Estadística descriptiva de los datos

Pita Fernández S, Pértega Díaz, S.
Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Juan Canalejo. A Coruña.
Actualización 06/03/2001.

Pita Fernández, S. Uso de la estadística y la epidemiología en atención primaria. En: Gil VF, Merino J, Orozco D, Quirce F. Manual de metodología de trabajo en atención primaria. Universidad de Alicante. Madrid, Jarpyo Editores, S.A. 1997; 115-161.

Introducción

Existen diferentes razones por las cuales los profesionales de la atención primaria deben conocer los fundamentos de la epidemiología y la estadística como instrumentos del trabajo cotidiano. Entre dichas razones señalamos las siguientes: los términos estadísticos y epidemiológicos invaden la literatura médica, la medicina es cada vez más cuantitativa, su conocimiento nos permitirá leer la bibliografía médica con más capacidad crítica para detectar errores potenciales y falacias. Nos será también útil para llegar a conclusiones correctas acerca de procedimientos para el diagnóstico y del resultado de las pruebas [1,2]. Su conocimiento nos permitirá a su vez valorar protocolos de estudio e informes remitidos para su publicación y participar, en definitiva, en la investigación médica. Resulta imprescindible, por lo tanto, conocer los conceptos básicos de estadística que nos faciliten la realización de estudios y conocer las posibilidades a desarrollar con ayuda de profesionales estadísticos para mejorar dicho análisis.

En este trabajo se pretende dar a conocer algunas nociones estadísticas que nos ayudarán a explorar y describir, en un primer momento, nuestros datos.

Poblaciones y muestras

Cuando se realiza un estudio de investigación, se pretende generalmente inferir o generalizar resultados de una muestra a una población. Se estudia en particular a un reducido número de individuos a los que tenemos acceso con la idea de poder generalizar los hallazgos a la población de la cual esa muestra procede. Este proceso de inferencia se efectúa por medio de métodos estadísticos basados en la probabilidad.

La población representa el conjunto grande de individuos que deseamos estudiar y generalmente suele ser inaccesible. Es, en definitiva, un colectivo homogéneo que reúne unas características determinadas.

La muestra es el conjunto menor de individuos (subconjunto de la población accesible y limitado sobre el que realizamos las mediciones o el experimento con la idea de obtener conclusiones generalizables a la población). El individuo es cada uno de los componentes de la población y la muestra. La muestra debe ser representativa de la población y con ello queremos decir que cualquier individuo de la población en estudio debe haber tenido la misma probabilidad de ser elegido.

Las razones para estudiar muestras en lugar de poblaciones son diversas y entre ellas podemos señalar [3]:

a. Ahorrar tiempo. Estudiar a menos individuos es evidente que lleva menos tiempo.
b. Como consecuencia del punto anterior ahorraremos costes.
c. Estudiar la totalidad de los pacientes o personas con una característica determinada en muchas ocasiones puede ser una tarea inaccesible o imposible de realizar.
d. Aumentar la calidad del estudio. Al disponer de más tiempo y recursos, las observaciones y mediciones realizadas a un reducido número de individuos pueden ser más exactas y plurales que si las tuviésemos que realizar a una población.
e. La selección de muestras específicas nos permitirá reducir la heterogeneidad de una población al indicar los criterios de inclusión y/o exclusión.

Tipos de datos

Lo que estudiamos en cada individuo de la muestra son las variables (edad, sexo, peso, talla, tensión arterial sistólica, etcétera). Los datos son los valores que toma la variable en cada caso. Lo que vamos a realizar es medir, es decir, asignar valores a las variables incluidas en el estudio. Deberemos además concretar la escala de medida que aplicaremos a cada variable.

La naturaleza de las observaciones será de gran importancia a la hora de elegir el método estadístico más apropiado para abordar su análisis. Con este fin, clasificaremos las variables, a grandes rasgos, en dos tipos [3-5]: variables cuantitativas o variables cualitativas.

a. **Variables cuantitativas**. Son las variables que pueden medirse, cuantificarse o expresarse numéricamente. Las variables cuantitativas pueden ser de dos tipos:

   - Variables cuantitativas continuas, si admiten tomar cualquier valor dentro de un rango numérico determinado (edad, peso, talla).
   - Variables cuantitativas discretas, si no admiten todos los valores intermedios en un rango. Suelen tomar solamente valores enteros (número de hijos, número de partos, número de hermanos, etc).

b. **Variables cualitativas.** Este tipo de variables representan una cualidad o atributo que clasifica a cada caso en una de varias categorías. La situación más sencilla es aquella en la que se clasifica cada caso en uno de dos grupos (hombre/mujer, enfermo/sano, fumador/no fumador). Son datos dicotómicos o binarios. Como resulta obvio, en muchas ocasiones este tipo de clasificación no es suficiente y se requiere de un mayor número de categorías (color de los ojos, grupo sanguíneo, profesión, etcétera).

