PROBABILIDAD Y ESTADÍSTICA
Aplicaciones y métodos

George C. Canavos
VIRGINIA COMMONWEALTH UNIVERSITY

Traducción:
Edmundo Gerardo Urbina Medal
Departamento de Ingeniería Eléctrica
UAM Ixtapalapa

Revisión Técnica:
Gustavo Javier Valencia Ramírez
Doctor en Matemáticas
Profesor Titular
Departamento de Matemáticas
Facultad de Ciencias
UNAM

MÉXICO • BUENOS AIRES • CARACAS • GUATEMALA
LISBOA • MADRID • NUEVA YORK • PANAMÁ • SAN JUAN
SANTAFÉ DE BOGOTÁ • SANTIAGO • SÃO PAULO
AUCKLAND • HAMBURGO • LONDRES • MILÁN • MONTREAL
NUEVA DELHI • PARÍS • SAN FRANCISCO • SINGAPUR
ST. LOUIS • SIDNEY • TOKIO • TORONTO

CAPÍTULO SIETE

Muestras aleatorias y distribuciones de muestreo

7.1 Introducción

En el capítulo uno se mencionó que para comprender la esencia de la inferencia estadística es necesario comprender la naturaleza de una población y de una muestra. Una población representa el "estado de la naturaleza" o la forma de las cosas con respecto a un fenómeno aleatorio en particular, mismo que puede identificarse a través de una característica medible X. La manera en que ocurren las cosas en relación con X puede definirse por un modelo de probabilidad que recibe el nombre de distribución de probabilidad de la población. Por otro lado, una muestra es una colección de datos que se obtienen al llevar a cabo repetidos ensayos de un experimento para lograr una evidencia representativa acerca de la población en relación con la característica X. Si la manera de obtener la muestra es imparcial y técnicamente buena, entonces la muestra puede contener información útil con respecto al estado de la naturaleza y a partir de ello se podrán formular inferencias. Ahora bien, estas últimas son inductivas y, por lo tanto, están sujetas a riesgo, dado que representan un razonamiento que va de lo particular a lo general.

En los capítulos cuatro, cinco y seis se examinaron con detalle algunas distribuciones de probabilidad que pueden servir como modelo para la distribución de una población de interés. En los capítulos restantes el principal objetivo es examinar distintas técnicas por medio de las cuales puede aplicarse el proceso inductivo de la inferencia estadística para proporcionar resultados útiles y confiables. La inferencia estadística se define como *la colección de técnicas que permiten formular inferencias inductivas y que proporcionan una medida del riesgo de éstas.* En este capítulo se establecerán algunos conceptos teóricos básicos con respecto al muestreo y a la inferencia estadística. La aplicación de estos conceptos se dará con gran detalle en capítulos posteriores.

7.2 Muestras aleatorias

Como la inferencia estadística se formula con base en una muestra de objetos de la población de interés, el proceso por medio del cual se obtiene será aquél que asegure

la selección de una buena muestra. En el capítulo uno se expuso que una manera de obtener una buena muestra resulta cuando el proceso de muestreo proporciona, a cada objeto en la población, una oportunidad igual e independiente de ser incluido en la muestra. Si la población consiste de N objetos y de éstos se selecciona una muestra de tamaño n, el proceso de muestreo debe asegurar que cada muestra de tamaño n tenga la misma probabilidad de ser seleccionada. Este procedimiento conduce a lo que comúnmente se conoce como una *muestra aleatoria simple*. En este contexto, la palabra "aleatorio" sugiere una total imparcialidad en la selección de la muestra.

La naturaleza de la inferencia inductiva demanda una muestra aleatoria porque la selección de ésta se lleva a cabo con el fin de proporcionar los medios adecuados para que pueda formularse una inferencia con respecto a alguna característica de la población de interés. Por ejemplo, pueden formularse inferencias de ciertas condiciones que se suponen válidas para la población si la muestra que se observó se encuentra o no dentro de la variación muestral, misma que prevalecerá si las condiciones son verdaderas. De esta forma la calidad de la aleatoriedad en una muestra asegura la aplicación correcta de la probabilidad para evaluar el riesgo inherente en un proceso inductivo.

