Inferencia Estadística.

Muestreos.

Nuestro conocimiento del mundo comienza con la observación[1]. Pero no podemos esperar a observar todos los casos relevantes antes de sacar conclusiones. Haría falta demasiado tiempo y demasiada paciencia. En este tema aprenderemos a valorar conclusiones extraídas a partir de información insuficiente. Nuestros resultados no serán «seguros», pero al menos sabremos qué probabilidad tienen de ser ciertos.

Así, por ejemplo, realmente *es posible* averiguar qué partido ganará las elecciones con un sondeo hecho a 1.000 personas, *si el sondeo está bien hecho*. Un ejemplo de sondeo mal hecho es el que se realizó para las elecciones norteamericanas tras la Segunda Guerra Mundial. Dio la victoria a los republicanos por amplio margen, pero luego ganaron los demócratas. ¿Qué pasó? Que habían hecho el sondeo por teléfono. En aquella época eran pocos (más bien los ricos) los que lo tenían, introduciendo un tremendo *sesgo* en la muestra. Por tanto, la muestra debe realizarse eligiendo los individuos de manera que representen bien al conjunto de la población.

Llamaremos **población** o **universo** al conjunto total de individuos, y **muestra** a la selección que hacemos de ellos. ¿Cómo elegir la muestra?

– Muestreo aleatorio simple: eligiendo totalmente al azar.

– Muestreo estratificado: determinamos una serie de características de los individuos que nos parecen relevantes (p.ej., el sexo, la edad o la renta) y procuramos elegir los individuos de la muestra de manera que las proporciones sean las mismas que en la población.

E1. (Muy facilito) Un investigador social sospecha que el patrón de voto depende de la edad, así que divide la población en cuatro grupos: menores de 30 (25%), entre 30 y 40 (20%), entre 40 y 55 (30%) y mayores de 55 (25%). Si su muestra es de 300 personas, ¿cómo debe elegir los individuos?

Sumas y Promedios en una Muestra.

[1] Ya oigo a los platónicos protestar...

Supongamos que una máquina produce clavos con una longitud de $\mu \pm \sigma$ centímetros. Interpretaremos eso como que la media de las longitudes es de μ cm, y su desviación típica de σ, aunque *no supondremos* que las longitudes sigan una distribución gaussiana. Supongamos que tomamos una muestra de N clavos y les medimos. Obtenemos N valores x_i. Si les sumamos, obtendremos una variable aleatoria:

$$X = \sum_{i=1}^{N} x_i$$

Al sumar variables aleatorias, se suman tanto las medias como las varianzas. Por tanto, la variable X tendrá:

$$\mu_X = N\mu \qquad \sigma_X^2 = N\sigma^2 \;\rightarrow\; \sigma_X = \sqrt{N}\sigma$$

¿Qué ocurrirá si considero, en lugar de la suma, el **promedio** de los valores?

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

Tanto la media como la desviación típica deberán ser divididas entre N:

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sqrt{N}\sigma}{N} = \frac{\sigma}{\sqrt{N}}$$

Por el teorema central del límite sabemos que la suma de una serie de variables aleatorias independientes converge a una gaussiana. También lo hace el promedio. Para N grande, las distribuciones se acercan a:

$$X \rightarrow \mathcal{N}(N\mu, \sqrt{N}\sigma) \qquad \bar{x} \rightarrow \mathcal{N}(\mu, \sigma/\sqrt{N})$$

¿Cuándo se puede suponer que la muestra es lo suficientemente grande como para que se pueda aplicar el teorema central del límite? Como siempre, hay una regla del dedo gordo: cuando $N \geq 30$.

Pongamos números. Supongamos que los clavos son producidos con una longitud de $8 \pm 0'4$ centímetros y extraemos una muestra de 100 clavos. Promediamos las longitudes de estos doscientos clavos. El resultado $\bar{x}$ corresponde con una distribución gaussiana (ya que $100 \gg 30$):

$$\mathcal{N}(8, 0'4/\sqrt{100}) = \mathcal{N}(0'8, 0'04)$$

Si nos preguntan por la probabilidad de obtener un valor mayor de $0'81$ para la media, tipificamos ese valor:

$$x = 0'81 \ \rightarrow \ z = \frac{0'81 - 0'8}{0'04} = 2'5$$

Así que la probabilidad pedida es $1 - F(2'5) \approx 1 - 0'99379 = 0'00621$, bastante baja.

E2. Tenemos unos ladrillos que miden 5 ± 1 centímetros. Si colocamos 100 de ellos en fila, ¿cuál es la probabilidad de que dicha fila mida más de $5'5$ metros?

E3. Los sueldos de los trabajadores de una empresa tienen un promedio de $600€$ y una desviación típica de $200€$. Elegimos al azar una muestra de 25 de ellos. Calcula la probabilidad de que la media de sus sueldos sea menor de $500€$.

Estimador de un Parámetro.

