2ª EDICIÓN

Métodos bioestadísticos

Luis A. Villarroel del Pino

Alfaomega

EDICIONES UC

Métodos Bioestadísticos

Segunda Edición

ALFAOMEGA
Empresas del Grupo
Colombia: Alfaomega Colombiana S.A.
Calle 62 No.20-46 esquina, Bogotá
Teléfono (57-1) 746 0102 Fax: (57-1) 210 0122
cliente@alfaomegacolombiana.com

México: Alfaomega Grupo Editor S.A. de C.V.
Calle Doctor Olvera No. 74, Colonia Doctores,
Delegación Cuaúhtemoc, Ciudad de México
C.P. 06720 • teléfono (52-55) 5089 7740
Fax (52-55) 5575 2420
Sin costo 01-800-020-4396
libreriapitagoras@alfaomega.com.mx

Argentina: Alfaomega Grupo Editor Argentino S.A.
Av. Córdoba 1215, Piso 10
Capital Federal, Buenos Aires
Teléfono/Fax: (54-11) 4811 7183 / 8352 / 0887
ventas@alfaomegaeditor.com.ar

Chile: Alfaomega Grupo Editor S.A.
Av. Providencia 1443. Oficina 24, Santiago
Teléfonos (56-2) 2235 4248 / 2947 9351 / 2235 5786
agechile@alfaomega.cl

www.alfaomega.com.co

EDICIONES UNIVERSIDAD
CATÓLICA DE CHILE
Vicerrectoría de Comunicaciones
Avenida Libertador Bernardo O'Higgins 390
Santiago, Chile
editorialedicionesuc@uc.cl
www.ediciones.uc.cl

Métodos Bioestadísticos
Segunda Edición
Bogotá, 2019

Primera edición, 2013

Diseño: Producciones gráficas ltda.
Portada: Camilo Umaña

ISBN: 978-958-778-483-1 (Edición Colombia)
ISBN: 978-956-14-2219-3 (Edición Chilena)

Impreso en Colombia
Printed and made in Colombia

FACULTAD DE MEDICINA

Métodos Bioestadísticos

Segunda Edición

Luis A. Villarroel del Pino

Alfaomega

EDICIONES UC

Dedicado a mis hijos,
Isidora, Ignacio y Luciano,
por ser siempre la luz del camino;
y a mi mujer, Paulina,
por todo su apoyo y motivación…
¡y para aportar con un grano de arena
a su afán cuantitativo!

Índice

Agradecimientos

Mis más sinceros agradecimientos a Angélica Domínguez de Landa y Andrea Villarroel Barrios, por todas las contribuciones que hicieron a los contenidos, tanto de forma como de fondo, y sobre todo por el apoyo y entusiasmo constantes, lo que hizo que escribir este texto fuera un verdadero agrado.

Introducción

Este texto está basado en una serie de apuntes que escribí originalmente para el curso MED104A-Bioestadística, para primer año de la carrera de medicina, de la Pontificia Universidad Católica de Chile. Paulatinamente, estos apuntes fueron utilizados por alumnos de otras carreras y programas, como los magísteres de epidemiología, de nutrición, de enfermería, carrera de odontología, entre otros.

Aunque inicialmente los apuntes fueron pensados como un resumen de partes específicas de la materia tratada en el curso, los mismos alumnos manifestaron lo importante que era para ellos contar con un texto que les diera un marco de referencia sobre los contenidos, y en más de una ocasión se acercaban a preguntar cuándo habría un apunte sobre materia como intervalos de confianza, modelos estadísticos, sobre el programa SPSS, etc.

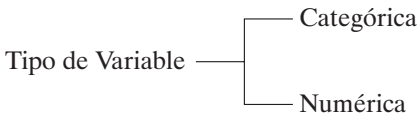
Esto fue lo que me motivó a convertir esos apuntes en un texto que abarcara todos los contenidos de un curso de bioestadística básica, y ahora en un libro que sea de utilidad para alumnos y profesionales de distintas disciplinas de la salud.

Debido a su génesis, este texto está escrito en un lenguaje poco matemático, pero sin sacrificar la rigurosidad en los conceptos o en la descripción de los métodos, y cada tema está ilustrado con muchos ejercicios resueltos, lo que facilitará la comprensión de cada materia y servirá de práctica a los alumnos que accedan a él.

Aunque originalmente el libro está dirigido a alumnos de pregrado y posgrado en ciencias de la salud, también espero que sea de mucha utilidad para los investigadores interesados en comprender mejor los métodos estadísticos que utilizan día a día.

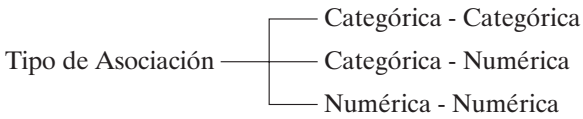
El texto se aproxima a los métodos bioestadísticos de la misma forma como un investigador o un estadístico se aproximaría a la solución de un problema de investigación: en primer lugar, observaría la o las variables incluidas en el estudio, determinaría el tipo al que estas pertenecen, y en base a esta clasificación, además del cumplimiento de algunos supuestos, determinaría el tipo de estadística descriptiva más adecuado para cada variable, el o los test estadísticos apropiados y el tipo de modelo estadístico que corresponde ajustar a los datos.

De esta manera, en el capítulo 1 se aborda el tema de la descripción de variables según su tipo (básicamente categórica o numérica), además de los conceptos de población, muestra y otros relacionados con la naturaleza de las variables.



Posteriormente, después de revisar algunos conceptos básicos de probabilidad en el capítulo 2 y de estudiar las distribuciones de probabilidad más comunes en el capítulo 3 (incluida la distribución normal), el tipo de variables nuevamente es importante en la construcción de intervalos de confianza, lo cual es revisado en el capítulo 4.