En este momento es importante estructurar el concepto de una muestra aleatoria simple empleando para ello los conceptos de probabilidad que se presentaron en los capítulos dos al seis. Para llevar a cabo lo anterior, primero se examinarán situaciones que se presentan, de manera frecuente, en los muestreos. La primera de éstas surge en muchos experimentos que involucran fenómenos aleatorios en la ingeniería y las ciencias físicas. En estos casos la población de interés no consiste en objetos tangibles a partir de los cuales se selecciona un cierto número para formar la muestra. Más bien, la población se considera constituida por un número infinito de posibles resultados para alguna característica medible de interés. Esta característica generalmente es una medición física como el nivel de concentración de un contaminante, la demanda de un producto o el tiempo de espera en un servicio. Sea X una característica medible y $f(x; \theta)$ la función de densidad de probabilidad de la distribución de la población. El siguiente procedimiento es una forma de muestreo para este tipo de población:

1. Se diseña un experimento y se lleva a cabo para proporcionar la observación X_1 de la característica medible X. El experimento se repite bajo las mismas condiciones proporcionando el valor X_2. El proceso se continúa hasta tener n observaciones $X_1, X_2, \ldots, X_n$ de la característica X.

En este procedimiento de muestreo, las observaciones muestrales se colectan a través de ensayos independientes que ocurren cada vez que el experimento se repite bajo condiciones idénticas para todos los factores que son controlables. En este contexto, cada observación del i-ésimo experimento se considera como una selección de la misma fuente que proporciona la observación de cualquier otro ensayo para X. En esencia, las observaciones bajo las mismas condiciones como resultado de repetidos ensayos independientes de un experimento, constituye lo que se denomina un *muestreo aleatorio con reemplazo*. De acuerdo con lo anterior, cada una de las observaciones $X_1, X_2, \ldots, X_n$ es una variable aleatoria cuya distribución de probabilidad es idéntica a la de la población.

Una situación diferente se presenta cuando se lleva a cabo una selección de objetos tangibles de una población que consiste en un número finito de objetos (seres humanos, animales, componentes mecánicos o eléctricos, etc.). La característica medible de interés puede ser un atributo, como el estado de un componente (defectuoso o no defectuoso), la opinión de una persona con respecto a cierto tema (a favor o en contra) o una medición cuantitativa como el CI de una persona o el tiempo de duración de un componente. Existen dos formas para obtener muestras aleatorias de este tipo de población:

2. Después de llevar a cabo una mezcla adecuada de los objetos de la población, se extrae uno y se observa la característica medible. Esta observación será X_1. El objeto se regresa a la población y ésta vuelve a mezclarse; después se extrae el segundo objeto. X_2 se constituye por la segunda observación. El proceso se continúa de esta forma hasta que se han extraído n objetos para tener una muestra de observaciones $X_1, X_2, \ldots, X_n$ de la característica X.
3. Después de una mezcla adecuada de los objetos que constituyen la población, n de éstos se seleccionan uno después de otro sin reemplazo. Este proceso proporciona una muestra de observaciones $X_1, X_2, \ldots, X_n$ de la característica X.

Nótese que la técnica 2 constituye un muestreo con reemplazo y la técnica 3 es un muestreo sin reemplazo. En el contexto general de una muestra aleatoria simple, la técnica recibe el nombre de aleatoria. Cuando los objetos se extraen después de una selección equitativa. Por consiguiente, la técnica de muestreo dos recibe el nombre de muestreo aleatorio con reemplazo, y la técnica tres el de muestreo aleatorio sin reemplazo. En la técnica dos, cada una de las observaciones $X_1, X_2, \ldots, X_n$ es una variable aleatoria cuya distribución de probabilidad es identica a la de la población, puesto que en cada extracción ésta tiene su forma original. En la técnica de muestreo tres, las observaciones $X_1, X_2 \ldots, X_n$ también son variables aleatorias cuyas distribuciones marginales son iguales a las de la población. Es decir, puede demostrarse que aun a pesar de que los objetos que se extraen de la población no sean reemplazados, la distribución no condicional de X_i es idéntica a la de la población, para toda $i = 1, 2, \ldots, n$.

