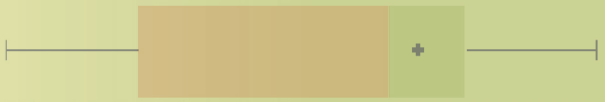
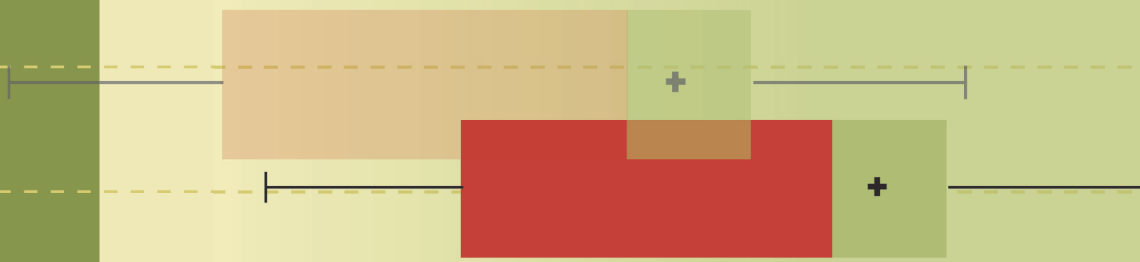
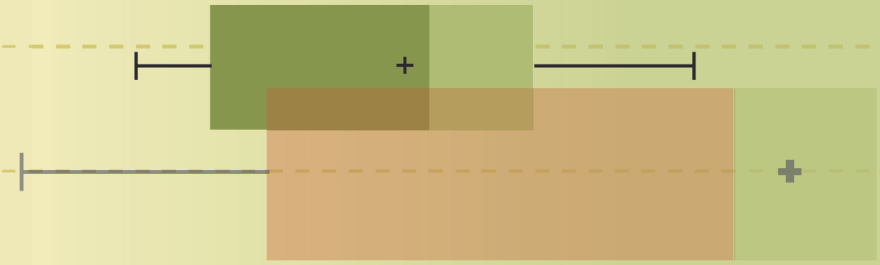
ediciones de la U
EDICIONES UNINORTE

HUMBERTO LLINÁS SOLANO





Estadística Inferencial



Estadística inferencial

Estadística inferencial

Humberto Llinás Solano

Barranquilla, 2010

Linás Solano, Humberto.

Estadística descritiva y distribuciones de probabilidad / Humberto Llinás; Carlos Rojas -- Barranquilla: Ediciones Uninorte, reimpr., 2010.

408 p.

ISBN: 978-958-8252-24-7

1. Estadística
I. Tít.

www.uninorte.edu.co
Km 5 vía a Puerto Colombia, A.A. 1569,
Barranquilla (Colombia)

http://edicionesdelau.com/
Transversal 42 #4 B-83
Bogotá (Colombia)

Primera edición, agosto de 2006
Primera reimpresión, agosto de 2009
Tercera reimpresión, noviembre de 2010

Coordinadora editoral
Zoila Sotomayor O.

Editor
Humberto Llinás Solano

Diseño de portada
Joaquín Camargo Valle

Impreso y hecho en Colombia
La Imprenta Editores
Bogotá
Printed and made in Colombia

El autor

Humberto LLinás Solano.

Licenciado en Ciencias de la Educación, con énfasis en Matemáticas, Física y Estadística de la Universidad del Atlántico. Magister en Matemáticas, convenio Universidad del Valle-Universidad del Norte. Doctor en Estadística (Dr. rer. nat.) de la Universidad Johannes Gutenberg de Mainz (Alemania). Desde 1998 se desempeña como profesor de tiempo completo de la Universidad del Norte y pertenece al grupo de investigación *Eureka* de dicha institución.

Contenido

Prefacio

Este libro fue compuesto a partir de un conjunto de notas de clases sobre la asignatura Estadística II, desarrollada en los programas de ingenierías y de administración de empresas de la Universidad del Norte. Sin embargo, está dirigido a un amplio público, ya que puede ser utilizado en cursos de estadística inferencial para ciencias sociales, ciencias biológicas, ciencias naturales y licenciatura en matemáticas. Se puede considerar como una continuación del texto *Estadística descriptiva y distribuciones de probabilidad*, también de mi autoría (véase la referencia [9]).

