masa corporal mediante un gráfico de sectores. El resultado obtenido con Epidat 4 es el siguiente:



Nótese que, ocasionalmente, la magnitud relativa de los sectores puede resultar difícil de captar visualmente. En este caso, por ejemplo, resulta difícil (si no se repara en los porcentajes) identificar cuál de las categorías entre peso normal y sobrepeso es más frecuente; esto ocurre con cierta frecuencia con este tipo de gráfico, por lo que es recomendable solicitar que figuren siempre los valores de las frecuencias o los porcentajes. Una buena alternativa es utilizar el diagrama de barras, que no da lugar a equívocos.

Por otra parte, la comparación por grupos de edad solo podría hacerse con gráficos separados, eligiendo la opción de segmentar por GEDAD; sin embargo, esta alternativa no sería la más apropiada para ese propósito, pues supone comparar 4 gráficos distintos que, de por sí, no son tan claros como el diagrama de barras.

1.5.3. Gráfico de líneas

El gráfico de líneas permite representar, mediante puntos unidos por un segmento, un conjunto de valores (eje Y) para cada una de las categorías de una variable cualitativa (eje X), generalmente períodos temporales con el objetivo de analizar tendencias (años, trimestres, …). Los valores que se representan pueden ser observaciones de una variable (número de defunciones diarias por gripe A) o estadísticos de resumen (prevalencia anual de fumadores en un período, tasa de mortalidad infantil o incidencia de tumores).

Como ya se comentó en la introducción, los gráficos de líneas y de intervalos de confianza de Epidat 4 no comparten el funcionamiento del resto de gráficos. Las diferencias se derivan de que, bajo estas opciones, el programa no resume la información de la base de datos para hacer el gráfico, sino que representa directamente los valores introducidos por el usuario. Por este motivo, también es posible cargar los datos de forma manual, además de importarlos a partir de un archivo en formato de Excel (*.xls, *.xlsx) o de OpenOffice (*.ods) mediante el asistente de datos. El hecho de que los datos de entrada estén ya resumidos hace que no sea posible definir filtros ni segmentar los resultados.

Epidat 4 permite representar más de una línea en el mismo gráfico, siempre que todas estén definidas en las mismas categorías. Un ejemplo podría ser la evolución anual de la

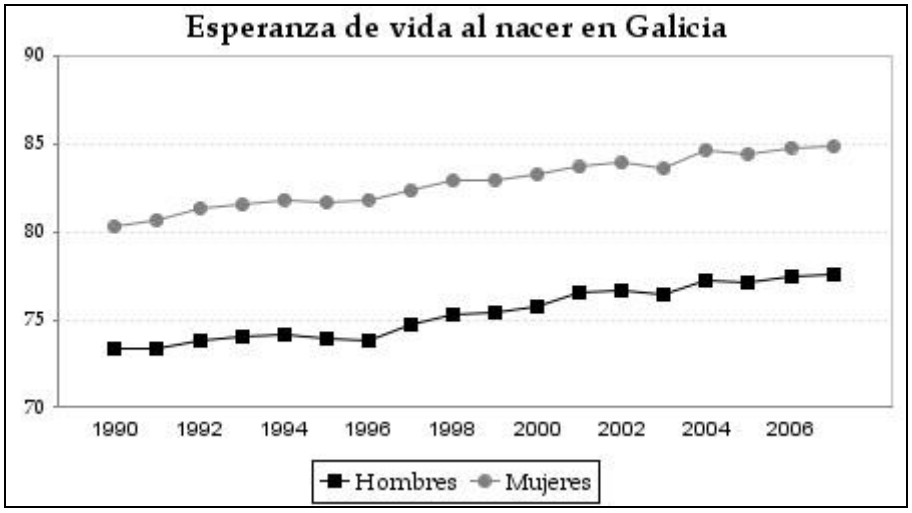
prevalencia de obesidad en hombres y en mujeres; en el gráfico se representarían dos líneas, una para cada sexo, y ambas series de prevalencias tendrían que estar definidas para los mismos años (categorías).

Para introducir los datos manualmente, es necesario especificar el número de líneas que se van a representar y el número de categorías del eje X, y completar la tabla resultante. El número de líneas está limitado a 50 y el número máximo de categorías es 1.000.