Cuando se aborda el tema de la asociación de dos variables, en el capítulo 5, una vez más se recurre a los tipos a los que pertenecen las variables involucradas en la asociación, lo que permite identificar más fácilmente el camino metodológico adecuado.



El tipo de variable también es importante en la construcción de modelos estadísticos, abordado en el capítulo 6, ya que construir un modelo para una variable numérica es muy distinto que crear uno para una variable categórica.

Finalmente, el modo en que se realizan todos estos análisis bioestadísticos usando SPSS se muestra en el capítulo 7 y usando Minitab en el capítulo 8.

Prefacio a la Segunda Edición

Esta segunda edición de Métodos Bioestadísticos corrige una serie de errores menores que tenía la edición anterior y actualiza el capítulo 7 dedicado a SPSS: ahora se describe la versión 24 (la más actualizada hasta el momento de editar este texto) en vez de la versión 19. El capítulo dedicado a MINITAB no fue modificado, ya que la versión 17 del programa (la más actual) no tiene cambios de importancia respecto a la versión 16 descrita en este libro.

[1]
Estadística descriptiva

1.1 Introducción

La estadística es la ciencia que estudia la recolección, organización, análisis e interpretación de un conjunto de datos, y sus conceptos generales pueden aplicarse a distintas disciplinas como la ingeniería, la agricultura, la economía (donde se denomina econometría) o la psicología (donde se denomina biometría). Cuando la estadística se aplica en las ciencias de la salud, se utiliza el término bioestadística.

En términos globales, la estadística puede dividirse en **descriptiva** y **analítica**. La estadística descriptiva, como su nombre lo indica, solo pretende describir o caracterizar un conjunto de datos. La estadística analítica, en cambio, plantea hipótesis respecto a una población, usando un subconjunto de datos disponible de esta.

Para llevar a cabo un **estudio descriptivo**, es necesario conocer los conceptos de población y muestra (aleatoria) y sus propiedades, los tipos de variables aleatorias posibles de encontrar en la práctica y la forma como se describen: tablas de frecuencias, medidas de tendencia central y de dispersión y percentiles. Todos estos conceptos los encontrará definidos en este capítulo.

Los conceptos necesarios para hacer un **estudio analítico** los hallará en el capítulo de Asociación de variables.

1.2 Población y muestra

Población

La **población** (también llamada **universo**) se define como el conjunto total de objetos o personas de interés en un estudio. Una característica relevante de la población es que todos sus elementos deben cumplir con un conjunto predefinido de características.

El conjunto de características debe permitir entender sin lugar a dudas cuál es la población en estudio. Por ejemplo, si un estudio plantea: "Se quiere determinar el porcentaje de personas de la ciudad de Valdivia que usa el detergente X", se está dando a entender que la población en estudio corresponde a todos los habitantes de Valdivia (¿los niños usarán el detergente X?). Quizás sería más adecuado plantear:

"Se quiere determinar el porcentaje de dueñas de casa de la ciudad de Valdivia que usa el detergente X".

En este mismo problema, si los investigadores contactaran a las dueñas de casa por teléfono para averiguar cuántas usan el detergente X, entonces sería necesario incorporar esta nueva característica (poseer teléfono) a nuestra definición de la población.

Lo habitual es que la población esté constituida por un gran número de personas u objetos, por lo que normalmente se hace inviable acceder a todos ellos. El proceso de recopilación de datos de toda la población se denomina **censo**. Aunque sería atractivo acceder a toda la población, existen varios problemas para llevar a cabo esta idea:

- Un censo usualmente requiere invertir mucho tiempo y recursos, mientras que los estudios en salud se hacen cumpliendo ciertos plazos y con recursos limitados.
- Como las poblaciones son dinámicas, el objeto en estudio puede ser distinto en los primeros individuos estudiados que en los últimos, sobre todo si estos son vistos mucho tiempo después que los primeros. Por ejemplo, si un investigador quiere determinar la prevalencia de estrés en la Octava Región, y recopila los datos entre enero y marzo de 2010, sus resultados se verán afectados por la ocurrencia del terremoto del 27 de febrero de ese año.
- En ocasiones no es posible identificar con facilidad a los sujetos que componen la población. Por ejemplo, la población de personas que viven con VIH, o el número de aves acuáticas que existen en una reserva ecológica.

Muestra (aleatoria)

Dados los inconvenientes que se presentan al estudiar una población, lo habitual es que las investigaciones científicas se basen en una muestra de la población de interés, es decir, en un subconjunto de los elementos de la población.

Para que lo averiguado en la **muestra** sea cierto para la población en su conjunto, la muestra debe cumplir con los siguientes requisitos:

- La muestra debe ser **aleatoria**. Esto es, los sujetos en la muestra deben ser escogidos al azar (mediante un sorteo), de modo que todas las personas u objetos de la población tengan una probabilidad mayor que cero de estar presentes en la muestra.
- La muestra debe ser de un **tamaño mínimo** adecuado. Se entenderá por "adecuado" que el número de individuos seleccionados al azar de la población (el tamaño de la muestra) permita obtener estimaciones con un margen de error acotado.

  Ejemplo 1.1. Supongamos que es de interés estimar el porcentaje de fumadores en cierta población, y que el porcentaje real de fumadores es de alrededor de 40%. Luego, si se quiere estimar con un margen de error de 5 puntos porcentuales, entonces el tamaño de la muestra debiera permitir obtener entre 35% y 45% de fumadores en la muestra.
- La muestra debe ser **representativa** de la población de la que procede. Esto se cumple cuando las características de la población relevantes para la investigación, están presentes en la misma proporción o promedio en la muestra. Por ejemplo, si la población tiene 30% de hombres, esta proporción se mantiene en la muestra estudiada. Si la edad promedio poblacional es 50 años, en la muestra se observa aproximadamente lo mismo, etc.

Sin embargo, es imposible determinar si efectivamente cada una de las características poblacionales está presente en la misma proporción o promedio en la muestra. En consecuencia, se asume que si una muestra es aleatoria y de tamaño mínimo adecuado, entonces esta es representativa de la población de interés.

