INTRODUCCIÓN A LA ESTADÍSTICA

Autor: Clara Laguna

1.1 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, televisión, etc., no nos aborde diariamente con cualquier tipo de información estadística.

Sólo cuando nos adentramos en un mundo más específico como es el campo de la investigación de las Ciencias Sociales: Medicina, Biología, Psicología,...percibimos que la Estadística se convierte en la única herramienta que 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.

Desde un punto de vista más amplio, podemos decir que la Estadística se utiliza como tecnología al servicio de las ciencias donde la variabilidad y la incertidumbre forman parte de su naturaleza.

La Estadística es la rama de las matemáticas aplicadas que permite estudiar fenómenos cuyos resultados son en parte inciertos. Al estudiar sistemas biológicos, esta incertidumbre se debe al desconocimiento de muchos de los mecanismos fisiológicos, a la incapacidad de medir todos los determinantes de la enfermedad y a los errores de medida que inevitablemente se producen. Así, al realizar observaciones en clínica o en salud pública, los resultados obtenidos contienen una parte sistemática o estructural, que aporta información sobre las relaciones entre las variables estudiadas, y una parte de “ruido” aleatorio. El objeto de la estadística consiste en extraer la máxima información sobre estas relaciones estructurales a partir de los datos recogidos.

Historia de la Estadística:
Su raíz: STATUS=cosas del estado.
Durante el siglo pasado, era considerada como la Ciencia del Estado.
Sus orígenes: El recuento.
Las civilizaciones antiguas recogían datos sobre población, producción agrícola y renta. Tal cantidad de información debía ser resumida en valores numéricos para su interpretación y uso en la toma de decisiones políticas.

Definición:

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.

"La estadística es la ciencia que permite tomar decisiones en situaciones de incertidumbre"

¿A qué nos referimos cuándo hablamos de variabilidad?

Cuando nos referimos a una determinada característica (p.ej. nivel de ácido úrico) en los individuos de una población nos estamos refiriendo a una distribución de valores. La observación de un determinado grupo de pacientes nos proporciona información acerca de esta distribución. Los resultados que observamos se relacionan con la distribución poblacional. Sin embargo, observaciones distintas proporcionan resultados distintos (aunque compatibles), por ejemplo un mismo tratamiento aplicado a grupos distintos de pacientes proporcionará resultados distintos.

La ***Bioestadística*** es la rama de la estadística que estudia la utilización de métodos estadísticos en problemas médicos y biológicos. Enseña y ayuda a investigar en todas las áreas de las Ciencias de la Vida donde la variabilidad no es la excepción sino la regla.

Podemos 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. Su tarea fundamental es la de hacer inferencias acerca de la población a partir de una muestra extraída de la misma.

El análisis de una base de datos siempre partirá de técnicas simples de resumen de los datos y presentación de los resultados. A partir de estos resultados iniciales, y en función del diseño del estudio y de las hipótesis preestablecidas, se aplicarán las técnicas de inferencia estadística que permitirán obtener conclusiones acerca de las relaciones entre las variables estudiadas. Las técnicas de estadística descriptiva no precisan de asunciones para su interpretación, pero la información que proporcionan no es fácilmente generalizable. La estadística inferencial permite esta generalización, pero requiere ciertas asunciones que deben verificarse para tener un grado razonable de seguridad en las inferencias.

1.2 CONCEPTOS BÁSICOS

Vamos a definir algunos conceptos básicos y fundamentales a los cuales haremos referencia continuamente:

Unidad estadística, **Individuo o Elemento**: personas u objetos que contienen cierta información que se desea estudiar y que pertenecen a la población en estudio.

Población: conjunto de individuos o elementos que cumplen ciertas propiedades y entre los cuales se desea estudiar un determinado fenómeno.

Muestra: subconjunto representativo de una población.

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

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

Estimador: función de los valores de una muestra que se elabora para indagar el valor de un parámetro de la población de la que procede la muestra.

Los parámetros poblacionales se denotan con letras del alfabeto griego, mientras que los estimadores muestrales se denotan con letras de nuestro alfabeto.

Así, por ejemplo, la media del colesterol en una población, que se denotaría por μ, es un parámetro que se estima a partir de la media de los valores de colesterol en una muestra obtenida en esa población, que se representaría por $\bar{x}$ *.*

Variables o caracteres: característica observable que varía entre los diferentes individuos de una población. Las variables pueden dividirse en cualitativas y cuantitativas.