Al optar por la entrada automática, se abre el asistente para la obtención de datos que permite, a través del botón "examinar", seleccionar el directorio y el archivo (OpenOffice o Excel) que contiene la tabla de valores. Es necesario recordar que Epidat 4 requiere que las tablas que han de importarse tengan una estructura fija. En este caso, la tabla debe contener tantas filas como número de categorías y tantas variables como líneas a representar en el gráfico.

Ejemplo

En Galicia, la esperanza de vida al nacer (EV) muestra una tendencia creciente en los últimos años, al igual que ocurre en el conjunto de España. El archivo EV-GALICIA.xls, incluido en Epidat 4, contiene la EV anual de Galicia en el período 1990-2007 para hombres y para mujeres. Para representar estos datos en un gráfico de líneas utilizando Epidat 4 hay que cargar los datos de forma automática utilizando el asistente, e identificar EV-HOMBRES y EV-MUJERES como variables para las líneas y AÑO como categorías del eje X. En el gráfico resultante no se visualizan las etiquetas correspondientes a los años, ya que el número de valores a mostrar es grande (17) y no caben todos los textos. Para verlos correctamente hay dos posibilidades: aumentar el tamaño del gráfico o indicar, en las opciones para el eje X, que se muestren cada 2 etiquetas. El resultado es el siguiente:



1.5.4. Gráfico de dispersión

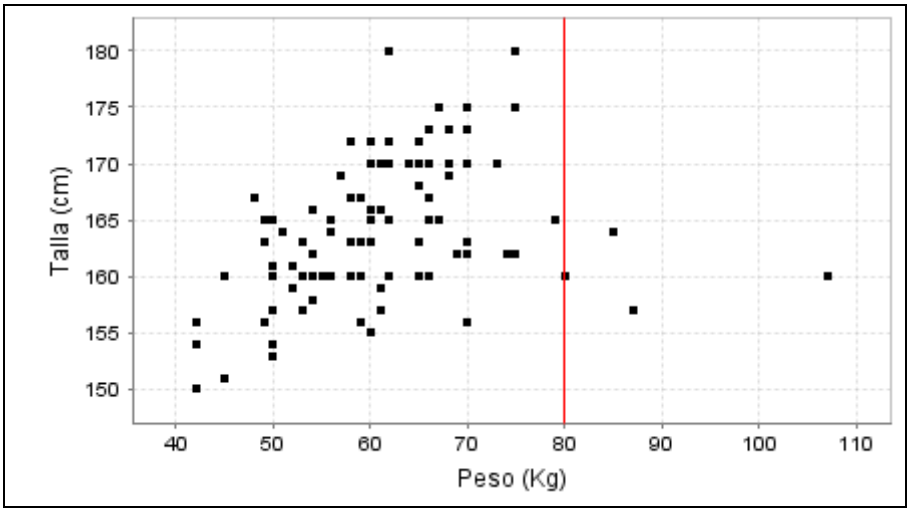
El gráfico o diagrama de dispersión se utiliza para describir visualmente la relación existente entre dos variables cuantitativas, como primer paso aconsejable antes de realizar otros análisis como calcular el coeficiente de correlación o ajustar un modelo de regresión lineal. Cada punto del diagrama representa un par de valores conformado por una observación de la primera variable (eje X) y una observación de la segunda (eje Y).

Para realizar un gráfico de dispersión en Epidat 4 hay que importar los datos individuales a partir de un archivo en formato de Excel (*.xls, *.xlsx) o de OpenOffice (*.ods) por medio del

asistente de datos. Deben identificarse dos variables cuantitativas, una para el eje X y otra para el eje Y y, opcionalmente, una variable categórica para segmentar los resultados.

Ejemplo

Para realizar el gráfico de dispersión mostrado en el apartado 1.4 (coeficiente de correlación) que representa la relación entre el peso y la talla en mujeres de 16 a 24 años del SICRI-2010, hay que cargar los datos de forma automática utilizando el asistente, e identificar la variable PESO para el eje X y la variable TALLA para el eje Y; también hay que definir un filtro con la condición "SEXO=1 y GEDAD=1", y añadir una línea vertical correspondiente a un peso de 80 Kg. usando la opción disponible en la pestaña "Eje X" del editor de gráficos. Se reproduce aquí el resultado:



1.5.5. Histograma

El *histograma* es uno de los gráficos más comunes para describir la distribución de frecuencias de una variable cuantitativa. En el eje horizontal, el histograma representa los intervalos en los que se dividen los valores de la variable; y en el eje vertical las frecuencias, porcentajes o densidades de cada uno de los intervalos, en forma de rectángulos o barras adyacentes [19].

