1

Conceptos asociados con la estadística

Históricamente, la estadística fue considerada una técnica asociada con el procesamiento de datos o de información proveniente de distintas fuentes: conteos de rebaños, producción agrícola, censos en diversas poblaciones y registros sobre el pago de impuestos, entre otras. Actualmente, la sociedad atraviesa un proceso de cambio profundo generado por la interacción de cuatro elementos: el conocimiento, la tecnología, la información y la comunicación (Mejía, 2011); en este contexto, la estadística se ha constituido en una herramienta poderosa que posibilita el procesamiento y el análisis de información, llegándose a consolidar como la "ciencia de los datos" y elemento fundamental en el método científico experimental (Aliaga & Gunderson, 2005; Batanero, 2001). Para afrontar los mencionados procesos de cambio, es razonable que los futuros ciudadanos, los docentes y profesionales en ejercicio adquieran los elementos básicos acerca de la estadística, que les permitirán tomar decisiones en situaciones de incertidumbre o fundamentados en una información procesada de forma adecuada, en distintos contextos de su actuación.

La estadística suele dividirse en dos ámbitos: la estadística descriptiva y la estadística inferencial; la primera se direcciona a describir los datos de una o más características pertenecientes a los individuos de una población o de una muestra (Valdivieso, 2011), y la segunda, a realizar afirmaciones o inferencias válidas para los individuos de una determinada población con base en la información proveniente de una muestra aleatoria (Gutiérrez & De la Vara, 2008). En este capítulo se abordan los principales conceptos asociados con la estadística, dando prioridad a los relacionados con la estadística inferencial, entre ellos: población, muestra, variable, variable aleatoria, muestra aleatoria, parámetros, estimadores y tipos de muestreo.

1.1 Algunos conceptos asociados con la estadística

A continuación, se abordan los conceptos de estadística descriptiva e inferencial, población, muestra, variable y de algunas escalas de medición; asimismo, se indican ejemplos alusivos, con el propósito de orientar al lector en el desarrollo de las actividades destinadas al estudio independiente.

1.1.1 Estadística descriptiva e inferencial

En estadística se ha de contar con un conjunto de elementos que, frecuentemente, se denominan individuos; estos pueden corresponder a objetos, personas, animales o acontecimientos reales o abstractos sobre los cuales interesa efectuar una investigación en un determinado contexto. Estos individuos conforman y delimitan la población objeto de estudio o universo; en ellos interesa investigar algunas características o variables. Cuando el proceso investigativo involucra a los individuos de una población o de una muestra, y el objetivo primordial es la descripción de algunas de sus variables, entonces se cae en el ámbito de la *estadística descriptiva;* en este caso, las actividades inherentes son: la recolección, la organización, la representación, el procesamiento, el análisis y la interpretación de información, a fin de generar conclusiones.

La recolección de información se realiza a través de instrumentos como el censo y la encuesta; el censo se hace sobre todos los individuos de la población objeto de estudio, y la encuesta, sobre los individuos que conforman la muestra. Tanto el censo como la encuesta son instrumentos conformados por un número determinado de preguntas, cada una referida a una de las características que interesa investigar. La organización de los datos suele efectuarse mediante la conformación de las denominadas tablas de frecuencia. La representación de la información se realiza a través de diagramas circulares o pasteles, diagramas de barras o histogramas, polígonos de frecuencia o de ojivas, entre otros. En el procesamiento de los datos se calculan ciertos valores, denominados parámetros o estimadores, ya sea que se traten de datos poblacionales o muestrales, por ejemplo: porcentajes, promedios, varianzas, desviaciones estándar, coeficientes de variación, medianas o coeficientes de correlación; estos valores se interpretan en el contexto del estudio, posibilitando así el análisis de la información y la obtención de conclusiones.

No obstante, la identificación de los individuos de toda la población es una actividad que resulta dispendiosa, imposible o costosa; en consecuencia, se hace necesario obtener una muestra representativa y aleatoria de esa población, y con base en ella inferir sobre las características de interés de toda la población,

de modo que ciertas afirmaciones o hipótesis se puedan comprobar y generalizar en ella; de este tipo de estudio se ocupa la *estadística inferencial*, llamada, algunas veces, inferencia estadística. Un proceso inferencial permite obtener conclusiones sobre los parámetros de una población por medio de una muestra probabilística. En este tipo de estadística se realizan actividades que guardan relación con el método científico experimental, la observación (muestreo), la formulación de hipótesis, la comprobación o prueba de hipótesis y la obtención de conclusiones. En general, se realizan dos tipos de procesos: estimación de parámetros y prueba de hipótesis.

1.1.2 Población

La *población,* o universo, corresponde a un conjunto de individuos que son objeto de estudio en un contexto bien determinado y que posibilitan observar e investigar algunas características comunes (Valdivieso, 2011). Cuando se conoce el número total de individuos se dice que la población es finita, y su tamaño se denota con N; en caso contrario, si es imposible identificar a todos los individuos involucrados en el estudio, se dice que la población es infinita. Un *individuo* es el elemento metodológico usado para estudiar una colectividad específica. Los individuos pueden ser objetos, animales, personas o entes concretos o abstractos.

Ejemplo 1.1. Frecuentemente, los individuos de la población de interés pueden ser materiales, productos ya terminados, partes o componentes de cierto tipo de electrodoméstico o los procesos que realiza una determinada empresa (Gutiérrez & De la Vara, 2008). En algunos casos, estas poblaciones se han de suponer grandes o infinitas. En empresas con producción en forma masiva es casi imposible medir cada pieza del material que será utilizado en una línea de fabricación o las propiedades de todos y cada uno de los productos terminados. Ahora, si la producción no se realiza masivamente, también conviene considerar tal proceso como una población grande, puesto que el flujo de proceso casi nunca se detiene, es decir, no se tiene el último artículo producido en tanto la fábrica esté operando. En casos como el mencionado, los procesos corresponden a poblaciones que pueden estudiarse por medio de muestras de artículos extraídos en algún momento del proceso.

Ejemplo 1.2. Se desea investigar el peso promedio de las unidades de chocolatina de tamaño mediano producidas por la máquina A por día en la empresa H, en la ciudad de Medellín, en el mes de febrero del 2016. En este caso, la cantidad de unidades es grande; sin embargo, se trata de una población finita de tamaño N.

1.1.3 Muestra

De manera intuitiva, una *muestra* es una parte representativa de la población; su tamaño suele denotarse con la letra *n* (Valdivieso, 2011). A fin de estudiar una realidad, conviene determinar un universo, o población, apropiadamente; para esto se requiere utilizar esquemas de representación de las características que los individuos comparten y que en determinado momento posibiliten seleccionar una muestra probabilística.

Un aspecto relevante consiste en obtener una muestra representativa, es decir, que contenga las características que se buscan analizar en los individuos de la población. Una manera de lograr la representatividad consiste en diseñar de forma pertinente un plan de muestreo aleatorio o al azar, cuya selección evite sesgos, en el sentido de no favorecer la inclusión de ciertos individuos particulares de forma intencionada; se busca realizar un procedimiento tendiente a garantizar que todos los individuos de la población tengan igual posibilidad de conformar la muestra (Botero, 2001; Gutiérrez & De la Vara, 2008). Existen diversos métodos para realizar un muestreo aleatorio, entre ellos, el aleatorio simple, el estratificado, el sistemático y por conglomerados; cada uno de estos posibilita la obtención de muestras representativas en correspondencia con los objetivos de investigación, de algunas eventualidades y características presentes en la población (Gutiérrez, 2005).