Enfoque

En este trabajo se asume, de manera básica, la aplicación e interpretación de los conceptos fundamentales de la estadística inferencial, sin dejar de lado la rigurosidad matemática en las distintas definiciones y teoremas que lo componen.

Descripción

Este texto se compone de:

- Seis capítulos. El capítulo 1 explica, muy brevemente, las diferentes técnicas de muestreo. En particular, tratamos (como base de la teoría desarrollada a lo largo del texto) el muestreo aleatorio simple para, luego, estudiar distribuciones muestrales de diversos estadísticos. En el 2, planteamos las estimaciones puntuales y por intervalos, mientras que, en el 3, explicamos los diferentes procedimientos de pruebas de hipótesis. El capítulo 4 desarrolla la técnica de análisis de varianza. El 5 presenta

diversos métodos relacionados con el análisis de datos categóricos. Por último, en el 6, estudiamos el modelo de regresión lineal simple y sus propiedades.

Cada capítulo, que se subdivide en secciones y subsecciones, comienza con una tabla de contenido del mismo, seguido de los objetivos y un ítem referente al empleo concreto de los conceptos a estudiar. Al final de cada sección, se incluyen numerosos ejercicios, que varían en grado de dificultad e involucran la aplicación de la teoría desarrollada en dicha sección. Así mismo, en cada capítulo proponemos una serie de ejercicios complementarios a fin de repasar todos los conceptos estudiados y entre los cuales se encuentran demostraciones de algunas propiedades matemáticas de los conceptos tratados. Estas últimas pueden ser de especial interés para los estudiantes de licenciatura en matemáticas.

Obviamente, algunas secciones y temas pueden ser omitidos de acuerdo con las circunstancias específicas, sin que esto haga perder continuidad. Ello, desde luego, está sujeto al criterio de la persona que dirija el curso.

- Tres apéndices. En el primero (apéndice A), presentamos una lista de las notaciones más usuales y especiales de nuestro texto. Así mismo, ofrecemos en éste, a manera de repaso, los conceptos teóricos, resultados y fórmulas más importantes del cálculo que se han utilizado. En el segundo (apéndice B), encontramos una guía rápida del uso del paquete estadístico *Statgraphics* en la Estadística descriptiva y una sección donde se explica el uso de la calculadora científica. Es importante señalar que al final de la sección relacionada con el uso del *Statgraphics* aparece una serie de ejercicios con el fin de poner en práctica lo explicado en dicha sección. En el tercero (apéndice C), aparecen las tablas estadísticas de uso frecuente, como normal, t de Student, F de Fisher, entre otras, así como diagramas (tablas) resumidos de distribuciones muestrales e intervalos de confianza.

- Una bibliografía, en la que enumeramos la lista de documentos y libros consultados, citados o no, que utilizamos como fuentes de información.

- Una sección que contiene las respuestas de algunos ejercicios de número impar.

- Un índice de los términos más importantes utilizados en el texto.

Características principales

Las características que marcan la diferencia entre nuestro texto y los de otros autores son:

1. *Énfasis en el análisis e interpretación de datos que presenta el computador*
   La revolución de los computadores personales ha modificado significativamente el

análisis de la información en los lugares de trabajo, así como la enseñanza de la estadística en las aulas. Pensamos que el uso de programas en forma de aplicaciones de hojas de cálculo (por ejemplo, *Statgraphics*) es parte integral del proceso de aprendizaje de la estadística. Nuestro enfoque privilegia el análisis de datos, la interpretación de los datos que se obtienen del programa *Statgraphic*s, además de una explicación detallada que indica cómo utilizar este programa, pero reduce la atención a los cálculos. Para implementarlo, hemos considerado una gran cantidad de salidas de datos (que se obtienen de los programas computacionales) y la hemos integrado en nuestro texto, dando mayor importancia a la interpretación de dichas salidas, no a los cálculos, que se realizan en forma manual.

2. *Secciones para el uso de Statgraphics*
   La disponibilidad de los computadores personales ha creado un ambiente de acceso relativamente sencillo a los programas estadísticos y las hojas de cálculo. Por ello, en lugar de apoyarnos en manuales suplementarios, utilizamos un enfoque pedagógico más conveniente al proporcionar una explicación del uso del programa. En la penúltima parte de los capítulos 2-6, hemos incluido la sección *Uso de statgraphics*, en la cual se muestra una guía sobre cómo usar este paquete estadístico en la solución de problemas.

