CAPÍTULO 1
INTRODUCCIÓN

1.1 OBJETO DE LA ESTADÍSTICA

La Estadística constituye una disciplina científica extremadamente amplia y que puede ser conceptualizada desde enfoques diferentes e incluso contrapuestos. No es raro, por tanto, que se hayan propugnado para ella distintas definiciones que, en el fondo, implican diferentes visiones sobre lo que constituye la característica esencial de esta ciencia.

Desde nuestro punto de vista, en un texto cuyo objetivo es la enseñanza de la Estadística a futuros ingenieros, una definición adecuada, que sintetiza las propuestas entre otros por Gnanadesikan y Bisgaard, podría ser la siguiente:

> *"La Estadística es la ciencia cuyo objeto es la obtención y el análisis de datos mediante el recurso a modelos matemáticos y a herramientas informáticas"*

En nuestra opinión es precisamente la palabra "**datos**" la clave en la definición de la Ciencia Estadística. Como afirma Joiner *"el foco de nuestra ciencia son los **datos**, no la variación aleatoria ni la probabilidad"*.

La Estadística no es una rama de las Matemáticas, aunque el recurso al lenguaje de las Matemáticas sea fundamental en ella, de la misma forma que la Mecánica o la Termodinámica tampoco son una rama de las Matemáticas pese al extensivo uso que en las mismas se hace de modelos matemáticos. Es un error considerar que la importancia real de las técnicas estadísticas es proporcional a su complejidad matemática. De hecho métodos gráficos y otras herramientas de análisis descriptivo son extremadamente útiles para la interpretación de datos reales, pese a la sencillez de su aparato matemático.

La definición que hemos avanzado también resalta el papel crucial que en la Estadística aplicada moderna desempeñan las herramientas informáticas. De hecho sin la existencia del software adecuado la mayor parte de la metodología estadística que utilizan los ingenieros (modelos de regresión múltiple, análisis multivariante, series temporales, análisis de fiabilidad, etc...) sería inaplicable. El recurso a paquetes estadísticos es imprescindible tanto para la estimación como para la validación de la mayoría de los modelos estadísticos avanzados. No es concebible, en nuestra opinión, que pueda explicarse una Estadística útil a futuros ingenieros sin recurrir a estas herramientas informáticas.

1.2 POBLACIONES

En la terminología estadística se denomina población al conjunto de todos los individuos o entes que constituyen el objeto de un determinado estudio y sobre los que se desea obtener ciertas conclusiones.

Ejemplo 1: en un estudio sobre la intención de voto de los ciudadanos españoles, la población la constituirá el conjunto de los aproximadamente 35 millones de españoles con derecho a voto.

Ejemplo 2: en un estudio sobre el desarrollo de la tristeza de los cítricos en la Comunidad Valenciana, la población estará formada por la totalidad de árboles de cítricos existentes en esta Comunidad.

Ejemplo 3: al realizar en una industria el control de calidad en recepción de una partida de piezas, la población estará constituida por la totalidad de las piezas que componen la partida.

Los ejemplos anteriores tratan en todos los casos de poblaciones con una existencia física real, constituidas por un número finito, aunque posiblemente muy elevado, de individuos.

Aunque pueda parecer sorprendente no es ésta la situación más frecuente en la práctica, sino que en general las poblaciones a estudiar son de carácter abstracto, fruto del necesario proceso de conceptualización que debe preceder al estudio científico de cualquier problema real.

Ejemplo 4: Un ejemplo trivial sacado de los juegos de azar sirve para ilustrar la idea anterior. Se desea estudiar si un dado es correcto o está trucado. ¿Qué quiere decir la afirmación de que el dado es correcto? En la práctica, que si se tira un número muy elevado de veces los seis resultados posibles saldrán aproximadamente con la misma frecuencia. Al abordar este problema nos referiremos a la población abstracta constituida por infinitos lanzamientos del dado en cuestión, población sobre la que deseamos estudiar si las frecuencias relativas con las que se presentan los seis resultados posibles son idénticas.