Ejemplo 1.3. Se han extraído de forma sistemática (cada 10 unidades), de la banda transportadora, 50 botellas de gaseosas de 200 mililitros, de la marca PP, a fin de analizar si el proceso de embotellado cumple con las características establecidas por la empresa productora de esta marca de refresco. En este caso se ha realizado un muestro sistemático.

1.1.4 Variable

Para estudiar las características de los individuos, correspondientes a una población o a una muestra, conviene utilizar variables o modelos que posibiliten representar y analizar tales características de forma razonable. De manera intuitiva, una *variable* es una representación de una característica que se quiere estudiar en esos individuos. Es conveniente aclarar que toda representación o modelo tiene sus limitaciones y genera diversas posibilidades para interpretar la información asociada con la(s) característica(s) que comparten los individuos. Las variables suelen denotarse con letras mayúsculas, ejemplo, *X, Y, Z*. Por otra parte, los datos son valores admisibles para una variable determinada y se denotan con letras minúsculas acompañadas de subíndices, la secuencia $x_1, x_2, ..., x_n$ indica

los n valores u observaciones correspondientes a una variable X asociada a una muestra; asimismo, $x_1, x_2, ..., x_N$ indica los N valores para una variable X que interese estudiarse en una población finita de tamaño N.

Ejemplo 1.4. X: género de unos estudiantes de la carrera profesional Administración de Empresas en el semestre I del año 2016, en la Universidad Pedagógica y Tecnológica de Colombia (UPTC), en la ciudad de Tunja. Los datos recolectados fueron: M, M, F, F, F, M, F, M, F, F, M, M, F, F, F, M, F, M, F, F, M, M, F, F, F, M, F, M, F, F. Donde M denota al género masculino, y F, al femenino. En este ejemplo, el tamaño de la muestra es n=30, que es el número de datos correspondiente a la variable género que se va a estudiar sobre esos individuos.

Ejemplo 1.5. Y: salario mensual en miles de pesos de los trabajadores de la Empresa H, en la ciudad de Tunja durante el año 2015. Los datos recolectados fueron: 700, 750, 690, 1000, 700, 700, 2000, 690, 800, 690, 700, 690, 800, 690, 700, 690, 800, 690, 700 y 690 miles de pesos. En este ejemplo, el tamaño de la población es N=20, este es el número de datos correspondiente a la variable salario mensual que se va a estudiar sobre los individuos de esta población de trabajadores.

En general, las variables pueden agruparse en cualitativas y cuantitativas. Las *cualitativas* son aquellas que permiten clasificar a los individuos en grupos disjuntos; por ejemplo, pueden representar características que responden la pregunta: ser o no ser; el ejemplo 1.4 corresponde a una variable cualitativa, puesto que cualquier individuo puede ser clasificado solo en uno de los dos grupos: género masculino o femenino. Por otra parte, si en las características que se requieren estudiar existe la noción de cantidad, intensidad o magnitud, entonces se representan a través de variables *cuantitativas* (Valdivieso, 2011); por ejemplo, la característica "peso" en kilogramos de los jugadores que conforman el equipo de baloncesto de la UPTC. Una variable cuantitativa es *continua* cuando admite cualquier valor en un intervalo de números reales, y es *discreta* cuando toma valores particulares en un intervalo dado o en un conjunto finito o numerable de números reales (Peña y Romo, 1997); el ejemplo, 1.5 involucra el trabajo con una variable cuantitativa.

Los datos de una variable corresponden a observaciones o mediciones ubicadas en alguna de las siguientes escalas: nominal, ordinal, de intervalo o de razón (Valdivieso, 2011). A continuación, se describen y ejemplifican cada una de ellas.

La escala nominal permite organizar a los individuos en grupos llamados "categorías", y los datos corresponden a variables cualitativas. Las categorías

organizan a los individuos en grupos exhaustivos y excluyentes, de forma que un individuo no pueda pertenecer a dos categorías al mismo tiempo.

Ejemplo 1.6. En la Institución Educativa del Este (IEE), en la ciudad A, para el año lectivo 2016 se ha registrado la siguiente matrícula por grado: quinto (5º), 25 estudiantes; cuarto (4º), 32; tercero (3º), 35; segundo (2º), 40, y primero (1º), 38 estudiantes. En total suman 170 estudiantes. En este contexto, la variable X (grado de escolaridad en la IEE) es cualitativa, se encuentra en escala nominal y presenta cinco categorías.

La escala ordinal posibilita la organización de los individuos al establecer un *orden* en las mediciones asociadas con una variable cuantitativa o al asignar un nivel de importancia en las observaciones de una variable cualitativa. El orden de los individuos se asigna iniciando desde aquel que presente menos cantidad o magnitud de la característica en estudio hasta quien tenga la mayor cantidad. En este caso, también se suelen usar las llamadas *variables con categorías ordenadas,* que se distinguen por tener un orden explícito de 8 o menos categorías, algunas de las cuales pueden responder a preguntas con las opciones: mucho, regular, poco, entre otras; también pueden corresponder a opciones de calificar una característica específica con números enteros de 1 a 5.

Ejemplo 1.7. Para la variable *Y*: asistencia de unos aficionados a partidos de fútbol del club A, en el segundo semestre del año 2015, los posibles valores de la variable se pueden asignar de acuerdo con el siguiente orden ascendente: (1) nunca, (2) pocas veces, (3) casi siempre, (4) siempre. En este caso, se trabaja con cuatro categorías ordenadas.

La escala de intervalo posibilita medir la cantidad de una característica usando la noción de "distancia" para hacer la diferencia entre dos datos cualesquiera. Esta escala presenta una unidad de medida común y constante que permite asignar un número real a todos los pares de individuos en un conjunto ordenado; la proporción de dos intervalos cualesquiera es independiente de la unidad de medida y del punto cero (Valdivieso, 2011); este punto no es indicativo de ausencia de la característica que se está midiendo, y tanto la unidad de medida como el punto cero son arbitrarios.

Ejemplo 1.8. El siguiente ejemplo fue adaptado de Siegel (1970). Se pretende estudiar la variable T: temperatura de distintos cuerpos inertes ubicados en la región A del departamento de Boyacá, en Colombia. Esta variable permite recoger datos medidos en una escala de intervalo; en este caso existen diversas escalas, como la de grados Celsius (C), Fahrenheit (F) y otras; las dos escalas

presentadas generan datos equivalentes, puesto que están relacionadas de forma lineal por medio de la siguiente ecuación: F = (9/5)C+32; en la escala de grados Celsius los puntos de congelamiento y de ebullición se alcanzan a 0 y 100 grados, respectivamente, en tanto que en la escala Fahrenheit se alcanzan a 32 grados y 212 grados; 10 grados Celsius equivalen a 50 grados *Fahrenheit*.