3. *Uso de la calculadora*
   En el apartado B.2 del apéndice, hemos incluido la sección *Uso de la calculadora en la estadística*, en la que presentamos una guía para obtener algunas medidas de centralización y de dispersión y realizar, así, cálculos de regresión con las calculadoras Casio fx-82MS, fx-83MS, fx-85MS, fx-270MS, fx-300MS y fx-350MS.

4. *Ayuda pedagógica*
   Este texto presenta características que facilitan el aprendizaje:

   - Escritura de índole conversacional.
   - Tabla de contenidos al comienzo de cada capítulo, seguidos de los objetivos del mismo.
   - Un ejemplo de "Empleo de la estadística", que muestra la aplicación de al menos uno de los métodos estadísticos (explicados en cada capítulo) en ingeniería, contabilidad, finanzas, administración o mercadotecnia.
   - Cuadros que resaltan la importancia de los conceptos.
   - Ejemplos que fortalecen los conceptos que se aprendieron.
   - Series de problemas con diferentes niveles de dificultad y complejidad.
   - Explicaciones e ilustraciones de las tablas estadísticas.

5. *Archivos de datos*
   Nuestro texto viene acompañado, además, de un disquete en el que se hallan los archivos de datos para algunos ejercicios que se deben resolver con ayuda del programa *Statgraphics*.

Signos convencionales utilizados en este texto

- En el texto se citan afirmaciones de la siguiente manera:

  ▷ Números de dos niveles y encerrados en paréntesis −por ejemplo, (5.11)− significan números de las ecuaciones. El primer número corresponde al capítulo donde está la ecuación; y el segundo, al número de la ecuación dentro del capítulo.

  ▷ Todos los números de dos niveles y sin paréntesis −por ejemplo, 4.3− hacen referencia a secciones, tablas y figuras. En este caso, el primer número alude al capítulo donde está la sección, tabla o figura; y el segundo, al número de la sección, tabla o figura dentro del capítulo.

  ▷ Todos los números de tres niveles −por ejemplo, 4.4.5− se refieren a definiciones, axiomas, teoremas y ejemplos del texto (como antes, el primer número corresponde al capítulo; el segundo, a la sección de ese capítulo; y el tercero, al número de la definición, axioma, teorema y ejemplo dentro de la sección).

  ▷ Todos los números de tres niveles y acompañados de una letra −por ejemplo, 4.4.5e− hacen referencia a una parte específica de una definición, axioma, teorema y ejemplo dentro del texto, como, por ejemplo, a la parte (e).

  ▷ Números sin paréntesis aluden a pies de páginas y números de ejercicios.

- Literaturas y referencias se citan con un número dentro de un corchete y, a veces, colocadas después del nombre del autor citado. Por ejemplo, Llinás [9]. En algunas ocasiones, las citas bibliográficas aparecen con más detalles. Por ejemplo, H. Llinás [9, pág. 41] significa que lo referenciado se encuentra en la página 41 de [9].

- Teoremas con una frase y/o literatura(s) entre paréntesis significan que dicho teorema se conoce con ese nombre y su correspondiente demostración se puede encontrar en la(s) literatura(s) citada(s).

- El símbolo ◀ indica el final de un ejemplo.

- Los ejercicios propuestos para ser resueltos con el paquete estadístico *Statgraphics* aparecen con el símbolo Ⓢ.

- Los ejercicios de demostraciones aparecen marcados con el símbolo ⋆.

Agradecimientos

Mi gratitud y reconocimiento a los profesores que, de alguna u otra forma, ayudaron en la revisión de este texto mediante sugerencias y críticas constructivas.

De igual manera, expreso sinceros agradecimientos a Ediciones Uninorte por darme la oportunidad de publicarlo.

De manera especial agradezco también a mi esposa, Greyci, por transcribir gran parte del material en el computador con ayuda del programa MiKTeX.

Finalmente, quiero agradecer a mi madre, esposa e hijos por su apoyo, paciencia, comprensión, amor y ayuda para hacer de este libro una realidad. Lo dedico a ellos. También lo dedico a los profesores Alberto Assa y Peter Paul Konder y a mi padre, que descansen en paz.