La aleatoriedad y el tamaño de una muestra son características que podemos controlar (el tamaño muestral se puede calcular y el investigador suele escoger, entre varios métodos de selección al azar, el que se adecúe mejor a su estudio). La representatividad, en cambio, es una cualidad de la muestra.

1.3 Parámetros y estimadores

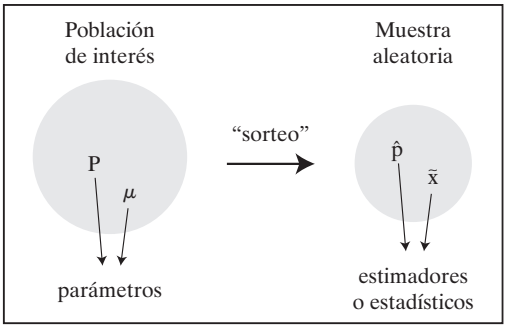
Llamaremos **inferencia estadística** a las deducciones que hacemos acerca de una población de interés, a partir de los resultados obtenidos mediante una muestra aleatoria de dicha población.

Por ejemplo, si en una muestra aleatoria se calcula que el promedio de edad es de 20 años, entonces se inferirá que el promedio de edad de la población de la cual procede la muestra debiera ser de aproximadamente 20 años, con un margen de error dado por el tamaño de la muestra. O bien, si se calcula que 38% de los individuos en la muestra es fumador, entonces se deducirá que el porcentaje de fumadores poblacional debiera ser aproximadamente 38%.

El promedio de edad y el porcentaje de fumadores poblacionales se denominan **parámetros poblacionales** (o simplemente **parámetros**). En general, un parámetro es cualquier función de los datos calculada en la población.

El promedio de edad y el porcentaje de fumadores calculados en la muestra, y utilizados para aproximar el verdadero valor poblacional, se denominan **estimadores muestrales** o **estadísticos**. Por lo común, un estimador puede ser cualquier función calculada con los datos muestrales y, como es un valor que representa a la muestra completa, suele llamarse también **medida resumen**.

[FIGURA 1.1]



Parámetros poblacionales y estimadores muestrales. P: proporción poblacional de individuos con alguna característica de interés; μ: promedio poblacional de una variable de interés.

La muestra debiera ser un buen reflejo de la población. De esta forma, el objetivo cuando se estiman parámetros poblacionales es que:

$$\hat{p} \approx P \quad y \quad \bar{x} \approx \mu$$

Esto solo es posible si la muestra escogida es representativa de la población de interés.

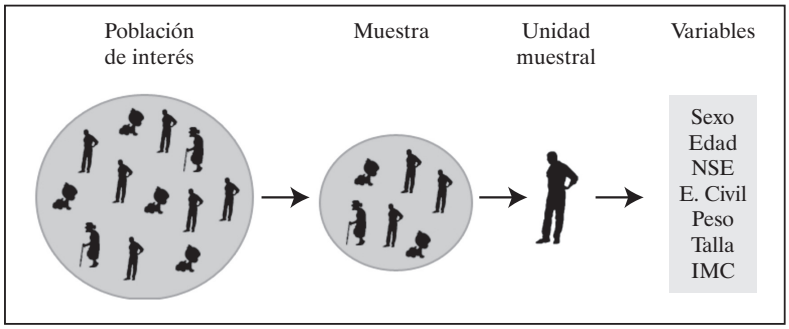
Un promedio o una proporción poblacional no es el único parámetro de interés en un estudio. Como puede ser cualquier función de los datos, podría interesar la mediana, varianza, desviación estándar, percentiles u otras funciones menos conocidas.

Por ejemplo, si se asume que la edad fértil es entre los 15 y 49 años, según el censo de 2002, el porcentaje de mujeres en edad fértil es 52%. Si se quiere hacer un estudio sobre el número de hijos promedio por mujer, en base a una muestra de 400 mujeres de Puente Alto, y se observan 190 mujeres en edad fértil, entonces 52% es el parámetro poblacional y 47,5% (190 mujeres de un total de 400) es el estimador de ese parámetro.

1.4 Variables aleatorias

Una vez que seleccionamos un conjunto de individuos de la población para que formen parte de la muestra aleatoria, cada uno de estos individuos es caracterizado por un conjunto de variables de interés en el estudio.

[FIGURA 1.2]



Población, muestra, unidad muestral y variables que pueden determinarse a partir de esta.

Se denomina **unidad muestral** a cada elemento susceptible de ser seleccionado. Habitualmente la unidad muestral corresponde a un individuo, aunque no siempre es así. Por ejemplo, en un estudio de contaminación intradomiciliaria la unidad muestral podría ser un hogar (y no los sujetos que viven en ella). En un estudio en que interesa analizar el cambio a través de los años en el número de aves acuáticas en una reserva ecológica, la unidad muestral será un número de aves en cada momento del tiempo.

Llamaremos **variable** a cualquier característica que tome dos o más valores en una población. Por ejemplo, edad, sexo, hábito tabáquico, presencia o ausencia de una patología, valores de colesterol total, HDL y triglicéridos en un examen de lípidos, etc. Nosotros estudiaremos **variables aleatorias**, para las cuales no es posible anticipar su resultado, aun cuando se intente controlar los factores que puedan afectarlas. Visto de otra forma, si al mantener constantes las condiciones experimentales no es posible predecir el valor de una variable, entonces se está frente a una variable aleatoria.

Nótese que si la característica toma solo un valor, entonces es una **constante** y no es de interés estadístico. Por ejemplo, en el estudio de las dueñas de casa que usan Detergente X, la ciudad de residencia es constante, por lo que no es útil para discriminar entre las mujeres que usan el detergente de las que no lo hacen.