La escala de razón hace posible establecer una "relación" entre las cantidades de la característica evaluada al realizar división entre los valores de la variable que la representa. En esta escala, el cero es absoluto e indica ausencia de la característica, en tanto que en las variables en escala de intervalo el cero es relativo y, en consecuencia, no ha de interpretarse como la ausencia de la característica.

Ejemplo 1.9. Se quiere estudiar la variable *I*: ingresos en millones de pesos por mes obtenidos por 12 pequeñas empresas ubicadas en la ciudad de Manizales, en Colombia, en marzo del año 2015; los datos recolectados fueron los siguientes: 15, 11, 9.6, 10.4, 9.8, 13.3, 9.5, 8.9, 6.4, 0, 2.8 y 13 millones de pesos. El valor 0 es un dato de la variable *I,* este indica ausencia de ingresos en una de las pequeñas empresas (quizá estuvo sin operar); en este caso, el cero es absoluto.

1.2 Variable aleatoria y muestra aleatoria

En este apartado se presentan de manera formal los conceptos de variable aleatoria, como una función medible, y de muestra aleatoria; estos conceptos se fundamentan en el desarrollo de los procesos de inferencia estadística, tanto en lo referente a la estimación como al contraste de hipótesis.

1.2.1 Variable aleatoria

Si $\Omega \neq \phi$ denota un espacio muestral provisto de una familia $\Im$ de subconjuntos de Ω que constituya un σ – álgebra sobre Ω y haga posible la definición de una medida de probabilidad *P,* entonces la terna $(\Omega,\Im,P)$ define un espacio de probabilidad sobre Ω. Si, además, se tiene el espacio medible (R, β), donde R corresponde al conjunto de los números reales, y β es el σ – álgebra de Borel en R, entonces una variable aleatoria X es una función medible definida desde el espacio muestral hacia el conjunto de los números reales (Burbano y Valdivieso, 2015; Papoulis, 1991; Shao, 1999):

$$X : \Omega \rightarrow R$$

tal que para todo evento E en el σ – álgebra de Borel se tiene que

$$X^{-1}(E) \in \Im$$

Si R_X denota el rango de la variable aleatoria X, entonces X es discreta si R_X es un conjunto finito o contable (discreto); mientras que X es continua si R_X es un conjunto no contable en R.

Para una variable aleatoria X, definida sobre el espacio de probabilidad $(\Omega,\Im,P)$ y con valores en el espacio (R,β,P_X), y el evento determinado por

$$\{X \in B\} = \{\omega \in \Omega : X(\omega) \in B\} \text{ con } B \in \beta$$

la medida siguiente:

$$P_X(B) = P(\{X \in B\})$$

para todo $B \in \beta$, se denomina medida de probabilidad inducida por la variable aleatoria X.

Ahora, si X es una variable aleatoria discreta, definida sobre el espacio de probabilidad $(\Omega,\Im,P)$, tal que para cada $x \in R$,

$$f(x) = P_X(\{x\})$$

entonces a la función f se le denomina función de probabilidad $(f.p.)$ de la variable aleatoria X, la cual ha de cumplir las siguientes dos condiciones:

i) $f(x) \geq 0$.

ii) $\sum_{x_i \in R_X} f(x_i) = 1$.

Para la variable aleatoria continua X, si existe una función real f que satisface las dos siguientes condiciones:

i) $f(x) \geq 0$

ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

entonces la función f recibe el nombre de función de densidad de probabilidad $(f.d.p.)$ para esa variable.

Sea X una variable aleatoria real definida sobre el espacio $(\Omega,\Im,P)$. Si X es una variable discreta, entonces la función de distribución de probabilidad se denota y se define así:

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

Si X es una variable aleatoria continua con función de densidad *f,* entonces la función de distribución de probabilidad se denota y se define así:

$$F_X(x) = P_X((-\infty, x]) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

1.2.2 Muestra aleatoria

Una muestra aleatoria es un conjunto de variables aleatorias $X_1, X_2, ..., X_n$ idénticamente distribuidas, es decir, todas y cada una de aquellas tienen la misma función de distribución de probabilidad y son independientes, donde *n* es el tamaño de la muestra (Canavos, 1988; Lindgren, 1993; Mayorga, 2003).

Para una realización $x_1, x_2, ..., x_n$ o conjunto de valores correspondientes a la muestra aleatoria $X_1, X_2, ..., X_n$, lo anterior significa que,

$$f_{X_1, X_2, ..., X_n}(x_1, x_2, ..., x_n) = f_{X_1}(x_1).f_{X_2}(x_2)...f_{X_n}(x_n)$$

Ejemplo 1.10. Para una variable aleatoria *X* con distribución normal de media μ y varianza σ^2 o desviación estándar σ, la función de densidad de probabilidad es

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad x \in R$$

La variable aleatoria *X* con distribución normal de media $\mu = 0$ y desviación estándar $\sigma = 1$ se denomina variable normal estándar, y se denota así: $X \sim N(0,1)$; su función de densidad de probabilidad es

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right], \quad x \in R$$

Para una variable aleatoria *X* con distribución normal de media 100 y desviación estándar de 4, la función de densidad de probabilidad es

$$f(x) = \frac{1}{4\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-100}{4}\right)^2\right], \quad x \in R$$

Una muestra aleatoria $X_1, X_2, ..., X_n$ para esta variable *X,* cuya realización es $x_1, x_2, ..., x_n$, satisface las siguientes igualdades:

$$f(x_1) = \frac{1}{4\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_1 - 100}{4} \right)^2 \right], \; x_1 \in R$$

$$f(x_2) = \frac{1}{4\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_2 - 100}{4} \right)^2 \right], \; x_2 \in R$$

Así sucesivamente hasta obtener:

$$f(x_n) = \frac{1}{4\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_n - 100}{4} \right)^2 \right], \; x_n \in R$$

Los valores $x_1, x_2, ..., x_n$ se pueden determinar mediante procesos de simulación de valores de una variable aleatoria con distribución normal con $\mu = 100$ y $\sigma = 4$ (ver Burbano, Valdivieso y Salcedo, 2014). Cuando en un análisis estadístico se involucran más de dos variables a fin de estudiar su efecto simultáneo, se realiza un análisis multivariante (Hair y Taham, 2008).

1.3 Parámetros y estimadores

En esta sección se describen ciertos valores denominados parámetros y algunas funciones que se definen a través de las variables aleatorias que conforman una muestra aleatoria. Asimismo, se presenta una versión del denominado teorema central del límite.

1.3.1 Parámetros

Los *parámetros* son ciertos valores considerados verdaderos y válidos para toda la población objeto de estudio; se calculan con los datos de una variable en los individuos de una población determinada. Algunas medidas descriptivas correspondientes a parámetros son: la media poblacional, la varianza poblacional, la desviación estándar poblacional, el coeficiente de variación poblacional y el porcentaje poblacional, entre otras.

