6

Estadística descriptiva. Representación de datos descriptivos

Alberto Rodríguez Benot
Rodolfo Crespo Montero

6.1. Introducción

Tal como vimos en la introducción, la estadística descriptiva comprende la presentación, organización y resumen de los datos de una manera científica. Mediante el estudio de ciertos estadísticos, nos permite conocer magnitudes que representan a la globalidad de los datos disponibles de forma resumida. Incluye diversos métodos de organizar y representar gráficamente los datos, para dar una idea de lo que nos muestran. Las tablas, los diagramas de barras o los gráficos sectoriales o "tartas" son algunos de los elementos de estadística descriptiva. También incluye varios parámetros numéricos (como la media aritmética) que resumen los datos con muy pocos números clave.

6.2. Síntesis de los datos

Una vez organizados los datos en tablas y representados gráficamente, es útil sintetizarlos o resumirlos en medidas o números que permitan trabajar cómodamente y que contengan el máximo de información. Existen dos tipos de medi-

das que describen las características de la distribución de frecuencias de los valores de una variable: las medidas de centralización y de dispersión.

6.2.1. Medidas de centralización

Las medidas de centralización definen los valores de la variable en torno a los cuales tienden a concentrarse las observaciones. Son: media, mediana, moda y los cuartiles, deciles y percentiles.

- *Media:* la media aritmética es la medida de centralización más conocida y utilizada. Se calcula sumando todos los valores observados y dividiendo por el número de observaciones de la muestra. Se representa como x. Su principal ventaja es su fácil manejo matemático y estadístico. Sin embargo, tiene la desventaja de ser muy sensible a los valores extremos en una muestra que no tenga una distribución normal (veremos más tarde que significa esto). Si por ejemplo analizamos los días de estancia hospitalaria de los 7 últimos trasplantados renales en nuestro Servicio, y tenemos: 3, 3, 4, 7, 9, 11 y 12 días. Puesto que son siete datos, x = (3 + 3 + 3 + 4 + 7 + 9 + 11 + 12) / 7 = 49/7 = 7; la estancia media de los pacientes es de 7 días. Pero si en lugar de 12 días un paciente permanece ingresado 89, la nueva media sería 18 días, muy alejada de la previa de 7 días. Esto se debe a que un valor extremo (89), muy distante del resto, influye negativamente en la media. En este caso, la mediana es una medida mejor de centralización.
- *La media geométrica* es un parámetro de centralización que se utiliza para datos exponenciales o del tipo de crecimiento de poblaciones. Se calcula multiplicando los datos entre sí y aplicando después la raíz de orden n. Se utiliza con mucha menor frecuencia que la media aritmética.
- *Mediana:* la mediana es la observación equidistante de los extremos, o lo que es lo mismo, el valor que, una vez ordenados los datos, deja igual número de observaciones por encima y por debajo. Siguiendo con el ejemplo anterior (3 + 3 + 3 + 4 + 7 + 9 + 11 + 12), la mediana es el valor 7; y cuando sustituimos el 12 por 89 días (3 + 3 + 3 + 4 + 7 + 9 + 11 + 89), la mediana sigue siendo el valor 7. Como vemos, la mediana es mucho menos sensible a los valores extremos que la media, y es la medida de centralización a emplear en las variables cualitativas ordinales, en las que es imposible calcular la media aritmética. Si n es par, la mediana es al media de los dos valores centrales. Por supuesto, se puede utilizar también con datos interválicos y proporcionales. Gráficamente, en el polí-

gono de frecuencias acumuladas, la mediana es el valor correspondiente al 50% de las observaciones en el eje de abscisas (eje x).

- *Moda:* la moda es el valor que se observa con más frecuencia, el más repetido. En el ejemplo anterior, (3 + 3 + 3 + 4 + 7 + 9 + 11 + 12) la moda es 3 por ser el valor más repetido. Si no se repite ningún valor, la muestra no tiene moda, es amodal. Si se repiten varios valores diferentes, puede ser bimodal, trimodal o multimodal. Gráficamente, la moda equivale al valor que alcanza la frecuencia máxima o pico en el polígono de frecuencias.
- *Cuartiles, Deciles, Percentiles:* son medidas de localización, pero no central, sino que localizan otros puntos de una distribución. Los cuartiles dividen los datos en cuatro partes iguales, los deciles en diez partes iguales y los percentiles, en cien partes iguales. Por definición, el cuartil 2 coincide con el decil 5 y con el percentil 50, y todos ellos con la mediana.