Dado un parámetro cualquiera de una población (que queramos medir) y una muestra de dicha población, se llama **estimador** del parámetro al valor que podríamos estimar a partir de dicha muestra. Así, por ejemplo, si queremos estimar la media de una población μ y sólo contamos con una muestra de 100 individuos x_i, consideraremos el «promedio» $\bar{x}$ como un estimador para la μ. Si la muestra fuera la población entera, el estimador coincidiría con el parámetro. Y, para muestras grandes, no deberían separarse mucho.

¿Cómo estimar la desviación típica de la población a partir de la muestra? Definiremos el «estimador de la desviación» s:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N - 1}}$$

Un momento... (oigo que alguien dice): lo entiendo si cambias el $N - 1$ por un simple N. En ese caso es la desviación típica de los datos de la muestra, que converge a la desviación de la población cuando la muestra es muy grande. Bueno (respondo yo)... eso sería si sustituyéramos $\bar{x}$ por μ. Pero μ no lo conocemos, sólo su estimador $\bar{x}$. Eso introduce un error que *de alguna manera* se compensa con el $N - 1$[2].

[2] La demostración del $(N - 1)$ es complicada y no la haremos aquí.

E4. Queremos averiguar la distribución de probabilidad para el número de asignaturas suspensas entre alumnos de 3º de ESO. Hacemos una muestra con 20 de ellos y obtenemos:

N asignaturas	0	1	2	3	4	5
N alumnos	7	3	4	2	2	2

A partir de la tabla, estima la media y la desviación típica de la distribución real.

Estimación de la Media.

Imaginad que nos dan una muestra y hemos calculado sus valores de $\bar{x}$ y s. A continuación nos piden estimar la media μ. Eso es fácil: $\mu \approx \bar{x}$. Pero además nos pueden pedir que demos un **intervalo de confianza** con una probabilidad dada p. ¿Qué significa esto? Pues un intervalo de valores de μ para los que la probabilidad de obtener $\bar{x}$ sea *mayor* que p. Veamos un ejemplo.

Obtener el intervalo de confianza para la media de una población con una probabilidad del 95% sabiendo que una muestra de 400 individuos ha dado un promedio $\bar{x} = 10$ y un estimador de la desviación $s = 2$.
Supondremos que $\sigma = s = 2$. Por el teorema central del límite, el promedio $\bar{x}$ proviene de una distribución normal de media μ y desviación $\sigma/\sqrt{N} = 2/\sqrt{400} = 0'1$.
Averiguamos el radio característico para el 95% de probabilidad. No tenemos la media de la distribución, pero tampoco nos hace falta. Siguiendo los pasos establecidos para intervalos característicos, sumamos al 95% la «cola izquierda», $0'95 + 0'025 = 0'975$. Buscamos el valor $0'975$ de manera inversa en la tabla de la normal y obtenemos un z de $1'96$. Como la desviación es de $0'1$, el radio característico es de $0'1 \cdot 1'96 = 0'196$.
Sabemos que μ tiene que ser tal que $\bar{x}$ esté como máximo a una distancia $0'196$ de él. El μ mínimo que cumple la condición es $\mu_- = 10 - 0'196 = 9'804$. El valor máximo será $\mu_+ = 10 + 0'196 = 10'196$. Por tanto, sabemos que $\mu \in [9'804, 10'196]$ con una confianza del 95%.

¿Por qué ese método? Si $\mu < \mu_-$, entonces la probabilidad de obtener el valor $\bar{x}$ que tenemos sería menor del 95%, y lo mismo ocurre si $\mu > \mu_+$.

Por tanto, estimar la media teniendo $\bar{x}$, N, σ ó s y la confianza p se reduce al final a:
• Calcular la desviación típica de la distribución de promedios: $s/\sqrt{N}$.

• Hallar el intervalo característico de la distribución $\mathcal{N}(\bar{x}, s/\sqrt{N})$ para la probabilidad p.

Q1. Cuando el grado de confianza que nos piden (p) crece, ¿qué ocurre con el radio del intervalo de confianza?

E5. Hemos realizado una muestra de 100 individuos extraídos de una población de la que deseamos conocer la media. Para la muestra, $\bar{x} = 2'8$ y $s = 5$. Obtén un intervalo de confianza del 80% para la media.

E6. Una muestra al azar de 200 cojinetes de bolas dieron una media de 2 cm. y una desviación típica de $0'1$ cm. Halla los intervalos de confianza del $68'26\%$, $95'44\%$ y $99'73\%$ para el diámetro medio de los cojinetes.

E7. Hemos analizado una muestra muy pequeña de una población y hemos obtenido $s = 10$. Queremos hacer una muestra mayor y obtener una estimación para la media que, con una probabilidad del 99%, esté contenida en un intervalo de anchura $0'01$. ¿Cuál es el tamaño que debería tener la segunda muestra?

E8. Hemos tomado una muestra de 81 personas de una ciudad. El 90% de ellas tiene una altura en el intervalo $[173'4, 175'8]$. Estima μ y σ para la población.

Q2. Hemos hecho una muestra de 25 individuos de una población y el radio de un determinado intervalo de confianza ha sido de ± 4. Si la muestra hubiera sido de 100 individuos, ¿cómo cambiaría dicho radio?