Observación final

Estimado lector:
Trabajé con mucha dedicación para que este libro resultara eficaz a nivel pedagógico y no tuviera errores. No obstante, si tiene preguntas, observaciones o sugerencias, por favor, póngase en contacto conmigo a través de la siguiente dirección: *hllinas@uninorte.edu.co*.

Humberto Llinás Solano

CAPÍTULO 1

Distribuciones fundamentales de muestreo

Contenido

☞ Objetivos del capítulo

1. *Desarrollar el concepto de distribución muestral.*
2. *Examinar el teorema central del límite.*
3. *Analizar la distribución muestral de la media, proporción, diferencia de dos medias, diferencia de dos proporciones, varianza y razón de dos varianzas.*

☞ Empleo de la estadística

≪Un fabricante de neumáticos ha desarrollado un nuevo producto que, comparado con la línea actual, tendrá, según cree, una mayor duración en relación con las millas recorridas. Para evaluar el nuevo neumático, los gerentes necesitan un estimado (o una estimación) de la media de las millas que dura el nuevo producto. El fabricante selecciona, entonces, una muestra de 120 neumáticos para probarlos, obteniendo como resultado una media de la muestra de 36.500 millas. En consecuencia, se obtuvo el valor de 36.500 como estimado de la media para la población de neumáticos nuevos.≫

Introducción

En este capítulo, dedicaremos gran parte de nuestra atención a analizar problemas con el objeto de estudiar las diversas distribuciones que, a su vez, nos permitan averiguar características de una población a partir de la información proporcionada por una muestra de dicha población. Este es el objetivo de la *estadística inferencial*. La razón principal para observar una muestra en lugar de la población completa consiste en que la recogida de toda la información resulta exageradamente costosa en la mayoría de las ocasiones.

Además del factor económico, una enumeración completa de la población, llamada CENSO, puede ser imposible por circunstancias como el tiempo, que puede ser insuficiente en determinadas condiciones o, también, debido a factores ambientales. Este último sería el caso, por ejemplo, de un censo cuyo objeto fuese la población marina del Océano Atlántico. Pero incluso en los casos en que se dispone de recursos suficientes para analizar la población completa, tal vez sea preferible dedicar esos recursos a un subconjunto pequeño de la población, con el fin de que tal concentración de esfuerzos produzca medidas más precisas.

A continuación enunciaremos los usos del muestreo en diversos campos:

- *Política.* Las muestras de las opiniones de los votantes se usan para que los candidatos midan la opinión pública y el apoyo en las elecciones.

- *Sociología.* El sociólogo que desea conocer las actitudes de los adolescentes frente al aborto, no emprende la tarea de entrevistar a todos los adolescentes que hay en el país, más bien elige una muestra de ellos y los entrevista.

- *Educación.* Las muestras de las calificaciones de los exámenes de estudiantes se usan para determinar la eficiencia de una técnica o programa de enseñanza.

- *Industria.* Muestras de los productos de una línea de ensamblaje sirven para el propósito de controlar la calidad.

- *Medicina.* Un fabricante de drogas que desea saber los resultados de algún medicamento para bajar la tensión en la sangre y compararlo con una droga de la competencia, no lleva a cabo un experimento con todos los pacientes conocidos que sufran de hipertensión.

- *Agricultura.* Las muestras del maíz cosechado en una parcela proyectan en la producción los efectos de un fertilizante nuevo.

- *Gobierno.* Una muestra de opiniones de los votantes se usaría para determinar los criterios del público sobre cuestiones relacionadas con el bienestar y la seguridad nacionales.

1.1 Errores y técnicas de muestreo

1.1.1 Errores muestrales y no muestrales

Cuando se usan valores muestrales (o estadísticos), para estimar valores poblacionales (o parámetros), pueden ocurrir dos tipos generales de errores: el *error muestral* y el *error no muestral* (o *sistemático*).

Errores muestrales

Es improbable, por ejemplo, que la media de la muestra fuera *idéntica* a la media de la población. Asimismo, tal vez la desviación estándar u otra medición que se calcule con base en la muestra no sea *exactamente igual* al valor correspondiente de la población. Así, es posible que existan ciertas diferencias entre las *estadísticas de la muestra*, como la media o la desviación estándar, y los *parámetros de la población* correspondientes.