Si se han recolectado los datos $x_1, x_2, ...x_N$, correspondientes a una variable cuantitativa X , para ser estudiada en todos los individuos de la población de tamaño *N*, se puede calcular el parámetro llamado *media poblacional*, que se

denota con μ y se define mediante la siguiente expresión:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

El parámetro denominado *varianza poblacional* se denota y define de la siguiente manera:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

El parámetro *desviación estándar poblacional* se denota y define así:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

El parámetro *coeficiente de variación poblacional* es un valor que se calcula de la siguiente forma:

$$CV = \frac{\sigma}{\mu}$$

Ahora, si x representa el número de individuos que presentan una característica A de interés en la población, el parámetro proporción poblacional se denota y define mediante la razón:

$$p = \frac{x}{N}$$

Ejemplo 1.11. Para la variable *X,* utilidades en millones de pesos por día de las cinco empresas más eficientes en la ciudad de Paipa, en Boyacá, Colombia, en el año 2014, se obtuvieron los siguientes datos 12, 11, 10, 9, 8 millones de pesos. En efecto, se trata de una población conformada por los cinco individuos correspondientes a las cinco empresas más eficientes en el año 2014 en la dicha ciudad. Se requiere calcular la media, la varianza, la desviación estándar y el coeficiente de variación poblacional. Asimismo, se necesita obtener el porcentaje o proporción poblacional de las empresas que obtuvieron por lo menos 10 millones de utilidad por día.

Se trata de una población con $N = 5$. En principio se calculará el promedio o media poblacional de las utilidades:

$$\mu = \frac{\sum_{i=1}^{5} x_i}{N} = \frac{12+11+10+9+8}{5} = \frac{50}{5} = 10$$

Este parámetro indica que la utilidad promedio fue de 10 millones de pesos por día. Este valor también sugiere que si el total (50 millones) se repartiera en partes iguales, entonces cada empresa debería tener una utilidad de 10 millones por día.

A continuación, se calcula la *varianza poblacional*:

$$\sigma^2 = \frac{\sum_{i=1}^{5} (x_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{(12-10)^2 + (11-10)^2 + (10-10)^2 + (9-10)^2 + (8-10)^2}{5}$$

$$\sigma^2 = \frac{(2)^2 + (1)^2 + (0)^2 + (-1)^2 + (-2)^2}{5} = \frac{4+1+0+1+4}{5} = 2$$

El valor anterior corresponde a dos (millones de pesos) cuadrados, unidades que en el mundo real no tienen sentido (pesos cuadrados); en consecuencia, hay necesidad de obtener la *desviación estándar poblacional* al extraer la raíz cuadrada, así:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{5} (x_i - 10)^2}{5}} = \sqrt{2} \cong 1.4142 \text{ millones de pesos}$$

Este parámetro indica que la dispersión de los datos con respecto a la utilidad promedio fue de 1.4141 millones de pesos. Este valor también proporciona una idea de cuánto se alejan los datos respecto a la utilidad promedio.

Luego el *coeficiente de variación poblacional* es:

$$CV = \frac{\sigma}{\mu} = \frac{1.4142}{10} = 0.14142 \cong 14.14\ \%$$

En general, si el CV es inferior al 8 %, se considera que los datos son homogéneos; si se ubica entre el 8 % y el 18 %, los datos son casi homogéneos; si va del 18 % hasta el 32 %, los datos son casi heterogéneos, y si es mayor al 32 %, los

datos son heterogéneos (Valdivieso, 2011). En este caso, el CV es un porcentaje comprendido entre el 8 % y el 18 %; en consecuencia, se interpreta que los datos correspondientes a la variable utilidades son casi homogéneos.

Si x representa el número de empresas que obtuvieron por lo menos 10 millones de pesos por día en esa población, entonces la proporción o porcentaje poblacional es:

$$p = \frac{x}{N} = \frac{3}{5} = 0.6 = 60\ \%$$

Ahora, si X es una variable aleatoria real definida sobre el espacio de probabilidad $(\Omega, \Im, P)$, entonces:

i) Si X es una variable aleatoria discreta con rango $R_X = \{x_1, x_2, ...\}$ y f es su función de probabilidad, entonces, el valor esperado de X, o media de la variable aleatoria, está dado por:

$$\mu = E(X) = \sum_{x_i \in R_X} x_i f(x_i) = \sum_{x_i \in R_X} x_i P(X = x_i),$$

siempre y cuando la anterior suma exista (Blanco, 2004; Burbano y Valdivieso, 2015).

ii) Si X es una variable aleatoria continua con función de densidad f_X, entonces el valor esperado de X, o media de la variable aleatoria, está dado por:

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx,$$

toda vez que la anterior integral exista.

La varianza de la variable aleatoria X se denota y define por:

$$\sigma^2 = Var(X) = E\left(X - E(X)\right)^2,$$

siempre y cuando el valor esperado de X exista (Blanco, 2004). La anterior expresión se puede escribir de la siguiente forma:

$$\sigma^2 = Var(X) = E\left(X^2\right) - \left(E(X)\right)^2$$

Al número

$$\sigma = \sqrt{Var(X)}$$

se le denomina la desviación estándar de la variable aleatoria X.

1.3.2 Estimadores

Un estimador es una variable aleatoria construida como una función de las variables que conforman una muestra aleatoria (Lindgren, 1993); esta no depende de parámetro alguno constitutivo de la expresión algebraica que identifica el modelo asumido para representar una variable en la población objeto de estudio (Mayorga, 2003).

Es decir, dada una muestra aleatoria $X_1, X_2, ..., X_n$ un estimador T es una función determinada de la siguiente manera:

$$T = f(X_1, X_2, ..., X_n)$$

Así, por ejemplo, la estadística o estimador denominado media muestral se denota con $\overline{X}$ y se define así:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Ahora, si los datos $x_1, x_2, ...x_n$ son las observaciones o valores de una variable cuantitativa X obtenidos como realizaciones de una muestra aleatoria de tamaño n, entonces una estimación o valor $\overline{x}$ de la estadística $\overline{X}$, denominada media muestral, se obtiene así:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

El estimador, o estadística, llamado varianza muestral y denotado con S^2, se define de la siguiente forma:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$$

De forma similar, si los datos $x_1, x_2, ...x_n$ son las mediciones de una variable cuantitativa X, obtenidos como realizaciones de una muestra aleatoria de tamaño n, entonces una estimación s^2 de la estadística S^2, denominada varianza muestral, se obtiene de la siguiente manera:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

El estimador llamado varianza corregida, o cuasivarianza, se denota con $\hat{S}^2$ y se define así:

$$\hat{S}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Una estimación o valor $\hat{s}^2$ de la estadística $\hat{S}^2$ puede calcularse mediante la siguiente expresión:

$$\hat{s}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Usando las expresiones anteriores se puede obtener la siguiente relación:

$$\hat{s}^2 = \frac{n}{n-1}s^2$$

Un valor s para la estadística S, denominada *desviación estándar* muestral, se denota y se define de la siguiente forma:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$

Un valor $\hat{s}$ para la estadística $\hat{S}$, llamada desviación estándar corregida o cuasidesviación estándar, es:

$$\hat{s} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Un valor del coeficiente de variación muestral se calcula de la siguiente manera:

$$cv = \frac{\hat{s}}{\overline{x}}$$

El *coeficiente de variación* es adecuado calcularlo, y tiene sentido solamente cuando se trabaja con variables cuantitativas en escala de razón.

