1.1. El concepto de Estadística.

¿Qué es y para qué sirve?

La Estadística se ocupa de la **recolección, agrupación, presentación, análisis e interpretación de datos.**

A menudo se llaman estadísticas a las listas de estos datos, cosa que crea una cierta ambigüedad, que no debería originarnos confusiones.

La Estadística no son sólo los resultados de encuestas, ni el cálculo de unos porcentajes, la Estadística es un **método científico que pretende sacar conclusiones a partir de unas observaciones hechas**.

¿Cuándo empezó la Estadística?

La Estadística actual es el resultado de la **unión de dos disciplinas** que evolucionaron de forma independiente hasta confluir en el siglo XIX:

- el **Cálculo de Probabilidades**, que nació en el siglo XVII como la teoría matemática de los juegos de azar.

  Correspondencia entre Pascal y Fermat (1654) para resolver *el problema de los puntos.* De Mère, amigo de Pascal y jugador empedernido, le pidió a Pascal que le ayudara en calcular cómo dividir las apuestas de una partida empezada que debía interrumpirse cuando llegaba la policía, pues el juego era ilegal en Francia.
- la **"Estadística"**, o **ciencia del Estado**, que estudia la descripción de datos, y que tiene unas raíces más antiguas, de hecho, tan antiguas como la humanidad (censos de población).

  Desde la antigüedad los Estados han recogido datos sobre sus habitantes con el objetivo principal de recaudar impuestos y tributos, y reclutar jóvenes para el ejército. Babilonia, Imperio romano, Carlomagno… España lleva a cabo uno de los primeros censos de población en el siglo XVI en Perú.

La interacción de ambas líneas de pensamiento da lugar a la **ciencia que estudia *cómo obtener conclusiones de la investigación empírica mediante el uso de modelos matemáticos*.**

El impulso fundamental para la aparición de la **Estadística** fue la necesidad de *estimar* cantidades desconocidas a partir de *muestras*. Este problema se planteó en el campo de la Astronomía (necesidad de verificar la teoría de **Newton**) donde, por falta de precisión de los instrumentos de medición, distintas mediciones de la misma cantidad conducían a resultados diferentes. El problema consistía en decidir cuál de estas mediciones era la mejor. Este problema lo solucionó Gauss, introduciendo la *ley normal* o *gaussiana* como modelo de los errores de medición.

La revolución que supuso en la Física Newton, se produjo en la Biología por la obra de **Darwin**. Así a finales del siglo XIX Galton (primo de Darwin) y Pearson inventan métodos para medir la relación entre dos variables y aparece así la idea de *regresión* y el *coeficiente de correlación*.

En el campo de las Ciencias Sociales, los primeros trabajos para encontrar *relación entre variables* son debidos a Adolfo Quetelet (1796-1874, astrónomo belga de curiosidad ilimitada) y Florence Nightingale (1820-1910, primera mujer estadística, enfermera británica condecorada con la orden del mérito británico) que aplicaron la Estadística a problemas demográficos, sociológicos y políticos.

Resumiendo:

La Estadística actúa como **disciplina puente entre los modelos matemáticos y los fenómenos reales**.

Un modelo matemático es una abstracción simplificada de una realidad más compleja y siempre existirá una cierta discrepancia entre lo que se observa y lo previsto por el modelo.

La Estadística proporciona una metodología para evaluar y juzgar estas discrepancias entre la realidad y la teoría.

1.2. Conceptos generales.

Algunas definiciones:

Población estadística: conjunto finito o infinito de elementos, denominados **individuos**, sobre los cuales se realizan observaciones. Ejemplos: *todos los habitantes de cierto lugar, todos los ejemplares de una determinada especie de tortugas, todos los microchips que fabrica una empresa, etc.*

Muestra: subconjunto finito de una población. El número de individuos que forman la muestra se denomina **tamaño muestral**.

Variable o **carácter**: cada una de las características que pueden observarse en un individuo de la muestra. Ejemplos: en una muestra de una población de seres humanos podemos medir: *la altura, la edad, el peso, el sexo, número de hermanos…*; en una muestra de una población de una especie de tortugas podemos medir: *la anchura del caparazón, la longitud del caparazón, la edad…;*

Tipos de variables:

Cualitativas, categóricas (o alfanuméricas):

Pueden tomar valores no cuantificables numéricamente.

Se denomina **categoría** a cada uno de los valores que toma la variable.

Nominales: si no existe ningún orden entre las categorías de la variable. Ejemplos: *el grupo sanguíneo (A ,B ,AB, O); el color de los ojos (azules, verdes, marrones, negros),…*
Hay que distinguir las variables **binarias**, aquéllas que sólo toman dos valores posibles (sí/no, presencia/ausencia de cierto carácter), dentro de las nominales. Ejemplo: *el sexo, ser fumador, tener carné de conducir, ser daltónico,…*

Ordinales: cuando existe un cierto orden entre las categorías de la variable. Ejemplo: *el nivel de estudios (sin estudios, básicos, medios, superiores), el grado de miopía (ausencia, bajo, medio, alto),…*

Cuantitativas (o numéricas):

Pueden tomar valores cuantificables numéricamente.