Ejemplo 5: En una estudio sobre el contenido de agua en los tanque de acrilonitrilo fabricados por una empresa, la población sobre la que interesa obtener resultados podrían constituirla todos los tanques que pueda fabricar en el futuro la empresa.

Ejemplo 6: En un estudio sobre la eficiencia de diversos algoritmos de encaminamiento de mensajes entre nudos en una red de procesadores, la población a investigar la constituirían todos los mensajes que puedan llegar a generarse en la red.

Como se desprende de los ejemplos anteriores los "individuos" que forman una población pueden corresponder a entes de naturaleza muy diversa (personas, árboles, piezas, lanzamientos de dados, parcelas, mensajes, etcétera...). En los casos de los tres primeros ejemplos dichos individuos tienen una existencia real, previa a la realización del estudio. En casos como los de los ejemplos 4 5, y 6 los individuos que constituyen la población pueden irse generando mediante la realización de un determinado proceso (lanzar un dado, fabrica un tanque de acrilonitrilo, emitir un mensaje desde un nudo,...). A estos procesos, que en sucesivas realizaciones pueden

ir generando los diferentes individuos de la población les denominamos experimentos aleatorios

1.3 VARIABLES ALEATORIAS

1.3.1 Concepto

¡En toda población real existe VARIABILIDAD! Unos españoles votan a ciertos partidos y otros a otros; unos naranjos tienen tristeza y otros no; una determinada dimensión varía algo de una pieza a otra; el número que sale al lanzar el dado varía de unas tiradas a otras; el rendimiento obtenido varía de unas parcelas a otras; unos mensajes tienen retardos más elevados que otros,...

A cualquier característica que puede constatarse en cada individuo de una población se le denomina característica aleatoria. Así el partido a que piensan votar los individuos (Ejemplo 1) la ausencia o presencia de tristeza en los árboles (Ejemplo 2), el contenido de agua en los tanques (Ejemplo 5) o el retardo de un mensaje (Ejemplo 6) son características aleatorias. Muchas características aleatorias se expresan numéricamente (como el número de puntos obtenidos al lanzar un dado, el contenido de agua en un tanque o el retardo de un mensaje); a este tipo de características aleatorias se las denomina variables aleatorias.

Cuando una característica aleatoria es de tipo cualitativo (como por ejemplo el partido político a votar) nada impide codificar numéricamente sus diferentes alternativas y tratarla como una variable aleatoria. Hay que tener, sin embargo, cuidado en estos casos, porque operaciones perfectamente legítimas con características intrínsecamente numéricas (como, por ejemplo, sumar y promediar los rendimientos de diferentes parcelas) carecerían de sentido en este caso.

Autoevaluación: ¿Qué sentido práctico tendría el resultado de sumar y promediar los códigos de los problemas considerados más importantes por un conjunto de individuos?

1.3.2 Variables discretas y variables continuas

Cuando el conjunto de los valores que podría tomar una determinada variable aleatoria es discreto (es decir, finito o infinito numerable) se dice que dicha variables es de tipo discreto (a veces a las variables de este tipo se les denomina también atributos), por oposición a aquellos casos en que dicho conjunto es un infinito continuo en los que la variable se denomina continua.

Ejemplos de variables discretas serían el número de puntos al lanzar un dado, el número de picadas de ceratitis en cada naranja de un huerto, el número de errores en un programa de ordenador y también cualquier variable que se origine al codificar las diferentes alternativas de una característica cualitativa (sexo, partido votado, etcétera..).

Ejemplos de variables continuas serían todas las características que se miden sobre una escala de naturaleza básicamente continua (estaturas, pesos, rendimientos, tiempos, resistencias, etcétera...)

1.3.3 Variables k-dimensionales

Cuando sobre cada individuo de la población se estudian K características diferentes (todas ellas expresables numéricamente) se tiene una variable aleatoria K-dimensional. Por ejemplo si en la población constituida por los estudiantes de la UPV se estudia el sexo (codificado como 1 ó 2), la edad, la estatura y el peso, estaremos ante una variable aleatoria de dimensión 4.