Ahora, si x representa el número de individuos que presentan una característica A de interés en una determinada muestra, una estimación de la proporción se denota y se calcula mediante la razón:

$$\hat{p} = \frac{x}{n}$$

Es conveniente aclarar que la siguiente expresión es una variable aleatoria; sin embargo, no corresponde a un estimador, puesto que involucra un parámetro desconocido:

$$\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n-1}$$

Ejemplo 1.12. Del conjunto de datos del ejemplo 1.11 se pueden seleccionar cinco muestras aleatorias de tamaño $n = 4$ mediante un muestreo aleatorio sin reemplazo; una de ellas está constituida por los datos: 11, 9, 10, 8; estos se utilizan para calcular las estimaciones que se indican a continuación:

Un valor $\overline{x}$ de la variable promedio muestral $\overline{X}$ o estimación de la media poblacional es:

$$\overline{x} = \frac{\sum_{i=1}^{4} x_i}{4} = \frac{11+9+10+8}{4} = \frac{38}{4} = 9.5$$

Un valor s^2 de la variable S^2 o estimación de varianza poblacional es:

$$s^2 = \frac{\sum_{i=1}^{4} (x_i - 9.5)^2}{4} = \frac{(11-9.5)^2 + (9-9.5)^2 + (10-9.5)^2 + (8-9.5)^2}{4}$$

$$s^2 = \frac{(1.5)^2 + (-0.5)^2 + (0.5)^2 + (-1.5)^2}{4} = \frac{2.25 + 0.25 + 0.25 + 2.25}{4} = \frac{5}{4} = 1.25$$

Otra estimación de la varianza poblacional es $\hat{s}^2$, valor particular de la variable $\hat{S}^2$; esta se obtiene así:

$$\hat{s}^2 = \frac{\sum_{i=1}^{4} (x_i - 9.5)^2}{4-1} = \frac{5}{3} \cong 1.6666$$

Se puede observar que el valor $\hat{s}^2$ se puede obtener usando la siguiente expresión:

$$\hat{s}^2 = \frac{n}{n-1}s^2 = \frac{4}{4-1}\left(\frac{5}{4}\right) = \frac{5}{3} \cong 1.6666$$

Un valor s para la variable aleatoria S, denominada *desviación estándar* muestral, es:

$$s = \sqrt{\frac{\sum_{i=1}^{4}(x_i - 9.5)^2}{4}} = \sqrt{1.25} \cong 1.118$$

Un valor $\hat{s}$ para la variable $\hat{S}$, desviación estándar corregida o cuasidesviación estándar, es:

$$\hat{s} = \sqrt{\frac{\sum_{i=1}^{4}(x_i - 9.5)^2}{4-1}} = \sqrt{1.6666} \cong 1.2909$$

Un valor del coeficiente de variación muestral es:

$$cv = \frac{\hat{s}}{\overline{x}} = \frac{1.2909}{9.5} \cong 0.1358 = 13.58\ \%$$

El anterior valor se encuentra entre el 8 % y el 18 %, en consecuencia, los datos de la variable en esta muestra son casi homogéneos.

Ahora, si x representa el número de empresas que obtuvieron por lo menos 10 millones de pesos por día en esta muestra, entonces la proporción o porcentaje muestral es:

$$\hat{p} = \frac{x}{n} = \frac{2}{4} = 0.5 = 50\ \%$$

1.3.3 Teorema central del límite

En concordancia con Mayorga (2003) y Blanco (2004), el teorema central del límite establece que la media aritmética de variables aleatorias independientes e igualmente distribuidas tiende a un normal estándar cuando el número de variables aleatorias involucradas es grande y cuando la varianza es finita y diferente de cero. Este teorema fue demostrado por primera vez en el año 1733 por el matemático De Moivre; una versión más general fue dada por Laplace en 1812, y la versión que se conoce actualmente fue presentada por Liapounoff en 1901. A continuación, se expresa de manera simbólica el teorema central del límite en la versión de Lindeberg-Lévy, obtenida de forma independiente por cada uno de estos matemáticos en la segunda década del siglo xx.

Si $X_1, X_2, ..., X_n$ es una muestra aleatoria de una población con valor esperado μ y varianza σ^2 finitos, considerando la variable aleatoria

$$Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \text{con} \quad \bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n},$$

entonces, la sucesión de variables aleatorias $\{Z_n\}$ converge en distribución a una variable aleatoria con distribución normal estándar.

Ahora, si $T = \bar{X}_n$, entonces se deduce que $E(T) = \mu$ y $Var(T) = \frac{\sigma^2}{n}$, debido a que la muestra aleatoria considerada está constituida por *n* variables aleatorias independientes que presentan la misma media μ y varianza constante σ^2; en este contexto, el teorema central del límite se puede expresar de la siguiente manera:

$$Z = \frac{T - E(T)}{\sqrt{Var(T)}} \sim N(0,1)$$

1.4 Algunos tipos de muestreo

Para escoger una muestra representativa de la población se deben utilizar técnicas o métodos que aseguren tal representatividad y permitan inferir acerca de las características poblacionales de interés. Los individuos que se van a observar en la muestra se pueden seleccionar usando métodos aleatorios (Gutiérrez, 2005) o considerando algunos criterios o necesidades (métodos no probabilísticos, por conveniencia, subjetivos o a juicio). Las muestras aleatorias o probabilísticas son las que permiten hacer uso de la teoría estadística para realizar inferencias con sustento científico. En esta sección se presentan algunos métodos para realizar un muestreo aleatorio, entre ellos, el muestreo aleatorio simple, el estratificado, el muestreo sistemático y por conglomerados; además, se describen algunos métodos no probabilísticos.

1.4.1 Muestreos aleatorios o probabilísticos

En el muestreo aleatorio se deben cumplir algunas condiciones para obtener una muestra probabilística; para el caso de una selección simple son las siguientes (Särndal, Swenson & Wretman, 1992):

i) Poder definir el conjunto total de muestras posibles,

$$S = \{S_1, S_2, ..., S_T\},$$

que pueden seleccionarse de la población de acuerdo con el procedimiento muestral.

ii) Conocer para cada una de las muestras posibles la probabilidad $\pi(S)$ de que sea seleccionada.

iii) El procedimiento utilizado debe dar a cada elemento de la población una probabilidad de selección diferente de cero.

iv) La selección, como se mencionó antes, debe ser aleatoria; esto es, el mecanismo de probabilidad diseñado para la selección debe ser tal que cada muestra posible S tenga la probabilidad de selección asignada previamente, $\pi(S)$.

Según Botero (2001), las condiciones *ii)* y *iv),* junto con las fórmulas de estimación correspondientes, determinan el diseño muestral. Cuando el muestreo consta de varias etapas, las condiciones anteriores se deben cumplir en cada una de ellas. En general, todo tipo de muestreo que no cumpla con alguna de las condiciones enunciadas anteriormente es un muestreo no probabilístico.

A continuación se indican algunos diseños muestrales básicos para realizar muestreo aleatorio en poblaciones finitas:

- Muestreo aleatorio simple sin reemplazo
- Muestreo aleatorio simple con reemplazo
- Muestreo estratificado aleatorio simple
- Muestreo por conglomerados
- Muestreo sistemático aleatorio

1.4.1.1 Muestreo aleatorio simple sin reemplazo

En el muestreo aleatorio simple sin reemplazo, también denominado irrestrictamente aleatorio, todas las muestras posibles de tamaño n tienen igual probabilidad de ser seleccionadas; en consecuencia, todos los individuos de la población también tienen la misma posibilidad de ser seleccionados. Cada individuo se selecciona una sola vez, es decir, un individuo es escogido y ya no regresa a la población para ser considerado nuevamente. Para garantizar que un procedimiento de selección permita obtener una muestra aleatoria se utilizan números aleatorios generados por computador o los indicados en tablas de libros de estadística.