6.2.2. Medidas de dispersión

Una vez definidos los valores de la variable en torno a los cuales tienden a concentrarse las observaciones, el siguiente planteamiento es describir cómo de agrupados o dispersos se encuentran los datos de la muestra en torno a esos valores, pues una medida de tendencia central es insuficiente para caracterizar una distribución. Esta información nos la ofrecen las medidas de dispersión: recorrido o rango, desviación media, varianza, desviación estándar y coeficiente de variación.

- *Recorrido o rango:* es la diferencia entre los valores máximo y mínimo de la variable. En el ejemplo 3, 3, 4, 7, 9, 11, 12, el rango es 12-3 = 9. Su principal ventaja es que se calcula con gran facilidad. Pero dado que no tiene en cuenta los valores intermedios, su utilidad es muy limitada. Es útil como media de dispersión en las variables cualitativas ordinales, o para indicar si nuestros datos tienen algunos valores extraordinarios.
- *Recorrido intercuartílico:* como consecuencia de los problemas que presenta el recorrido, en particular su inestabilidad al considerar muestras diferentes o bien cuando se añaden nuevos individuos, a veces se usa otro índice de dispersión con datos ordinales, el recorrido intercuartil, también llamado media de dispersión. Se calcula dividiendo en primer lugar los datos (previamente ordenados) en cuatro partes iguales, obteniendo así los cuartiles Q1, Q2 y Q3; la diferencia entre el cuartil Q3 y el Q1 es el recorrido intercuartil, y abarca el 50% de los datos. Recordemos que Q2 = mediana. Como el recorrido intercuartil se refiere sólo

al 50% central de los datos, se afecta en mucha menor medida por los valores extremos que el recorrido propiamente dicho, lo que la convierte en una medida mucho más útil.

- *Desviación media, Varianza (S^2) y desviación estándar (S o DE):* son las medidas de dispersión más frecuentemente utilizadas en biomedicina. Se basan en cálculos de la diferencia entre cada valor y la media aritmética (x-x). Al calcular esta diferencia, debe prescindirse del signo negativo o positivo de cada resultado, por lo que la medida de dispersión se muestra como "±" desviación. La principal diferencia entre las tres medidas es cómo se prescinde del signo negativo: en la desviación media, se toman los valores absolutos /x-x/; en la varianza (S^2 para muestras y $\sigma 2$ para poblaciones) se eleva al cuadrado la diferencia: (x-x)2.

Como en la varianza los datos están al cuadrado, para regresar a las unidades originales basta tomar la raíz cuadrada de la varianza. Obtenemos así la desviación típica o estándar (DE), S para muestras y σ para poblaciones.

Cuanto más dispersos estén los valores de la media, mayor será la desviación estándar. Es la medida de dispersión más importante y utilizada.

- Coeficiente de variación: se emplea para comparar la variabilidad relativa de diferentes distribuciones, partiendo del problema de que las desviaciones estándar no son comparables al estar referidas a distintas medias. Este sería el caso de querer comparar la variabilidad de la presión arterial de un grupo de pacientes con su edad. Se usa con frecuencia para comparar métodos de medida, y es un valor adimensional. Se calcula dividiendo la DE por la media, multiplicando después por 100.

6.2.3. Medidas para variables cualitativas

La mayoría de las medidas anteriores no son aplicables a las variables cualitativas, ya que sus valores no son numéricos, sino que representan recuentos o frecuencias de ocurrencia de un suceso. Existen tres formas básicas de presentar estos datos:

1. Proporción o frecuencia relativa, que es el número de casos que se presenta una característica (a) dividido por el número total de observaciones (a+b): a/(a+b). Su valor oscila entre 0 y 1. Si multiplicamos una proporción por 100, obtenemos un porcentaje.
2. Razón o cociente, que es el número de casos que presentan una característica (a) dividido por el número de casos que no la presentan (b): (a/b).
3. Tasa, que es similar a la proporción, pero multiplicada por una cifra (por

ejemplo 1000, 10000, 100000) y se calcula sobre un determinado período de tiempo.

6.3. Representación gráfica

Una vez obtenidos los datos es preciso mostrarlos de una forma ordenada y comprensible. La forma mas sencilla es colocarlos en una *Tabla*, donde se muestran las variables, las categorías de cada variable y el número de eventos de cada categoría. En ciertas ocasiones, especialmente cuando trabajamos con un gran número de datos, las tablas no son prácticas y es necesario una mejor visión de los datos con una mirada rápida. Esto se consigue con los gráficos. La selección del gráfico dependerá del tipo de datos empleados. Comenzaremos con los gráficos para datos cuantitativos.

