ESTADÍSTICA

Jesús García de Jalón de la Fuente

1. Introducción

La Estadística trata de describir colectividades formadas por un gran número de objetos. El conjunto de los objetos que se estudian se denomina población. En ocasiones, el estudio se hace a partir de un cierto número de objetos tomados de la población (muestra).

Sobre la población o sobre una muestra se mide una magnitud. Los valores que toma esta magnitud forman la variable estadística. Si la variable estadística toma valores numéricos se dice que es cuantitativa. Si no es así (por ejemplo si se estudia la raza de una población de gatos) la variable es cualitativa.

Una variable estadística cuantitativa puede tomar un número finito de valores o los infinitos valores comprendidos en un cierto intervalo. En el primer caso hablaremos de variable estadística discreta y en el segundo de variable continua. En realidad la variable nunca es estrictamente continua en el sentido explicado pues la precisión de los instrumentos de medida no permite apreciar infinitos valores. En la práctica, la variable será continua cuando pueda tomar un número muy elevado de valores; en este caso, los valores de la variable esstadística se agrupan en intervalos.

2. Frecuencias

La frecuencia o frecuencia absoluta de un valor x de la variable estadística es el número de objetos de la población que presentan ese valor. Representaremos esta frecuencia por f. La frecuencia de un determinado valor mdividido por el número de elementos de la población, esto es, la proporción de elementos de la población que presenta este valor es la frecuencia relativa que representaremos por h. Evidentemente se cumple que:

$$h = \frac{f}{N}$$

donde N es el número de objetos de la población.

La frecuencia acumulada F de un resultado x es el número de elementos de la población en los que la variable toma valores menores o iguales que x. Dividiendo por el número de elementos de la población se obtiene la frecuencia acumulada relativa H.

De las definiciones se deducen algunas condiciones que deben cumplir estos valores: la suma de todas las frecuencias debe ser igual al número de objetos de la población y la última frecuencia acumulada también debe ser igual a este número. La suma de las frecuencias relativas debe ser 1 y también la última frecuencia acumulada relativa. También debe cumplirse que, por ejemplo:

$$F_4 = f_1 + f_2 + f_3 + f_4$$

es decir a la suma de las frecuencias absolutas anteriores. O también:

$$F_4 = f_4 + F_3$$

o sea, la frecuencia correspondiente más la frecuencia acumulada anterior. Relaciones similares deben cumplirse para las frecuencias relativas.

Los valores de la variable estadística y las correspondientes frecuencias se representan en las llamadas tablas de frecuencias, que tienen la forma (se presentan dos tablas, una para variable discreta y otra para variable continua):

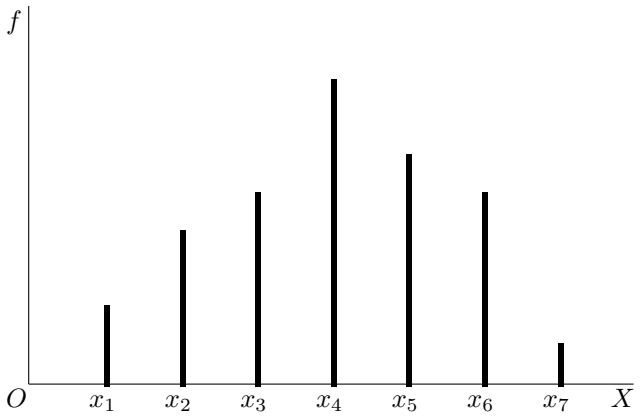
x	f	h	F	H
x_1	f_1	h_1	F_1	H_1
x_2	f_2	h_2	F_2	H_2
x_3	f_3	h_3	F_3	H_3
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_n	f_n	h_n	F_n	H_n

x	f	h	F	H
$[x_0, x_1)$	f_1	h_1	F_1	H_1
$[x_1, x_2)$	f_2	h_2	F_2	H_2
$[x_2, x_3)$	f_3	h_3	F_3	H_3
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$[x_{n-1}, x_n)$	f_n	h_n	F_n	H_n

