Capítulo 1

Introducción

En la actualidad, vivimos constantemente bombardeados por cifras, datos e información numérica de todo tipo. La interpretación de esta información puede resultar una tarea difícil, si además recordamos cuántas veces se nos pretende manipular con ella. Para la obtención, interpretación y evaluación de toda esta información resulta indispensable el conocimiento de los métodos estadísticos.

La Estadística pretende determinar formas eficientes de obtener información sobre un fenómeno o población y cómo analizar dicha información para hacer inferencias sobre la población, siempre tomando en cuenta la presencia de perturbaciones originadas por el azar, ya sean inherentes al objeto de estudio o debidas a errores de medición. Los métodos y conceptos que desarrolla la estadística pueden aplicarse, con las modificaciones adecuadas, en muchos campos diferentes, lo cual representa uno de los principales atractivos de la materia.

La Estadística actúa como puente entre los modelos matemáticos y los fenómenos reales. Un modelo matemático es una abstracción simplificada de una realidad más compleja, y por lo tanto siempre existirán discrepancias entre la información suministrada por el modelo y las observaciones reales, las cuales consideraremos producto del azar. La Estadística proporciona una metodología para evaluar esas discrepancias y determinar la bondad del modelo.

1.1 Algunos problemas que resuelve la Estadística.

Entre los problemas que trata de resolver la Estadística podemos mencionar los siguientes:

- **Descripción de datos:**
  Es muy poca la información útil que podemos obtener simplemente observando una tabla de datos. Necesitamos, entonces, procedimientos para resumir eficientemente la información, ya sean de tipo gráfico o numérico. En este caso, suele hablarse de *Estadística Descriptiva.*

- **Elección y análisis de muestras:**
  Al estudiar una población o fenómeno, en general es imposible (o muy costoso) obtener información sobre todos los individuos de la población o repetir un número grande de veces un experimento. Por consiguiente, nos vemos en la necesidad de usar información parcial, y deseamos que ésta sea lo mejor posible. Nos enfrentamos, entonces, al problema de cómo elegir un subconjunto significativo de la población (*Teoría de Muestreo*) o de cómo diseñar un experimento que nos proporcione la mayor cantidad posible de información sobre el fenómeno de interés (*Diseño Experimental*). Así mismo, deseamos utilizar esta información parcial para obtener inferencias sobre el total de la población o fenómeno estudiado en base a los resultados de la muestra. Para ello, suponemos que el azar afecta los resultados que hemos obtenido y empleamos *Modelos Probabilísticos.*

- **Contraste de Hipótesis:**
  Cuando se desea probar la validez de alguna hipótesis, es necesario recabar información que sea pertinente a dicha hipótesis y que nos permita observar el fenómeno particular de interés. Para ello, es necesario en general el *Diseño de Experimentos* con el fin de obtener información relevante al problema. De la misma manera, una vez recabada dicha información, es necesario disponer de métodos que permitan la interpretación de los resultados, es decir, que permitan decidir si la información obtenida apoya o contradice la hipótesis planteada.

- **Medición de relaciones entre variables.**
  En muchos casos es interesante estudiar las relaciones entre las diferentes variables que intervienen en el problema. Por ejemplo, cómo se relaciona el precio del dólar con el índice de inflación, o cuál es el efecto de la alimentación sobre el incremento de peso de los pollos en una granja avícola. De esta manera, al conocerse una de las variables puede predecirse el valor de la otra. Para hacer esto, recurrimos al ajuste de *Modelos Probabilísticos*, pues suponemos la presencia de perturbaciones en la información, las cuales atribuímos al azar.

- **Predicción.**
  En muchas ocasiones deseamos predecir eventos futuros, como por ejemplo cuál será el precio de un barril de petróleo durante el próximo año, o cuánto alcanzará la inflación. La estadística desarrolla metodologías para realizar estas predicciones dentro de ciertos márgenes de error conocidos.

- **Decisión.**
  Ante cualquier situación, distintas decisiones producirán ganancias o pérdidas diferentes. ¿Cómo podemos elegir aquella decisión o línea de acción que produzca la mayor ganancia esperada? A esta rama de la Estadística suele denominársele *Teoría de Decisión.*

1.2 Fases de un Análisis Estadístico.

Cuando realizamos un análisis estadístico para un determinado problema, podemos identificar varias etapas fundamentales en el proceso:

1. **Planteamiento del problema:**
   En esta primera etapa, es necesario definir claramente los objetivos del estudio a realizar, con el fin de poder relacionarlos con valores numéricos de variables observables. Por ejemplo, si el objetivo de nuestro estudio es relacionar el desempeño de un estudiante de la USB con respecto a su nota de admisión, debemos fijar de antemano cómo vamos a medir ese desempeño: índice académico, índice de aprovechamiento, etc. Vale destacar que no siempre es fácil conseguir medidas adecuadas para el problema específico que se desea estudiar, y en muchos campos se dedican grandes esfuerzos de investigación a la búsqueda de medidas numéricas que permitan cuantificar ciertas situaciones (desnutrición, gravedad de enfermedades, etc.)

