Estadística Descriptiva

Introducción

¿Qué es la Estadística?

Cuando coloquialmente se habla de Estadística, se suele pensar en una relación de datos numéricos presentada de forma ordenada y sistemática. Esta idea es la consecuencia del concepto popular que existe sobre el término y que cada vez está más extendido debido a la influencia de nuestro entorno, ya que hoy día es casi imposible que cualquier medio de difusión, periódico, radio o televisión, no nos aborde diariamente con cualquier tipo de información sobre accidentes de tráfico, índices de crecimiento de población, turismo, tendencias políticas, etc. Sólo cuando nos adentramos en un mundo más específico como es el campo de la investigación de las Ciencias Sociales, Medicina, Biología, Psicología,... empezamos a percibir que la Estadística no sólo es algo más, sino que se convierte en la única herramienta que, hoy en día, permite dar luz y obtener resultados, y por tanto beneficios, en cualquier tipo de estudio, cuyos movimientos y relaciones, por su variabilidad intrínseca, no puedan ser abordadas desde la perspectiva de las leyes deterministas. Podríamos, desde un punto de vista más amplio, definir la Estadística como *la ciencia que estudia cómo debe emplearse la información y cómo dar una guía de acción en situaciones prácticas que entrañan incertidumbre.*

La Estadística se ocupa de los métodos y procedimientos para recoger, clasificar, resumir, hallar regularidades y analizar los datos, siempre y cuando la variabilidad e incertidumbre sea una causa intrínseca de los mismos; así como de realizar inferencias a partir de ellos, con la finalidad de ayudar a la toma de decisiones y en su caso formular predicciones.

Podríamos por tanto clasificar la Estadística en Descriptiva, cuando los resultados del análisis no pretenden ir más allá del conjunto de datos, e Inferencial cuando el objetivo del estudio es derivar las conclusiones obtenidas a un conjunto de datos más amplio.

Estadística Descriptiva: Describe, analiza y representa un grupo de datos utilizando métodos numéricos y gráficos que resumen y presentan la información contenida en ellos.

Estadística Inferencial: Apoyándose en el cálculo de probabilidades y a partir de datos muestrales, efectúa estimaciones, decisiones, predicciones u otras generalizaciones sobre un conjunto mayor de datos.

Definiciones Básicas

Se establecen a continuación algunas definiciones de conceptos básicos como son: elemento, población, muestra, caracteres, variables, etc., a las cuales se hace referencia continuamente a lo largo del curso.

Elementos. Población. Caracteres

Individuos o elementos: personas u objetos que contienen cierta información que se desea estudiar.

Población: conjunto de individuos o elementos que cumplen ciertas propiedades comunes.

Muestra: subconjunto representativo de una población.

Parámetro: función definida sobre los valores numéricos de características medibles de una población.

Estadístico: función definida sobre los valores numéricos de una muestra.

Con relación al tamaño de la población, ésta puede ser:

Finita, como es el caso, por ejemplo, del número de personas que se conectan a un servidor de Internet en un día;

Infinita, si, por ejemplo, se estudia el mecanismo aleatorio que describe la secuencia de caras y cruces obtenida en el lanzamiento repetido de una moneda al aire.

Caracteres: propiedades, rasgos o cualidades de los elementos de la población. Estos caracteres se pueden dividir en cualitativos y cuantitativos.

Modalidades: diferentes situaciones posibles de un carácter. Las modalidades deben ser a la vez exhaustivas y mutuamente excluyentes: cada elemento posee una y sólo una de las modalidades posibles.

Clases: conjunto de una o más modalidades en el que se verifica que cada modalidad pertenece a una y sólo una de las clases.

Ejemplo

Consideramos la población formada por todos los estudiantes de la Universidad Carlos III (finita). La altura media de todos los estudiantes es el parámetro μ. El conjunto formado por los alumnos del *Grado en Ecología* es una muestra de dicha población y la altura media de esta muestra, $\overline{x}$, es un estadístico.