Definición 1.1.1 *El* ERROR MUESTRAL *es la diferencia entre un estadístico de la muestra y el parámetro correspondiente de la población.*

En general, el error muestral se refiere a la variación natural existente, entre muestras tomadas de la misma población, cuando una de ellas no es copia exacta de la población.

Ejemplo 1.1.2 *Se toman muestras de tamaño 2 de una población consistente en tres valores: 2, 4 y 6. Supongamos que el muestreo se hace con reemplazo (es decir, el número elegido se reemplaza antes de escoger el siguiente) y que se seleccionan muestras ordenadas.*[1] *Hállese la media poblacional, todas las muestras, la media de cada muestra y los errores muestrales.*

SOLUCIÓN:
La media poblacional equivale a

$$\mu = \frac{2+4+6}{3} = 4.$$

La tabla 1.1 *contiene una lista de todas las muestras ordenadas de tamaño 2 que es posible escoger con reemplazo de la población de valores 2, 4 y 6. También contiene las medias muestrales y los correspondientes errores muestrales.*

Tabla 1.1: Muestras ordenadas de tamaño 2 de la población de valores 2, 4 y 6 ◄

Muestras ordenadas	Media muestral $\overline{x}$	Error muestral $e = \overline{x} - \mu$
(2,2)	2	$2 - 4 = -2$
(2,4)	3	$3 - 4 = -1$
(2,6)	4	$4 - 4 = 0$
(4,2)	3	$3 - 4 = -1$
(4,4)	4	$4 - 4 = 0$
(4,6)	5	$5 - 4 = 1$
(6,2)	4	$4 - 4 = 0$
(6,4)	5	$5 - 4 = 1$
(6,6)	6	$6 - 4 = 2$

Aun si hemos tenido gran cuidado para asegurar que dos muestras del mismo tamaño sean representativas de una cierta población, no esperaríamos que las dos sean idénticas en todos sus detalles. El error es un concepto importante que nos ayudará a entender mejor la naturaleza de la estadística inferencial.

[1] En una muestra ordenada, el orden en que se escogen las observaciones es importante. Por ejemplo, la muestra ordenada (2,4) es distinta de la muestra ordenada (4,2). En la muestra (4,2), se escogió primero 4 y luego 2.

Errores no muestrales o sistemáticos

En los análisis prácticos, existe la posibilidad de que aparezca un error que no esté relacionado con el procedimiento de muestreo usado. Estos errores aparecerían también si se tomara un censo de la población completo. Se conocen como ERRORES NO MUESTRALES o SISTEMÁTICOS. En un estudio particular, existen potenciales errores no muestrales por varias causas, como muestran los ejemplos 1.1.3, 1.1.4 y 1.1.6.

Ejemplo 1.1.3 (La población de la que realmente se muestrea no es la relevante) *Un célebre ejemplo es el estudio de las actitudes de varios millones de personas, realizado por el* LiteraryDigest, *un periódico popular en ese entonces, para predecir al ganador de la presidencia en 1936, cuando el republicano Alfred Landon competía contra el demócrata Franklin Roosevelt. Los nombres de las personas que se incluyeron en la encuesta se obtuvieron del directorio telefónico y de otras listas, tales como la de suscriptores de la revista y los registros de automóviles. Estas fuentes no representaban en absoluto a las clases más pobres, puesto que mucha gente que prefería votar por Roosevelt no tenía teléfono y no se suscribía a periódicos. La mayoría de los entrevistados mostraron su preferencia por Landon y, en consecuencia, el periódico predijo que este candidato ganaría por un gran margen. Pero, Landon perdió. La moraleja de la historia es que, si uno quiere realizar inferencia sobre una población (en este caso, el electorado de Estados Unidos), es importante muestrear de la población y no de algún subgrupo de ella, aunque la segunda opción parezca conveniente.* ◄

Ejemplo 1.1.4 (Los individuos bajo estudio dan respuestas inexactas o inciertas) *Esto podría pasar si las preguntas se redactasen de manera que fuesen difíciles de entender o de forma que una respuesta particular pareciera más aceptable o más deseable. Además, hay preguntas que pueden ser delicadas y, en tal caso, sería temerario esperar respuestas uniformemente sinceras. Supongamos, por ejemplo, que el director de una fábrica quiere valorar las pérdidas anuales de la compañía debidas a robos de los empleados. En principio, podría seleccionarse una muestra aleatoria de empleados y preguntarles: "¿Qué ha robado usted de esta fábrica en los últimos doce meses?" Claramente, ¡ésta no es la mejor forma de proceder para obtener la información deseada! De hecho, ya hemos hablado de una posibilidad para abordar este problema. Para obtener una descripción y una ilustración de este procedimiento (llamado el método de respuesta aleatorizada*[2]*) se puede acudir al ejemplo 2.1.17 en* Llinás *[9].* ◄

Otro tipo de error muestral es el denominado sesgo de las muestras.

