Introducción a la Estadística

Hugo S. Salinas

Fuente: probabilidadyestadistica_isi@yahoo.com.ar

Objetivo de la Estadística

Diariamente los medios de comunicación "bombardean" con datos. Las "estadísticas" se nutren de los números generados por espacios informativos, publicidad, resultados de eventos deportivos, sondeos de opinión, debates públicos, etc.. Las organizaciones modernas tienen gran variedad de datos en sus archivos de documentos y en las computadoras. Cientos o miles de valores se agregan a este total todos los días.

Algunos de los datos nuevos se generan normalmente durante el registro de las actividades; otros son el resultado de estudios e investigaciones especiales.

Sin los procedimientos estadísticos, ninguna organización podría transformar en información útil la gran cantidad de datos generados por su actividad.

*El **análisis estadístico** nos provee un conjunto de principios y procedimientos para manipular, resumir e investigar datos con el fin de obtener información útil en la toma de decisiones.*

El tratamiento estadístico de los datos requiere el uso de computadoras. Este material de trabajo proporciona numerosos ejemplos de salidas de computadora, resueltos con el software estadístico R (software gratis que puedes bajar desde la página oficial The *R* Project for Statistical Computing).

Bajar desde aquí: http://dirichlet.mat.puc.cl/bin/windows/base/R-2.9.2-win32.exe

Áreas de Aplicación de la Estadística

En todas las profesiones es importante la recolección y el estudio de datos:

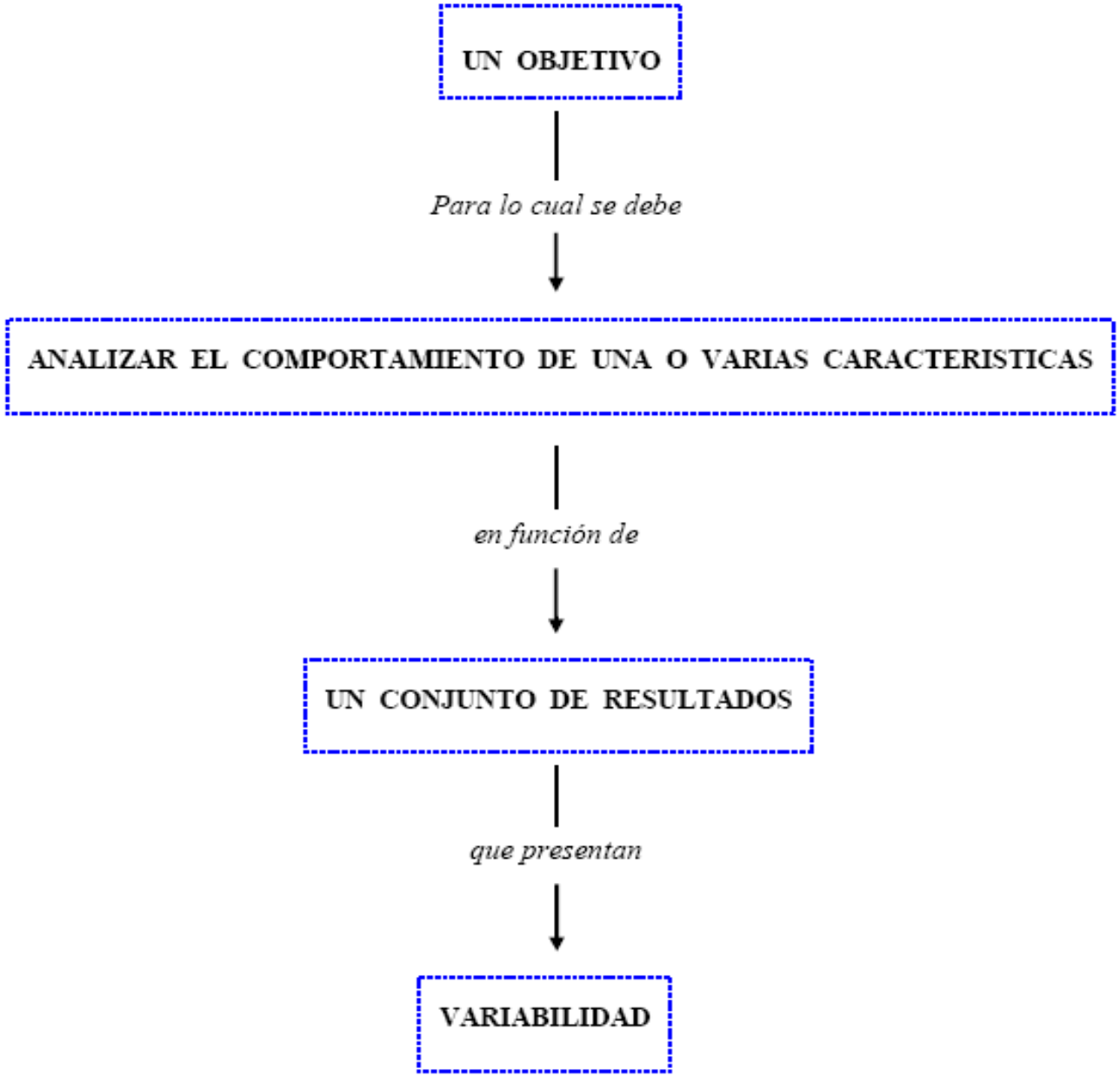
- Los **ingenieros** de control de calidad recopilan datos sobre la fiabilidad de partes y productos fabricados, calidad de procesos, etc. para mejoramiento del producto.
- Las oficinas de estadística del **gobierno** publican cada mes nueva información numérica sobre la inflación y el desempleo, a través de índices de precios, tasa de desempleo, etc.
- Quienes se dedican a realizar previsiones, los **economistas**, los asesores financieros y los que determinan las políticas de una empresa, industria y del gobierno estudian estos datos para tomar decisiones basadas en la información obtenida.
- Con el fin de ofrecer un tratamiento adecuado a sus pacientes, los dentistas, los **médicos** y en general el personal de un centro de salud, deben entender la información estadística de las investigaciones que se publican en las revistas médicas sobre efectos de nuevas drogas, tratamientos de enfermedades, etc.