En el proceso de medición de estas variables, se pueden utilizar dos escalas:

- **Escalas nominales**: ésta es una forma de observar o medir en la que los datos se ajustan por categorías que no mantienen una relación de orden entre sí (color de los ojos, sexo, profesión, presencia o ausencia de un factor de riesgo o enfermedad, etcétera).
- **Escalas ordinales**: en las escalas utilizadas, existe un cierto orden o jerarquía entre las categorías (grados de disnea, estadiaje de un tumor, etcétera).

Estadística descriptiva

Una vez que se han recogido los valores que toman las variables de nuestro estudio (datos), procederemos al análisis descriptivo de los mismos. Para variables categóricas, como el sexo o el estadiaje, se quiere conocer el número de casos en cada una de las categorías, reflejando habitualmente el porcentaje que representan del total, y expresándolo en una tabla de frecuencias.

Para variables numéricas, en las que puede haber un gran número de valores observados distintos, se ha de optar por un método de análisis distinto, respondiendo a las siguientes preguntas:

a. ¿Alrededor de qué valor se agrupan los datos?
b. Supuesto que se agrupan alrededor de un número, ¿cómo lo hacen? ¿muy concentrados? ¿muy dispersos?

a. Medidas de tendencia central

Las medidas de centralización vienen a responder a la primera pregunta. La medida más evidente que podemos calcular para describir un conjunto de observaciones numéricas es su valor medio. La **media** no es más que la suma de todos los valores de una variable dividida entre el número total de datos de los que se dispone.

Como ejemplo, consideremos 10 pacientes de edades 21 años, 32, 15, 59, 60, 61, 64, 60, 71, y 80. La media de edad de estos sujetos será de:

$$\overline{X} = \frac{21+32+15+59+60+61+64+60+71+80}{10} = 52.3\ años$$

Más formalmente, si denotamos por (X_1, X_2,...,X_n) los n datos que tenemos recogidos de la variable en cuestión, el valor medio vendrá dado por:

$$Media(X) = \frac{\sum_{j=1}^{n} X_j}{n}$$

Otra medida de tendencia central que se utiliza habitualmente es la **mediana**. Es la observación equidistante de los extremos.

La mediana del ejemplo anterior sería el valor que deja a la mitad de los datos por encima de dicho valor y a la otra mitad por debajo. Si ordenamos los datos de mayor a menor observamos la secuencia:

15, 21, 32, 59, 60, 60,61, 64, 71, 80.

Como quiera que en este ejemplo el número de observaciones es par (10 individuos), los dos valores que se encuentran en el medio son 60 y 60. Si realizamos el cálculo de la media de estos dos valores nos dará a su vez 60, que es el valor de la mediana.

Si la media y la mediana son iguales, la distribución de la variable es simétrica. La media es muy sensible a la variación de las puntuaciones. Sin embargo, la mediana es menos sensible a dichos cambios.

Por último, otra medida de tendencia central, no tan usual como las anteriores, es la moda, siendo éste el valor de la variable que presenta una mayor frecuencia.

En el ejemplo anterior el valor que más se repite es 60, que es la **moda**.

b. Medidas de dispersión

Tal y como se adelantaba antes, otro aspecto a tener en cuenta al describir datos continuos es la dispersión de los mismos. Existen distintas formas de cuantificar esa variabilidad. De todas ellas, la **varianza** (S^2) de los datos es la más utilizada. Es la media de los cuadrados de las diferencias entre cada valor de la variable y la media aritmética de la distribución.

$$S_X^2 = \frac{\sum_{j=1}^{n} \left(X_j - Media(X)\right)^2}{n}$$

Esta varianza muestral se obtiene como la suma de las de las diferencias de cuadrados y por tanto tiene como unidades de medida el cuadrado de las unidades de medida en que se mide la variable estudiada.

En el ejemplo anterior la varianza sería:

$$S_x^2 = \frac{(15-52.3)^2 + (21-53,2)^2 \ldots (80-53,2)^2}{10} = 427,61$$

La **desviación típica** (S) es la raíz cuadrada de la varianza. Expresa la dispersión de la distribución y se expresa en las mismas unidades de medida de la variable. La desviación típica es la medida de dispersión más utilizada en estadística.

$$S_X = \sqrt{\frac{\sum_{j-1}^{n}\left(X_j - Media(X)\right)^2}{n}}$$

Aunque esta fórmula de la desviación típica muestral es correcta, en la práctica, la estadística nos interesa para realizar inferencias poblacionales, por lo que en el denominador se utiliza, en lugar de n, el valor n-1.