Modalidades o categorías: posibles valores de una variable. Las modalidades deben ser a la vez exhaustivas y mutuamente excluyentes (cada elemento posee una y sólo una de las modalidades posibles). Las modalidades pueden agruparse en *clases* (intervalos).

1.2.1 Tipos de variables

Variables cualitativas:
Se usan con datos que **representan categorías** que son mutuamente excluyentes, aunque se utilicen números para cada categoría no representan cantidades. Para su medición usamos escalas:

- **Nominales**: no hay relación entre las categorías. Una escala nominal sólo permite clasificar (no jerarquizar ni ordenar).
  Ej.: Género, nacionalidad, situación geográfica

  Distinguimos dos tipos de variables cualitativas o categóricas nominales:
  - *Dicotómicas o binarias*: sano/enfermo, hombre/mujer
  - *Policotómicas* (con varias categorías): grupo sanguíneo (A/B/AB)

- **Ordinales**: sus posibles categorías se encuentran jerarquizadas y ordenadas.
  Ej.: Mejoría a un tratamiento, satisfacción de un usuario, grado de dolor

Es buena idea codificar las variables cualitativas asignando un código numérico a cada categoría ("etiqueta") para poder procesarlas con facilidad.
¡Ojo! Aunque se codifiquen como números, debemos recordar siempre el verdadero tipo de las variables con las que estamos trabajando y su significado cuando vayamos a usar programas estadísticos.

¡No todo está permitido con cualquier tipo de variable!

Variables cuantitativas:
Se usan con datos que se expresan mediante **cantidades numéricas** que permiten hacer operaciones matemáticas. Existen los siguientes tipos:

- **Discretas:** sólo puede tomar valores enteros.
  Ej.: Nº de hijos, nº de intervenciones

- **Continuas:** Si sus posibles valores están en un conjunto infinito. Las podemos categorizar en intervalos (trataremos este tema en la clase práctica con SPSS).
  Ej.: Edad, peso, tensión arterial

En la tabla se resumen los distintos tipos de variables y su utilidad:

TIPO DE VARIABLE	SIRVE PARA
Cualitativa nominal	CLASIFICAR
Cualitativa ordinal	JERARQUIZAR
Cuantitativa discreta	CONTAR
Cuantitativa continua	MEDIR

Figura 1.1

Aquí tenéis un ejemplo de base de datos en SPSS:

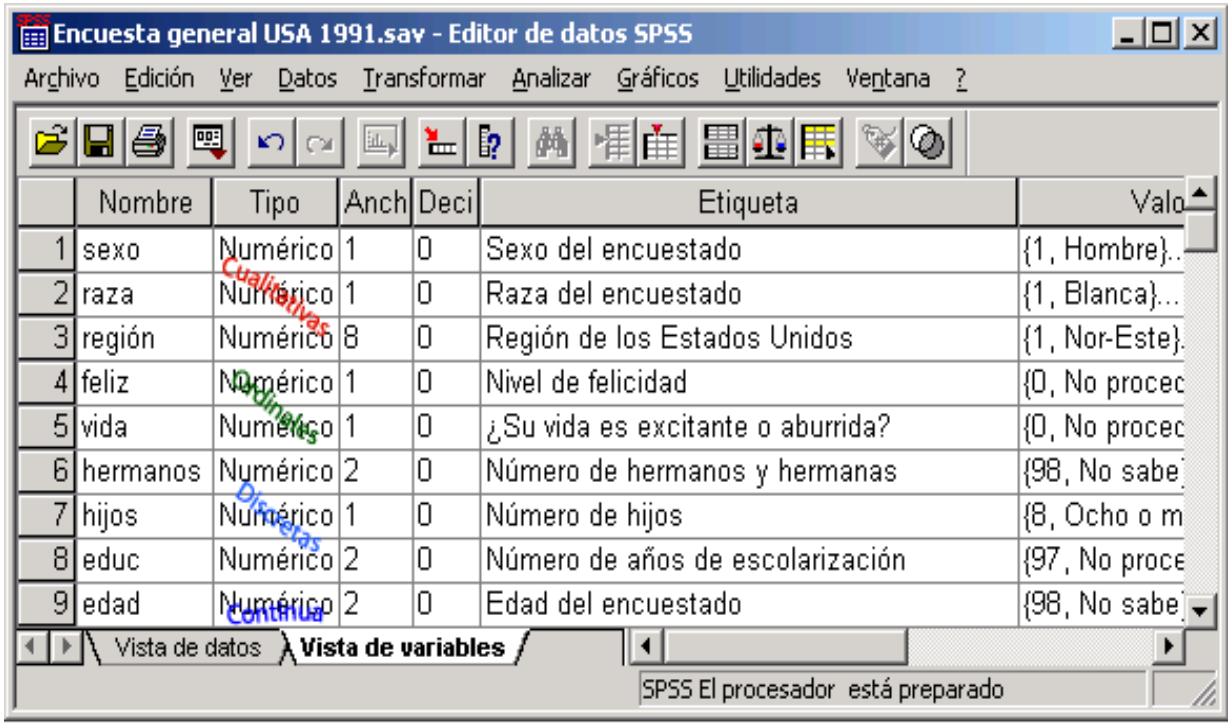

Figura 1.2

1.3 MUESTREO

Cuando se decide cuantificar sólo una parte de las unidades de una población y a partir de esta información estimar sus parámetros, entonces estamos planteando un problema de muestreo.

La estadística habitualmente estudia sólo una muestra de individuos. Se entiende por **muestra** al subconjunto de una población de mayor tamaño. Se entiende por **población** o **universo** al conjunto de todos los individuos o elementos (*unidades de análisis*) que cumplen ciertas características. Al proceso de extracción de una muestra a partir de una población se le denomina **muestreo**. A la interpretación del tratamiento estadístico de unos datos que acaba generalizándolos a toda la población se le llama **inferencia**.

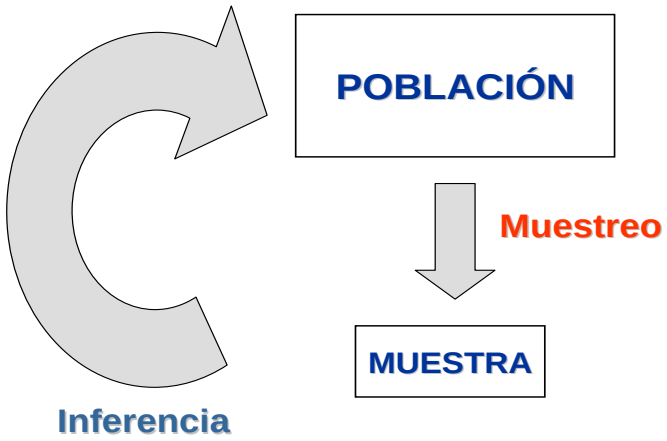

Figura 1.3

El *muestreo* es una herramienta de la investigación científica, su función básica es determinar qué parte de la realidad en estudio (población o universo) debe de examinarse con la finalidad de hacer inferencias sobre el todo de la que procede.

Un proceso inductivo (que va de lo particular a lo general) se asocia inevitablemente a la posibilidad de cometer errores. El error es mayor en la medida que la muestra sea más pequeña, y sobre todo, cuando la muestra no refleja o "representa" la realidad sobre la que recaen las conclusiones de la inferencia.

El error que se comete debido al hecho de que se sacan conclusiones sobre cierta población, a partir de la observación de sólo una parte de ella, se denomina **error de muestreo.**

Dentro del proceso de selección de una muestra, la población suele dividirse en **unidades de muestreo** las cuales deben de cubrir por entero a la población, en otras palabras, todo miembro de la población o unidad de análisis pertenece a una y sólo una unidad de muestreo.

Una unidad de muestreo puede contener un conjunto de *unidades de análisis*, incluso, un conjunto de unidades de muestreo correspondientes a una etapa posterior. La lista de las unidades de muestreo, recibe el nombre de **marco muestral**.

Ejemplo 1.1.

Supongamos que se estudia una población de escolares a fin de conocer la prevalencia de cierta dolencia. Cada escolar es una unidad de análisis pero, en lugar de contar con un listado de escolares, se tiene un listado de colegios (alguno de los cuales se elegirán para el estudio).Una vez hecha esta selección, se toman algunas aulas de los colegios elegidos. Finalmente, dentro de estas últimas se eligen alumnos que integran la muestra definitiva.
El listado de colegios primero, el de las aulas que tiene cada colegio seleccionado y el de niños, correspondiente a cada aula elegida dentro de los colegios de la muestra, constituyen el ***marco muestral*** *del estudio.*

Puesto que hay tres procesos escalonados de selección, existen en este caso las llamadas unidades de muestreo de primera, segunda y tercera etapa.