La densidad de un rectángulo es el cociente entre la frecuencia relativa del intervalo correspondiente y su amplitud; de este modo, el área del rectángulo (base=amplitud del intervalo × altura=densidad) coincide con su frecuencia relativa, y el área total del histograma es 100%.

A la hora de interpretar un histograma, Oliveras [20] recomienda identificar primero el patrón general que lo caracteriza y, a continuación, las desviaciones que puede haber respecto a ese patrón. Este autor ilustra con algunos ejemplos la interpretación de distintos histogramas. A modo de resumen, pueden darse las siguientes situaciones:

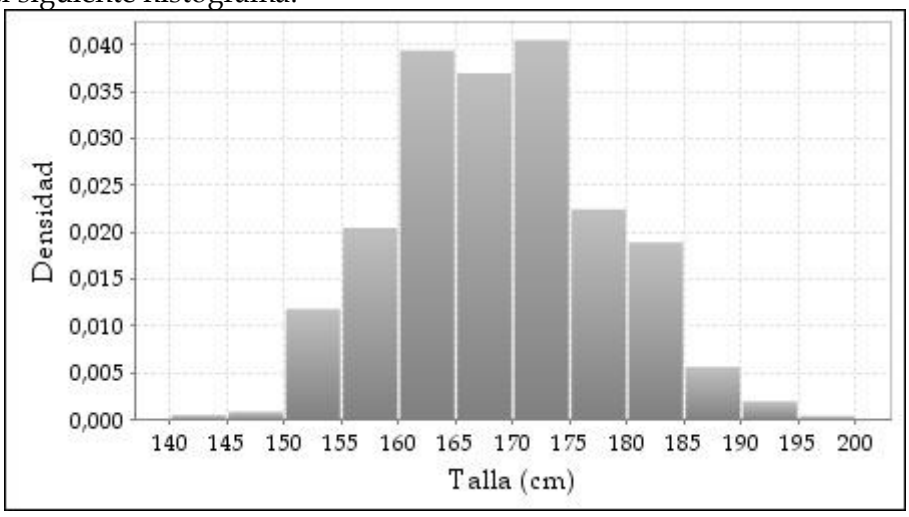
- Presencia de datos anómalos: el histograma permite identificar fácilmente los valores anormalmente altos o bajos en relación al resto de las observaciones.
- Simetría: la forma del histograma indica visualmente cómo se distribuyen los valores de la variable a ambos lados de la media, de forma que permite identificar si la distribución es más o menos simétrica o tiene cierto grado de asimetría a la derecha o la izquierda.

- Histograma con varios picos: cuando la distribución tiene más de una moda, el histograma presenta varios picos. Normalmente esto se debe a la superposición de varias poblaciones que tienen medias distintas y deberían analizarse por separado.
- Histograma dentado: está asociado normalmente al sistema de medición de datos, por ejemplo, la tendencia al redondeo con datos autodeclarados.

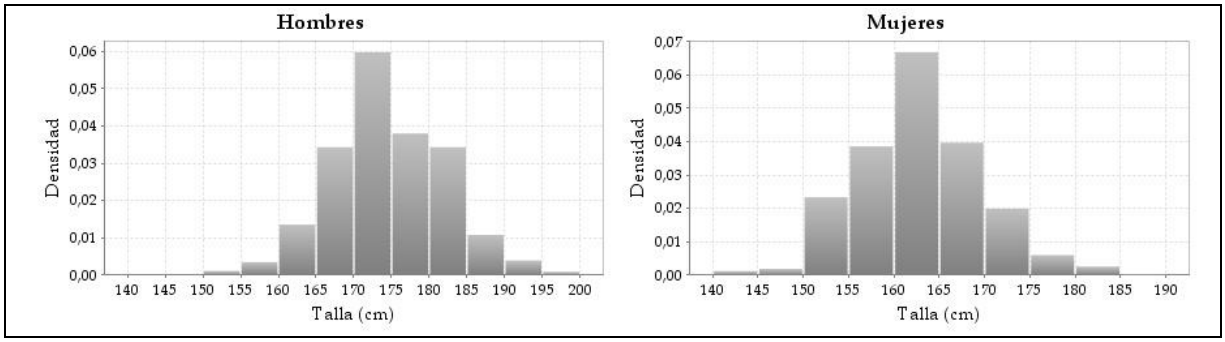
Para realizar un histograma en Epidat 4 hay que importar los datos individuales a partir de un archivo en formato de Excel (*.xls, *.xlsx) o de OpenOffice (*.ods) por medio del asistente de datos. Debe identificarse una variable cuantitativa y, opcionalmente, una variable categórica para segmentar los resultados. Cuando las barras representan densidades, el programa ofrece la posibilidad de mostrar la curva normal con la misma media y desviación estándar de los datos. Además, el usuario puede personalizar los intervalos o dejar que el programa los calcule de forma automática. Esta última opción es recomendable como un primer paso cuando se desconoce cómo se distribuyen los datos y, posteriormente, pueden modificarse los intervalos si es necesario.