Determinar cuáles variables aleatorias deben ser medidas a cada unidad muestral es de vital importancia para el estudio. Por ejemplo, si interesa investigar factores de riesgo de infarto al miocardio, no puede dejar de medirse la edad, el hábito tabáquico o el consumo de alcohol, ya que todos son factores que se asocian con el fenómeno en estudio.

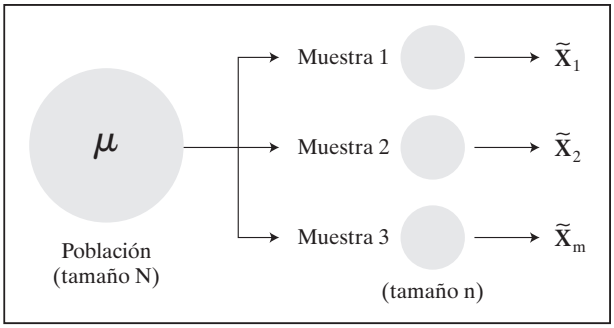
1.5 Variabilidad muestral

Cuando tomamos una muestra aleatoria de una población, lo que hacemos es observar una de muchas posibles muestras aleatorias de la población de interés.

Por ejemplo, si la población está compuesta de N = 50 individuos y decidimos tomar una muestra de n = 5 de ellos, entonces el número de muestras posibles de obtener es 2.118.760. Aunque el número de muestras posibles puede ser muy grande, en la práctica nosotros solo tenemos acceso a una de ellas.

En consecuencia, si es de interés calcular el promedio muestral, lo que obtenemos es uno de muchos promedios muestrales posibles de conseguir.

[FIGURA 1.3]



De una población se pueden extraer muchas muestras diferentes de tamaño ***n***. A partir de cada muestra se obtiene un promedio muestral distinto.

Claramente, si tomamos distintas muestras, el estimador será siempre diferente. Esto es conocido como **variabilidad muestral**.

Luego, ¿cómo podemos saber si nuestro promedio muestral es un buen estimador de μ?

La respuesta a esta pregunta está dada por el tamaño de la muestra y el método de selección. Para ilustrarlo, consideremos el siguiente ejemplo.

Ejemplo 1.2. (Se agradece la colaboración del Dr. Guillermo Marshall R., Profesor Titular de la Facultad de Matemáticas, por su aporte de este ejemplo). La siguiente hoja muestra las edades en que 350 personas enfermaron de cáncer al pulmón en cierta comunidad (asumamos que esta es la población completa). La edad media de los 350 pacientes de cáncer al pulmón es μ **= 61.9 años**.

[TABLA 1.1]

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	64	66	46	71	65	73	61	70	27	80	52	61	39	76
2	75	58	90	73	85	75	44	74	52	80	50	65	45	78
3	64	76	73	50	59	54	74	60	42	74	83	60	83	73
4	84	65	41	73	57	73	69	91	70	47	54	29	51	55
5	73	59	63	66	48	60	55	62	55	63	75	80	67	92
6	79	75	93	45	72	60	78	72	47	65	77	57	50	64
7	63	73	75	49	61	41	70	72	43	64	69	43	63	57
8	71	42	45	71	62	38	79	50	50	49	54	67	65	49
9	76	44	72	65	64	49	60	71	61	71	59	59	62	58
10	51	50	**73**	78	58	76	53	71	44	53	70	74	72	66
11	49	63	**68**	62	71	67	60	80	63	30	81	81	39	81
12	51	63	**59**	67	33	62	61	63	51	45	56	43	49	79
13	65	38	**40**	80	63	57	67	42	57	71	46	58	92	53
14	68	76	**81**	65	50	79	42	81	47	79	46	77	69	62
15	49	63	**72**	62	62	53	86	69	60	66	70	53	86	65
16	84	59	**40**	57	67	48	54	74	54	44	65	52	58	49
17	60	67	**70**	44	52	68	76	69	63	86	62	82	61	56
18	68	47	**59**	73	63	61	59	43	58	65	48	50	51	50
19	63	63	**72**	95	61	61	86	60	63	58	46	82	57	72
20	33	52	63	69	51	53	54	45	71	45	39	53	46	73
21	53	62	61	71	59	45	79	70	63	51	51	67	53	56
22	67	85	84	52	42	68	49	56	69	66	63	66	68	39
23	73	57	67	77	66	56	48	61	49	51	75	64	68	63
24	25	56	65	67	88	63	60	68	69	52	70	56	67	48
25	57	49	62	61	49	52	70	68	59	51	55	88	58	61

Planilla con edad de 350 personas. En el recuadro se observa una muestra de n = 10 casos.

Para observar el efecto del tamaño muestral en la estimación de μ (es decir, en el cálculo de $\overline{x}$), consideremos muestras de n = 10 casos consecutivos, como la que se observa en el recuadro de la **Tabla 1.1**. Se obtuvieron 40 muestras de 10 casos, y se calculó la edad promedio para cada muestra, obteniéndose los siguientes resultados:

[FIGURA 1.4]

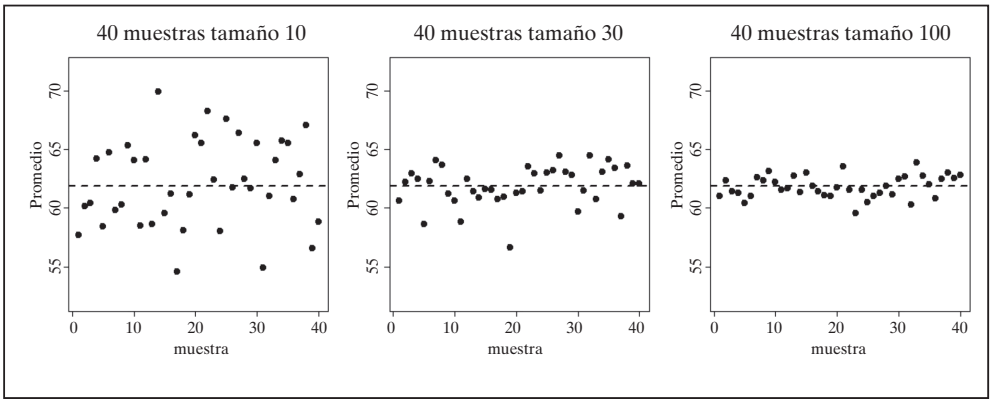
Promedio de los 10 casos del recuadro									
63,7	62,4	56,5	66,9	61,7	55,7	67,4	58,9	62,7	59,1
61,6	70,1	58,8	59,6	57,3	59,3	59,0	60,6	61,6	65,8
65,2	57,9	53,6	65,0	59,5	57,1	66,3	57,2	66,2	57,7
68,0	65,0	65,3	65,5	57,3	63,1	60,1	66,0	59,8	60,5