En estos casos es frecuente utilizar los valores de aquellas componentes cuya naturaleza intrínseca es cualitativa (por ejemplo, el sexo) para dividir la población inicial en subpoblaciones (en nuestro caso: chicos y chicas) entre las cuales interesa estudiar las diferencias en las pautas de variabilidad existentes en las otras componentes de la variable aleatoria (ejemplo: ¿cómo difieren las pautas de variabilidad del peso o la estatura entre chicos y chicas en la UPV?).

Es importante darse cuenta de la diferencia entre una variable aleatoria K-dimensional, en la que las K variables se miden sobre los individuos de una única población, y un conjunto de K variables aleatorias unidimensionales, definidas sobre K poblaciones distintas.

Autoevaluación: El contenido en zumo y el calibre de las naranjas de un huerto ¿constituyen una variable aleatoria bidimensional? ¿Y el número de líneas de código y el número de errores en los programas preparados por una empresa de software? ¿Y el contenido de leucocitos en la sangre de individuos alcohólicos y no alcohólicos? ¿Y las estaturas del marido y de la mujer en los matrimonios jóvenes de un país? (Ver respuesta en el Anejo al final del Tema)

Autoevaluación: la definición clara de la población sobre la que se desea obtener conclusiones es el primer paso de cualquier estudio. El alumno deberá plantearse 3 problemas que le interesen de su vida cotidiana y definir en cada caso, con la mayor precisión posible, la población implicada y la(s) variable(s) aleatoria(s) implicada(s), analizando su naturaleza discreta o continua.

Autoevaluación: En el estudio de insecticidas se define la LD50 (Dosis Letal 50) de un producto como aquella dosis mínima que administrada a ratas provoca la muerte al 50% de las mismas. Al estudiar la LD50 de un determinado producto: ¿Cuál es la población implicada, y cual la variable aleatoria considerada? (Ver respuesta en el Anejo al final del Tema)

Autoevaluación: En una factoría interesa cuantificar, con el fin de controlar el consumo de energía (utilizada en su mayor parte en la climatización de las naves), la relación existente entre el consumo diario de electricidad y la temperatura media del día correspondiente. ¿Cuál es en el contexto anterior la población implicada y la variable aleatoria considerada? (Ver respuesta en el Anejo al final del Tema)

1.4 MUESTRAS. DATOS ESTADÍSTICOS

En general no resulta posible estudiar la totalidad de los individuos de una población para obtener información sobre ésta. Incluso cuando esta posibilidad existe técnicamente, como es el caso al tratar de poblaciones reales finitas, dicho procedimiento suele ser impracticable por consideraciones económicas.

En consecuencia para obtener información sobre una población hay que limitarse a analizar sólo un subconjunto de individuos de la misma. A éste subconjunto se le denomina muestra.

La forma de seleccionar los individuos que han de constituir la muestra tiene, como es lógico, una importancia capital para garantizar que ésta permita obtener conclusiones

que puedan extrapolarse validamente a la población de la que la muestra procede. No hay que olvidar nunca que el objeto final del estudio es siempre la población y que la muestra es sólo un medio para obtener información sobre ésta.

Con el fin de permitir inferir conclusiones válidas sobre una población la muestra debe ser "representativa" de ésta. En teoría la única forma de garantizar la representatividad de una muestra es seleccionando al azar los individuos que la van a componer, de forma que todos los individuos de la población tengan "a priori" una probabilidad idéntica de pertenecer a la muestra. Aunque ésta forma de proceder rara vez sea aplicable de forma estricta en la práctica, siempre hay que extremar las precauciones para que la forma real de obtener la muestra sea lo más parecida posible a la ideal.

En realidad en muchos casos un conocimiento previo sobre la población es indispensable para decidir si una muestra puede considerarse o no representativa de la misma.