Organización de los datos

Variables estadísticas

Cuando hablemos de variable haremos referencia a un símbolo $(X, Y, A, B, \ldots)$ que puede tomar cualquier modalidad (valor) de un conjunto determinado, que llamaremos *dominio* de la variable o *rango*. En función del tipo de dominio, las variables las clasificamos del siguiente modo:

Variables cualitativas cuando las modalidades posibles son de tipo nominal. Por ejemplo, una variable de color $A \in \{$"rojo", "azul", "verde"$\}$

Variables cuantitativas ordinales son las que, aunque sus modalidades son de tipo nominal, es posible establecer un orden entre ellas. Por ejemplo, si estudiamos la llegada a la meta de un corredor en una competición de 20 participantes, su clasificación C es tal que $C \in \{1^{\text{o}}, 2^{\text{o}}, 3^{\text{o}}, \ldots, 20^{\text{o}}\}$.

Otro ejemplo de variable cuantitativa ordinal es el nivel de dolor, D, que sufre un paciente ante un tratamiento médico: $D \in \{$" inexistente", "poco intenso", "moderado", "fuerte"$\}$.

Variables cuantitativas son las que tienen por modalidades cantidades numéricas con las que podemos hacer operaciones aritméticas. Dentro de este tipo de variables podemos distinguir dos grupos:

Discretas, cuando no admiten siempre una modalidad intermedia entre dos cualesquiera de sus modalidades. Un ejemplo es el número de caras X, obtenido en el lanzamiento repetido de una moneda. Es obvio que cada valor de la variable es un número natural $X \in \mathbb{N}$.

Continuas, cuando admiten una modalidad intermedia entre dos cualesquiera de sus modalidades, por ejemplo, el peso X de un niño al nacer. En este caso, los valores de las variables son números reales, es decir, $X \in \mathbb{R}$.

Ocurre a veces que una variable cuantitativa continua por naturaleza, aparece como discreta. Este es el caso en que hay limitaciones en lo que concierne a la precisión del aparato de medida

de esa variable, por ejemplo, si medimos la altura en metros de personas con una regla que ofrece dos decimales de precisión, podemos obtener $C \in \{\ldots, 1.50, 1.51, 1.52, 1.53, \ldots\}$. En realidad lo que ocurre es que con cada una de esas mediciones expresamos que el verdadero valor de la misma se encuentra en un intervalo de radio 0,005.

Por tanto cada una de las observaciones de X representa más bien un intervalo que un valor concreto.

Tal como hemos citado anteriormente, las modalidades son las diferentes situaciones posibles que puede presentar la variable. A veces, éstas son muy numerosas (por ejemplo, cuando una variable es continua) y conviene reducir su número, agrupándolas en una cantidad inferior de clases. Estas clases deben ser construidas de modo que sean exhaustivas e incompatibles, es decir, cada modalidad debe pertenecer a una y sólo una de las clases.

Tablas Estadísticas

Consideremos una población estadística de n individuos, descrita según un carácter o variable C cuyas modalidades han sido agrupadas en un número k de clases, que denotamos mediante $c_1, c_2, \ldots, c_k$. Para cada una de las clases c_i, $i = 1, \ldots, k$, se pueden considerar las siguientes magnitudes:

Frecuencia absoluta de la clase c_i es el número, n_i, de observaciones que presentan una modalidad perteneciente a esa clase.

Frecuencia relativa de la clase c_i es el cociente, f_i, entre las frecuencias absolutas de dicha clase y el número total de observaciones, es decir,

$$f_i = \frac{n_i}{n}$$

Obsérvese que f_i es el tanto por uno de observaciones que están en clase c_i. Multiplicado por 100 representa el porcentaje en % de la población que comprende esa clase.