Promedios obtenidos al repetir 40 veces el experimento de tomar una muestra n = 10.

Posteriormente, se seleccionaron **n = 30** casos consecutivos y se calculó la edad media, y esta operación se realizó 40 veces. El mismo procedimiento se siguió con muestras de **n = 100** casos consecutivos.

En las figuras siguientes se observan los promedios muestrales obtenidos en cada grupo de experimentos **(Figura 1.5)**. La línea horizontal representa el promedio poblacional (61,9 años). En estas se observa que la variabilidad muestral es menor en la medida en que el tamaño muestral aumenta. Después, para obtener un buen estimador de un parámetro poblacional, lo recomendable es tomar una muestra lo más grande posible.

[FIGURA 1.5]



Promedios de edad obtenidos al extraer 40 muestras de distintos tamaños de una población de 350 (se observa que a medida que el tamaño de la muestra es mayor, los promedios obtenidos se aproximan más al promedio poblacional).

1.6 Tipos de muestreo

La selección de una muestra de la población de interés es de vital importancia para la obtención de resultados válidos. Si la muestra no es representativa de la población de la que procede, todos los cálculos que se hagan serán válidos solo para la muestra, sin posibilidad de extrapolar estos resultados a los individuos que no fueron incluidos en ella.

En general, estaremos interesados en muestras aleatorias, las cuales implican una selección al azar (mediante un sorteo) de los individuos que componen la muestra, en alguna etapa del proceso de muestreo. Este tipo de muestreo se denomina muestreo probabilístico.

Los principales tipos de muestreo aleatorio son el muestreo aleatorio simple, el muestreo estratificado y el muestreo sistemático. Además, actualmente adquieren mayor importancia tipos de muestreo más complejos, como el muestreo por conglomerados.

1.6.1 Muestreo aleatorio simple

Es un método de selección en que todos los elementos de la población tienen la misma probabilidad de ser elegidos en la muestra. En este tipo de muestreo se asume que la población en estudio es homogénea respecto a las variables que afectan al fenómeno estudiado.

Para aplicar este método es necesario tener un registro de todos los sujetos poblacionales (por ejemplo, un listado de los RUT, del número de ficha clínica, etc.).

La selección de los individuos muestrales podría hacerse con métodos tan simples como una bolsa con papeles numerados o con una tómbola (si la población fuera muy pequeña), hasta el uso de tablas de números aleatorios o la generación de números aleatorios mediante un computador.

Ejemplo 1.3. Supongamos que se quiere seleccionar 10 individuos de una población de 1.000. Para obtener esta muestra con Excel, podemos usar la función **Aleatorio()**, que genera números al azar entre 0 y 1. Por lo tanto, si la función se multiplica por 1.000 (que corresponde al tamaño de la población), se generan números entre 0 y 1.000.

Realizando lo anterior diez veces, se obtuvieron los siguientes números:

317,8 957,4 143,6 132,8 720,8 948,6 152,6 421,4 316,8 5,0

Luego, la muestra aleatoria simple está compuesta por los individuos:

5, 133, 144, 153, 317, 318, 421, 721, 949 y 957.

No obstante, si se hiciera nuevamente el proceso de usar la función Aleatorio(), se obtendrá una muestra distinta a la descrita.

1.6.2 Muestreo estratificado

Cuando la población es heterogénea respecto a una o más variables que afecten al fenómeno estudiado, seleccionar los datos mediante muestreo aleatorio simple podría resultar en una muestra no representativa de la población. En este caso, las conclusiones derivadas del análisis de los datos serían inválidas.

Por ejemplo, si las variables de interés tienen un comportamiento distinto según nivel socioeconómico (NSE), un muestreo aleatorio simple podría resultar en una proporción de individuos en cada NSE distinta a la observada en la población y por lo tanto los estimadores serían incorrectos.

Por ello, el investigador puede segmentar la población en **estratos**, los que corresponden a subconjuntos heterogéneos entre sí, pero que agrupan unidades homogéneas.

El muestreo estratificado es un método de selección en que se obtiene una muestra aleatoria simple de cada estrato por separado y se calculan los estimado-

res de parámetros (medias, proporciones, etc.) para cada estrato. Finalmente, se calcula un promedio ponderado de los estimadores de los estratos para obtener la medida resumen de interés.

Algunos problemas de investigación en los que podría ser útil usar muestreo estratificado son los siguientes:

- Interesa determinar el gasto promedio en alimentación de los hogares de cierta ciudad. Como el nivel de gasto es una característica que depende fuertemente del nivel socioeconómico familiar (NSE), conviene hacer estratos de la ciudad agrupando los hogares con niveles socioeconómicos semejantes. Así, la ciudad se podría dividir en zonas de NSE bajo, medio y alto, formando tres estratos. Al interior de cada estrato se toma una muestra aleatoria simple de hogares y se cuantifica el gasto en alimentación de cada hogar.
- En un muestreo para estimar la cosecha total de café en un país centroamericano, se sabe que la región ecológica donde se ubican los árboles influye mucho en su productividad. Después, sería conveniente estratificar las regiones según altura sobre el nivel del mar, nivel de vientos y temperatura antes de seleccionar los predios y determinar la productividad.