Autoevaluación: se desea estudiar la relación que existe entre la estatura y el peso en la juventud española. El conjunto de los alumnos matriculados en Estadística en 3º en la ETSIA de Valencia ¿puede considerarse una muestra representativa de la población a efectos del estudio en cuestión?
Dicho conjunto ¿puede considerarse una muestra representativa para estudiar las tendencias políticas en la juventud española? ¿Y para estudiar el nivel cultural? ¿Y para estudiar la característica aleatoria color de los ojos?

Cuando la población estudiada es real (Ejemplos 1, 2 y 3 del Apartado 2.1) la muestra se forma, como hemos señalado, seleccionando de la forma más aleatoria posible un conjunto de individuos de la misma. Cuando se muestrea una población abstracta, del tipo de las mencionadas en los Ejemplos 4, 5 y 6, la forma de "extraer" una muestra no es más que realizar un cierto número de veces el experimento aleatorio que genera los individuos de la población (por ejemplo: lanzar varias veces el dado, plantar unas cuantas parcelas con la variedad en estudio o generar un conjunto de mensajes en la red de multiprocesadores).

Autoevaluación: plantear cómo se podría obtener una muestra representativa en cada una de las tres poblaciones definidas por el alumno en la segunda autoevaluación del Apartado 1.3..3.

1.5 ESTADÍSTICA DESCRIPTIVA E INFERENCIA ESTADÍSTICA

Los valores observados para la variable aleatoria en los individuos que forman la muestra constituyen los **datos estadísticos**.

El tratamiento de dichos datos con el fin de poner de manifiesto sus características más relevantes es el objeto de la **Estadística Descriptiva**.

Con esta finalidad la Estadística Descriptiva utiliza herramientas sencillas como:

- Tabulaciones adecuadas
- Cálculo de parámetros que sinteticen las distintas características de las pautas de variabilidad observada (posición, dispersión, asimetría, grado de relación, etcétera...)
- Representaciones gráficas (histogramas, diagramas Box-Whisker, diagramas de dispersión, ...)

En estos análisis meramente descriptivos no se pretende obtener conclusiones de carácter general sobre la población de la que procede la muestra, sino simplemente

sintetizar o visualizar, de forma que puedan captarse con claridad, los aspectos esenciales de las pautas de variabilidad existentes en los datos y que quedan frecuentemente enmascarados bajo un alud de números ("los árboles no dejan ver el bosque")

El carácter elemental de muchas de las técnicas de Estadística Descriptiva, no debe llevar a minusvalorar su gran importancia práctica. Un análisis descriptivo adecuado de los datos es siempre el primer paso en cualquier buen estudio estadístico.

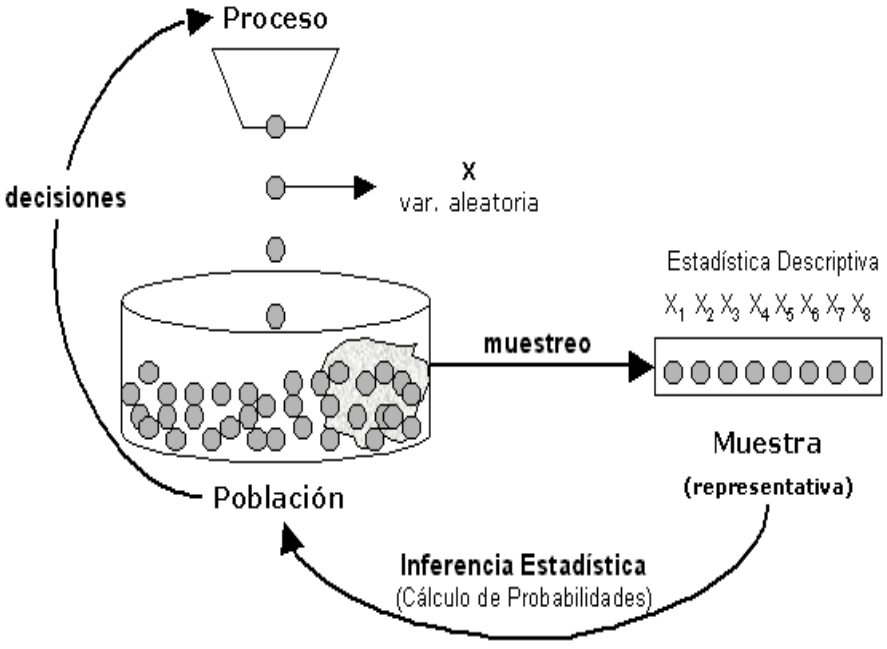
En general, sin embargo, el objetivo final de un estudio estadístico es la población muestreada, y la muestra es sólo un medio para llegar a conocerla. El análisis de los datos con el fin de obtener conclusiones que sean aplicables a la población de la procede la muestra (y, en contextos industriales, al proceso que genera los individuos de dicha población) constituye el objeto de la **Inferencia Estadística**.