Definición 1.1.5 *El* SESGO MUESTRAL *es la tendencia sistemática a favorecer la selección de ciertos elementos de una muestra en lugar de otros.*

Ejemplo 1.1.6 (Una forma de esta posibilidad surge de la no respuesta) *Si ésta es importante puede inducir a errores muestrales y sistemáticos adicionales. En otros casos, los errores*

[2]Ver, por ejemplo, M. D. Geurts, "Using a randomized response research design to eliminate nonresponse biases in business research", *Journal of Academy of Marketing Science*, 8 (1980), 83-90.

muestrales surgen como consecuencia de la disminución inesperada de la muestra. También, como se ha visto, pueden presentarse si la población muestreada no es la población de interés. En este sentido, los resultados obtenidos pueden considerarse como una muestra aleatoria de la población de los individuos que responderían. Pero es dable que las personas seleccionadas sean distintas de la población general de alguna manera importante. Si esto es así, inducirán un sesgo en las estimaciones resultantes.

Si se sospecha que el sesgo de la no respuesta tiene un carácter molesto, hay tres posibilidades abiertas. Primero, el investigador puede solicitar información mediante un mecanismo del que se sepa que produce una proporción de respuestas altas. Segundo, hasta donde sea posible, deben compararse las características de los individuos que responden y de los que no, en aspectos tales como sexo, edad y raza, para comprobar si hay diferencias obvias entre los dos grupos. Finalmente, se debe buscar contacto con los individuos que no respondieron, algunos de los cuales pueden estar bien dispuestos para contestar a unas pocas preguntas claves. Si sus respuestas difieren significativamente de las de los individuos que respondieron al principio, debe hacerse una corrección del sesgo de la no respuesta. ◄

Es importante señalar que el sesgo muestral se refiere a una tendencia sistemática inherente a un método de muestreo, lo cual produce estimaciones de un parámetro que son, en promedio, menores (sesgo negativo) o mayores (sesgo positivo), que el parámetro real. Los ejemplos 1.1.3 y 1.1.7 ilustran situaciones para errores que resultan de colecciones de datos que caen en esta categoría.[3]

Ejemplo 1.1.7 *Si buscamos información relativa a las actitudes hacia el aborto y obtenemos una muestra que consta preponderadamente de hombres, podríamos encontrar un sesgo muestral.* ◄

Los errores que resultan de la acumulación de datos o de su procesamiento se clasifican también como errores no muestrales, como se ilustra en el siguiente ejemplo.

Ejemplo 1.1.8 *Al recabar datos pueden generarse errores no muestrales cuando los instrumentos usados para realizar las mediciones están fuera de ajuste o mal calibrados. Además, pueden ocurrir errores de procesamiento si los datos están mal colocados, si se pierden al registrarlos o si las respuestas proporcionadas por las personas durante el estudio no son verdaderas. Este último caso puede darse, en concreto, con preguntas relativas a la edad, en las que mucha gente miente por vanidad.* ◄

No existe un procedimiento general para identificar y analizar errores sistemáticos. No obstante, los efectos de estos errores pueden ser muy importantes. La principal recomendación es que el investigador ponga cuidado en cosas tales como identificar la población relevante, diseñar el cuestionario y tratar la no respuesta de manera que minimice su importancia. En el resto de este capítulo, asumiremos que se han tomado estas precauciones y nuestra exposición se centrará en el tratamiento de los errores muestrales.

[3]En el ejemplo 1.1.3, la muestra estaba fuertemente sesgada a favor de Landon.