Respecto al número de individuos por seleccionar de cada estrato, existen dos criterios principales:

- **Asignación proporcional**. El número de individuos por seleccionar de cada estrato es proporcional al tamaño poblacional del estrato. Por ejemplo, si 25% de los habitantes de cierta ciudad son de nivel socioeconómico bajo, 65% de nivel medio y 10% de nivel alto, y se quiere una muestra estratificada de n = 120 casos, entonces, usando asignación proporcional, se debieran muestrear 30, 78 y 12 casos de cada NSE, respectivamente.
- **Asignación óptima**. El número de individuos por seleccionar de cada estrato es proporcional a la variabilidad de la característica en estudio al interior del estrato. Por ejemplo, si el gasto en alimentación presenta el doble de variación en el NSE alto que en los niveles medio y bajo, entonces se podría muestrear el doble de casos del NSE alto que de los otros dos niveles.

1.6.3 Muestreo sistemático

Este método de selección aleatoria es aplicable cuando los elementos de la población están ordenados físicamente y no existe un registro escrito o computacional, que permita hacer una selección por muestreo aleatorio simple. Por ejemplo, las fichas clínicas de un hospital, ordenadas según fecha de hospitalización en un estante, sería una situación adecuada para usar este tipo de muestreo.

Si la población tiene N elementos y se quiere una muestra aleatoria sistemática de n elementos, el procedimiento es el siguiente:

i) Calcule el tamaño del salto sistemático k = N/n. Si k tiene decimal, use la parte entera del número.

ii) Elegir un número entero al azar, r, entre 1 y k.

iii) Seleccionar de la población ordenada los elementos en la posición

r, r + k, r + 2k,…, r + (n-1)k

Al final del proceso, se tendrá una muestra de *n* elementos seleccionados sistemáticamente.

Ejemplo 1.4. Si tenemos 478 fichas clínicas y necesitamos seleccionar 55 para una encuesta de calidad de atención, tenemos:

N = 478 (tamaño de la población)

n = 55 (tamaño de la muestra)

k = 478 / 55 = 8.69. Usaremos saltos de 8 unidades.

Puede usar la función Aleatorio() de Excel para elegir un número al azar entre 1 y 8 ("= aleatorio()*8"). Supongamos que se obtiene r = 7.

Los números seleccionados serán 7, 15, 23, 31,…, 439 y 447.

Este método tiene la ventaja de que es fácil de aplicar. Sin embargo, se asume que el orden de los elementos de la población no afectará la estimación del parámetro de interés. Una desventaja importante es que este método es cada vez menos aceptado en publicaciones científicas.

1.6.4 Muestreo por conglomerados

Cuando es de alto costo realizar un muestreo aleatorio simple o cuando este último es inaplicable debido a que los individuos que componen la población no están identificados, un muestreo por conglomerados puede ser un método de selección adecuado.

Los conglomerados son divisiones de la población en que los elementos al interior de cada uno son heterogéneos, pero existe homogeneidad entre estas agrupaciones. Es decir, se quiere que haya "diversidad" al interior de cada conglomerado, pero que no importe cuáles conglomerados están presentes en la muestra, ya que entre ellos no hay mucha diferencia.

Esto es opuesto a lo que ocurre con los estratos, ya que aquí interesa que los individuos al interior de cada uno sean homogéneos entre sí y haya heterogeneidad entre los estratos.

El muestreo por conglomerados es un sistema de selección al azar que consta de dos fases principales: primero se eligen conglomerados al azar y luego se seleccionan elementos al interior de estos mediante un muestreo aleatorio simple.

Ejemplo 1.5. Si se quiere una muestra de 600 viviendas de una ciudad, podría ser de alto costo hacer muestreo aleatorio simple, ya que con seguridad se tendría que recorrer toda la ciudad. Si se toma una muestra por conglomerados, se podrían seleccionar al azar 20 zonas de la ciudad (entendiendo por zona un conjunto de varias manzanas), luego seleccionar 10 manzanas de cada zona y por último 3 viviendas de cada manzana, teniéndose una muestra total de 600 viviendas.

1.6.5 Selección con y sin reposición

Se asume en los tipos de muestreo anteriores que estos se realizan **sin reposición**. Es decir, sin devolver el elemento seleccionado a la población, después de ser observado. En este caso, la probabilidad de observar nuevamente el mismo elemento es cero, y la probabilidad de observar cualquier otro elemento se ve afectado por las observaciones anteriores.

Un muestreo es **con reposición** cuando cada elemento seleccionado es devuelto a la población de la cual procede después de ser observado. En este caso, la población siempre contiene los mismos elementos, por lo que todos conservan su probabilidad inicial de ser observados.

Nótese que, aunque en el muestreo sin reposición se altera la probabilidad de seleccionar un elemento, cuando ya han sido seleccionados otros previamente, si la población es lo suficientemente grande, esta probabilidad se puede considerar constante.

1.7 Tipos de variables

Como se mencionó antes (ver **punto 1.4**) cada uno de los individuos seleccionado en la muestra es caracterizado por un conjunto de variables de interés en el estudio. Estas variables podrían ser registradas, por ejemplo, en una planilla Excel. La planilla siguiente muestra el sexo, edad, nivel socioeconómico (NSE), estado civil y peso de seis individuos:

[TABLA 1.2]

Sexo	Edad	NSE	Estado civil	Peso
F	26	Alto	1	46,5
F	34	Medio	2	55,3
F	21	Medio	1	50,0
M	44	Bajo	2	73,1
F	32	Medio	3	54,7
M	30	Alto	2	68,5

Segmento de una base de datos con cinco variables (sexo, edad, NSE, estado civil y peso) para seis casos.

Cada variable registrada se puede clasificar en uno de los siguientes tipos:

- **Variable nominal.** Es aquella en que podemos clasificar sus valores en clases o categorías, sin establecer un ordenamiento sugerido por la magnitud de sus valores.