La diferencia básica entre las dos técnicas es la noción de independencia. En la técnica dos, las observaciones $X_1, X_2, \ldots, X_n$ constituyen un conjunto de variables aleatorias independientes e idénticamente distribuidas (IID) dado que, por el proceso de reemplazo, ninguna observación se ve afectada por otra. En la técnica tres, a pesar de que las observaciones $X_1, X_2, \ldots, X_n$ poseen la misma distribución, no son independientes.

Recuérdese que, para la técnica uno, el muestreo se lleva a cabo con reemplazo a pesar de que la población no se encuentre constituida por objetos tangibles. De hecho, la técnica de muestreo dos es un caso especial de la primera, dado que la población no se afecta después de cada extracción. Sin embargo, es interesante notar que puede preferirse el muestreo aleatorio sin reemplazo si el tamaño de la población es relativamente pequeño*. En estos casos, si el muestreo se lleva a cabo con re-

* El lector recordará que esto es precisamente lo que constituye una distribución hipergeométrica tal como se discutió en la sección 4.4.

emplazo es muy probable que el mismo objeto sea seleccionado más de una vez. Es por esta razón que en las encuestas de preferencia el muestreo se hace sin reemplazo. Por otro lado, si el número de objetos en la población es muy grande, es irrelevante si el muestreo se lleva a cabo con reemplazo o sin éste. Conforme crece el tamaño de la población, el muestreo aleatorio sin reemplazo es, en todos los intentos y para cualquier propósito, igual al muestreo aleatorio con reemplazo.

Al hablar de la inferencia estadística se supondrá la existencia de una muestra aleatoria, como la descrita por la técnica de muestreo 1, y que se define de manera formal de la siguiente manera:

Definición 7.1 Si las variables aleatorias $X_1, X_2, \ldots, X_n$ tienen la misma función (densidad) de probabilidad que la de la distribución de la población y su función (distribución) conjunta de probabilidad es igual al producto de las marginales, entonces $X_1, X_2, \ldots, X_n$ forman un conjunto de n variables aleatorias independientes e idénticamente distribuidas (IID) que constituyen una *muestra aleatoria* de la población.

Cuando el objetivo es formular una inferencia estadística, debe hacerse un intento honesto para obtener una muestra aleatoria que porporcione la base teórica necesaria para la inferencia. Desde un punto de vista práctico, lo anterior no siempre es fácil. Por ejemplo, en muchas ocasiones es difícil decidir cuándo se están manteniendo condiciones idénticas durante el proceso de reunir datos en experimentos científicos. Esto es especialmente cierto si los factores ambientales crean condiciones heterogéneas. Sin embargo, es responsabilidad del experimentador decidir cuándo una muestra observada de datos es, en gran medida, aleatoria.

Para ilustrar el proceso de muestreo en un experimento científico, supóngase que se tiene interés en la concentración de cierto contaminante en un depósito de agua. Se coloca una boya que contiene un instrumento para medir el nivel de concentración en el sitio de interés. El instrumento registra el nivel de concentración cada n intervalos. De esta forma, las observaciones $X_1, X_2, \ldots, X_n$ constituyen una muestra del nivel de concentración en el sitio de interés. Antes de que el instrumento registre el nivel de concentración para el i-ésimo periodo, la observación X_i es una variable aleatoria para $i = 1, 2, \ldots, n$. El valor registrado x_i (el valor numérico correspondiente a la observación X_i) es una *realización* de la variable aleatoria. Al final de los n intervalos las mediciones $x_1, x_2, \ldots, x_n$ que registra el instrumento son las realizaciones, o datos muestrales, de las correspondientes variables aleatorias $X_1, X_2, \ldots, X_n$. Sin embargo, es válido preguntarse si la anterior es veraderamente una muestra aleatoria. Nadie puede proporcionar una respuesta legítima sin tener información adicional. Por ejemplo, ¿está el investigador consciente de todos los sucesos que durante el periodo de muestreo podría causar un cambio significativo en el nivel de concentración del contaminante? ¿Consideró el lapso de muestreo adecuado o existen algunas fluctuaciones temporales que deben ser consideradas? ¿Es probable que el error en el instrumento sea mayor conforme transcurre el tiempo? Preguntas como las anteriores deben contestarse antes de dar un juicio definitivo sobre la aleatoriedad de la muestra.

