ESTADÍSTICA DESCRIPTIVA: UNA VARIABLE

Julián de la Horra
Departamento de Matemáticas U.A.M.

1 Introducción

Cuando estamos interesados en estudiar alguna característica de una población (peso, longitud de las hojas, indicadores de contaminación, etc) lo más completo es, evidentemente, estudiar la población entera. Pero esto suele requerir demasiado tiempo y demasiado dinero. Otras veces, el estudio de un elemento es destructivo, con lo cual es imposible hacer un análisis de toda la población (nos quedaríamos sin población). Por tanto, normalmente, nos conformaremos con un conocimiento parcial de la población. Esto lo conseguiremos observando unos cuantos elementos y viendo cómo es o cuánto vale en ellos esa característica que nos interesa. Este conjunto de elementos que observamos es lo que llamaremos una muestra de la población.

El objetivo básico de la Estadística Descriptiva para una variable es hacer una descripción lo más sencilla posible de los resultados obtenidos en la muestra. Esta descripción se hará mediante representaciones gráficas y mediante resúmenes numéricos. Este capítulo está dedicado a hacer un estudio descriptivo de lo obtenido en una muestra concreta, cuando nos interesamos en una sola característica, es decir, en una sola variable estadística o variable respuesta. Estas variables pueden ser de dos tipos: cualitativas y cuantitativas.

2 Variables cualitativas

Una variable respuesta es cualitativa cuando sólo puede clasificarse en categorías no numéricas. Ejemplos de variables cualitativas son el color de los ojos de las personas de una ciudad, la Facultad o Escuela en la que están matriculados los estudiantes de una Universidad, etc. En este caso sólo podemos hacer representaciones gráficas. Su objetivo es dar una idea visual sencilla de la muestra obtenida. Naturalmente, hay una gran variedad de representaciones gráficas: diagramas de barras, diagramas de sectores,... Todas ellas son muy sencillas de comprender y de interpretar.

3 Variables cuantitativas

Una variable respuesta X es cuantitativa cuando toma valores numéricos. Son las más interesantes ya que con ellas podemos obtener resúmenes numéricos que no tenían sentido para las variables cualitativas. Es muy habitual distinguir dos tipos de variables cuantitativas que indicamos a continuación:

Discretas: Sólo pueden tomar un conjunto finito o numerable de valores (generalmente valores enteros).
Continuas: Pueden tomar cualquier valor en un intervalo (finito o infinito).

Sin embargo, es conveniente resaltar que para la mayoría de las cosas que vamos a hacer es irrelevante si la variable es discreta o continua. Utilizaremos la siguiente notación, tanto para variables discretas como para variables continuas:

n: Tamaño de la muestra = Número de elementos observados.

$x_1, \ldots, x_n$: Representan los n valores de la variable respuesta obtenidos en la muestra (puede haber repeticiones).

A veces, al estudiar variables continuas, no disponemos de los datos originales sino que nos dan los datos agrupados en una serie de intervalos o clases $A_1, \ldots, A_k$. En este caso, la notación sería:

n: Tamaño de la muestra = Número de elementos observados.

$x_1, \ldots, x_k$: Representantes de las clases $A_1, \ldots, A_k$ (generalmente, los puntos medios de los intervalos).

$n_1, \ldots, n_k$: Número de observaciones dentro de cada clase (frecuencias absolutas).

$f_1, \ldots, f_k$: Frecuencias relativas dentro de cada clase ($f_i = n_i/n$).

Por supuesto, si se puede, es preferible utilizar los datos originales a usar los datos agrupados en unas clases artificiales. Intuitivamente, los datos originales contienen más información que los datos agrupados.

4 Resúmenes numéricos

Definición.- La **media muestral** es una medida de centralización que se define como:

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

Cuando se trate de una variable continua con los datos agrupados, usaremos:

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

Es decir, es como si el valor x_i hubiera aparecido n_i veces. Pero insistimos en que si los datos están sin agrupar, no tiene mucho sentido agruparlos. •

Definición.- La **mediana muestral** es otra medida de centralización cuya idea es la siguiente:

La mediana, M, es el valor de la muestra que deja el 50% de los datos por debajo (son menores) y el 50 % de los datos por encima (son mayores).

Por tanto, para hallar la mediana de una muestra ordenamos las observaciones de menor a mayor y tenemos dos posibilidades:

Si el número de observaciones es impar, la mediana es el valor central.

Si el número de observaciones es par, la mediana es el punto medio de los dos valores centrales.

Si se trata de una variable continua con los datos agrupados, se puede hallar el intervalo mediana, es decir, la clase en la que se encuentra la mediana. Después, se puede hacer una interpolación, con el objetivo de hallar el valor aproximado de la mediana. •

La idea de la mediana se puede extender a los cuartiles:

Definición.- El **primer cuartil**, Q_1, es el valor de la muestra que deja el 25% de los datos por debajo (son menores) y el 75% de los datos por encima (son mayores).

El **tercer cuartil**, Q_3, es el valor de la muestra que deja el 75% de los datos por debajo (son menores) y el 25% de los datos por encima (son mayores).

El método para hallar Q_1 y Q_3 es análogo al empleado para hallar la mediana. •

Definición.- La **moda muestral** de una variable discreta es una medida de centralización que se define como el valor que aparece más repetido en la muestra.•

La moda es menos interesante como medida de centralización por varias razones: no tiene sentido para variables continuas (habría que agrupar), puede no ser un valor central, puede haber una moda en cada extremo, etc.