Frecuencia absoluta acumulada N_i, se calcula sobre variables cuantitativas o cuantitativas ordinales, y es el número de elementos de la población cuya modalidad es inferior o equivalente a la modalidad c_i:

$$N_i = n_1 + n_2 + \ldots + n_i = \sum_{j=1}^{i} n_j$$

Frecuencia relativa acumulada, F_i, se calcula sobre variables cuantitativas o cuantitativas ordinales, siendo el tanto por uno de los elementos de la población que están en alguna de las clases y que presentan una modalidad inferior o igual a la c_i, es decir,

$$F_i = \frac{N_i}{n} = \frac{n_1 + n_2 + \ldots + n_i}{n} = f_1 + f_2 + \ldots + f_i = \sum_{j=1}^{i} f_j,$$

como todas las modalidades son exhaustivas e incompatibles ha de ocurrir que

$$\sum_{j=1}^{k} n_j = n_1 + n_2 + \ldots + n_k = n,$$

o lo que es lo mismo,

$$\sum_{j=1}^{k} f_j = \sum_{j=1}^{k} \frac{n_j}{n} = \frac{\sum_{j=1}^{k} n_j}{n} = \frac{n}{n} = 1.$$

Llamaremos **distribución de frecuencias** al conjunto de clases junto a las frecuencias correspondientes a cada una de ellas. Una **tabla estadística** sirve para presentar de forma ordenada las distribuciones de frecuencias. Su forma general es la siguiente:

Modalidades	**Frec. Absolutas**	**Frec. Relativas**
C	n_i	f_i
c_1	n_1	$f_1 = \frac{n_1}{n}$
$\ldots$	$\ldots$	$\ldots$
c_j	n_j	$f_j = \frac{n_j}{n}$
$\ldots$	$\ldots$	$\ldots$
c_k	n_k	$f_k = \frac{n_k}{n}$
	n	1

Modalidades	**Frec. Abs. Acum.**	**Frec. Rel. Acum**
C	N_i	F_i
c_1	$N_1 = n_1$	$F_1 = \frac{N_1}{n} = f_1$
$\ldots$	$\ldots$	$\ldots$
c_j	$N_j = n_1 + \ldots + n_j$	$F_j = \frac{N_j}{n} = f_j$
$\ldots$	$\ldots$	$\ldots$
c_k	$N_k = n$	$F_k = 1$

Ejemplo

Calcular los datos que faltan en la siguiente tabla:

l_{i-1} — li	n_i	f_i	N_i
0 – 10	60	f_1	60
10 – 20	n_2	0,4	N_2
20 – 30	30	f_3	170
30 – 100	n_4	0,1	N_4
100 – 200	n_5	f_5	200
	n		

Solución:

Sabemos que la última frecuencia acumulada es igual al total de observaciones, luego $n = 200$.

Como $N_3 = 170$ y $n_3 = 30$, entonces $N_2 = N_3 - n_3 = 170 - 30 = 140$.

Además al ser $n_1 = 60$, tenemos que $n_2 = N_2 - n_1 = 140 - 60 = 80$.

Por otro lado podemos calcular n_4 teniendo en cuenta que conocemos la frecuencia relativa correspondiente:

$$f_4 = \frac{n_4}{n} \Longrightarrow n_4 = f_4 \cdot n = 0{,}1 \cdot 200 = 20.$$

Así:

$N_4 = n_4 + N_3 = 20 + 170 = 190$.

Este último cálculo nos permite obtener $n_5 = N_5 - N_4 = 200 - 190 = 10$.

Al haber calculado todas las frecuencias absolutas, es inmediato obtener las relativas:

$$\begin{aligned} f_1 &= \frac{n_1}{n} = \frac{60}{200} = 0{,}3 \\ f_3 &= \frac{n_3}{n} = \frac{30}{200} = 0{,}15 \\ f_5 &= \frac{n_5}{n} = \frac{10}{200} = 0{,}05 \end{aligned}$$

Escribimos entonces la tabla completa:

l_{i-1} — l_i	n_i	f_i	N_i
0 – 10	60	0,3	60
10 – 20	80	0,4	140
20 – 30	30	0,15	170
30 – 100	20	0,1	190
100 – 200	10	0,05	200
	200		

Elección de las clases

En cuanto a la elección de las clases, deben seguirse los siguientes criterios en función del tipo de variable que estudiemos:

Cuando se trate de variables cualitativas o cuantitativas ordinales, las clases c_i serán de tipo nominal. En el caso de variables cuantitativas, existen dos posibilidades.

Si la variable es discreta, las clases serán valores numéricos $x_1, \ldots, x_k$.