En el contexto de la definición 7.1, la función (densidad) conjunta de probabilidad de $X_1, X_2, \ldots, X_n$ es la función de verosimilitud de la muestra dada por

$$L(\underline{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta), \tag{7.1}$$

en donde $\underline{x} = \{x_1, x_2, \ldots, x_n\}$ denota los datos muestreados. Cuando las realizaciones $\underline{x}$ se conocen, $L(\underline{x}; \theta)$ es una función del parámetro desconocido θ. La utilidad de la función de verosimilitud para estimar parámetros se examinará en el capítulo ocho.

Ejemplo 7.1 Se ilustrará el concepto de muestra aleatoria dado en la definición 7.1 mediante lo siguiente: sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de n variables aleatorias IID de una población cuya distribución de probabilidad es exponencial con densidad

$$f(x; \theta) = \frac{1}{\theta} \exp(-x/\theta), \qquad 0 < x < \infty.$$

Cuando se observa X_1 y se registra su realización x_1,

$$f(x_1; \theta) = \frac{1}{\theta} \exp(-x_1/\theta), \qquad 0 < x_1 < \infty.$$

Ahora se observa X_2 y se registra su realización x_2. Dado que X_1 y X_2 son estadísticamente independientes y tienen las mismas densidades marginales,

$$f(x_2|x_1) = f(x_2; \theta) = \frac{1}{\theta} \exp(-x_2/\theta), \qquad 0 < x_2 < \infty.$$

La función de densidad conjunta de X_1 y X_2 es

$$f(x_1, x_2; \theta) = f(x_1; \theta) f(x_2; \theta) = \frac{1}{\theta^2} \exp[-(x_1 + x_2)/\theta], \qquad 0 < x_i < \infty, i = 1, 2.$$

Por lo tanto, se desprende que para una muestra aleatoria de tamaño n

$$L(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{\theta^n} \exp[-(x_1 + x_2 + \cdots + x_n)/\theta], \qquad 0 < x_i < \infty, i = 1, 2, \ldots, n.$$

7.3 Distribuciones de muestreo de estadísticas

En los comentarios introductorios del capítulo uno se mencionó de manera breve que las características muestrales denominadas "estadísticas" se emplean para hacer inferencias con respecto a las características de la población, las que reciben el nombre de "parámetros". El objetivo de esta sección será el de examinar con detalle el papel que desempeñan las estadísticas en relación con la inferencia. En particular, se desa-

rrollará la noción de una distribución de muestreo de una estadística, que es uno de los conceptos más importantes en inferencia estadística.

Para colocar a las estadísticas en una mejor perspectiva se debe definir y analizar, de manera formal, un parámetro de población.

Definición 7.2 Un *parámetro* es una caracterización numérica de la distribución de la población de manera que describe, parcial o completamente, la función de densidad de probabilidad de la característica de interés. Por ejemplo, cuando se especifica el valor del parámetro de escala exponencial θ, se describe de manera completa la función de densidad de probabilidad

$$f(x; \theta) = \frac{1}{\theta} \exp(-x/\theta).$$

La oración "describe de manera completa" sugiere que una vez que se conoce el valor de θ entonces puede formularse cualquier proposición probabilística de interés. A manera de ilustración, si $\theta = 2$, entonces:

$$P(X > 4) = \frac{1}{2} \int_4^\infty \exp(-x/2)dx = 0.1353.$$

Por otra parte, si se especifica un valor del parámetro de forma α, de la distribución gama, la función de densidad de probabilidad

$$f(x; \alpha, \theta) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} \exp(-x/\theta)$$

no se encuentra especificada de manera completa, ya que no se ha hecho ninguna mención con respecto al valor del párametro de escala θ.