Por tanto, la medida que se utiliza es la cuasidesviación típica, dada por:

$$S_X = \sqrt{\frac{\sum_{j-1}^{n}\left(X_j - Media(X)\right)^2}{n-1}}$$

Aunque en muchos contextos se utiliza el término de desviación típica para referirse a ambas expresiones.

En los cálculos del ejercicio previo, la desviación típica muestral, que tiene como denominador n, el valor sería 20.678. A efectos de cálculo lo haremos como n-1 y el resultado seria 21,79.

El haber cambiado el denominador de n por n-1 está en relación al hecho de que esta segunda fórmula es una estimación más precisa de la **desviación estándar** verdadera de la población y posee las propiedades que necesitamos para realizar inferencias a la población.

Cuando se quieren señalar valores extremos en una distribución de datos, se suele utilizar la amplitud como medida de dispersión. La amplitud es la diferencia entre el valor mayor y el menor de la distribución.

Por ejemplo, utilizando los datos del ejemplo previo tendremos 80-15 =65.

Como medidas de variabilidad más importantes, conviene destacar algunas características de la varianza y desviación típica:

- Son índices que describen la variabilidad o dispersión y por tanto cuando los datos están muy alejados de la media, el numerador de sus fórmulas será grande y la varianza y la desviación típica lo serán.
- Al aumentar el tamaño de la muestra, disminuye la varianza y la desviación típica. Para reducir a la mitad la desviación típica, la muestra se tiene que multiplicar por 4.
- Cuando todos los datos de la distribución son iguales, la varianza y la desviación típica son iguales a 0.
- Para su cálculo se utilizan todos los datos de la distribución; por tanto, cualquier cambio de valor será detectado.

Otra medida que se suele utilizar es el **coeficiente de variación** (CV). Es una medida de dispersión relativa de los datos y se calcula dividiendo la desviación típica muestral por la media y multiplicando el cociente por 100. Su utilidad estriba en que nos permite comparar la dispersión o variabilidad de dos o más grupos. Así, por ejemplo, si tenemos el peso de 5 pacientes (70, 60, 56, 83 y 79 Kg) cuya media es de 69,6 kg. y su desviación típica (s) = 10,44 y la TAS de los mismos (150, 170, 135, 180 y 195 mmHg) cuya media es de 166 mmHg y su desviación típica de 21,3. La pregunta sería: ¿qué distribución es más dispersa, el peso o la tensión arterial? Si comparamos las desviaciones típicas observamos que la desviación típica de la tensión arterial es mucho mayor; sin embargo, no podemos comparar dos variables que tienen escalas de medidas diferentes, por lo que calculamos los coeficientes de variación:

$$\text{CV de la variable peso} = \frac{10,44}{69,6} = 15\%$$

$$\text{CV de la variable TAS} = \frac{21,30}{166} = 12,8\%$$

A la vista de los resultados, observamos que la variable peso tiene mayor dispersión.

Cuando los datos se distribuyen de forma simétrica (y ya hemos dicho que esto ocurre cuando los valores de su media y mediana están próximos), se usan para describir esa variable su media y desviación típica. En el caso de distribuciones asimétricas, la mediana y la amplitud son medidas más adecuadas. En este caso, se suelen utilizar además los **cuartiles y percentiles**.

Los cuartiles y percentiles no son medidas de tendencia central sino **medidas de posición**. El percentil es el valor de la variable que indica el porcentaje de una distribución que es igual o menor a esa cifra.

Así, por ejemplo, el percentil 80 es el valor de la variable que es igual o deja por debajo de sí al 80% del total de las puntuaciones. Los cuartiles son los valores de la variable que dejan por debajo de sí el 25%, 50% y el 75% del total de las puntuaciones y así tenemos por tanto el primer cuartil (Q1), el segundo (Q2) y el tercer cuartil (Q3).

Bibliografía

1. Sackett, D.L., Haynes, R.B., Guyatt, G.H., Tugwell, P. Epidemiología clínica. Ciencia básica para la medicina clínica. 2ª ed. Madrid: Médica Panamericana; 1994.

2. Fletcher RH., Fletcher SW., Wagner E.H. Epidemiología clínica. 2ª ed. Barcelona: Masson, Williams & Wilkins; 1998.

3. Dawson-Saunders B, Trapp RG. Bioestadística Médica. 2ª ed. México: Editorial el Manual Moderno; 1996.

4. Milton JS, Tsokos JO. Estadística para biología y ciencias de la salud. Madrid: Interamericana M_cGraw Hill; 2001.

5. Martín Andrés A, Luna del Castillo JD. Bioestadística para las ciencias de la salud. 4ª ed. Madrid: NORMA; 1993.