Ejemplo

La distribución de la talla de los 2.000 encuestados en el SICRI-2010 (archivo SICRI-2010.xls) se muestra en el siguiente histograma:

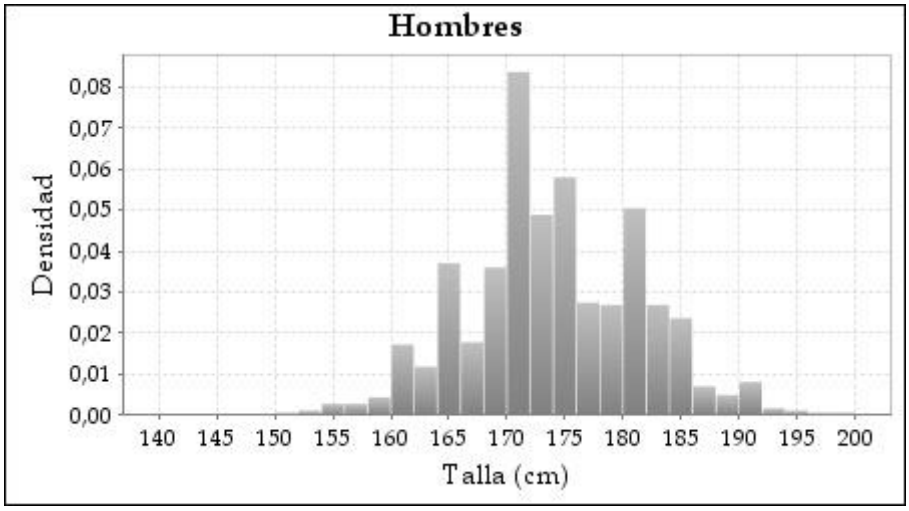


En la figura se aprecian dos *picos*, que corresponden a los intervalos 160-165 y 170-175, y que están identificando los valores más frecuentes en mujeres y hombres, respectivamente. Si se representa este histograma segmentando por SEXO se obtiene el siguiente resultado:



Ahora las dos distribuciones obtenidas son unimodales. Además, puede observarse también una ligera asimetría hacia la derecha en la talla de los hombres, mientras que la talla de las mujeres tiene una distribución más simétrica.

Por otra parte, estos histogramas están construidos con intervalos de amplitud 5 cm. Si se repite con intervalos de amplitud 2 cm, el resultado en hombres, por ejemplo, es el siguiente:



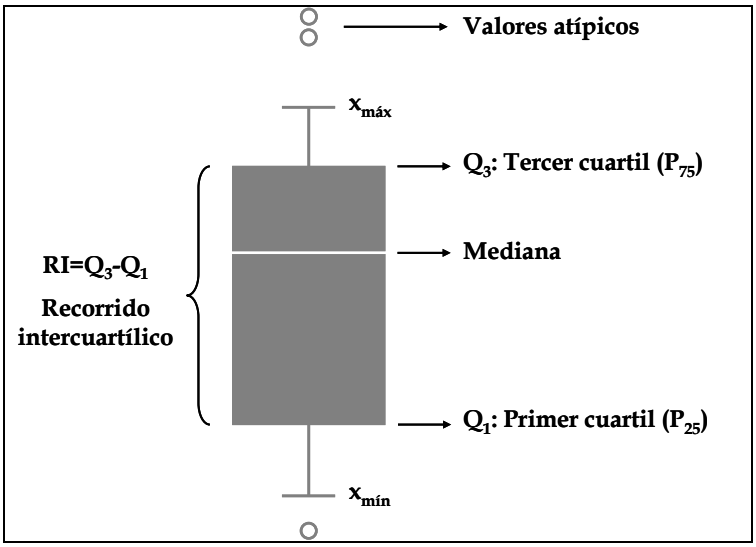
Ahora se obtiene claramente lo que anteriormente se denominó un histograma dentado, debido a la tendencia de los encuestados a declarar la talla en múltiplos de 5. En este caso, el problema se soluciona al considerar intervalos de amplitud 5 cm, que darían lugar a un histograma más adecuado.