La esencia de todo lo anterior es que, dado que los parámetros son prácticamente inherentes a todos los modelos de probabilidad, es imposible calcular las probabilidades deseadas sin un conocimiento del valor de éstos. Es por esta razón que la noción de una estadística y su distribución de muestreo es muy importante en inferencia estadística. Esto es, los parámetros o sus funciones se estiman con base en estadísticas que, a su vez, se obtienen a partir de la información contenida en una muestra aleatoria.

Antes de dar la definición de una estadística, debe notarse que desde un punto de vista clásico (no bayesiano), un parámetro se considera como una constante fija cuyo valor se desconoce. Desde una perspectiva bayesiana un parámetro siempre es una variable aleatoria con algún tipo de distribución de probabilidad. Se considerará a los parámetros, principalmente desde el punto de vista clásico, aunque también se dará el punto de vista bayesiano, a fin de dar una perspectiva apropiada.

Definición 7.3 Una *estadística* es cualquier función de las variables aleatorias que se observaron en la muestra de manera que esta función no contiene cantidades desconocidas.

Considérese la muestra $\underline{X} = \{X_1, X_2, \ldots, X_n\}$ que consiste de n variables aleatorias IID con una función de densidad de probabilidad $f(x; \theta)$ que depende de un parámetro desconocido θ. Supóngase que se definen funciones como

$$T_1(\underline{X}) = (X_1 + X_2 + \cdots + X_n)/n,$$

$$T_2(\underline{X}) = (X_1^2 + X_2^2 + \cdots + X_n^2)/n,$$

$$T_3(\underline{X}) = X_1 + X_2,$$

y así sucesivamente. Todas ellas son estadísticas porque se determinan de manera completa por las variables aleatorias que contiene la muestra. De manera general, denótese una estadística por $T = u(\underline{X})$. Dado que T es una función de variables aleatorias, es en sí misma una variable aleatoria, y su valor específico $t = u(\underline{x})$ puede determinarse cuando se conozcan las realizaciones $\underline{x}$ de $\underline{X}$. Si se emplea una estadística T para estimar un parámetro desconocido θ, entonces T recibe el nombre de *estimador* de θ, y el valor específico de t como un resultado de los datos muestrales recibe el nombre *estimación* de θ. Esto es, un estimador es una estadística que identifica al mecanismo funcional por medio del cual, una vez que las observaciones en la muestra se realizan, se obtiene una estimación.

Una estadística es, sustancialmente, diferente de un parámetro. Un parámetro es una constante pero una estadística es una variable aleatoria. Además, un valor del parámetro descrito describe de manera completa un modelo de probabilidad (suponiendo una distribución uniparamétrica); ningún valor de la estadística puede desempeñar tal papel si cada uno de éstos depende del valor de las observaciones de las muestras. Y dado que las muestras se toman en forma aleatoria, ninguna muestra es más válida que cualquier otra que se haya tomado con el mismo fin.

Para ilustrar el concepto de una estadística se dará solución al siguiente problema: supóngase que se tiene interés en la duración promedio de cierta clase de batería miniatura. Se asegura que el proceso de manufactura de ésta es el mismo y que se emplean materiales idénticos. Se decide seleccionar aleatoriamente cinco pilas diarias durante 20 días. Para cada muestra diaria, las cinco baterías se someten a una prueba de duración que consiste en registrar el tiempo de operación. La prueba termina cuando todas dejan de funcionar. Como se supone que el proceso de fabricación es el mismo durante el periodo de muestreo, este esquema proporciona 20 muestras aleatorias distintas, donde cada una contiene cinco variables aleatorias independientes y distribuidas de manera idéntica. Sea $\underline{X}_j = \{X_{1j}, X_{2j}, \ldots, X_{5j}\}$ el conjunto de variables aleatorias de la $j$ésima muestra para $j = 1, 2, \ldots, 20$, y $\underline{x}_j = \{x_{1j}, x_{2j}, \ldots, x_{5j}\}$ los correspondientes tiempos de duración observados. Considérese la estadística.