Esto significa que los valores con que se identifica cada nivel de la variable son arbitrarios. Por ejemplo, la variable sexo es nominal, ya que podemos identificar sus niveles mediante M (Masculino) y F (Femenino); o bien H (Hombre) y M (Mujer); o mediante 1 (Mujer) y 2 (Hombre), etc.

Otras variables nominales son: estado civil, causa de muerte, ciudad de residencia, tipo de parto, etc.

- **Variable ordinal.** Es un tipo de variables en la que sus valores o clases se pueden ordenar. Incluye variables con categorías (como gravedad de una enfermedad definida como leve, moderada o severa) y scores (como el test de Apgar del recién nacido).

En las ciencias de la salud, se generan muchas variables ordinales que intentan cuantificar características difíciles o imposibles de medir directamente, como la gravedad cardiaca medida usando scores como APACHE o TISS; el desarrollo puberal medido con escala de Tanner; el estado nutricional que se puede definir como bajo peso, normal, sobrepeso y obeso, etc.

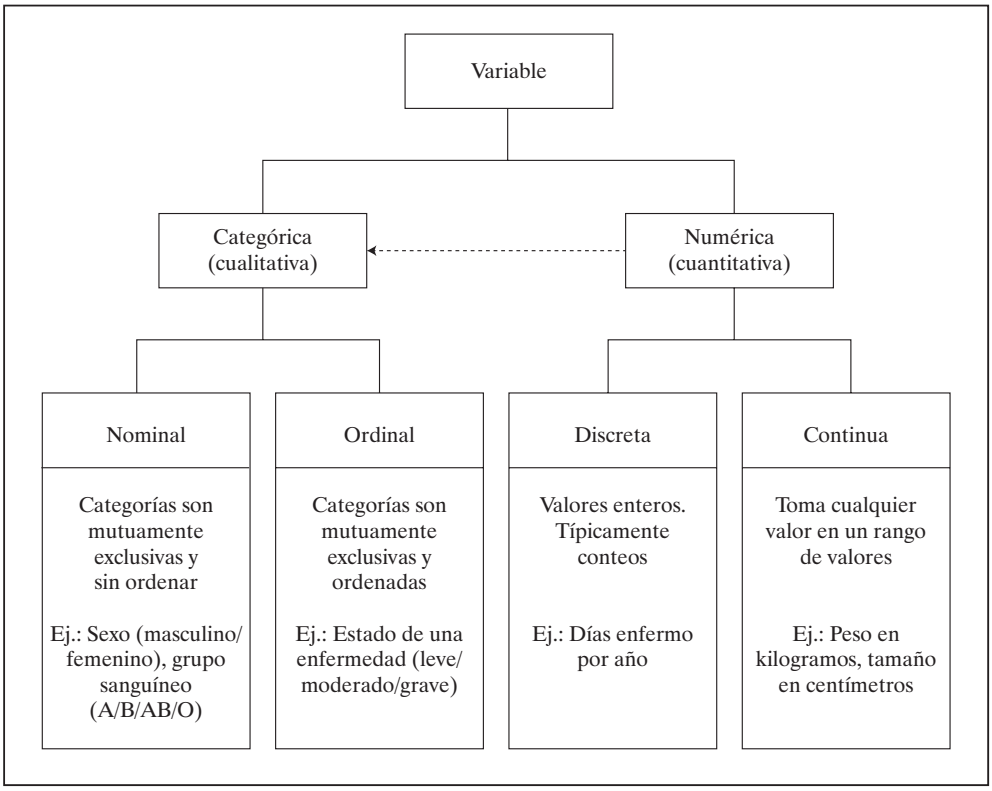
Como se observa, las variables ordinales no tienen unidad de medida. Tampoco tiene sentido cuantificar la diferencia o la razón entre dos valores ordinales. Por ejemplo, si una persona tiene un puntaje de gravedad igual a 30 y otra tiene un puntaje de gravedad igual a 60 (asumiendo que mayor puntaje significa mayor gravedad), no podemos decir que la segunda tenga el doble de gravedad que la primera; solo podemos decir que la segunda está más grave.

- **Variable intervalar.** Es una variable cuantificable de manera objetiva, por lo que posee un orden natural en sus valores y es posible medir la diferencia entre dos valores. Generalmente tiene unidad de medida.

Una variable intervalar se denomina discreta cuando no puede tomar decimales, como en las variables de conteo (número de hijos, número de caries, días de hospitalización, etc.). Se denomina continua cuando toma cualquier valor en un intervalo (como el peso, talla, índice de masa corporal, triglicéridos, etc.).

El esquema siguiente resume los tipos de variables:

[FIGURA 1.6]



Clasificación de tipos de variables.

Los dos tipos de variable de interés estadístico

Para la mayoría de las descripciones y análisis estadísticos basta con identificar dos tipos de variables:

- **Variables categóricas.** Son aquellas para las cuales no es posible y no tiene sentido obtener su promedio. Incluye las nominales (como sexo o estado civil), las ordinales con pocos niveles (como nivel socioeconómico o severidad de una enfermedad) y las intervalares en rangos (como grupos etarios o peso de recién nacido en rangos).
- **Variables numéricas.** Son aquellas para las cuales tiene sentido obtener su promedio. Incluye las intervalares (como peso o número de hijos) y las ordinales que toman un rango amplio de valores (como puntaje Apgar o score Apache de gravedad cardiaca).

Nótese que una variable numérica puede transformarse en categórica construyendo rangos. Asimismo un conjunto de variables categóricas puede transformarse en una variable numérica construyendo scores.

1.8 Notación para variables aleatorias y sus mediciones

Por lo general se utiliza la letra N mayúscula para referirse al tamaño de una población (asumiendo que es una población finita) y la letra n minúscula para referirse al tamaño de una muestra.