1.5.6. Diagrama de cajas

El *diagrama de cajas* (en inglés, box-plot) es un gráfico útil para resumir y comparar grupos de datos procedentes de una variable continua, o bien de una variable discreta con un amplio recorrido de valores. Este gráfico utiliza la mediana, los cuartiles y los valores mínimo y máximo para reflejar el nivel, la dispersión y la simetría de una distribución de valores; también permite identificar valores atípicos [21][22].

Los extremos de las cajas son el primer y el tercer cuartil, de modo que la amplitud de una caja es el recorrido intercuartílico, y dentro de ella se resalta el lugar que ocupa la mediana. Por tanto, dentro de la caja se encuentran el 50% central de los datos, un 25% entre la mediana y cada cuartil, lo que permite ver la forma en que se distribuyen las observaciones: por ejemplo, si la mediana está en el centro de la caja, ello indica que la distribución es simétrica.

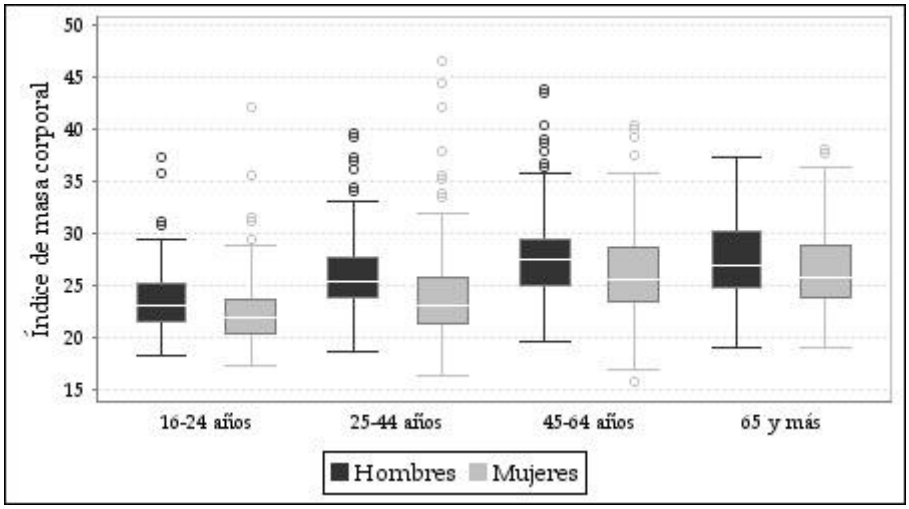
Las líneas que se proyectan fuera de la caja (patillas) se extienden hasta los denominados **valores adyacentes**, que son los valores mínimo y máximo de las observaciones una vez eliminadas las observaciones atípicas. Se consideran valores atípicos aquellos que están a una distancia de los extremos de la caja superior a 1,5 veces el recorrido intercuartílico, es decir, los que caen fuera del intervalo (Q_1-1,5RI, Q_3+1,5RI), donde Q_1 es el primer cuartil, Q_3 es el tercer cuartil y RI es el recorrido intercuartílico. La siguiente figura describe los distintos elementos de una caja:



Algunos paquetes estadísticos, como SPSS, distinguen dentro de los valores atípicos los llamados atípicos extremos, como aquellos con un valor inferior a Q_1-3R o superior a Q_3+3R. Sin embargo, Epidat no hace esta distinción. Lo que permite el programa es no mostrar en el gráfico los valores atípicos.

Para realizar un diagrama de cajas en Epidat 4 hay que importar los datos individuales a partir de un archivo en formato de Excel (*.xls, *.xlsx) o de OpenOffice (*.ods) mediante el asistente de datos. Hay dos posibilidades para identificar las variables necesarias:

- Opción 1: una variable cuantitativa para resumir y una variable cualitativa para definir las categorías del eje X; opcionalmente, se podrían identificar variables cualitativas para definir grupos (es decir, series de cajas) y/o segmentar los resultados. Esta opción se usaría, por ejemplo, para hacer un diagrama de cajas del índice de masa corporal (*Resumir*) por grupos de edad (*Categorías del eje X*) en cada sexo (*Definir grupos*). El resultado, con los datos del SICRI-2010, sería el siguiente:



- Opción 2: dos o más variables cuantitativas para resumir y, opcionalmente, una variable categórica para segmentar los resultados; las otras dos variables (*Categorías del eje X* y *Definir grupos*) no se permiten. Esta opción es útil para comparar la distribución de distintas variables medidas en los mismos individuos, por ejemplo, el peso antes y después de una dieta, o los valores de presión arterial en distintos momentos del día.