Figura 1.4

Podemos resumir en estos cuatro puntos las **ventajas** que la **utilización de muestras** presenta con respecto a la enumeración completa de la población:

- Coste reducido. Si los datos se obtienen de una pequeña fracción del total, los gastos de recogida y tratamiento de los datos se reducen. Incluso si la obtención de información en toda la población es factible, suele ser mucho más eficiente la utilización de técnicas de muestreo.
- Mayor rapidez. Los datos pueden ser más fácilmente recolectados y estudiados si se utiliza una muestra que si se emplean todos los elementos de la población. Por tanto, el uso de técnicas de muestreo es especialmente importante cuando se necesita la información con carácter urgente.

- Mayor flexibilidad y mayores posibilidades de estudio. La disponibilidad de registros completos es limitada. Muy a menudo, la única alternativa posible para la realización de un estudio es la obtención de datos por muestreo.

- Mayor control de calidad del proceso de recogida de datos. Al recoger datos en un número menor de efectivos, resulta más fácil recoger un número mayor de variables por individuo, así como tener un mejor control de la calidad del proceso de recogida de datos.

La teoría de muestreo persigue un **doble objetivo**. Por un lado, estudia las técnicas que permiten obtener muestras representativas de la población de forma eficiente. Por otro lado, indica cómo utilizar los resultados del muestreo para estimar los parámetros poblacionales, conociendo a la vez el grado de incertidumbre de las estimaciones.

Así, la teoría de muestreo pretende *dar respuesta a varias preguntas de interés*:

- ¿Cómo se eligen a los individuos que componen la muestra?
- ¿Cuántos individuos formarán parte de la muestra?
- ¿Cómo se cuantifican las diferencias existentes entre los resultados obtenidos en la muestra y los que hubiéramos obtenido si el estudio se hubiera llevado a cabo en toda la población?

1.3.1 Técnicas de muestreo

La característica más importante de una **muestra** es que debe ser **representativa** de la población objeto de estudio para poder extrapolar los resultados a la población total.

Las nociones de muestra representativa y de muestra probabilística suelen identificarse erróneamente como una y la misma. Como consecuencia puede ocurrir que, al admitir que la muestra no fue seleccionada por vía del azar, el investigador sienta que su estudio carece del rigor científico necesario.
La noción que mejor sintetiza la idea de representatividad es la siguiente:
"Lo que debe procurarse es que la muestra tenga internamente el mismo grado de diversidad que la población"

Los métodos para seleccionar una muestra representativa son numerosos, podemos clasificarlos en:

- **Probabilístico**: Todos los individuos tienen la *misma probabilidad* de ser elegidos para formar parte de la muestra (principio de equiprobabilidad).

- **No Probabilístico:** La elección de los individuos *no depende de la probabilidad*, sino del proceso de toma de decisiones del investigador (las muestras seleccionadas por decisiones subjetivas tienden a estar sesgadas).

MUESTREO PROBABILISTICO

Los procedimientos probabilísticos reducen la carga subjetiva que podría influir en la elección de las unidades que se van a estudiar, y sobre todo, permiten medir el grado de precisión con que se realizan las estimación de los parámetros poblacionales.

El azar no necesariamente inyecta representatividad a cada muestra que se obtenga, sino lo que realmente asegura es la imparcialidad en la conducta del investigador.

Se considera que el método de selección de la muestra tiene un carácter estadísticamente riguroso cuando su diseño cumple las siguientes condiciones:

- A cada elemento de la población, se le otorgue una probabilidad **conocida** de pertenecer a la muestra.
- Y por supuesto, esta probabilidad no sea nula.

Vamos a describir brevemente los principales **procedimientos probabilísticos** de selección de muestras:

- Muestreo aleatorio simple
- Muestreo sistemático
- Muestreo aleatorio estratificado
- Muestreo por conglomerados
- Muestreo polietápico

Muestreo aleatorio simple (m.a.s.)

Es el más sencillo y conocido de los distintos tipos de muestreo probabilístico. Supongamos que se pretende **seleccionar una muestra de tamaño *n* a partir de una población de *N* unidades**. Un muestreo aleatorio simple es aquel en el que cada unidad de muestreo de la población tiene la misma probabilidad de ser seleccionado.