Discretas: si solamente toman valores aislados (generalmente enteros). Suelen corresponder a contajes. Ejemplos: *el número de hermanos, el número de cafés/día, el número de multas/año,…*

Continuas: potencialmente puede tomar cualquier valor numérico dentro de un intervalo o de una unión de intervalos. Ejemplos: *el tiempo de reacción a un cierto medicamento, el peso de un individuo, la longitud del caparazón de una tortuga,…*

1.3. Métodos de muestreo.

¿Por qué seleccionamos una muestra?

En la práctica no va a ser posible estudiar todos los elementos de la población, por varias razones:

- El estudio puede implicar la **destrucción del elemento** (*estudio de la vida media de una partida de bombillas, estudio de la tensión de rotura de unos cables…*)
- Los elementos pueden existir **concep-tualmente**, pero no en realidad (*población de piezas defectuosas que producirá una máquina en su vida útil*).
- Puede ser **inviable económicamente** estudiar a toda la población.
- El estudio llevaría tanto tiempo que sería **impracticable** e incluso las propiedades de la población podrían variar con el tiempo.

Tipos de muestreo

Muestreo aleatorio simple.
Muestreo estratificado.
Muestreo por conglomerados.

1.3.1. Muestreo aleatorio simple.

¿Cuándo de utiliza?

Cuando los elementos de la población son homogéneos respecto de la variable de estudio, es decir, cuando *a priori* **no disponemos de información adicional sobre la población**.

Definición:

Una muestra es aleatoria simple cuando:
1) cada elemento de la población tiene la misma probabilidad de ser escogido,
2) las observaciones se realizan con reposición, de manera que la población es idéntica en todas las extracciones.

Comentarios:

La condición (1) asegura la representatividad.
La condición (2) se impone por simplicidad: si el tamaño de la población ***N*** es grande con respecto al tamaño muestral ***n***, es prácticamente indiferente realizar el muestreo con o sin reposición.
(**Atención**: Si el cociente ***n/N>0.1*** los métodos que estudiaremos aquí son aproximados y deberían tenerse en cuenta las correcciones pertinentes).

¿Cómo se realiza?

Se utilizan las *tablas de números aleatorios*: se enumeran los elementos de la población del 1 al N y se toman números aleatorios de tantas cifras como tenga N. **El valor del número aleatorio indicará el elemento a seleccionar**.

1.3.2. Muestreo estratificado.

Disponemos de **información adicional** sobre la población e interesa que la muestra tenga una **composición análoga** a de la población.

Los elementos de la población **se dividen en clases o estratos** según los valores de alguna otra variable (*por ejemplo, según el sexo, la edad, la profesión,...*)

¿Cómo se realiza?

- se asigna un número o **cuota** de miembros a cada estrato,
- dentro de cada estrato se seleccionan los elementos por muestreo aleatorio simple.

Si hay k estratos de tamaños $N_1,\ldots,N_k$, de manera que $N= N_1+\ldots+N_k$, la composición de la muestra será $n= n_1+\ldots+n_k$, donde las cuotas n_i, se pueden determinar de dos formas distintas:

1) proporcionalmente al tamaño de cada estrato:

$$n_i = n\ \frac{N_i}{N}$$

2) proporcionalmente a la variabilidad de cada estrato:

$$n_i = n\ \frac{\sigma_i\, N_i}{\sum_{i=1}^{k} \sigma_i\, N_i}$$

donde σ_i es una medida de la variabilidad del estrato i-ésimo.

1.3.3. Muestreo por conglomerados.

Hay situaciones en que ni el muestreo aleatorio simple ni el estratificado son aplicables.

En estos casos es habitual que los elementos de la población se encuentren agrupados en **conglomerados**, de los cuales sí que se sabe cuántos hay. (*Por ejemplo, la población se distribuye en provincias, los habitantes de una ciudad se distribuyen en barrios,…*)

¿Cómo se realiza?

Si puede suponerse que **cada conglomerado es una muestra representativa de la población total** respecto de la variable de estudio, podemos:

- seleccionar al azar **algunos** de estos conglomerados,
- dentro de cada conglomerado, analizar:
  a) **todos** sus elementos,
  b) **una muestra aleatoria simple** de sus elementos.

Inconveniente:

Si los conglomerados son heterogéneos entre ellos, puesto que sólo se analizan algunos de ellos, **la muestra final puede ser no representativa de la población**.

Las ideas de **estratificación** y **conglomerado** son opuestas:

- La estratificación funciona mejor cuánto **mayor sean las diferencias entre estratos**, pero es necesario que los estratos sean **homogéneos internamente**.
- Los conglomerados funcionan mejor cuánto **menores sean las diferencias entre ellos**, pero deben ser muy **heterogéneos internamente**, es decir, dentro de cada conglomerado debe estar incluida toda la variabilidad de la población.

Conclusión:

La regla general que se aplica a todos los procedimientos de muestreo es que *cualquier información previa tiene que utilizarse para subdividir la población y asegurar una mayor* ***representatividad de la muestra***.

Una vez se tienen las subpoblaciones homogéneas, la selección dentro de ellas debe realizarse por muestreo aleatorio simple.