Vocabulario básico de inferencia estadística. Estadística unidimensional

CURSO	TEMA	WWW.DANIPARTAL.NET
1ºBach	Estadística 01	Colegio Marista "La Inmaculada" de Granada

INFORMACIÓN GENERAL

Definición de conceptos: población, individuos, muestra, tamaño de la población, tamaño de la muestra, muestreo aleatorio y estratificado. Parámetros estadísticos unidimensionales.

Vídeo asociado:

https://youtu.be/aZRzTiDCYdQ

¿QUÉ ES INFERIR? CONCEPTO DE VARIABLE ESTADÍSTICA

Inferir supone obtener conclusiones generales a partir de datos particulares. Por ejemplo: estudiar la altura de un grupo de niños y niñas de 2ºESO de un colegio y obtener datos generales para toda la población de niños y niñas de ese curso en el país.

Los niños del colegio serían una **muestra**, mientras que todos los niños de España que cursan 2ºESO sería la **población**. Una muestra, por lo tanto, es un conjunto de datos individuales tomados dentro de una población mucho más amplia.

Supongamos que medimos, en una clase de 2ºESO de un colegio escogido al azar, la altura de los alumnos en centímetros. Es lógico pensar que habrá alumnos más altos, otros más bajos, alumnos que medirán igual, etc. El valor de la altura x es una **variable**, que en cada alumno tomará un valor concreto x_i.

Los alumnos escogidos para realizar las medidas se llama **muestra de la población** (representan una parte de toda la población). El número de alumnos de la muestra es el **tamaño de la muestra**. Igualmente, el número de alumnos de la población se denomina **tamaño de la población**.

Los factores que afectan a la variable altura son muchos: la genética, la alimentación, el clima, el estado de salud, etc. El valor de la altura x_i de cada alumno varía (de ahí el nombre de variable) dentro de la población. El conjunto de valores de la altura x_i de cada alumno de la muestra nos ofrece información estadística (en término medio) del valor de la altura en la población.

La Estadística nos ayuda a entender el comportamiento de una población muy numerosa a partir de los datos recogidos en una muestra de menor tamaño.

Si la variable x solo puede tomar valores enteros (por ejemplo, altura en centímetros), hablaremos de **variable discreta**. Y si la variable x puede tomar valores reales con decimales (por ejemplo, densidad de la sangre de una persona en gramos por centímetro cúbico), hablaremos de **variable continua**.

TIPOS DE MUESTREO

¿Cómo elegimos a los individuos de la muestra para que podamos realizar estimaciones correctas sobre la población?

Una opción es un **muestreo aleatorio simple**: todos los individuos de la población tienen las mismas posibilidades de ser elegido para la muestra. Tomaremos muestras de n elementos al azar con reemplazamiento.

Otra opción es el **muestreo estratificado proporcional**: dividir la población en estratos (subgrupos de tamaño N_1, N_2, N_3,...) y elegir muestras aleatorias simples dentro de cada estrato (de tamaño n_1, n_2, n_3,…) que son proporcionales al tamaño de cada estrato: $N_1/n_1 = N_2/n_2 = N_3/n_3 = \ldots$

PARÁMETROS ESTADÍSTICOS

Si trabajamos con una única variable estadística estaremos en el campo de la estadística unidimensional, donde se definen una serie de **parámetros estadísticos** que vamos a estudiar a partir de la siguiente tabla.

Altura en cm en un grupo de alumnos de 2ºESO
159
176
164
157
166
181
143
149
164
166
170

Tamaño de la muestra: número de medidas realizadas. Se representa por la letra n. En la tabla anterior, el tamaño es N=11.

Recorrido: Diferencia entre el mayor y el menor valor obtenido en los valores. En nuestro ejemplo, el rango es 181-143=38.

Media aritmética: Suma de todos los valores dividido por el tamaño de la muestra.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{N}$$

En nuestro ejemplo:

$$\bar{x} = \frac{1795}{11} = 163{,}18$$

Frecuencia absoluta del valor: Número de veces que aparece el valor x_i. Se representa como n_i. La suma de todas las frecuencias absolutas coincide con el tamaño de la muestra.

$$\sum_{i=1}^{n} n_i = N$$

Para mostrar la frecuencia absoluta suele crearse una tabla de dos columnas. En la primera columna se muestran los valores de la variable ordenados de menor a mayor, sin repetición. En la segunda columna se indican las frecuencias absolutas de cada valor.

Variable altura en cm	n_i
143	1
149	1
157	1
159	1
164	2
166	2
170	1
176	1
181	1
	$N=\sum_{i=1}^{n} n_i=11$

También podemos calcular la media con las frecuencias absolutas:

$$\overline{x}=\frac{\sum_{i=1}^{n} n_i \cdot x_i}{N}$$

Aplicado a nuestro ejemplo:

$$\overline{x}=\frac{1\cdot 143+1\cdot 149+1\cdot 157+1\cdot 159+2\cdot 164+2\cdot 166+1\cdot 170+1\cdot 176+1\cdot 181}{11}=163{,}18$$

La frecuencia absoluta n_i también recibe el nombre de peso (es decir, cómo de importante es el valor x_i según el valor de n_i).