Puede probarse que el m.a.s. es un procedimiento equiprobabilístico; es decir, todas las unidades de la población tienen la **misma probabilidad *n/N*** de ser elegidas en la muestra.
A la probabilidad que tiene cada individuo de pertenecer a la muestra se le denomina *fracción de muestreo*: f= n/N

Para la selección de una m.a.s., se enumeran previamente las unidades de la población de 1 a *N* y a continuación se seleccionan *n* números distintos entre 1 y *N* utilizando algún procedimiento aleatorio (mediante una tabla de números aleatorios o un generador de números aleatorios por ordenador).

Ejemplo 1.2

- *Elegir una muestra aleatoria de 5 estudiantes en un grupo de estadística de 20 alumnos.*
- *Extraer una muestra a partir de 37.488 historias clínicas del Servicio de Planificación Familiar de un Hospital.*

Muestreo sistemático

Cuando los elementos de la población están ordenados en una lista, podemos muestrear de la siguiente forma:

- En primer lugar, se calcula la *constante de muestreo k=N/n*.
- Se elige aleatoriamente un número de arranque *r* entre 1 y *k*, donde *k* es la parte entera de *N/n*
- Se le suma a *r* (primera unidad elegida) la constante *k* sucesivamente hasta completar el tamaño de la muestra.

Ejemplo 1.3

De un conjunto de 1.000 unidades queremos seleccionar 200.

La constante de muestreo será K = 1000/200 = 5, por tanto se escogerá a una de cada cinco.
La primera será sorteada entre los números del 1 al 5; si el elegido es el 2, el siguiente sería el 7 (2+k),… y así hasta completar la muestra.

Muestreo estratificado

Cuando se desea asegurar la representatividad de determinados subgrupos o estratos de la población, la alternativa más sencilla es seleccionar por separado distintas submuestras dentro de cada estrato. Los **estratos** han de definir subgrupos de población que sean *internamente homogéneos* con respecto a la característica o parámetro de interés y, por tanto, **heterogéneos entre sí**.

En la práctica, los estratos se definen en función de variables fáciles de medir previamente y relevantes para el tema objeto de estudio (edad, sexo, área geográfica de residencia). En general, el número de estratos *L* ha de ser reducido (rara vez resulta eficiente utilizar más de 5 estratos) y el tamaño por estrato no debe ser muy pequeño.

Para la selección de una muestra estratificada de tamaño *n*, la población de *N* unidades se divide en *L* estratos de tamaños $N_1, N_2,..., N_L$, cuya suma es igual a *N*.

1. Decidir el *número de variables* elegidas para la estratificación.
2. Elegir las *variables de la estratificación* e indicar el orden de estas variables, escogiendo como primera la que más discrimina.
3. Distribuir la muestra en cada estrato, ***AFIJACIÓN***.
   Que la muestra sea representativa de cada estrato y que cada estrato esté suficientemente representado para poder tomar como válidos los resultados obtenidos.

En el muestreo estratificado, es necesario determinar cómo se distribuye el tamaño muestral total *n* entre los distintos estratos; es decir, la asignación de los tamaños muestrales $n_1,..., n_L$. El procedimiento utilizado con mayor frecuencia es la **afijación proporcional**: el tamaño de la muestra de cada estrato es proporcional al tamaño del estrato correspondiente con respecto a la población total.

Ejemplo 1.4

Tenemos una población con N= 10.000 y queremos distribuir una muestra de tamaño n= 600 personas en L= 3 estratos.

La distribución de la población por edades es la siguiente:
Grupo A: 1.500 habitantes <18 años
Grupo B: 6.500 habitantes 18-60 años
Grupo C: 2.000 habitantes >60 años

*Asignamos por **Afijación Proporcional** a cada estrato su tamaño muestral:*
$n_L = (n/N)xN_L$

Estrato 1. Grupo A: 600x (1.500 / 10.000) = 90 hab.
Estrato 2. Grupo B: 600x (6.500 / 10.000) = 390 hab.
Estrato 3. Grupo C: 600x (2.000 / 10.000) = 120 hab.

Muestreo por conglomerados

La aplicación de los diseños muestrales anteriores requiere de la enumeración u ordenación de todos los elementos de la población. Sin embargo, a menudo no se dispone de una lista completa o, aun disponiendo de tal lista, resulta muy costoso obtener información de las unidades muestreadas. Por ejemplo, si se seleccionara una muestra aleatoria simple de 1000 individuos de una gran ciudad, los individuos seleccionados estarían muy dispersos y la recogida de información sería extraordinariamente laboriosa. En tales circunstancias, una alternativa consiste en clasificar a la población en grupos o conglomerados, para así seleccionar una muestra de estos conglomerados y después tomar a todas o a una parte de las unidades incluidas dentro de los conglomerados seleccionados.