2. **Construcción del modelo:**
   A continuación, debemos establecer qué tipo de relación creemos que poseen entre sí las variables que se especificaron en el paso anterior. Esta relación nos permitirá plantear un *modelo*, es decir, una relación matemática entre las variables con la cual podremos describir el fenómeno bajo estudio. En general, consideraremos modelos que poseen una parte sistemática (componente de señal) y una parte aleatoria (desviaciones a partir de la parte sistemática o ruido), es decir:

$$\text{Y} = \text{Parte Sistemática} + \text{Parte Aleatoria}$$

donde Y es la variable de interés. Cuando un modelo carece de parte aleatoria, nos encontramos ante un modelo *determinístico* (por ejemplo, las leyes de la física o la química). Cuando consideramos la componente aleatoria, hablamos de un modelo *probabilístico*. Son estos últimos los modelos de interés para la Estadística.

3. **Recolección de la información muestral:** Una vez definido el problema y planteado el modelo, necesitamos recolectar información que nos permita estudiar su validez. Esta información puede provenir de una población finita, en cuyo caso es necesario diseñar una muestra que represente lo mejor posible las características de la población, o bien puede provenir de un experimento, el cual debe ser diseñado de tal manera que la información obtenida sea pertinente al fenómeno estudiado y al modelo que se ha establecido previamente.

   Llegado este punto, es necesario alertar contra la práctica usual de obtener primero la información y establecer a posteriori el modelo que se desea usar. Esto puede tener dos consecuencias. La primera de ellas es que los datos que se han obtenido no permitan verificar la validez de las hipótesis que se han planteado, obligando a repetir el proceso de recolección de información. La segunda es el análisis del problema únicamente en base a los datos obtenidos. En este caso, es probable plantear modelos que no se ajusten realmente a la situación estudiada, y obtener así conclusiones erróneas. La moraleja de esta historia es: *antes de recolectar los datos, ya se debe saber cómo van a ser analizados, y no lo contrario.*

4. **Depuración de la muestra:**
   Por mucho cuidado que se haya puesto en la obtención de los datos, es en general inevitable que se presenten errores de medición, transcripción, etc. Una regla empírica afirma que entre el 2% y el 5% de los datos tienen este tipo de errores. Es necesario entonces realizar un estudio previo de los datos con el fin de depurar la muestra. Para ello, suelen emplearse técnicas de estadística descriptiva, y en general, lo que se conoce como Análisis Exploratorio de Datos (EDA).

5. **Estimación de los parámetros del modelo:**
   Todo modelo involucra cantidades desconocidas o *parámetros*. Es necesario disponer de valores, al menos aproximados, de estos parámetros con el fin de poder usar este modelo para hacer predicciones o tomar decisiones. Diversas técnicas permiten, en base a los datos obtenidos, obtener *estimados* para los parámetros del modelo, es decir, usar ciertas funciones de los datos o *estimadores* con el fin de obtener valores aproximados para los parámetros involucrados en el modelo.

6. **Hipótesis de simplificación.**
   Dentro de la misma familia de modelos, podemos usar modelos más simples, o más complejos. Sin embargo, mientras más complejo sea un modelo, más difícil serán su interpretación y su uso. Por tanto, preferiremos el modelo más sencillo con el cual sea posible representar adecuadamente el fenómeno en estudio. Este es el *Principio de Parsimonia*: un modelo complejo se preferirá a otro más sencillo únicamente si los datos proporcionan una evidencia abrumadora en favor del modelo complejo.

7. **Crítica y diagnosis del modelo**
   Una vez escogido el mejor modelo dentro de la familia planteada, debemos ver si este modelo realmente representa la situación estudiada, es decir, debemos confirmar si las hipótesis en las cuales nos basamos para escoger esa familia de modelos son ciertas, y si el modelo que hemos elegido explica un alto porcentaje de la variación de los datos. Si esto no fuera cierto, es necesario regresar al paso 2 y estudiar nuevamente el problema para plantear un nuevo modelo o familia de modelos. El proceso se repite hasta obtener resultados satisfactorios.

Es necesario tener siempre en mente que cualquier modelo estadístico es, en algún sentido, *provisional*, y que puede ser alterado por la obtención de nueva información sobre la situación en estudio.