Si la variable es continua las clases vendrán definidas mediante lo que se denomina **intervalos**. En este caso, las modalidades que contiene una clase son todos los valores numéricos posibles contenidos en el intervalo, el cual viene normalmente definido de la forma

$[l_{i-1}, l_i) = \{x : l_{i-1} \leq x < l_i\}$ o bien $(l_{i-1}, l_i] = \{x : l_{i-1} < x \leq l_i\}$.

En estos casos llamaremos **amplitud del intervalo** a las cantidades $a_i = l_i - l_{i-1}$ y **marca de clase** c_i, a un punto representativo del intervalo. Si éste es acotado, tomamos como marca de clase al punto más representativo, es decir, el punto medio del intervalo, $c_i = \frac{l_i + l_{i-1}}{2}$. La marca de clase no es más que una forma abreviada de representar un intervalo mediante uno de sus puntos. Por ello hemos tomado como representante al punto medio del mismo. Esto está plenamente justificado si recordamos que cuando se mide una variable continua como el peso, la cantidad con cierto número de decimales que expresa esta medición, no es el valor exacto de la variable, sino una medida que contiene cierto margen de error, y por tanto representa a todo un intervalo del cual ella es el centro.

En el caso de variables continuas, la forma de la tabla estadística es la siguiente:

	M. clase	**Frec. Abs.**	**Frec. Rel.**	**F. Abs. Ac.**	**F. Rel. Ac.**
	C	n_i	f_i	N_i	F_i
$l_0 - l_1$	c_1	n_1	$f_1 = n_1/n$	$N_1 = n_1$	$F_1 = f_1$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$l_{j-1} - l_j$	c_j	n_j	$f_j = n_j/n$	$N_j = N_{j-1} + n_j$	$F_j = F_{j-1} + f_j$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$l_{k-1} - l_k$	c_k	n_k	$f_k = n_k/n$	$N_k = n$	$F_k = 1$
		n	1		

Elección de intervalos para variables continuas

A la hora de seleccionar los intervalos para las variables continuas se plantean varios problemas, como son el número de intervalos a elegir y sus tamaños respectivos. La notación más común que usaremos para un intervalo será $l_{j-1} - l_j \stackrel{def}{\equiv} (l_{j-1}, l_j]$

El primer intervalo, $l_0 - l_1$, podemos a cerrarlo en el extremo inferior para no excluir la observación más pequeña, $l_0 : l_0 - l_1 \stackrel{def}{\equiv} [l_0, l_1]$.

Éste es un convenio que tomaremos en las páginas que siguen. El considerar los intervalos por el lado izquierdo y abrirlos por el derecho no cambia de modo significativo nada de lo que expondremos. El número de intervalos, k, a utilizar no está determinado de forma fija y por tanto tomaremos un k que nos permita trabajar cómodamente y ver bien la estructura de los datos. Como referencia nosotros tomaremos una de los siguientes valores aproximados:

$$N^o \text{ intervalos } = k \approx \begin{cases} \sqrt{n} & \text{si } n \text{ no es muy grande} \\ 1 + 3{,}22 \log(n) & \text{en otro caso} \end{cases}$$

Por ejemplo, si el número de observaciones que tenemos es $n = 100$, un buen criterio es agrupar las observaciones en $k = \sqrt{100} = 10$ intervalos. Sin embargo si tenemos $n = 1{,}000{,}000$, será más razonable elegir $k = 1 + 3{,}22 \log n \approx 20$ intervalos, que $k = \sqrt{1000000} = 1000$.

La amplitud de cada intervalo $a_i = l_i - l_{i-1}$ se suele tomar constante, considerando la observación más pequeña y y más grande de la población (respectivamente $l_0 = x_{\text{mín}}$ y $l_k = x_{\text{máx}}$) para calcular la amplitud total, A, de la población $A = l_k - l_0$ de forma que la amplitud de cada intervalo sea: $a_i = a \ \forall i = 1, \ldots, k$ donde $a = A/k$. Así la división en intervalos podría hacerse tomando:

$l_0 = x_{min}$

$l_1 = l_0 + a$

.....