El número total de muestras de tamaño n que son posibles al seleccionarlas sin reemplazo de una población de tamaño N está dado por:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

La probabilidad de seleccionar una muestra aleatoria cualquiera de tamaño n de una población de tamaño N se calcula de la siguiente forma:

$$\frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}$$

La probabilidad de que un individuo cualquiera de la población esté presente en la muestra se calcula dividiendo el número de muestras posibles que contendrían al individuo por el número posible de muestras, es decir:

$$\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

Ejemplo 1.13. Se tiene una población conformada por 12 individuos, en los cuales interesa estudiar el ingreso mensual en miles de pesos. Se quiere seleccionar una muestra aleatoria de tamaño 3 usando muestreo aleatorio simple sin reemplazo. El número total de muestras distintas que se obtiene es:

$$\binom{12}{3} = \frac{12!}{3!(12-3)!} = \frac{12!}{3!9!} = 220$$

La probabilidad de seleccionar una muestra compuesta por tres individuos determinados, es:

$$\frac{1}{220} = 0.004545$$

La probabilidad de que un individuo cualquiera de la población pertenezca a la muestra es:

$$\frac{3}{12} = 0.25$$

1.4.1.2 Muestreo aleatorio simple con reemplazo

En esta clase de muestreo, todas las muestras de tamaño n tienen igual probabilidad de ser seleccionadas, cada individuo de la población tiene igual probabilidad de ser escogido. Para realizar el muestreo aleatorio simple con reemplazo, cualquier individuo de la población es susceptible de escogerse más de una vez para formar parte de la muestra, dado que el individuo es escogido, observado y regresado nuevamente a la población con la posibilidad de ser escogido nuevamente.

El número total de muestras posibles es N^n, la probabilidad de selección de una sucesión específica de n unidades es $\frac{1}{N^n}$. La probabilidad de que cualquier individuo de la población sea seleccionado al menos una vez es:

$$1-\left(\frac{N-1}{N}\right)^n$$

Ejemplo 1.14. Se tiene una población conformada por 10 individuos con el interés de estudiar sus gastos diarios en miles de pesos. Se quiere seleccionar una muestra aleatoria de tamaño 4 usando muestreo aleatorio simple con reemplazo. El número total de muestras posibles de tamaño 4 con repetición es:

$$10^4 = 10\ 000$$

La probabilidad de que una sucesión cualquiera conformada por 4 individuos sea seleccionada es:

$$\frac{1}{10^4} = \frac{1}{10000} = 0.0001$$

La probabilidad de que un individuo específico sea seleccionado, al menos una vez, para conformar la muestra es:

$$1-\left(\frac{10-1}{10}\right)^4 = 1-\left(\frac{9}{10}\right)^4 = 1-\left(\frac{6561}{10000}\right) = 0.3439$$

1.4.1.3 Muestreo estratificado

La población se divide en grupos disjuntos denominados *estratos,* de tal forma que entre los individuos de cada grupo no existan diferencias importantes en

lo referente a las características que interesa estudiar (los individuos presentan características similares dentro de cada estrato), pero los grupos entre sí son muy diferentes. En muchas ocasiones, la población se divide en estratos con facilidad, y en otras ya se encuentra dividida convenientemente. Una vez identificados los estratos (mediante una variable auxiliar), se toma una muestra aleatoria simple de cada uno de ellos. El número de individuos en cada estrato se determina de dos maneras:

i) Usando muestreo estratificado proporcional, es decir, utilizando la expresión siguiente:

$$n_i = \frac{x_i}{N} n$$

Donde x_i es el número de elementos del $i-$ésimo estrato, n es el tamaño de la muestra y N es el tamaño de la población.

Ejemplo 1.15. Se tiene una población de 1 700 individuos dividida en 4 estratos, como se indica en la Figura 1.1; se quiere seleccionar una muestra aleatoria de tamaño 80.

El tamaño de la población es N =1 700, el tamaño de la muestra es n = 80

$$n_1 = \frac{500}{1700} \star 80 = 23.52$$

$$n_2 = \frac{300}{1700} \star 80 = 14.11$$

$$n_3 = \frac{200}{1700} \star 80 = 9.41$$

$$n_4 = \frac{700}{1700} \star 80 = 32.94$$

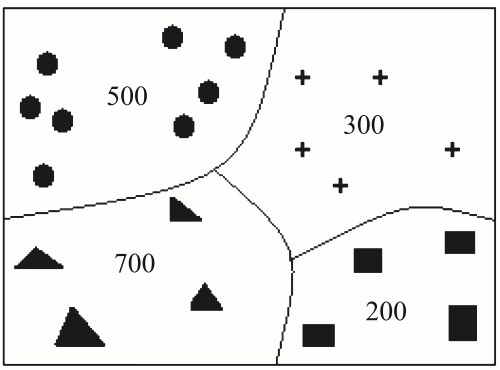

Figura 1.1 Población conformada por cuatro estratos

El tamaño de la muestra, en este caso, se determina de la siguiente forma:

$$n = n_1 + n_2 + n_3 + n_4 = 24+14+9+33=80$$

Por lo cual, 24 individuos se han de seleccionar mediante muestreo aleatorio simple del estrato 1; asimismo, 14 individuos del estrato 2; 9 individuos del estrato 3 y 33 individuos del estrato 4.

ii) Utilizando muestreo aleatorio no proporcional; en este caso se toma igual número de elementos en cada estrato, usando la siguiente expresión:

$$n_i = \frac{n}{k}$$

Donde k es el número de estratos y n es el tamaño de la muestra.

En el ejemplo anterior se deberían seleccionar, mediante muestreo aleatorio simple, $n_i = \frac{80}{4} = 20$ individuos de cada uno de los cuatro estratos.

1.4.1.4 Muestreo sistemático o en serie

Los individuos que conformarán la muestra se seleccionan a intervalos iguales (siguiendo una determinada frecuencia), pero escogiendo un individuo inicial (primer elemento que servirá para referenciar a los demás) de manera aleatoria; para seleccionar una muestra de tamaño n de una población de tamaño *N,* se toma como intervalo de muestreo (*c*) el valor inverso de la fracción de muestreo, es decir:

$$c = \frac{1}{\frac{n}{N}} = \frac{N}{n}$$

Luego, se toma un número λ tal que $0 < \lambda < c$ de manera aleatoria; este se convierte en el código para el primer elemento que conformará la muestra; los siguientes códigos o individuos para la muestra se obtienen agregando un valor entero próximo *c* al número λ, hasta que el n-ésimo individuo se encontrará en la posición $\lambda + (n-1)c$.

Ejemplo 1.16. Si se tiene una población con 1 700 individuos y se desea seleccionar una muestra de 80 individuos, entonces se ordenan los datos correspondientes a los individuos y se calcula:

$$c = \frac{N}{n} = \frac{1700}{80} = 21,25$$

El número λ se toma de manera aleatoria como un número menor que 21 y servirá como punto de partida. Si después de realizar un procedimiento aleatorio para elegir el valor de λ, el valor resultante fuera 10, entonces el segundo individuo será aquel que se encuentre en la posición 31 (10+21), el tercero el de la posición 52 (10+21+21) y así sucesivamente hasta que el individuo 80 ocupará la posición 10+(80-1)21=1 669.