Dado que sus conclusiones se basan en el análisis de sólo una parte de los individuos de la población, los resultados de cualquier inferencia estadística llevan siempre asociados un determinado margen de incertidumbre. El análisis estadístico permite, sin embargo, conocer y acotar este margen de incertidumbre.

Para realizar sus inferencias, la Ciencia Estadística se basa en los modelos matemáticos desarrollados en la Teoría del Cálculo de Probabilidades, que constituye por tanto una herramienta esencial para la Estadística.

Más allá de estos fundamentos matemáticos, el requisito esencial para que cualquier inferencia estadística sea correcta es que la muestra analizada sea **representativa** de la población sobre la que se desea inferir conclusiones. La obtención de muestras representativas, salvo en los casos en los que es físicamente posible seleccionar realmente al azar los individuos de la población que formarán la muestra, exige en general un buen conocimiento de la población a estudiar o del proceso que la genera.

En el siguiente esquema se sintetizan las ideas que acabamos de exponer



1.6 ENCUESTA

Con el fin de disponer de un conjunto de datos reales que puedan ser utilizados en diversos ejercicios, se responderá de forma anónima a las siguientes preguntas. Cada respuesta se realizará escribiendo el dígito o número correspondiente en el espacio previsto a la derecha.

1-SEXO (1-Varón 2-Mujer).....____

2-EDAD (en años).....____

3-MES DE NACIMIENTO (1 a 12).....____

4-ESTATURA (en centímetros).....____

5-PESO (en kgs).....____

6-POLITICAMENTE TE CONSIDERAS UNA PERSONA DE:

1-Derechas 2-Centro 3-Izquierda 4-Pasas del tema.....____

7-ESCRIBE UN DIGITO AL AZAR DE 0 A 9.....____

8-LUGAR DE RESIDENCIA DURANTE EL CURSO:

1-Hogar familiar
2-Colegio Mayor o residencia
3-Piso con compañeros
4-Pensión
5-Otra solución
.....____

9-¿COMO VIENES HABITUALMENTE A LA UNIVERSIDAD?

1-En tu coche
2-En tu moto o bici
3-Andando
4-En un coche de un compañero
5-Transporte público
.....____

10-¿CUANTOS MINUTOS HAS TARDADO HOY EN VENIR A LA UPV?..____

11-¿CUAL DE LOS SIGUIENTES PROBLEMAS CONSIDERAS MAS IMPORTANTE EN LA ESPAÑA ACTUAL?

1-Drogas
2-Paro juvenil
3-Terrorismo
4-Desigualdad social
5-Perdida de valores morales
.....____

Nota: las respuestas dadas a este cuestionario por los alumnos de un curso de la UPV en el curso académico 1989/90 están guardadas en el archivo Statgraphics curs8990.sf3[1]. Algunas de estas respuestas serán utilizadas como ejemplos a lo largo de este texto en diversas ocasiones

[1] Todos los archivos de datos que se mencionan en este texto pueden bajarse libremente de la URL http://personales.upv.es/rromero/descargas. Los alumnos de la Universidad Politécnica de Valencia pueden también bajárselos del Poliformat de la asignatura.