1.5.7. Gráfico de intervalos de confianza

El *gráfico de intervalos de confianza* permite representar un conjunto de estimaciones puntuales de una medida de resumen (medias, tasas de incidencia, riesgos relativos, prevalencias, etc), junto con sus intervalos de confianza.

Un caso particular de este gráfico es el conocido como *forest plot* [23], que se utiliza en meta-análisis para representar las medidas de efecto de los estudios individuales, junto con su intervalo de confianza. En este gráfico, los nombres de los estudios se representan en el eje vertical y los intervalos en el eje horizontal, y se incluye también el resultado del meta-análisis. Además, se destaca la línea correspondiente a un efecto nulo (por ejemplo, 1 para odds ratios o riesgos relativos, 0 para diferencia de riesgos o medias).

Como ya se comentó, los gráficos de líneas y de intervalos de confianza de Epidat 4 no comparten el funcionamiento del resto de gráficos. Las diferencias se derivan de que, en estas opciones, el programa no resume la información de la base de datos para hacer el gráfico, sino que representa directamente los valores introducidos por el usuario. Por este motivo, también es posible cargar los datos de forma manual, además de importarlos a partir de un archivo en formato de Excel (*.xls, *.xlsx) o de OpenOffice (*.ods) mediante el asistente de datos. El hecho de que los datos de entrada estén ya resumidos imposibilita tanto definir filtros como segmentar los resultados.

Epidat 4 permite representar más de una serie de intervalos en el mismo gráfico, siempre que todas estén definidas en las mismas categorías. Un ejemplo podrían ser las prevalencias de obesidad por Comunidad Autónoma, con sus respectivos intervalos de confianza, para hombres y para mujeres. El programa también permite elegir la orientación de los intervalos: horizontales o verticales.

Para introducir los datos manualmente, es necesario especificar el número de variables que se van a representar y el número de categorías del eje X, y completar la tabla resultante. El número de variables está limitado a 100 y el número máximo de categorías es 30.

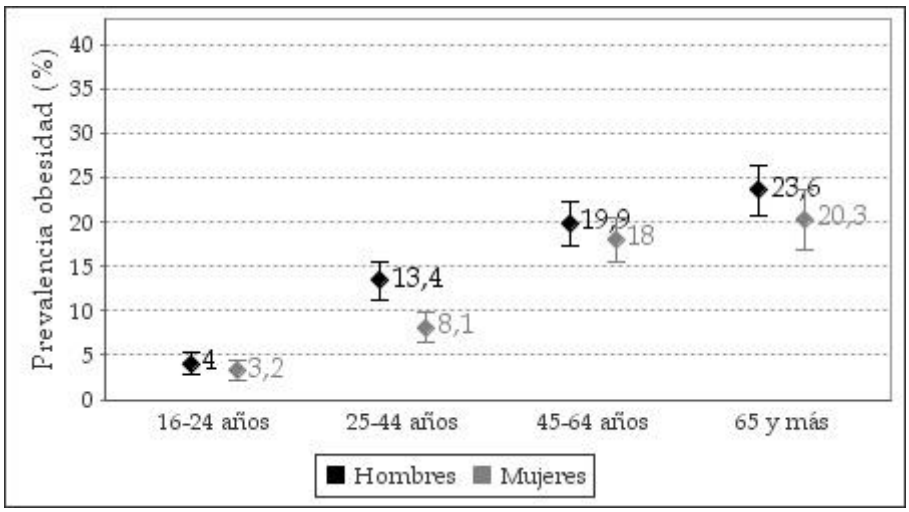
Al optar por la entrada automática se abre el asistente para la obtención de datos que permite, a través del botón "examinar", seleccionar el directorio y el archivo (OpenOffice o Excel) que contiene la tabla de valores. Es necesario recordar que Epidat 4 requiere que las tablas que han de importarse tengan una estructura fija. En este caso, la tabla debe contener tantas filas como número de categorías y tres variables (estimación puntual, límite inferior y límite superior) para cada serie de intervalos a representar en el gráfico.