1.4.1.5 Muestreo por conglomerados

Un conglomerado ha de entenderse como un subconjunto de la población cuyos individuos son generalmente heterogéneos; dentro de este es posible que aparezca casi todo el rango de la característica que se desea estudiar. Estos subconjuntos entre sí resultan altamente similares. Para usar este método de muestreo se procede de manera similar al muestreo estratificado, aunque en muchas investigaciones es suficiente tomar un conglomerado cualquiera como muestra.

Ejemplo 1.17. Si se tiene una población con 1000 individuos y se han identificado 4 conglomerados, como se indica en la Figura 1.2, se desea seleccionar una muestra de 100 individuos; en este caso se procede de la siguiente forma:

$N = 1000$
$n = 100$

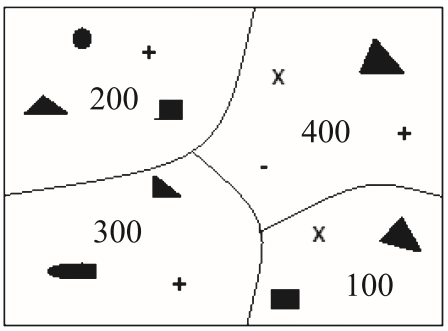

Figura 1.2. Población conformada por cuatro conglomerados

$$n_1 = \frac{200}{1000} \star 100 = 20 \qquad n_2 = \frac{400}{1000} \star 100 = 40$$

$$n_3 = \frac{100}{1000} \star 100 = 10 \qquad n_4 = \frac{300}{1000} \star 100 = 30$$

El tamaño de la muestra, en este caso, se determina de la siguiente forma:

$$n = n_1 + n_2 + n_3 + n_4 = 20+40+10+30=100$$

Sin embargo, también resulta adecuado tomar cualquiera de los conglomerados, por ejemplo, el tercer conglomerado, y de allí seleccionar mediante muestreo aleatorio simple una muestra de tamaño 80.

1.4.2 *Otros métodos de muestreo*

En algunas investigaciones es posible que se tenga que recurrir a los denominados métodos de muestreo determinísticos (no probabilísticos, no aleatorios); en estos, el investigador usa determinado juicio para seleccionar los individuos de una muestra. La argumentación más frecuente para utilizar estos métodos es que algunos individuos, quizá, ofrecerían mejor información acerca de la característica que se busca estudiar en la población, o que la probabilidad de seleccionar un individuo o una muestra es incalculable. Entre los métodos determinísticos están: el muestreo por criterio, el muestreo por conveniencia y el muestreo teórico.

1.4.2.1 Muestreo por criterio

La selección de los individuos se realiza bajo la hipótesis de que los individuos que se escogen para conformar la muestra son los más representativos de la población y tal vez suministren una información específica.

Ejemplo 1.18. En la realización de un estudio socioeconómico en el sector rural de la región A, del departamento B, es posible que se opte por incluir en la muestra al presidente de la Junta de Acción Comunal, al inspector de policía, al guardabosque, al representante del sector lechero, al de los paperos y al de los jornaleros, entre otros.

1.4.2.2 Muestreo por conveniencia

En diversas circunstancias, el acceso a los individuos de toda la población es difícil y se opta por obtener la información de quienes resulten de más fácil consecución. Este problema es frecuente cuando, por ejemplo, la investigación abarca sectores geográficos extensos, se requieren urgentemente los resultados o las personas pertenecen a círculos sociales cerrados con respecto a la característica que se quiere investigar, entre otros.

Ejemplo 1.19. Se desea investigar un problema de adicción a un nuevo fármaco en la población A; en este caso, es posible usar la técnica denominada bola de nieve, mediante la cual se identifica un individuo clave que quiera proporcionar información sobre este problema, y, a través de él, identificar un segundo individuo, y luego, a un tercer individuo y así sucesivamente, hasta conformar una muestra no probabilística con *n* individuos.

1.4.2.3 Muestreo teórico

De manera general, la investigación científica que se realiza a través de variables tanto cuantitativas como cualitativas utiliza métodos cuantitativos asociados con los procesos de inferencia estadística; no obstante, también existe la denominada *investigación cualitativa (IC)*, de la cual existen varios tipos; la *IC* utiliza métodos y diseños propios para indagar sobre la ocurrencia de un fenómeno de interés sin necesidad de recurrir al uso de variables en el sentido cuantitativo; este tipo de estudio involucra elementos como: conceptos, ideas y categorías relacionadas con una determinada "teoría" o con el conocimiento científico en general; esta clase de indagación se usa con frecuencia en campos de las ciencias sociales, humanas y de la educación (Campos, 2009). En *IC* se utilizan diseños y metodologías propios, entre ellos: el estudio de casos, el estudio de grupos focales, la investigación-acción (IA), la investigación acción participación (IAP), la fenomenología, la teoría fundamentada y la cartografía, solo por mencionar algunas.

En esta clase de estudios se sugiere realizar un muestreo "teórico", que es un muestreo de conceptos, no de individuos: "significa que el muestreo, más que predeterminado antes de comenzar la investigación, evoluciona durante el proceso; se basa en conceptos que emergen del análisis y que parecen ser pertinentes para la teoría que se está construyendo" (Straus y Corbin, 2002, p. 220), hasta alcanzar el *punto de saturación teórica*, "en el cual ya no emergen propiedades, dimensiones o relaciones nuevas durante el análisis" (p. 157). Las "categorías" emergentes de conocimiento permiten *comprender* de manera adecuada el fenómeno de interés que se esté estudiando (Strauss y Corbin, 1998).

En concordancia con Flores, Gómez & Jiménez (1999), Guba & Lincon (1994) y Simons (2011), en el enfoque cualitativo de investigación se considera la realidad de forma global y dinámica, elaborada por medio de procesos interactivos entre el sujeto y aquella; la *IC* sigue un camino de corte inductivo y considera la realidad como el punto inicial del proceso investigativo; los datos textuales recogidos posibilitan la generación de pre-teorías, la presentación de categorías emergentes y la pesquisa de nuevos datos que den cuenta de las particularidades detectadas en una situación determinada.

Ejemplo 1.20. De acuerdo con Simons (2011), Stake (1998), Yin (2009) y Yin (2014), un *estudio de caso* se puede conceptualizar como el estudio de la particularidad y la complejidad de un caso singular, para llegar a comprender su actividad en circunstancias importantes; en consecuencia, posibilita "el examen detallado, comprehensivo, sistemático y en profundidad del objeto de interés" (Flores, Gómez & Jiménez, 1999). Como ilustración, se quiere estudiar el caso de la Institución Educativa del Sur (IES), ubicada en la ciudad de Tunja, en Boyacá, Colombia, a fin de comprender el fenómeno de la deserción escolar y su relación con la violencia en el entorno escolar manifiesta en las acciones de sus estudiantes. En este contexto, los métodos cualitativos posibilitan realizar un estudio en profundidad de tal fenómeno, identificar categorías conceptuales y generar soluciones pertinentes.

