MEDIA Y VARIANZA MUESTRALES
COMO VARIABLES ALEATORIAS

GALIBERT, Ma. SILVIA
PANO, CARLOS O.

A efectos puramente didácticos imaginaremos que estamos interesados en estudiar una población compuesta por sólo cuatro sujetos que reciben terapia psicológica semanal: Laura, Inés, Claudia y Matías. La característica de interés es el número de sesiones semanales de análisis. Laura e Inés reciben una sesión semanal, Claudia dos y Matías tres.

Población de individuos: conjunto de las personas mencionadas.

Laura	Matías
Inés	Claudia

Supongamos que elegimos aleatoriamente una persona de esta población, por ejemplo sorteando uno de estos cuatro nombres, y registramos el número de sesiones. Así queda definida la variable aleatoria

X = número de sesiones semanales de análisis que recibe la persona elegida.

Luego, la **población de valores de X** es {1, 1, 2, 3}

1	3
1	2

Media de la población (o media de X)

La media de la población de valores de X (o media de X) es

$$\mu = \frac{1+1+2+3}{4} = \frac{7}{4} = 1{,}75$$

La hemos designado con la letra μ para diferenciarla de $\bar{X}$ que es la media de una muestra de la población.

Varianza de la población (o varianza de X) = V(X)

La designaremos con la letra σ^2. Cuando la población es finita -como en este caso- se calcula como el promedio de los cuadrados de los desvíos respecto de la media.

$$V(X) = \sigma^2 = \\ = \frac{(1-1{,}75)^2 + (1-1{,}75)^2 + (2-1{,}75)^2 + (3-1{,}75)^2}{4} = \\ =0{,}6875$$

μ y σ^2 son dos características numéricas de la población, es decir, dos **parámetros**. Supongamos que ignoramos tales parámetros y queremos estimarlos a partir de muestras de tamaño 2. Vamos a considerar el siguiente método:

Muestreo aleatorio con reposición

Se sortea al azar una de las cuatro personas, se registra su número de sesiones semanales de análisis y se devuelve la bolilla con la que se sorteó al bolillero. Nuevamente se elige una persona al azar entre las cuatro y se registra su número de sesiones semanales. Así se realiza lo que llamamos un muestreo al azar con reposición para obtener una muestra aleatoria de tamaño dos.

Este tipo de muestreo es útil es desde un punto de vista teórico pues la teoría estadística desarrollada sobre él es mucho más sencilla que la basada sobre el muestreo sin reposición, más razonable en la práctica. Sin embargo los resultados que se deducen del método de muestreo con reposición pueden ser usados como una muy buena aproximación en el muestreo sin reposición siempre que el tamaño de muestra sea pequeño en relación al tamaño poblacional (el presente ejemplo no es tal caso pues la muestra es el 50% de la población pero, como dijimos al principio, tomamos una población reducida para mayor simplicidad didáctica).

Las muestras que se pueden obtener en este ejemplo por el método antedicho son:

1era. elección	2da. elección
Laura	Laura
Laura	Inés
Laura	Claudia
Laura	Matías
Inés	Laura
Inés	Inés
Inés	Claudia
Inés	Matías
Claudia	Laura
Claudia	Inés
Claudia	Claudia
Claudia	Matías
Matías	Laura
Matías	Inés
Matías	Claudia
Matías	Matías

Llamando X_1 y X_2 al número de sesiones correspondientes a las personas elegidas en primero y segundo lugar respectivamente para integrar la muestra, tenemos, en correspondencia con la lista anterior,

X_1	X_2
1	1
1	1
1	2
1	3
1	1
1	1
1	2
1	3
2	1
2	1
2	2
2	3
3	1
3	1
3	2
3	3

Para cada una de las 16 muestras posibles podríamos calcular la media $\bar{x}$ y la varianza s^2 con lo que obtendríamos

$\bar{x}$	s^2
1	0
1	0
1,5	0,5
2	2
1	0
1	0
1,5	0,5
2	2
1,5	0,5
1,5	0,5
2	0
2,5	0,5
2	2
2	2
2,5	0,5
3	0

Así vemos que $\bar{X}$ y S^2 son variables aleatorias pues sus valores varían con las muestras, las cuales son seleccionadas al azar. Para designarlas hemos usado letras mayúsculas a fin de distinguirlas de cada valor en particular. Además

$$\bar{X} = \frac{X_1 + X_2}{2} \quad \text{y} \quad S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2}{1}$$

por lo que $\bar{X}$ y S^2 están en función de los valores muestrales X_1 y X_2; por eso $\bar{X}$ y S^2 son **estadísticos**. Como cada valor $\bar{x}$ de $\bar{X}$ es una estimación de μ y cada valor s^2 de S^2 es una estimación de σ^2 decimos que $\bar{X}$ y S^2 son, además, **estimadores**.

Un estimador se dice **insesgado** cuando su media coincide con el parámetro que pretende estimar.

Promediando los valores de $\bar{X}$ obtenemos

$$\frac{1 \times 4 + 1{,}5 \times 4 + 2 \times 5 + 2{,}5 \times 2 + 3}{16} = \frac{28}{16} = \frac{7}{4} = 1{,}75$$

que coincide con μ. Por eso decimos que

$\bar{X}$ es un estimador insesgado de μ.

Análogamente, promediando los valores de S^2 obtenemos

$$\frac{0 \times 6 + 0{,}5 \times 6 + 2 \times 4}{16} = \frac{11}{16} = 0{,}6875$$

que coincide con σ^2 por lo que

S^2 es un estimador insesgado de σ^2.

Observemos que la varianza muestral que hemos usado es la que se obtiene de dividir por el tamaño muestral menos 1; esto se hace justamente para obtener un estimador insesgado de la varianza poblacional.

La varianza de la población de valores de $\overline{X}$ es $V(\overline{X}) =$

$$= \frac{(1-1{,}75)^2 \times 4 + (1{,}5-1{,}75)^2 \times 4 + (2-1{,}75)^2 \times 5 + \ldots}{16}$$

$$\frac{\ldots + (2{,}5-1{,}75)^2 \times 2 + (3-1{,}75)^2}{16} = \frac{5{,}5}{16} = 0{,}34375$$

que es la mitad de la varianza de X por provenir de una muestra de tamaño 2.

En general, si la muestra es de tamaño n la varianza de la variable $\overline{X}$ se reduce a la n-ésima parte de la varianza de la variable X:

$$V(\overline{X}) = \frac{V(X)}{n} = \frac{\sigma^2}{n}$$

de donde el desvío estándar de $\overline{X}$ es

$$DS(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

Observemos a partir de esta expresión que cuanto mayor es el tamaño muestral menor es la variabilidad de la media.

Distribución de la media muestral $\overline{X}$

Cuando el tamaño muestral es suficientemente grande, la distribución de $\overline{X}$ es aproximadamente normal (tanto más normal cuanto mayor el tamaño de muestra) con media μ y varianza σ^2 / n.

Esto es un corolario del Teorema Central del Límite, muy importante en estadística.

Estandarizando $\overline{X}$ obtenemos el estadístico

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

que sigue aproximadamente la distribución normal estándar y se usa en inferencia estadística para probar hipótesis acerca de la media poblacional.