Áreas de Aplicación de la Estadística cont.

- En **política**, los funcionarios que ocupan cargos directivos consideran las estadísticas de la opinión pública para definir la legislación que quieren sus votantes.
- Las **empresas** basan sus decisiones en estudios de mercado sobre los patrones de compra de los consumidores, pruebas de nuevos productos, etc.

- De acuerdo con la experiencia, virtualmente toda persona involucrada en la toma de decisiones necesita conocimientos de **análisis estadístico**. Muy frecuentemente, en compañías grandes, se utiliza la **estadística** en forma habitual. Cuando se solicita personal para esos trabajos, se piden conocimientos sólidos de **análisis estadístico**.
- En cualquiera de estos u otros ejemplos se puede observar que tanto el registro de los datos que interesan, como su manejo o utilización, no siempre es simple y se necesitan procedimientos adecuados para llevarlos a cabo.

Áreas de Aplicación de la Estadística cont.



Ejemplo de Aplicación

- Antes de cada acto electoral se efectúan encuestas de la opinión pública a fin de obtener información sobre **la proporción de población que votará por cada candidato (objetivo).**
- Consultar a todos los votantes para lograr este objetivo, es obvio que sería una labor imposible; como única alternativa se investiga una muestra de ellos con la expectativa de que la proporción de votos para cada candidato en la muestra, se aproxime lo más posible a la correspondiente proporción en la población.
- Este es un ejemplo típico de **inferencia estadística: a partir de la proporción muestral se** infiere la correspondiente proporción poblacional.
- Como lo advertiría cualquier investigador de la opinión pública se trata de un trabajo **incierto.** Para tener **seguridad respecto a la proporción de votos de cada candidato en la población es** preciso esperar hasta que se cuenten todos los votos el día de la elección.
- Sin embargo, si el muestreo se realiza en forma imparcial y adecuada es **probable que la proporción muestral se aproxime a la proporción poblacional.**