Ejemplo

La prevalencia de obesidad en la población gallega de 16 años y más (fuente: SICRI-2010) aumenta con la edad, y es mayor en hombres que en mujeres. Las estimaciones de las prevalencias por grupos de edad y sexo, junto a los intervalos de confianza del 95%, se presentan en la siguiente tabla:

	Hombres			Mujeres		
	P(%)	IC(95%)		P(%)	IC(95%)	
16-24 años	4,0	2,7	5,3	3,2	2,1	4,4
25-44 años	13,4	11,3	15,6	8,1	6,4	9,9
45-64 años	19,9	17,4	22,3	18,0	15,5	20,6
65 y más	23,6	20,7	26,4	20,3	16,9	23,7

Para representar gráficamente estos datos, en Epidat 4 hay que hacer un gráfico de intervalos de confianza. Los datos pueden introducirse manualmente en una tabla con 2 variables y 4 categorías; la estimación 1 corresponde a los datos de los hombres y la estimación 2 a los de las mujeres. El gráfico resultante, que se muestra a continuación, muestra claramente la tendencia creciente de la prevalencia de obesidad con la edad, tanto en hombres como en mujeres.



Bibliografía

1 López-Moreno S, Garrido-Latorre F, Hernández-Ávila M. Desarrollo histórico de la epidemiología: su formación como disciplina científica. Salud Pública Méx. 2000;42(2):133-43.

2 Freedman D. From association to causation: some remarks on the history of statistics. Statistical Science. 1999;14(3):243-58.

3 Morabia A. Pierre-Charles-Alexandre Louis and the evaluation of bloodletting. J R Soc Med. 2006;99:158-60.

4 Silva LC. Cultura estadística e investigación científica en el campo de la salud: una mirada crítica. Madrid: Díaz de Santos; 1997.

5 Salas-Salvadó J, Rubio MA, Barbany M, Moreno B, Aranceta J, Bellido D et al. Consenso SEEDO 2007 para la evaluación del sobrepeso y la obesidad y el establecimiento de criterios de intervención terapéutica. Med Clin (Barc). 2007;128(5):184-96.

6 Pagano M, Gauvreau K. Fundamentos de bioestadística. 2ª ed. México: Thomson Learning; 2001.

7 Altman DG, Bland JM. Statistics notes: quartiles, quintiles, centiles, and other quantiles. BMJ. 1994;309:996.

8 Sobradillo B, Aguirre A, Aresti U, Bilbao A, Fernández-Ramos C, Lizárraga A et al. Curvas y tablas de crecimiento (estudios longitudinal y transversal). Instituto de Investigación sobre Crecimiento y Desarrollo. Bilbao: Fundación Faustino Orbegozo Eizaguirre; 2004.

9 Fernández-Abascal H, Guijarro MM, Rojo JL, Sanz JA. Cálculo de probabilidades y estadística. Barcelona: Editorial Ariel; 1994.

10 Armitage P. Correlation. En: Armitage P, Colton T, editores. Encyclopedia of Biostatistics Vol. 1. Chichester: John Wiley & Sons; 1998. pp. 971-6.

11 Benach J. Notas de metodología y estadística: las manos y la escritura (correlación no equivale a causa). JANO. 1996;L(1163):1306.

12 Pita-Fernández S. Notas de metodología y estadística: correlación frente a causalidad. JANO. 1996;LI(1174):243-4.

13 Rosner B. Fundamentals of biostatistics. 5ª ed. Belmont, CA: Duxbury Press; 2000.

14 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;i:307-10.

15 González-Alastrué JA, Jover L. Los gráficos en la comunicación y el razonamiento científicos: ¿instrumento u ornamento?. Med Clin (Barc). 2004;122(Supl 1):3-10.

16 Molinero LM [página en Internet]. Gráficos de datos estadísticos en medicina. Disponible en: www.seh-lelha-org/stat1.htm

17 Tufte ER. The visual display of quantitative information. Cheshire: Graphics Press; 1983.

18 Tufte ER. The visual display of quantitative information. 2ª ed. Connecticut: Graphics Press; 2009.

19 Oliveras KG. El histograma (I). Qué es y para qué sirve. JANO. 1997;LII(1204):1070.

20 Oliveras KG. El histograma (II). Objetivo: entender los datos. JANO. 1997;LII(1205):1171-2.

21 Simpson RJ, Johnson TA, Amara IA. The box-plot: an exploratory analysis for biomedical publications. Am Heart J. 1988;116 (6 Part 1):1663-5.