$l_k = x_{max} = l_0 + ka$

Observación:

Podría ocurrir que la cantidad a fuese un número poco cómodo a la hora de escribir los intervalos (ej. $a = 10{,}325467$). En este caso, es recomendable variar simétricamente los extremos, $l_0 < x_{\text{mín}} < x_{\text{máx}} < l_k$, de forma que se tenga que a es un número más simple (ej. $a = 10$).

Ejemplo

Sobre un grupo de $n = 21$ personas se realizan las siguientes observaciones de sus pesos, medidos en kilogramos:

$X \sim x_1, x_2, \ldots, x_{21}$						
58	42	51	54	40	39	49
56	58	57	59	63	58	66
70	72	71	69	70	68	64

Agrupar los datos en una tabla estadística.

Solución:

En primer lugar hay que observar que si denominamos X a la variable "peso de cada persona" ésta es una variable de tipo cuantitativa y continua. Por tanto a la hora ordenar los resultados en una tabla estadística, esto se ha de hacer agrupándolos en intervalos de longitud conveniente. Esto nos lleva a perder cierto grado de precisión. Para que la pérdida de información no sea muy relevante seguimos el criterio de utilizar $k = \sqrt{21}$ intervalos (no son demasiadas las observaciones). En este punto podemos tomar bien $k = 4$ o bien $k = 5$. Arbitrariamente se elige una de estas dos posibilidades. Por ejemplo, vamos a tomar $k = 5$.

Lo siguiente es determinar la longitud de cada intervalo, a_i $\forall i = 1, \ldots, 5$. Lo más cómodo es tomar la misma longitud en todos los intervalos, $a_i = a$ (aunque esto no tiene por qué ser necesariamente así), donde

$$\begin{aligned} l_0 &= x_{\text{mín}} = 39 \\ l_5 &= x_{\text{máx}} = 72 \\ A &= l_5 - l_0 = 72 - 39 = 33 \\ a &= \frac{A}{5} = \frac{33}{5} = 6{,}6 \end{aligned}$$

Entonces, tomaremos $k = 5$ intervalos de longitud $a = 6{,}6$ comenzando por $l_0 = x_{\text{mín}} = 39$ y terminando en $l_5 = 72$:

	$l_{i-1} - l_i$	c_i	n_i	f_i	N_i	F_i
$i = 1$	39 – 45,6	42,3	3	0,1428	3	0,1428
$i = 2$	45,6 – 52,2	48,9	2	0,0952	5	0,2381
$i = 3$	52,2 – 58,8	55,5	6	0,2857	11	0,5238
$i = 4$	58,8 – 65,4	62,1	3	0,1428	14	0,6667
$i = 5$	65,4 – 72	68,7	7	0,3333	21	1
			21	1		

Otra posibilidad a la hora de construir la tabla, y que nos permite que trabajemos con cantidades más simples a la hora de construir los intervalos, es la siguiente. Como la regla para

elegir l_0 y l_5 no es muy estricta podemos hacer la siguiente elección:

$$\begin{aligned} a' &= 7 \\ A' &= a' \cdot 5 = 35 \\ d &= A' - A = 35 - 33 = 2 \\ l_0 &= x_{\text{mín}} - \frac{d}{2} = 39 - 1 = 38 \\ l_5 &= x_{\text{máx}} + \frac{d}{2} = 72 + 1 = 73 \end{aligned}$$

ya que así la tabla estadística no contiene decimales en la expresión de los intervalos, y el exceso d, cometido al ampliar el rango de las observaciones desde A hasta A', se reparte del mismo modo a los lados de las observaciones menores y mayores:

	Intervalos	**M. clase**	**f.a.**	**f.r.**	**f.a.a.**	**f.r.a.**
	$l_{i-1} - l_i$	c_i	n_i	f_i	N_i	F_i
$i = 1$	38 – 45	41,5	3	0,1428	3	0,1428
$i = 2$	45 – 52	48,5	2	0,0952	5	0,2381
$i = 3$	52 – 59	55,5	7	0,3333	12	0,5714
$i = 4$	59 – 66	62,5	3	0,1428	15	0,7143
$i = 5$	66 – 73	69,5	6	0,2857	21	1
			21	1		