Cuando nos referimos a una variable aleatoria en forma genérica (la variable Sexo, la variable Peso, etc.), usaremos letras X, Y o Z mayúsculas.

Al referirnos a los valores que toma una variable aleatoria X en una muestra de tamaño n, usaremos la letra x minúscula con subíndices: $x_1, x_2,\ldots, x_n$. Donde x_1 es el valor de X en el primer sujeto muestral, x_2 el valor en el segundo sujeto y así sucesivamente.

Cuando aludimos a los valores muestrales ordenados de la variable aleatoria X, usaremos la notación: $x_{(1)}, x_{(2)},\ldots, x_{(n)}$. De modo que $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Nótese que $x_{(1)}$ es el mínimo valor muestral de X y $x_{(n)}$ es el máximo.

Usaremos el símbolo Σ para referirnos a la suma de un conjunto de valores. Por ejemplo:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

Emplearemos el símbolo π para aludir al producto de un conjunto de valores. Por ejemplo:

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \ldots \times x_n$$

Ejemplo 1.6. Consideremos la Tabla 1.2, que muestra los datos de algunas variables para n = 6 individuos.

Sea X la variable Edad. Luego, los valores muestrales de X son:

$$x_1 = 26,\ x_2 = 34,\ x_3 = 21,\ x_4 = 44,\ x_5 = 32 \text{ y } x_6 = 30.$$

Los valores muestrales ordenados son:

$$x_{(1)} = 21,\ x_{(2)} = 26,\ x_{(3)} = 30,\ x_{(4)} = 32,\ x_{(5)} = 34 \text{ y } x_{(6)} = 44.$$

La suma de los *n* valores muestrales es:

$$\sum_{i=1}^{n} x_i = 26 + 34 + 21 + 44 + 32 + 30 = 187$$

El producto de los n valores muestrales es:

$$\prod_{i=1}^{n} x_i = 26 * 34 * 21 * 44 * 32 * 30 = 784143360$$

1.9 Descripción de variables categóricas

Las variables categóricas se describen principalmente mediante dos medidas resumen: el **número** y el porcentaje de casos en cada nivel de la variable. Como veremos más adelante, la **proporción** de casos en cada categoría también es una medida resumen de interés, ya que estima la probabilidad de ocurrencia de un evento en la población.

Los resultados obtenidos para una variable categórica se muestran en una **tabla de frecuencias**, que es la forma en que habitualmente los programas estadísticos entregan el resumen de una variable categórica.

Verbigracia, la tabla siguiente resume los resultados obtenidos para la edad en que enferman de cáncer al pulmón los 350 casos descritos en el **Ejemplo 1.2.**

[TABLA 1.3]

Grupos de edad	Número de casos	Frecuencia relativa	Porcentaje	Porcentaje acumulado
<30	3	0,009	0,9	0,9
30-39	9	0,026	2,6	3,4
40-49	56	0,160	16,0	19,4
50-59	78	0,223	22,3	41,7
60-69	109	0,311	31,1	72,9
70-79	66	0,189	18,9	91,7
80-89	24	0,069	6,9	98,6
90 +	5	0,014	1,4	100,0
Total	**350**	**1,000**	**100,0**	

Tabla de frecuencias de 350 casos según el grupo de edad al cual pertenecen.

La interpretación de las columnas de la tabla de frecuencias es la siguiente:

- La primera columna da cuenta de los niveles observados en la muestra para la variable tabulada.
- La segunda columna indica el número de individuos en cada nivel de la variable. La última fila de esta columna presenta el total de casos.
- La tercera columna indica la proporción de sujetos en cada nivel (número de casos en el nivel dividido por el total de casos tabulados). La suma de estas siempre debe ser 1.

- La cuarta columna indica el porcentaje de casos en cada nivel (frecuencia relativa * 100). El total siempre debe sumar 100%.
- La última columna indica el porcentaje de casos hasta el nivel que se esté observando (por ejemplo, el porcentaje de casos que tiene menos de 80 años es 91,7%).

Por ejemplo, de la tabla de frecuencias podemos observar que:

- El 22,3% de la muestra tiene edad entre 50 y 59 años.
- El 41,7% tiene menos de 60 años.
- Si suponemos que los datos tabulados provienen de una población de tamaño 10.000, entonces podemos decir que existen aproximadamente 2.230 sujetos en la población con edad entre 50 y 59 años.
- El porcentaje de personas con 60 o más años es 100-41,7 = 58,3%.

1.10 Presentación gráfica de variables categóricas

Generalmente las variables categóricas se representan mediante **gráficos de barras** y **gráficos circulares** (también llamados tortas o pies).

Gráfico de barras

Un gráfico de barras es aquel en que los niveles de la variable se representan por barras verticales, cuya altura indica el número de casos, el porcentaje o la proporción de individuos en cada nivel.

[FIGURA 1.7]

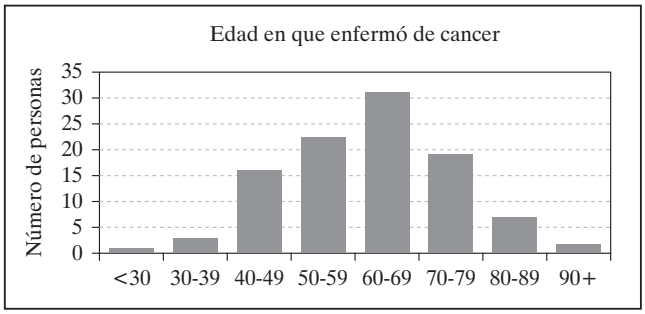


Gráfico de barras de datos en Tabla 1.3.

La figura anterior muestra el número de casos en cada nivel de la variable, por lo que es adecuado para hacer comparaciones entre los grupos graficados. Sin embargo, si se grafica el porcentaje es posible, además, hacer estimaciones a la población (los porcentajes muestrales estiman a los poblacionales). Igualmente,