Ejemplo de Aplicación

Ante este problema nos podemos preguntar:

- ¿Cómo obtener una muestra imparcial y adecuada?
- ¿Qué error se puede estar cometiendo al inferir sobre la población muestreada a partir de la información que nos da la muestra?
- ¿Qué seguridad tenemos de estar en lo cierto?

- Este problema representa la esencia del curso y se trabajará específicamente a lo largo de los capítulos.

Definiciones

- **Población:** es el grupo total de objetos (elementos, personas, registros, instituciones, períodos de tiempo, etc.) acerca del cual se obtienen conclusiones. En cuanto al tamaño, una población puede ser **finita** o **infinita**.

- **Variable:** es la característica de interés que interesa observar en la población en relación al objetivo de estudio. Se puede clasificar en:

  - **Variable cuantitativa** (o simplemente variable)**:** es aquella cuyos valores surgen naturalmente como cantidades numéricas. Ej.: salario, edad, diámetro, número de clientes, etc. A su vez se clasifica en:

    - **discreta:** cuando sólo puede asumir valores aislados (asociada generalmente a situaciones de conteo).
    - **continua:** cuando puede asumir cualquier valor en el intervalo real.

  - **Variable cualitativa** (o atributo): sólo puede clasificarse o a lo sumo jerarquizarse. Ej.: sexo, raza, nivel de instrucción, etc.

- **Unidad de análisis:** es cada uno de los objetos sobre los que se realiza la observación de una o más variables.

- **Censo:** es un intento de medir todos los elementos de una población de interés. En muchos casos el censo es impracticable, ya sea porque la población es infinita, porque la observación implica la destrucción de la unidad, por razones de costos, etc.

Definiciones cont.

- **Parámetro:** es una medida que resume información de una característica o variable. Se calcula a partir de todas las unidades de la población. Por ej.: promedio y proporción poblacional.

- **Muestra:** es una parte de la población que se usa como información.

- **Estadístico:** es una medida que resume información de una variable, pero calculada con los datos de la muestra. Por ej.: promedio y proporción muestral.

- **Inferencia estadística:** es el proceso de extraer conclusiones sobre la población basándose en la información de una muestra extraída de esa población.

Muestra

Muy frecuentemente es necesario seleccionar una **muestra** de unidades de la población, para extraer conclusiones respecto de la población en base a las observaciones muestrales.

La selección de una **muestra representativa** es un problema importante en las investigaciones estadísticas ya que ésta puede proporcionar una visión útil de la naturaleza de la población que se estudia, mientras que una muestra no representativa puede sugerir conclusiones totalmente erróneas sobre la población.

Muestra cont.

El punto esencial en el muestreo es tratar de que los elementos de la muestra representen a la población tan fielmente como se pueda. Por lo general, esta tarea es más difícil de lo que parece. Con frecuencia debe dedicarse mucho tiempo y atención al proceso de selección, ya que una vez medidos los elementos se supondrá que la muestra es representativa de la población.

Para ello, es importante que la selección de las unidades de análisis que intervengan en la muestra no esté influenciada por cuestiones de conveniencia o favoritismo.

La alternativa adecuada es utilizar el azar. Las muestras seleccionadas en forma aleatoria son **muestras probabilísticas**. En el curso se trabajará con **muestras aleatorias simples**:

> *Una **muestra aleatoria simple** se obtiene cuando se seleccionan **n** elementos de una población, de manera que todas las combinaciones posibles de n elementos de la población tienen igual posibilidad de ser elegidas.*

La **tabla de números aleatorios** proporciona listas de números generados al azar que pueden usarse para elegir muestras aleatorias.