Definición.- La **varianza muestral** es una medida de dispersión que se define como:

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

Para calcular la varianza suele ser más cómodo usar la siguiente expresión que obtenemos desarrollando el cuadrado:

$$\begin{aligned} v_x &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \end{aligned}$$

Si se trata de una variable continua con los datos agrupados, usaremos la expresión:

$$v_x = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 = ... = \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i^2 - n\bar{x}^2\right) \qquad \bullet$$

Observación: La definición que se ha dado de varianza muestral corresponde a la idea natural de medir la *dispersión cuadrática media* y, por este motivo, dividimos por n (número total de datos). Es muy frecuente encontrar textos y paquetes informáticos en los que, en la definición de varianza muestral, se divide por $n-1$, en vez de por n. Esto tiene su justificación en la Inferencia Estadística (cuando se buscan estimadores insesgados), pero no en la Estadística Descriptiva. Por supuesto, si n es grande, la diferencia entre dividir por n ó por $n-1$ es muy pequeña.

Definición.- La **desviación típica (o desviación standard) muestral** es una medida de dispersión que se define como la raíz cuadrada positiva de la varianza muestral. •

Con la desviación típica medimos la dispersión en las unidades originales, ya que la varianza nos da la media de los cuadrados de las desviaciones a la media muestral.

5 Representaciones gráficas

Se pueden hacer distintas representaciones gráficas con los datos de una variable cuantitativa X. También son sencillas de comprender, aunque requieren algo más de explicación que las representaciones gráfica de variables cualitativas. Veremos algunas de las más interesantes, comenzando por los diagramas de tallos y hojas:

Definición.- El procedimiento para construir un **diagrama de tallos y hojas** es como sigue:

1. Redondear los datos a un número conveniente de cifras significativas, de modo que el perfil que obtengamos sea informativo.
2. Colocarlos en una tabla con dos columnas separadas por una línea, de la siguiente forma:
   (a) Todas las cifras menos la última se escriben a la izquierda de la línea (forman el tallo).
   (b) La última cifra se escribe a la derecha (forma la hoja).
3. Cada tallo define una clase y se escribe sólo una vez. El número de hojas representa la frecuencia de dicha clase. •

Otra representación sencilla muy utilizada es el diagrama de caja y bigotes (box-plot):

Definición.- En primer lugar, obtenemos la mediana, M, el primer cuartil, Q_1, el tercer cuartil, Q_3, y los valores mínimo y máximo de las observaciones. La versíón mas sencilla de **diagrama de cajas y bigotes** consiste en dos cosas:

a) Un rectángulo vertical (caja) que comienza en Q_1, termina en Q_3, y tiene una línea central en M.

b) Dos líneas (bigotes) que parten de Q_1 y Q_3 y llegan, respectivamente, al mínimo y al máximo. •

Este diagrama nos da una idea rápida de la concentración y de la simetría de los datos.

Otra representación interesante para variables cuantitativas continuas con los datos agrupados es el histograma:

Definición.- Disponemos de los n datos agrupados en k intervalos, cada uno con una anchura a_i, $i = 1, ..., k$. El **histograma** consiste en construir sobre cada intervalo un rectángulo cuya área represente la frecuencia (absoluta o relativa) de dicho intervalo. De ese modo, si pensamos por ejemplo en frecuencias absolutas, la altura, h_i, de cada rectángulo sería:

$$\text{Área} = n_i = a_i\ h_i \qquad \Rightarrow \qquad h_i = \frac{n_i}{a_i} \quad \bullet$$

6 Ejemplo

En 1778, H. Cavendish realizó una serie de 29 experimentos con objeto de medir la densidad de la tierra. Sus resultados, tomando como unidad la densidad del agua, fueron:

5'50	5'61	4'88	5'07	5'26	5'55	5'36	5'29	5'58	5'65
5'57	5'53	5'62	5'29	5'44	5'34	5'79	5'10	5'27	5'39
5'42	5'47	5'63	5'34	5'46	5'30	5'75	5'68	5'85	

Queremos analizar estos datos descriptivamente.

En primer lugar, podemos representar los datos sobre la variable estadística X="Densidad de la tierra" con un diagrama de tallos y hojas:

48	8
49	
50	7
51	0
52	6997
53	64940
54	4276
55	05873
56	15238
57	95
58	5

Ordenando los datos de menor a mayor, la mediana sería el dato que ocupa el puesto decimoquinto: $M = 5,46$.

Procediendo de manera análoga, el primer cuartil, Q_1, sería el punto medio de los datos que ocupan los puestos séptimo y octavo:

$$Q_1 = (5,29 + 5,3)/2 = 5,295$$

Análogamente, el tercer cuartil, Q_3, sería el punto medio de los datos que ocupan los puestos 22 y 23:

$$Q_3 = (5,61 + 5,62)/2 = 5,615$$

Podemos calcular también la media y la desviación típica:

$$\bar{x} = \frac{1}{n}\sum x_i = 5,448; \qquad \text{Desviación típica} = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2} \simeq 0,22$$

Podemos observar que la media y la mediana son muy similares; esto es consecuencia de la simetría que se puede apreciar en el diagrama de tallos y hojas.

Podemos abordar también el análisis descriptivo que haríamos en el caso de que nos hubieran dado los datos agrupados en una serie de clases. Por ejemplo, supongamos que la información que tenemos viene resumida en la siguiente tabla:

Clases	x_i	n_i
$\leq 5,3$	5,2	8
(5,3;5,4]	5,35	4
(5,4;5,5]	5,45	5
(5,5;5,6]	5,55	4
$> 5,6$	5,7	8

Con esta información agrupada, tendríamos:

$$\bar{x} = \frac{1}{n}\sum n_i x_i = 5,45;$$

$$\text{Desviación típica} = \sqrt{\frac{1}{n}\sum n_i (x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\left(\sum n_i x_i^2 - n\bar{x}^2\right)} = 0,1930.$$

Lógicamente, existen pequeñas diferencias con respecto a lo que se obtuvo con los datos sin agrupar.