Actividades para el estudio independiente Capítulo 1

1.1 Una vez se haya hecho una lectura comprensiva de los temas y ejemplos del capítulo 1, complementar en los espacios en blanco.

a) De algunas actividades como: recolección, organización, representación, interpretación y análisis de información, referidas a una o más variables, de su estudio en una población o en una muestra determinada se ocupa la estadística _____

b) Cuando se selecciona una muestra aleatoria y con base en ella se infiere acerca de la presencia o ausencia de características de estudio o de parámetros en toda la población, se está trabajando con la estadística_____

c) Una_____es un subconjunto de individuos representativo de la población.

d) Cada una de las características que interesa estudiar en los individuos de una población o de una muestra se denominan _____

e) Aquellas características que permiten clasificar a los individuos en grupos o clases se les denomina variables_____

f) Aquellas características que incluyen la noción de cantidad, intensidad o magnitud se llaman variables_____

g) Una variable_____ es aquella que admite cualquier valor en un intervalo de números reales.

h) La variable X :número de trabajadores por empresa en la ciudad de Tunja en Colombia en el año 2015, corresponde a una variable _____

1.2 Clasificar cada una de las siguientes variables y determinar la escala de medición.

a) X : grado de escolaridad de las madres cabeza de familia de los estudiantes matriculados en el semestre I del año 2016 en la Universidad Nacional de Colombia, sin interesar el orden. _____

b) Y : preferencia por el producto H que está promocionando la empresa TT en la ciudad de Manizales en Colombia en el año 2016, calificadas así: 1: nada,

3: poco, 5: mucho. _____

c) I: ingreso mensual de los trabajadores de la empresa T&R en el mes de enero del año 2016. _____

d) T : temperatura ambiente de los salones de clase en el bloque R de la Universidad Pedagógica y Tecnológica de Colombia (UPTC) medida entre las 11 a.m. y las 11:30 a.m. en los últimos 15 días hábiles del mes de noviembre del año 2015. _____

e) P : peso en kilogramos de cada uno de los vacunos con 40 semanas de vida, de la granja T&H. _____

1.3 En correspondencia con cada una de las siguientes variables, complementar o responder a las preguntas formuladas.

La variable *C* : color del cabello de una muestra de estudiantes en la universidad A; los siguientes datos: N, N, B, N, N, N, N, B, R, R, B, N, R, N, N, N, N, R, R, N, N, N, R, N, N, N, R, R, R, N, N, con codificación: B=blanco, N=negro, R=Rubio.

a) El tamaño de la muestra es_____

b) La proporción muestral de los estudiantes con cabello blanco es_____

c) ¿Es adecuado calcular el promedio con los datos de la variable anterior? ____ Justifique su respuesta_____
_____

Para la variable *X*: peso en kilogramos de unos estudiantes que cursaron Cálculo I en el programa de Ingeniería en el semestre II del año 2015 en la UPTC, se han obtenido los siguientes datos: 62, 63.8, 65.4, 62, 58, 70, 65, 65, 63.8, 62, 63.8, 65.4, 62, 58, 70, 65, 65, 63.8, 62, 63.8, 65.4, 62, 58, 70, 65, 65, 63.8, 62, 63.8, 65.4, 62, 58, 70, 65, 65.

d) La media muestral o promedio en la muestra del peso de los estudiantes que cursaron Cálculo I es _____

e) La desviación estándar _____

f) El coeficiente de variación es _____ e indica que los datos son_____

1.4 Del conjunto de datos del ejemplo 1.11 es posible seleccionar 10 muestras aleatorias de tamaño $n = 3$ mediante un muestreo aleatorio sin reemplazo; una de ellas está constituida por los datos 9, 10, 8; en esta muestra, obtener el promedio, la varianza, la cuasivarianza o varianza corregida, la desviación estándar, la desviación estándar corregida, el coeficiente de variación y el porcentaje de las empresas que han obtenido por lo menos 10 millones por día de utilidades.

1.5 Se tiene una población conformada por 100 individuos, interesa estudiar sus gastos semanales en miles de pesos. Se quiere seleccionar una muestra aleatoria de tamaño 5 usando muestreo aleatorio simple sin reemplazo. Además, determinar el número total de muestras distintas y posibles de obtener, la probabilidad de seleccionar una muestra compuesta por cinco individuos específicos y la probabilidad de que un individuo cualquiera de la población pertenezca a la muestra.

1.6 Se tiene una población conformada por 20 individuos, el interés se centra en estudiar la relación peso-talla a fin de implementar una dieta para disminuir el peso. Se quiere seleccionar una muestra aleatoria de tamaño 3 usando muestreo aleatorio simple con reemplazo; determinar el número total de muestras de tamaño 3 con repetición posible, la probabilidad de que una sucesión cualquiera conformada por 3 individuos sea seleccionada y la probabilidad de que un individuo específico sea seleccionado al menos una vez para conformar la muestra.

1.7 Se tiene una población de 5 000 individuos, dividida en 4 estratos, como se indica en la Figura 1.3; seleccionar una muestra aleatoria de tamaño 100, usando: *i)* muestreo proporcional y *ii)* por cuotas iguales o muestreo no proporcional.

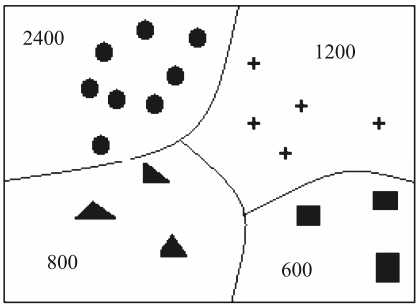

Figura 1.3 Población conformada por cuatro estratos

1.8 Si se tiene una población con $N = 2\,500$ individuos y se desea seleccionar una muestra de $n = 90$ individuos, describa el procedimiento para seleccionar esa muestra usando un muestreo aleatorio sistemático.

Ejercicios para el capítulo 1

1.1 Mencionar tres diferencias entre estadística descriptiva y estadística inferencial.

1.2 ¿Es posible transformar una variable cuantitativa en una cualitativa? De ser así, proporcione un ejemplo. ¿Es posible transformar una variable cualitativa en una cuantitativa? Explique.

1.3 Proporcionar 2 ejemplos de variables:

- Cualitativas
- Discretas
- Continuas

1.4 Escribir 2 ejemplos de variables

- En escala ordinal
- En escala de razón
- En escala nominal
- En escala de intervalo

1.5 ¿Es posible calcular el coeficiente de asimetría para una variable aleatoria? De ser así, ¿cuál es la expresión para calcularlo?

1.6 ¿Es posible calcular el coeficiente de curtosis para una variable aleatoria? De ser así, ¿cuál es la expresión para calcularlo?

1.7 Del conjunto de datos del ejemplo 1.11 es posible seleccionar 10 muestras aleatorias de tamaño $n = 2$ mediante un muestreo aleatorio sin reemplazo; una de ellas está constituida por los datos 11, 10; en esta muestra, obtener el promedio, la varianza, la cuasivarianza o varianza corregida, la desviación estándar, la desviación estándar corregida, el coeficiente de variación y el porcentaje de las empresas que han obtenido por lo menos 10 millones de utilidades por día.