Prólogo

Estadística aplicada a la investigación con R está dirigido a la comunidad académica y científica que se encuentre desarrollando y aportando conocimiento a la ciencia y que, de alguna manera, requiere la aplicación de fundamentos de estadística. Es así como el propósito de este libro es facilitar las herramientas necesarias para el desarrollo en los procesos investigativos y que contribuya al logro de las metas planteadas por los interesados. El compendio del texto se encuentra con una fundamentación teórica estadística básica acompañado con una diversidad de ejemplos aplicados propios de las disciplinas del saber e implementados con el uso del *software* estadístico de R, permitiendo dar más claridad a la fundamentación teórica y profundidad a las aplicaciones planteadas. Cada tema cuenta con una serie de ejemplos aplicados de las diferentes disciplinas del saber, de tal manera que el lector visualice el amplio panorama de aplicación de la estadística. Los ejemplos y ejercicios con que cuenta el texto son de fuente propia y han sido seleccionados cuidadosamente de tal manera que le permiten al investigador afianzar los conocimientos de estadística y observar su aplicación directa.

El primer capítulo del libro induce al lector a identificar los fundamentos con que todo estudio estadístico debe contar antes de profundizar en los diferentes temas propios de la investigación. El segundo conduce al interesado a identificar y utilizar los diferentes métodos gráficos existentes para la representación de variables. El tercer capítulo presenta la clasificación de los métodos numéricos existentes para la representación de los datos cuantitativos. El cuarto hace referencia a las funciones y operaciones básicas con el uso de la herramienta estadística de Rstudio, su instalación, el uso de la sintaxis para su aplicación, una introducción al álgebra de matrices con su aplicación en R, la visualización de gráficos, así como la construcción de un *data frame* para su ejecución y la generación de la estadística descriptiva. También encontrará el mecanismo de cargue de diversas librerías y bases de datos, así como la exportación de bases a otros formatos. El capítulo quinto está dedicado a la predicción o justificación del comportamiento y tendencia de una población a través de la estimación puntual o por intervalo de un parámetro partiendo del uso de un conjunto de datos de una muestra. El capítulo seis se enfoca en la inferencia acerca del comportamiento de los parámetros de una población por medio de la aplicación de las diferentes pruebas de hipótesis. El capítulo séptimo está enfocado en la representación estadística del comportamiento de los fenómenos naturales de

tal manera que se pueda llevar a cabo su predicción y proyección por medio del modelamiento de la regresión lineal. Finalmente, el capítulo octavo presenta una ampliación del anterior al presentar la aplicación de los diferentes fenómenos de la naturaleza o del comportamiento humano por medio de un modelo de regresión múltiple.

Cualquier inquietud o sugerencia sobre el tema, por favor, contactarme a través del correo: **ingwicava@gmail.com**

Capítulo 1

Introducción

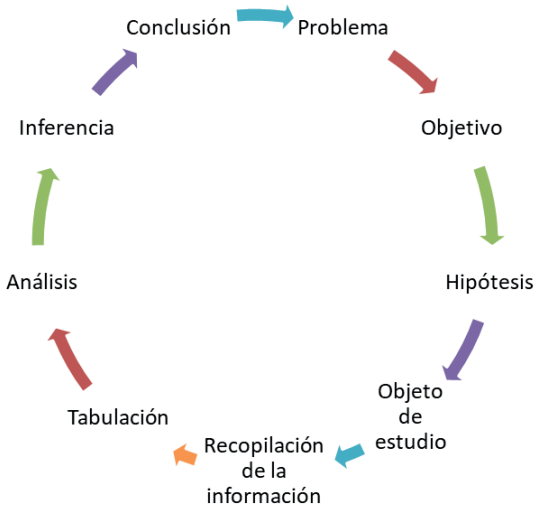
La aplicación de la ciencia de la estadística a un problema específico debe tener presente una serie de características tanto internas como externas que permitan al investigador tomar la decisión más idónea de acuerdo a los objetivos y a la hipótesis que se plantea. Para que un estudio estadístico sea confiable, se debe realizar de manera ***eficiente*** y ***eficaz*** en la medida que se cuente con los ***recursos*** y el ***tiempo*** necesarios para llevarlo a cabo. Eficiente, por cuanto se deben realizar los cálculos con las variables adecuadas que respondan a los requerimientos inicialmente planteados en el estudio y, por el otro lado, se busca que los cálculos sean confiables en los hallazgos o predicciones realizadas, es decir, que sea eficaz. Adicionalmente, se deben tener presentes los recursos con que se cuente para llevar a cabo el estudio en un periodo de tiempo establecido para tal fin y que permita lograr con éxito los resultados esperados; esto es ser eficaz.