$$T_j = (X_{1j} + X_{2j} + \cdots + X_{5j})/5$$

como un estimador del tiempo de duración promedio de las baterías. Si se supone que los tiempos observados son los que aparecen en la tabla 7.1, entonces para la j-ésima muestra existe una realización t_j para la estadística T_j. Es decir, cada muestra diaria proporciona una estimación de la duración promedio de las baterías.

Nótese que las estimaciones que aparecen en la tabla para la duración promedio tienen una variación que se encuentra entre 140.8 y 157.2 horas. De esta forma, existe una variabilidad inherente entre estas estimaciones. Además, para cualquier estadística se espera una variabilidad de muestra a muestra, dado que una estadística es una variable aleatoria. De hecho, para cada estadística existe lo que se conoce como su distribución de muestreo, la cual toma en cuenta la variabilidad inherente y proporciona los medios necesarios por medio de los cuales puede evaluarse la estadística. Se definirá la distribución de muestreo de una estadística con base en muestras aleatorias, de acuerdo con la definición 7.1.

Definición 7.4 La *distribución de muestreo* de una estadística T es la distribución de probabilidad de T que puede obtenerse como resultado de un número infinito de muestras aleatorias independientes, cada una de tamaño n, provenientes de la población de interés.

Dado que se supone que las muestras son aleatorias, la distribución de una estadística es un tipo de modelo de probabilidad conjunta para variables aleatorias independientes, en donde cada variable posee una función de densidad de probabilidad igual a la de las demás. De manera general, la distribución de muestreo de una estadística no tiene la misma forma que la función de densidad de probabilidad en la distribución de la población.

Para ilustrar lo anterior, considérese la distribución de muestreo de una estadística para los 20 promedios muestrales dados en la tabla 7.1. Mediante el empleo de los métodos del capítulo uno, se agrupan las 20 realizaciones en cinco clases y se obtienen las frecuencias relativas que aparecen en la tabla 7.2.

TABLA 7.1 Tiempos de duración (en horas) observados para una muestra aleatoria de baterías

Número de muestra	1	2	3	4	5	6	7	8	9	10
	163	159	150	136	136	138	155	158	135	166
	132	144	125	157	146	145	145	150	144	142
	154	139	139	168	158	150	151	153	148	156
	152	146	134	158	154	138	154	151	150	154
	148	144	156	167	156	158	141	138	148	160
Promedio de la muestra	149.8	146.4	140.8	157.2	150.0	145.8	149.2	150.0	145.0	155.6
Número de muestra	11	12	13	14	15	16	17	18	19	20
	150	154	148	149	150	147	158	164	153	135
	152	150	166	158	138	151	147	136	160	150
	163	141	148	139	153	161	141	143	156	164
	161	159	149	146	151	142	130	137	142	152
	139	153	154	136	161	149	147	152	156	144
Promedio de la muestra	153.0	151.4	153.0	145.6	150.6	150.0	144.6	146.4	153.4	149.0

TABLA 7.2 Grupos y frecuencias relativas para las 20 medias muestrales

Límites de clase	*Frecuencia de la clase*	*Frecuencia relativa*
140.6–144.0	1	0.05
144.1–147.5	6	0.30
147.6–151.0	7	0.35
151.1–154.5	4	0.20
154.6–158.0	2	0.10
Total	20	1.00

A partir de estas frecuencias relativas es evidente que la más alta concentración de tiempos de duración promedio se encuentra entre 147.6 y 151 horas, en donde los tiempos de duración promedio por debajo de 144 horas o por encima de 154.6 tienen una probabilidad muy pequeña. La distribución de muestreo de una estadística hace posible este tipo de análisis de probabilidad, esencial para valorar el riesgo inherente cuando se formulan inferencias.