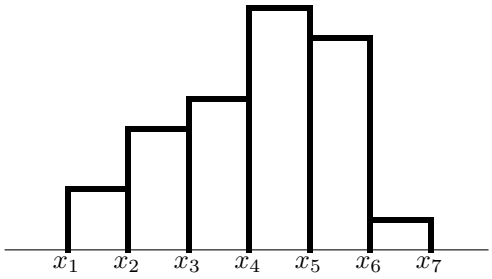
3. Gráficos

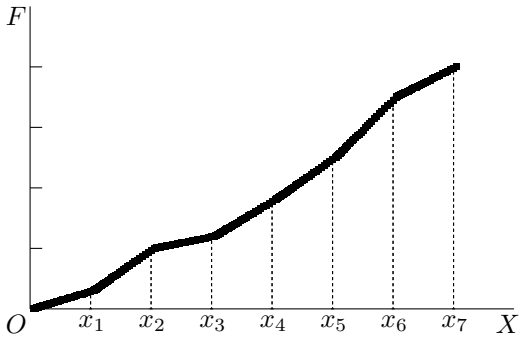
Los valores de la variable estadística y sus frecuencias pueden representarse gráficamente de muchas maneras. Consideraremos solamente los más comunes.

Para variable discreta se utilizan los diagramas de barras. Los valores de la variable se indican sobre el eje de abscisas y sobre ellos se dibuja una barra de altura proporcional a la frecuencia. Pueden representarse de esta forma tanto las frecuencias absolutas como las frecuencias relativas o las frecuencias acumuladas:



Si la variable estadística es continua se utilizan los **histogramas** y los **polígonos de frecuencias acumuladas**. Un histograma consiste en representar los intervalos en que hemos dividido la variable sobre el eje de abscisas y, sobre él, se dibuja un rectángulo de área proporcional a la frecuencia correspondiente:





Como en el caso anterior pueden construirse histogramas de frecuencias absolutas o relativas. Si los intervalos tienen la misma longitud, la altura de los rectángulos resulta proporcional a la frecuencia. Si no es así, como lo que debe resultar proporcional a la frecuencia es el área de los rectángulos, la altura es proporcional a la **densidad de frecuencia**, esto es, a la frecuencia dividida por la longitud del intervalo.

Los polígonos de frecuencias acumuladas (absolutas o relativas) se obtiene tomando como ordenada soble el extremo derecho del intervalo la frecuencia acumulada correspondiente y uniendo los puntos así obtenidos mediante segmentos:

4. Media y desviación típica

la media o media aritmética de una variable estadística se define como la suma de todos los valores de la variable dividido por el número de elementos de la población:

$$\overline{x} = \frac{\Sigma x}{N}$$

La suma de todos los valores de la variable estadística se puede expresar mediante la suma de cada uno de los valores que toma por sus correspondientes frecuencias. Así:

$$\overline{x} = \frac{\Sigma f_i x_i}{N} = \Sigma h_i x_i$$

donde se ha hecho uso de la relación $\frac{f_i}{N} = h_i$. En caso de que los datos aparezcan agrupados en intervalos, tomaremos como valor de la variable el punto medio del intervalo

La media es, como hemos visto, un número que cumple que $\Sigma x = N\overline{x}$, es decir, si todos los valores de la variable fuesen iguales a la media, su suma sería la misma. La media nos permite comparar dos poblaciones sobre las que se ha medido la misma magnitud pero no nos permite saber si los valores de la variable están próximos a la media o no. Por ejemplo, una media de cinco se puede obtener con dos cincos o con un diez y un cero.

Para saber cómo están distribuidos los valores en torno a la media son precisos otros parámetros. Estos son la desviación media y sobre todo la varianza y la desviación típica.