De manera práctica, se define la ciencia de la estadística como aquella que se encarga de la selección, clasificación, ordenamiento y tabulación de los datos para su respectivo análisis, descripción e interpretación. Estos datos pueden provenir de una población o de una muestra seleccionada a la que se le aplica su análisis estadístico numérico y gráfico; esto es lo que se define como estadística descriptiva. Cuando a estos datos se les aplica los modelos estadísticos con sus leyes respectivas, se habla de estadística probabilística y cuando se infiere el comportamiento de los parámetros de una población, se está hablando de estadística inferencial (Mendenhall & Sincich, 2002).

1.1 Etapas en todo estudio estadístico

Un estudio estadístico está conformado por varias etapas, desde la concepción del problema hasta el hallazgo de sus resultados, que se pueden enmarcar dentro de las conclusiones.

Figura 1.1. Etapas en todo estudio estadístico



1.1.1 Problema

Tener conocimiento de la problemática existente y para la cual se requiere de la aplicación estadística.

En una comunidad, se han identificado numerosas muertes en adultos jóvenes por causa del cáncer de pulmón.

1.1.2 Objetivo

Definir de manera sucinta y breve las características de la problemática que se quiere identificar o comprobar.

Identificar las causas de muerte en los adultos jóvenes en la población durante el periodo XX.

1.1.3 Hipótesis

Proponer la afirmación que se quiere corroborar o contradecir respecto a una población (objeto de estudio).

Las personas fumadoras tienen mayor probabilidad de padecer cáncer de pulmón.

En una prueba de hipótesis:

H_0: todas las personas que fuman tienen cáncer.

H_1: al menos una persona fumadora no padece cáncer.

1.1.4 Definir el objeto de estudio

¿Cuál es mi universo de análisis?

Se define como el conjunto de elementos a los cuales se les desea realizar el estudio.

Conjunto de ciudadanos que habitan la región de interés de estudio.

¿Cuál es la población de estudio?

Se define como el conjunto de elementos que son susceptibles de medir y por el cual está interesado el estudio.

Población mayor de edad que habita un pueblo determinado y que se encuentra en condiciones de responder a una encuesta.

1.1.5 Realizar un censo o muestreo

Cuando se lleva a cabo el levantamiento de información a una población determinada, es necesario definir si se realiza un censo (a toda la población) o una muestra (a una parte de la población). La decisión depende de varios factores, como la disponibilidad de todos los elementos, la limitación de los recursos y el tiempo para realizar el estudio, entre otros (Wu & Thompson, 2020).

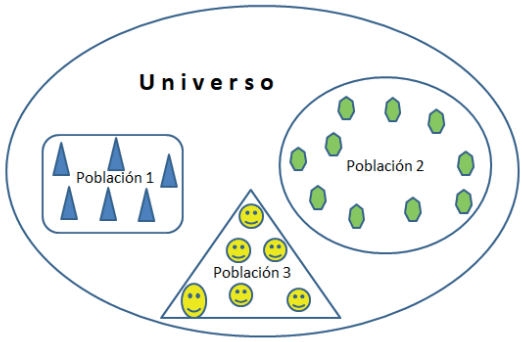
1.1.6 Selección de la muestra

Existen diversos factores que se deben tener en cuenta a la hora de seleccionar una muestra: conocer las características de la población, si los elementos que la componen poseen características muy homogéneas, la dificultad de obtener información de los miembros o elementos que forman parte de la población, los costos elevados en que se puede incurrir o las pruebas destructivas que se necesiten llevar a cabo, entre otros. Son factores que obligan a realizar un muestreo en vez de un censo (Ospina, 2001). Luego el muestreo se define como un subconjunto de datos seleccionados de la población con características de interés en común, que mantiene una representatividad sobre la población.

Conjunto de personas de una comunidad que puede ser seleccionada para diligenciar una encuesta determinada.
Selección de muestras médicas con el propósito de evidenciar algún tipo de patrón o comportamiento.
En el proceso de manufactura, se selecciona una muestra para el control de calidad de las bandas de disco para el uso en el frenado de los carros.

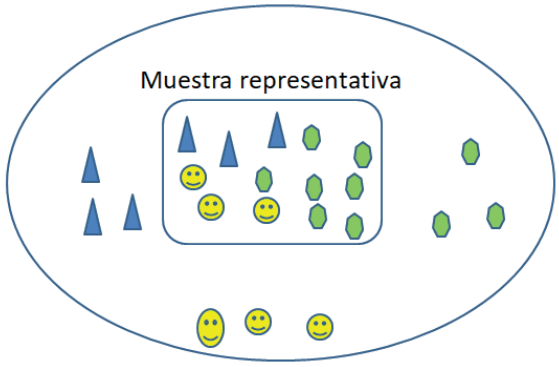
También puede dar el caso de que se tenga un universo de análisis compuesto por una población con características disímiles, pero, a su vez, con grupos que guardan cierto comportamiento o patrón por el que puedan ser agrupados (figura 1.2).