La mayoría de las calculadoras manuales y casi todos los paquetes de computadora generan listas de números aleatorios que pueden usarse para seleccionar muestras aleatorias.

Además de la **muestra aleatoria simple,** existen otras técnicas de muestreo probabilístico apropiadas a distintas situaciones que no serán analizadas en el presente curso.

¿Por qué se toman muestras?

Se utilizan muestras y no se estudia la población total por cualquiera de las razones siguientes:

- ✓ Recursos limitados
- ✓ Datos disponibles limitados.
- ✓ Prueba destructiva
- ✓ Mas exactitud

1. La limitación de los **recursos** (tiempo, dinero, etc.) desempeña siempre un papel importante que justifica el uso de muestras. Si la **población es grande**, el censo ocasiona un costo elevado y muchas veces, aunque económicamente se pudiera realizar, llevaría tanto tiempo que la información no resultaría de interés. En este mundo tan cambiante, el muestreo permite conseguir la **información rápidamente** en un momento determinado.

2. A veces, independientemente de los recursos, **sólo existe una pequeña muestra**. Por ejemplo, se puede tener a prueba una máquina que se supone más eficiente que otras, para decidir si se compran unidades semejantes. El gerente de control de calidad sencillamente no puede esperar hasta observar la población completa de los productos de esta máquina, en lugar de ello, debe observar una muestra de productos de dicha máquina y basar su **decisión en una inferencia** que hace a partir de dicha muestra.

¿Por qué se toman muestras? cont.

3. El muestreo puede implicar una prueba **destructiva**. Por ejemplo, suponga que se desea conocer el promedio de vida de los focos producidos por una fábrica determinada. Sería insensato esperar a que todos los focos se quemaran para conocer su promedio de vida.
4. Un censo no ofrece garantía absoluta de calidad. La observación de toda la población puede ser una tarea enorme que lleve a cometer muchos más errores que cuando se observa una muestra cuidadosamente diagramada. Por ejemplo, una gran cantidad de personal poco capacitado puede cometer **errores de medición** que no cometería una menor cantidad de personal mejor capacitado.

Errores de las muestras

En el ejemplo de las encuestas previas a la elección, puede suceder que la proporción de votos obtenida por cada uno de los candidatos en la muestra, quizás represente muy mal a la correspondiente en la población, por distintas razones:

a) Independientemente de lo bien dirigido y diseñado que esté el procedimiento de **muestreo**, puede ocurrir que se obtenga una muestra de votantes "no representativa" de la población. Estos casos de mala suerte son **posibles pero no probables.**
b) El **muestreo** puede estar mal diseñado. Por ejemplo, cuando se muestrea una población de votantes es erróneo tomar sus nombres de una **guía telefónica**, puesto que quedarán excluidos los votantes que no poseen teléfono.

Recolección de Datos

Los datos se pueden obtener por observación o por experimentación. Si simplemente se observa la característica de interés sin intervenir en el proceso en estudio, se está ante un **estudio observacional. En cambio si se interviene en el proceso en estudio imponiendo algún** tratamiento en forma deliberada sobre las unidades de análisis a fin de observar las respuestas, se está ante un **experimento.**

Según **la fuente**, los datos pueden ser primarios o secundarios. Los **datos primarios se** recogen específicamente para el análisis deseado. Los **datos secundarios ya se han** compilado y están disponibles para el análisis estadístico.

La ventaja de usar datos secundarios para una **investigación estadística** es que ya se dispone de ellos y no es necesario recogerlos para un proyecto específico. Incluso la compra de los datos a una compañía comercial es por lo general menos costosa que obtener datos primarios.

La desventaja de los datos secundarios es que estas fuentes no siempre cubren las necesidades específicas del análisis y además **no siempre son confiables**. Esta es la razón por la que muchos investigadores requieren obtener datos primarios orientados específicamente al asunto que se está investigando.