6.3.1. Histograma

Se utiliza para variables cuantitativas continuas. En el eje x se muestran los datos de la variable, que por ser continuos requieren ser agrupados previamente en intervalos, y en el eje y se representa la frecuencia con la que aparece cada dato. La anchura del intervalo y la altura que alcanza determinan el área de cada intervalo, que es proporcional a la frecuencia de cada intervalo. Da una idea muy aproximada de la forma de la distribución que sigue la variable (figura 19).

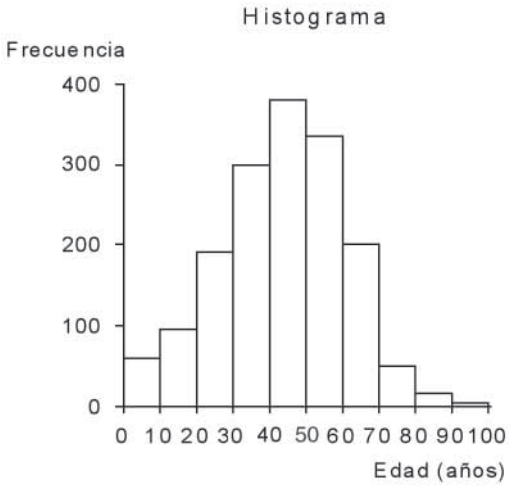

Figura 19. Histograma para representar variables continuas

6.3.2. Polígono de frecuencias

Utiliza la misma escala que el histograma, y se construye uniendo los puntos medios de la zona más alta de los rectángulos (figura 20). También aquí lo más importante es el área existente debajo del polígono, que es igual al área del histograma correspondiente. En el polígono de frecuencias acumuladas, la línea representa la frecuencia de cada intervalo sumada a la de los intervalos anteriores (figura 21). Es un método práctico para determinar percentiles.

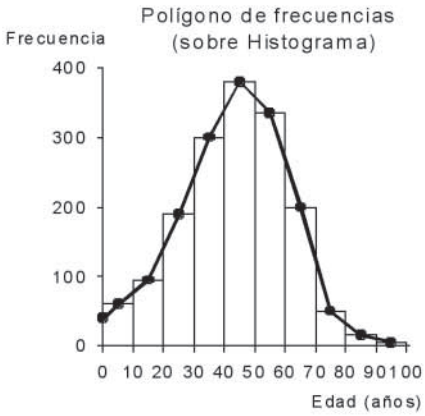

Figura 20. Polígono de frecuencias sobre histograma para representar variables continuas

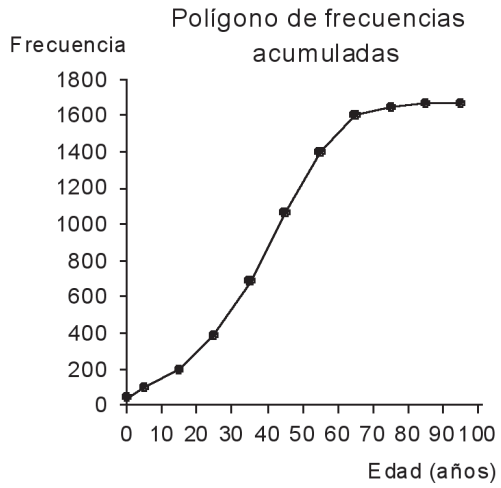

Figura 21. Polígono de frecuencias acumuladas para representar variables continuas

6.3.3. Nube de puntos

Es un gráfico donde se muestran dos variables cuantitativas, una en el eje x y otro en el y, mostrando los valores mediante puntos o símbolos (figura 22).

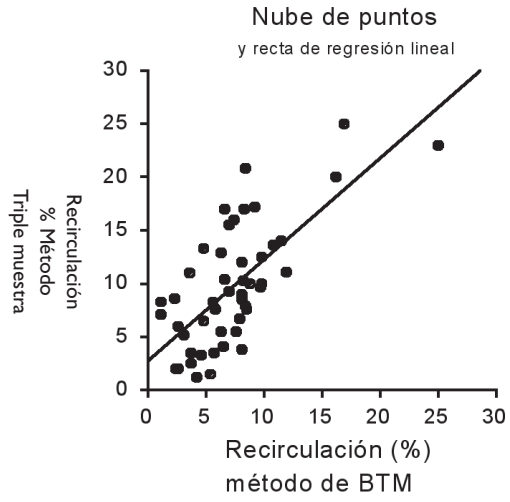

Figura 22. Nube de puntos para describir la relación entre dos variables continuas

Para los datos cualitativos, los más utilizados son:

6.3.4. Diagrama de barras

Se utiliza para variables cualitativas y cuantitativas discretas, y se construyen de forma similar al histograma, pero las barras están separadas entre sí (indicando que la variable no ocupa todo el eje de abscisas, precisamente por ser discreta o cualitativa). El diagrama de barras compuesto representa dos o más variables en el mismo gráfico (figura 23).