1.1.2 Técnicas de muestreo aleatorio

El sesgo muestral puede suprimirse o minimizarse, usando el PRINCIPIO DE ALEATORIZACIÓN. Este principio se refiere a cualquier proceso de selección de una muestra de la población en el que la selección es imparcial o no está sesgada. Una muestra elegida con procedimientos aleatorios se llama *muestra aleatoria*. Los tipos más comunes de técnicas de muestreo aleatorio son el *muestreo aleatorio simple*, *el muestreo estratificado*, *el muestreo por conglomerados* y *el muestreo sistemático*. Ahora, explicaremos brevemente cada uno de ellos.

Muestreo aleatorio simple

Como ya se ha dicho, para evitar el sesgo muestral y lograr inferencias válidas acerca de la población, es importante que el proceso de selección de la muestra esté basado en el principio de aleatorización. La forma más sencilla para conseguir esto es diseñar un mecanismo de selección en el cual todas las muestras de un tamaño dado tengan la misma probabilidad de ser elegidas. Esto conduce a la siguiente definición.

Definición 1.1.9 *Un procedimiento de* MUESTREO ALEATORIO SIMPLE *es aquel en el que todas las posibles muestras del mismo tamaño tienen la misma probabilidad de ser escogidas. A las muestras obtenidas por procedimientos de este tipo se las denomina* MUESTRAS ALEATORIAS SIMPLES. *Matemáticamente, se dice que las variables aleatorias* X_1, X_2, ..., X_n *forman una* MUESTRA ALEATORIA (SIMPLE) *de tamaño n si se cumplen las dos condiciones siguientes:*

(a) Las variables X_1, X_2, ..., X_n *son independientes.*

(b) Toda X_i *tiene la misma distribución de probabilidad.*

*Este método se usa con tanta frecuencia que, en muchos casos, el adjetivo "*simple*" se elimina de ambos términos definidos anteriormente.*

Las condiciones (a) y (b) se pueden "unir" diciendo que las variables X_i son independientes e idénticamente distribuidas (i.i.d). Si el muestreo[4] es con reemplazo o de una población infinita, las condiciones (a) y (b) se satisfacen exactamente. Estas condiciones se satisfacen de manera aproximada si el muestreo es sin reemplazo, pero el tamaño de la

[4]El muestreo aleatorio simple se puede llevar a cabo de dos maneras: *con reemplazo* o *sin reemplazo*. Cuando el muestreo es SIN REEMPLAZO, solamente se permite a un individuo dado de la población aparecer una vez en la muestra. Cuando el muestreo es CON REEMPLAZO, no hay ningún límite para el número de veces que un individuo dado de la población pueda aparecer en la muestra. En las aplicaciones prácticas se usa el muestreo sin reemplazo.

muestra n es mucho menor que el tamaño N de la población. En la práctica si $n/N \leq 0,05$ (es decir, a lo sumo 5% de la población se muestrea), podemos suponer que las X_i forman una muestra aleatoria.

Ejemplo 1.1.10 *Una cadena nacional de comidas rápidas desea seleccionar aleatoriamente y sin importar el orden, 5 de los 10 estados de un país para tomar muestras sobre el gusto de los consumidores. Una muestra aleatoria simple garantizará que las* $\binom{10}{5} = 252$ *muestras de tamaño 5 tengan la misma probabilidad de ser utilizada en el estudio. En este caso, la probabilidad de escoger una muestra aleatoria simple de tamaño 5 será*

$$P(\text{escoger una muestra de tamaño } 5) \;=\; \frac{1}{\binom{10}{5}} \;=\; \frac{1}{252} \;\approx\; 0,00397.$$

Análogamente, la probabilidad de escoger una muestra aleatoria simple de tamaño 7 será

$$P(\text{escoger una muestra de tamaño } 7) \;=\; \frac{1}{\binom{10}{7}} \;=\; \frac{1}{120} \;=\; 0,00833. \qquad \blacktriangleleft$$

El proceso de muestreo aleatorio simple puede llevarse a cabo introduciendo los miembros de la población en una caja y mezclándolos entre sí, para luego extraer, digamos, n de ellos. No obstante, en la práctica, para el caso de una población finita, (digamos, con N individuos) no es necesario hacerlo de este modo; pues también pueden usarse *tablas de números aleatorios* para conseguir el mismo resultado.