22 Williamson DF, Parker RA, Kendrich JS. The box plot: a simple visual method to interpret data. Ann Intern Med. 1 Jun 1989;110(11):916-21.

23 Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ. 2001;322:1479-80.

Anexo 1: Fórmulas del módulo de análisis descriptivo

Esquema del módulo

1. Tablas de frecuencias
2. Tablas de contingencia
3. Estadísticos descriptivos
4. Correlación
5. Gráficos
   - 5.1. Barras
   - 5.2. Sectores
   - 5.3. Líneas
   - 5.4. Dispersión
   - 5.5. Histograma
   - 5.6. Diagrama de cajas
   - 5.7. Intervalos de confianza

3.- ESTADÍSTICOS DESCRIPTIVOS

Se tienen n observaciones x_1, x_2, …, x_n de una variable numérica X.

Medidas de tendencia central [Rosner (2000, p. 9-16)]:

Suma:

$$S = \sum_{i=1}^{n} x_i$$

Media:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Mediana: es el percentil 50 (ver percentiles).

Moda: es la observación u observaciones más frecuentes.

Media geométrica:

$$\bar{x}_g = \text{Exp}\left\{\frac{1}{n}\sum_{i=1}^{n} \ln(x_i)\right\}$$

Medidas de dispersión [Rosner (2000, p. 18-24)]:

Desviación típica:

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

Varianza:

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

Coeficiente de variación:

$$CV = \frac{s}{\bar{x}}$$

Recorrido:

$$R = \text{Máx}\{x_i\} - \text{Mín}\{x_i\}$$

Recorrido intercuartílico:

$$RI = Q_3 - Q_1$$, donde Q_1 y Q_3 son el primer y el tercer cuartil, respectivamente

Percentiles [Altman & Bland (1994), Mood & Graybill (1963, p. 408)]:

Percentil de orden k:

$$P_k = (1-f)x_r + fx_{r+1}$$

Cuartiles:

$$Q_1=P_{25}, \quad Q_2=P_{50}, \quad Q_3=P_{75}$$

Deciles:

$$D_k=P_k$$, con k=10, 20, 30, 40, 50, 60, 70, 80, 90

Donde:

- $x_1, x_2, \ldots, x_n$ es la muestra ordenada de valores,
- $R = \frac{(n+1)k}{100}$,
- $r = [R]$ es la parte entera de R, $0 \leq r \leq n$,
- $f = R - r$ es la parte fraccionaria de R,
- $x_0=x_1$ y $x_{n+1}=x_n$,
- n es el número de observaciones,
- k es el orden del percentil.

Medidas de forma de la distribución [Fernández-Abascal (1994, p. 273-274)]:

Asimetría:

$$A = \frac{m_3}{m_2^{3/2}}$$

Curtosis:

$$K = \frac{m_4}{m_2^2} - 3$$

Donde:

- $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$ es el momento central de orden k, k=2, 3, 4.

4.- CORRELACIÓN

Coeficiente de correlación de Pearson [Rosner (2000, p. 451-55)]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Coeficiente de correlación de Spearman [Rosner (2000, p. 497)]:

$$r = \frac{\sum_{i=1}^{n}(r_{x,i} - \bar{r}_x)(r_{y,i} - \bar{r}_y)}{\sqrt{\sum_{i=1}^{n}(r_{x,i} - \bar{r}_x)^2 \sum_{i=1}^{n}(r_{y,i} - \bar{r}_y)^2}}$$

Donde:

- $x_1, x_2, \ldots x_n$ son las n observaciones de la variable X,
- $y_1, y_2, \ldots y_n$ son las n observaciones de la variable Y,
- $r_{x,i}$ es el rango de la observación x_i, es decir, es la posición que ocupa x_i en la muestra ordenada. En caso de empate, a los valores iguales se les asigna la media de sus rangos.

Bibliografía

- Altman DG, Bland JM. Statistics notes: quartiles, quintiles, centiles, and other quantiles. BMJ. 1994;309:996.

- Fernández-Abascal H, Guijarro MM, Rojo JL, Sanz JA. Cálculo de probabilidades y estadística. Barcelona: Editorial Ariel; 1994.

- Mood AM, Graybill FA. Introduction to the theory of statistics. New York: McGraw-Hill; 1963.

- Rosner B. Fundamentals of biostatistics. 5ª ed. Belmont, CA: Duxbury Press; 2000.