Los conglomerados acostumbran a ser agrupaciones naturales de individuos como hogares, hospitales, colegios, provincias, etc. A diferencia de la estratificación, las **diferencias dentro de cada conglomerado** deben ser **máximas**: en cada conglomerado debe haber unidades representativas de toda la población, de lo contrario se perdería información al seleccionar únicamente algunos de ellos. Los resultados no varían si se selecciona uno u otro conglomerado. El número de conglomerados es típicamente elevado, de los cuales suele seleccionarse un número relativamente pequeño para resolver el problema de la dispersión muestral.

Ejemplo 1.5
Si queremos extraer una muestra para un estudio epidemiológico en niños escolarizados en Zaragoza, con edades comprendidas entre 5 y 14 años, a partir de una lista completa de las aulas de todos los centros escolares, podemos elegir aleatoriamente un cierto número de aulas (conglomerados) de manera que la muestra estaría formada por todos los niños de las aulas seleccionadas.

Muestreo polietápico

Los diseños muestrales empleados en la práctica se realizan combinando las técnicas descritas anteriormente. En muchas situaciones, resulta más apropiado obtener la muestra final en diferentes etapas o pasos.
En un muestreo polietápico, la población se divide en grupos exhaustivos y mutuamente excluyentes, que constituyen las llamadas *unidades de primera etapa*; cada una de ellas se desagrega a su vez en subgrupos o *unidades de segunda etapa*, y así sucesivamente, hasta llegar en una *última etapa* a los elementos o *unidades de análisis*.
La selección de unidades en cada una de las etapas se realiza mediante una técnica de muestreo diferente y la muestra final será la resultante de aplicar sucesivamente cada una de estas técnicas.

Ejemplo 1.6
Para obtener una muestra de pacientes diabéticos ingresados en nuestro país, en una primera etapa se escoge una muestra de hospitales, y en la segunda etapa, una muestra de pacientes diabéticos ingresados en los hospitales elegidos.

MUESTREO NO PROBABILISTICO

Los métodos de muestreo no probabilístico son aquellos en los que las unidades de análisis se recogen según criterios del investigador y no utilizando métodos en los que interviene el azar, de modo que no es posible estimar la probabilidad que tiene cada elemento de ser incluido en la muestra y no todos los elementos tienen posibilidad de ser incluidos. No garantizan la representatividad de la muestra y por lo tanto no permiten realizar estimaciones inferenciales sobre la población.

Al igual que en los anteriores existen diferentes tipos de muestro no probabilístico: por cuotas, por conveniencia, método bola de nieve,…

1.3.2 Tamaño muestral

Una vez abordados los diferentes diseños de muestreo es decir, como se obtiene la muestra, uno de los puntos que preocupa a la mayoría de los investigadores es cuantos individuos debe tener la muestra es decir que tamaño.
Buscamos una fórmula que nos de un "número mágico" de sujetos que formen nuestra muestra. No vamos a entrar en fórmulas de cálculo de tamaño muestral (ver material de apoyo) puesto que para comprender su desarrollo necesitamos tener algunos conceptos estadísticos que todavía no hemos desarrollado.

Sin embargo, vamos a tener en cuenta algunas consideraciones:

- El tamaño de la muestra estará en función de cuan frecuente sea lo que deseamos medir. Si lo que quiero estudiar es muy frecuente, necesitaré una muestra más pequeña, que si se da con menor frecuencia. Sin embargo, nosotros calculamos el tamaño muestral para conocer algo que desconocemos y, sin embargo, debemos de partir de un conocimiento de su valor (en la mayoría de los casos por otros estudios) para saber que tamaño de muestra elegimos.

- El tamaño de la muestra estará en función del máximo error de muestreo que se esté dispuesto a admitir al estimar un parámetro (a menor error necesitamos mayor muestra). Se supone, por tanto, que hay un error máximo, lo cual no siempre es fácil de determinar a priori y, en cualquier caso, se trata de una decisión esencialmente subjetiva.

- Cuanto más complejo es el diseño que utilizamos, mayor será la muestra que necesitaremos, puesto que el efecto del diseño nos hace aumentar el tamaño de la muestra para conseguir el mismo grado de precisión.