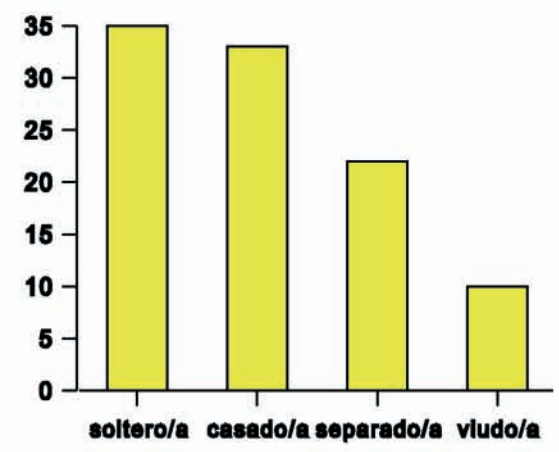

Figura 23. Diagrama de barras para variables categóricas

6.3.5. Gráfico circular o de sectores

Es otro método empleado con frecuencia para datos cualitativos, en el que un círculo representa el total, y un segmento o porción del pastel es la proporción o porcentaje de cada categoría de la variable (figura 24). Es el gráfico adecuado para variables con categorías mutuamente excluyentes (no se puede estar soltero y casado a la vez).

Figura 24. Gráfico de pastel para variables categóricas

6.3.6. Gráfico de caja

Sirve para representar variables cualitativas en escala ordinal y cuantitativas discretas. Se construye un rectángulo de altura igual al recorrido intercuartílico, dentro se traza un segmento en el punto correspondiente a la mediana y se define los valores adyacentes o bigotes: el valor adyacente inferior es el valor más pequeño de la distribución. El valor adyacente superior es la observación más grande de la serie (figura 25). Los valores de la distribución que sean menores que el valor adyacente inferior o mayores que el superior se denominan observaciones extremas o *"outliers"*.

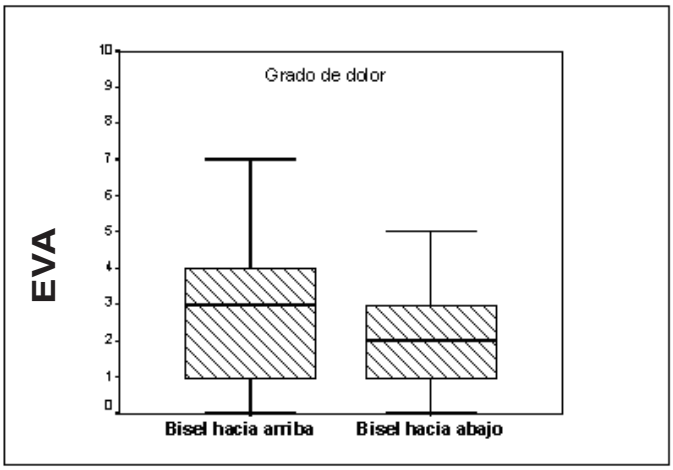

Figura 25. Diagramas de caja (boxplot) para variables cualitativas ordinales y cuantitativas discretas

6.4. Consideraciones importantes

Podemos considerar la estadística descriptiva como el conjunto de técnicas para ordenar y representar los datos en tablas, y resumirlos mediante el cálculo de diferentes medidas. Por tanto, podemos distinguir tres apartados:

-Tabulación, que consiste en ordenar los datos originales y presentarlos de forma que, sin perder información, sea más fácil conocer la distribución de los mismos. El resultado final es una tabla donde se muestran los valores de la variable que se tabula y sus frecuencias.

-Cálculo de medidas para resumir la distribución. Pueden ser de tendencia central, que indican alrededor de que valores se agrupan los datos observados; y de variabilidad o dispersión, que indican si los valores de la variable están muy dispersos o concentrados.

-Representación gráfica, que facilita un análisis visual de los datos y permite sacar conclusiones acerca de las características globales de la distribución.