La distancia de un valor de la variable x_i a la media es $|x_i - \overline{x}|$. La media de estas cantidades se llama **desviación media**:

$$DM = \frac{\Sigma f_i |x_i - \overline{x}|}{N} = \Sigma h_i |x_i - \overline{x}|$$

Más que la desviación media se utilizan la **varianza**, definida por:

$$\sigma^2 = \frac{\Sigma f_i (x_i - \overline{x})^2}{N} = \Sigma h_i (x_i - \overline{x})^2$$

y su raíz cuadrada o **desviación típica**:

$$\sigma = \sqrt{\frac{\Sigma f_i (x_i - \overline{x})^2}{N}} = \sqrt{\Sigma h_i (x_i - \overline{x})^2}$$

La media y la desviación típica tienen las siguientes propiedades:

- Si se suma el mismo número a todos los valores de la variable, la media queda incrementada en esa cantidad pero la desviación típica no varía.
- Si todos los valores de la variable se multiplican por el mismo número, la media y la desviación típica quedan multiplidados por ese número. La multiplicación de todos los valores por un número puede interpretarse como un cambio de unidades. Esta propiedad dice que la media y la desviación típica se expresan en las nuevas unidades.

El cociente de la desviación típica y la media se llama **coeficiente de variación**:

$$CV = \frac{\sigma}{\overline{x}}$$

Para comparar un valor de la variable estadística con el resto de los valores obtenidos en una determinada población se utilizan las **puntuaciones típicas**. En estas se toma como valor cero el de la media y como unidad la desviación típica. El paso de la variable x al valor típico z se hace mediante la fórmula:

$$z = \frac{x - \overline{x}}{\sigma}$$

o, despejando $x = \overline{x} + z\sigma$.

5. Mediana, cuartiles y percentiles

Supongamos que todos los valores obtenidos de la variable estadística se ordenan de menor a mayor. La **mediana** será entonces el valor central, esto es, el valor que deja el mismo número de términos a su izquierda y a su derecha. Si el número de términos es par entonces se tomará como mediana la media de los valores centrales.

La mediana se puede obtener fácilmente a partir de la tabla de frecuencias relativas acumuladas. Si en la tabla aparece la frecuencia acumulada $0,50$ (o sea el $50\,\%$) entonces la mediana es la media entre el valor de la variable correspondiente a ese $0,50$ y el siguiente. Si no aparece en la tabla el valor $0,50$, entonces es el valor de la variable correspondiente al primer valor de la frecuencia acumulada relativa superior a $0,50$.

Si la variable es continua, esto es, si aparece dividida en intervalos, se puede localizar el intervalo mediano tal como se ha expuesto en el párrafo anterior. Una vez conocido este intervalo se tomará como mediana el valor de la variable correspondiente al $50\,\%$ en el polígono de frecuencias acumuladas relativas.

Si el intervalo mediano es (x_1, x_2) y a los extremos del intervalo les corresponden unas frecuencias acumuladas relativas H_1 y H_2, el valor de la mediana está dado por:

$$Me = x_1 + \frac{0,50 - H_1}{H_2 - H_1}(x_2 - x_1)$$

De forma similar, se llaman primero, segundo y tercer cuartil, los valores de la variable correspondientes a frecuencias acumuladas de $0,25$, $0,50$ y $0,75$, es decir, aquellos que dividen al conjunto de valores obtenidos en cuatro partes con el mismo número de términos. Se representan por Q_1, Q_2 y Q_3. El segundo cuartil coincide con la mediana. Pueden obtenerse por fórmulas similares a la mediana:

$$Q_1 = x_1 + \frac{0,25 - H_1}{H_2 - H_1}(x_2 - x_1) \quad Q_2 = x_1 + \frac{0,50 - H_1}{H_2 - H_1}(x_2 - x_1) \quad Q_3 = x_1 + \frac{0,75 - H_1}{H_2 - H_1}(x_2 - x_1)$$

donde (x_1, x_2) representa el intervalo en que se encuentra el cuartil y F_1 y F_2 las frecuencias relativas acumuladas en los extremos del intervalo.

Los percentiles P_1, P_2, ..., P_{99}, dividen todos los valores de la variable estadística en 100 partes con el mismo número de términos. Conocido el intervalo (x_1, x_2) en que se encuentra el percentil P_c y los valores de las frecuencias relativas acumuladas en los extremos del intervalo, se calcula su valor por:

$$P_c = x_1 + \frac{c - H_1}{H_2 - H_1}(x_2 - x_1)$$