Figura 1.2. Universo de análisis: compuesto por poblaciones con diferentes características



Por ejemplo, una empresa quiere lanzar al mercado un producto específico y desea conocer la percepción que se tiene del producto de consumo en particular. Para tal efecto, pretende encuestar a un grupo potencial de consumidores para tener una percepción objetiva de la población objeto de estudio; se deben analizar las características del universo de análisis (figura 2) y seleccionar una muestra que sea representativa. Es decir, que su tamaño represente las características de la población (figura 3) y que identifique las cualidades por las cuales se desea encuestar.

Figura 1.3. Muestra representativa



1.1.7 Tipo de muestreo a aplicar

Cuando se realiza un estudio estadístico, la veracidad y confiabilidad de los resultados dependen en primera instancia del origen de extracción de la información, es decir, de cómo y a quién se le realiza el levantamiento de la información.

Llevar a cabo un muestreo no es tarea fácil y, para ello, existen diversos tipos de muestreos (Lohr, 2000a). En consecuencia, se habla de dos tipos de muestreo: el probabilístico y el no probabilístico. El primero se basa en la selección de la muestra, que se pueda definir el conjunto de las muestras posibles, es decir, conocer para cada una de las muestras posibles la probabilidad de selección, y el muestreo no probabilístico o por conveniencia es aquel donde no se aplica alguna de las consideraciones anteriores y se fundamenta en la experticia del investigador para la selección de la muestra (Ospina, 2001).

Existen diversos tipos de muestreo probabilístico (Lohr, 2000b), entre los que se destacan:

- **Muestreo aleatorio simple sin reemplazo:** se realiza cuando todas las muestras posibles tienen la misma probabilidad de ser seleccionadas y al elegir cualquiera de ellas ya no se tiene en cuenta para la selección de la siguiente. Cuando es con reemplazo, la que se selecciona se ingresa de nuevo a la población para que haga parte en la siguiente selección de la muestra, la cual puede tener la posibilidad de ser elegida nuevamente.
- **Muestreo estratificado:** se lleva a cabo cuando se divide a la población en estratos (que poseen características similares) y en cada uno de ellos se selecciona una muestra aleatoria. Es importante que estos estratos guarden una homogeneidad entre sí.
- **Muestreo sistemático:** al tener una población ordenada, se procede a seleccionar la muestra a partir de un elemento elegido y sistemáticamente se van seleccionando los demás.
- **Muestreo por conglomerados:** se seleccionan grupos de la población que guardan alguna característica en común y los elementos del grupo se seleccionan todos.
- **Muestreo en varias etapas:** es aquel que se aplica a poblaciones en donde es necesario llevar diversos tipos de muestreo para la selección de la muestra representativa.

1.1.8 ¿Qué es lo que se desea estudiar?

Responder a esta pregunta es identificar el problema que es objeto de estudio. Para ello, estadísticamente es necesario definir las variables involucradas y aquellas que conformarán la encuesta si se va a recolectar información por medio de instrumentos. Cada variable debe dar aporte y contribuir en la consecución de la respuesta al objetivo de la investigación. De aquí se desprende la importancia de un buen diseño del instrumento de recolección de información: encuesta, base de datos que como medio permite llegar a las variables de interés. La encuesta debe ser precisa, objetiva, que mida lo que debe medir: “requerir información a un

grupo socialmente significativo de personas acerca de los problemas en estudio para luego, mediante un análisis de tipo cuantitativo, sacar las conclusiones que se correspondan con los datos recogidos" (Sabino, 1992).

1.1.9 Recopilar la información

En este paso, es importante que el investigador defina la metodología a utilizar para la recolección de la información. El uso de las herramientas tecnológicas permite el ahorro de tiempo y la disminución sustancial de los errores que se cometen si solamente se llevara a cabo de manera manual.

1.1.10 Tabulación de la información

La eficiencia en la tabulación de la información permite al investigador lograr los alcances que quiera en el desarrollo de los análisis estadísticos y obtener la profundidad que desee. Para esto, ya se cuenta con *software* especializados para tal fin, como PYTHON, R, SPSS, STATA, SAS, hoja de cálculo de Excel (este último no es un *software* estadístico).

1.1.11 Análisis de la información

Análisis descriptivo: este paso es muy importante debido a la necesidad de utilizar la estadística descriptiva para el análisis de las variables; se debe tener la habilidad de no redundar en la repetición o en la obtención de información innecesaria para el estudio (Faraway, 2005).

1.1.12 Interpretación de la información

Lo más sensible en la estadística es su interpretación. Gracias a los resultados que se obtienen en la aplicación de sus modelos, se requiere del conocimiento y experticia por parte del investigador para realizar la interpretación adecuada y no caer en vicios o errores que se pueden cometer por desconocimiento.

1.1.13 Informe estadístico

De acuerdo a las necesidades y el alcance que el investigador logre, el informe debe ser lo más claro y sencillo posible, debe reflejar el logro de los objetivos dando respuesta al planteamiento de la pregunta y debe evidenciar la metodología utilizada, así como las fuentes de información utilizadas.