Frecuencia relativa: Es el cociente de cada frecuencia absoluta entre el número de muestras tomadas.

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

El valor de la frecuencia relativa viene dada en tanto por 1. Si lo multiplicamos por 100 lo tendremos en tanto por ciento (%). La suma de todas las frecuencias relativas es igual a la unidad:

$$\sum_{i=1}^{n} f_i=1$$

Podemos añadir a la tabla una columna con los valores de las frecuencias relativas.

Variable altura en cm	n_i	$f_i=\frac{n_i}{N}$
143	1	0,091
149	1	0,091
157	1	0,091
159	1	0,091
164	2	0,182
166	2	0,182
170	1	0,091
176	1	0,091
181	1	0,091
	$N=\sum_{i=1}^{n} n_i=11$	$\sum_{i=1}^{n} f_i=1$

Frecuencia absoluta acumulada: Suma de las frecuencias absolutas de los valores de la variable menores o iguales a x_i.

$$N_i = n_1 + n_2 + n_3 + ... + n_i = \sum_{k=1}^{i} n_k$$

La frecuencia absoluta acumulada del último valor de x_i de la muestra coincide con el tamaño de la muestra. En una cuarta columna añadimos los valores de las frecuencias absolutas acumuladas.

Variable altura en cm	n_i	$f_i = \frac{n_i}{N}$	$N_i = \sum_{k=1}^{i} n_k$
143	1	0,091	1
149	1	0,091	2
157	1	0,091	3
159	1	0,091	4
164	2	0,182	6
166	2	0,182	8
170	1	0,091	9
176	1	0,091	10
181	1	0,091	11
	$N = \sum_{i=1}^{n} n_i = 11$	$\sum_{i=1}^{n} f_i = 1$	

Frecuencia relativa acumulada del valor: Es la suma de las frecuencias relativas de los valores de la variable menores o iguales a x_i.

$$F_i = f_1 + f_2 + f_3 + ... + f_i = \sum_{k=1}^{i} f_k$$

El valor de la frecuencia relativa acumulada viene dado en tanto por 1 o en tanto por ciento (%). La frecuencia relativa acumulada del último valor de x_i de la muestra es igual a la unidad. Añadimos una columna de frecuencia relativa acumulada a la tabla:

Variable altura en cm	n_i	$f_i = \frac{n_i}{N}$	$N_i = \sum_{k=1}^{i} n_k$	$F_i = \sum_{k=1}^{i} f_k$
143	1	0,091	1	0,091
149	1	0,091	2	0,182
157	1	0,091	3	0,273
159	1	0,091	4	0,364
164	2	0,182	6	0,546
166	2	0,182	8	0,728
170	1	0,091	9	0,819
176	1	0,091	10	0,910
181	1	0,091	11	1
	$N = \sum_{i=1}^{n} n_i = 11$	$\sum_{i=1}^{n} f_i = 1$		

Moda: Valor de la variable con mayor frecuencia absoluta (mayor número de repeticiones). En nuestro ejemplo tenemos dos modas: 164 y 166.

Mediana: El menor valor de la variable que acumula el 0,5 (50%) de la frecuencia relativa acumulada en %. Si el número de datos es impar, la mediana es el valor que deja por debajo y por arriba el mismo número de datos (valor central del conjunto de datos). Si el número de datos es par, la mediana es la media de los valores consecutivos que dejan por debajo y por encima el mismo número de datos.

En nuestro ejemplo (tamaño N=11 impar) la mediana es 164.

La mediana también recibe el nombre de percentil 50 o segundo cuartil: acumula el 50% de las observaciones.

El percentil 25 o primer cuartil es el valor que acumula el 25% de las observaciones. Y el percentil 75 o tercer cuartil es el valor que acumula el 75% de las observaciones.

Varianza: La media aritmética de los cuadrados de las desviaciones respecto a la media. La varianza siempre es un valor positivo o nulo. Una varianza muy pequeña indica que el valor medio $\bar{x}$ es un buen representante del conjunto de los valores particulares, y viceversa. Se representa por s^2 o por σ^2.

$$s^2=\frac{n_1(x_1-\overline{x})^2+n_2(x_2-\overline{x})^2+...+n_n(n_n-\overline{x})^2}{n_1+n_2+...+n_n}=\frac{\sum_{i=1}^{n} n_i(x_i-\overline{x})^2}{N}=\sum_{i=1}^{n} f_i(x_i-\overline{x})^2$$

En nuestro ejemplo, la varianza es 109,97.

Desviación típica (o desviación estándar): es la raíz cuadrada de la varianza. Nuevamente, un valor bajo de la desviación típica indica que la mayor parte de los valores de la muestra están distribuidos cerca de la media. Se representa por s o por σ.

$s=\sqrt{s^2}$

En nuestro ejemplo, la desviación típica es 10,49.

Coeficiente de variación: cociente entre la desviación típica y la media, multiplicado por 100%. Permite comparar el grado de dispersión en dos muestras distintas (a mayor coeficiente de variación, mayor dispersión respecto al valor de la media).

$CV=\frac{s}{\bar{x}}\cdot 100\%$

En nuestro ejemplo, el coeficiente de variación resulta 6,43%.