Definición 1.1.11 *Una* TABLA DE NÚMEROS ALEATORIOS *consiste en una tabla de números que se hace y se presenta en tal forma que cada uno de los números 0 a 9 aparecen en ella con una frecuencia aproximadamente igual. Es decir, cada uno de estos números aparecen en la tabla con la misma probabilidad.*

Las tablas están construidas de forma que el proceso descrito en la definición 1.1.11 tiene las mismas propiedades que el muestreo aleatorio simple. Una de las posibles formas de construir una tabla de números aleatorios consistiría en meter en un caja 10 bolas numeradas de 0 a 9. Después de haberlas mezclado bien, se extrae una de las bolas y se anota su número. A continuación se devuelve esta bola a la caja y se repite el proceso. Puede repetirse el procedimiento para obtener números con tantas cifras como se precisen. Este proceso tiene la propiedad de que cada uno de los posibles números tiene la misma probabilidad, y las elecciones sucesivas son independientes unas de otras. El problema es que resulta extremadamente tedioso.

En la práctica, pueden generarse números aleatorios de manera mucho más rápida con la ayuda de un computador, ya que existen mecanismos que imitan de forma efectiva el procedimiento que acabamos de describir. La tabla del apéndice es una página de números aleatorios, tomados de una tabla que contiene un millón de dígitos aleatorios. Expliquemos el procedimiento de sacar una muestra aleatoria simple por medio de un ejemplo.

Ejemplo 1.1.12 *Hay 180 estudiantes de primer año en un colegio rural. Con el fin de obtener información acerca de la costumbre de ver televisión, un consejero de orientación desea seleccionar una muestra aleatoria simple de diez estudiantes para llenar un cuestionario. En la oficina del rector se encuentra una lista alfabética de los estudiantes numerados consecutivamente de 1 a 180. El consejero utiliza la tabla del apéndice para determinar qué estudiantes formarán la muestra.*

Como el número de estudiantes de la población es de 180 (un número de tres dígitos) es conveniente pensar en los números de 1 a 180 como los números 001, 002, 003, . . ., 180. Solamente se aprovecharán los números de tres dígitos que queden entre 001 y 180.

El consejero selecciona al azar un punto de partida en la página de los números aleatorios cerrando los ojos y tocando con la punta de su lápiz. El número que quede más cerca a la punta de su lápiz es el punto de partida. La punta del lápiz toca el papel en un punto más cercano al número 1, ubicado en la intersección de la fila 36 y la columna 7, que a cualquier otro (véase la tabla 1.2a*).*

Tabla 1.2: Una parte de tabla de número aleatorios

⋮	⋮	
66790	72193	⋯
16427	71681	⋯
63988	0 1 319	⋯
67468	22553	⋯
⋮	⋮	

(a) El 1 está en la fila 36 y la columna 7.

⋮	⋮	
66790	72193	⋯
16427	71681	⋯
63988	0 131 9	⋯
67468	22553	⋯
⋮	⋮	

(b) El primer número de tres dígitos es 131.

⋮	⋮	
66790	72193	⋯
16427	71681	⋯
63988	01319	⋯
67468	2 255 3	⋯
⋮	⋮	

(c) El siguiente número a 131 es 255.

Como el primer número de tres dígitos que hay en esta posición es 131 (véase la tabla 1.2b*), el estudiante número 131 de la lista queda incluido en la muestra. El consejero mueve hacia abajo (la dirección del movimiento es arbitraria y pudo haber sido hacia arriba, hacia la diagonal, etc.) el lápiz hasta el siguiente número de tres dígitos que, como es 255 (véase la tabla* 1.2c*), no se puede utilizar.*

Siguiendo hacia abajo, los siguientes números utilizables son 063 y 120 (véase la tabla 1.3a*). Por tanto, los estudiantes 63 y 120 quedan incluidos en la muestra. Cuando el consejero llegue hasta el final de la página, simplemente mueve hacia la derecha un dígito, que según la tabla* 1.3b*, sería 302. Como este número no es utilizable, tiene en cuenta los números de tres dígitos que van hacia arriba*[5] *y que son utilizables como, por ejemplo, el 065 (véase la tabla* 1.3c*). Al final, el*

[5]Nuevamente, la dirección es arbitraria. Por ejemplo, el consejero pudo haber corrido el lápiz hacia la izquierda o empezar en la parte superior de la página.

Tabla 1.3: Una parte de tabla de número aleatorios

⋮	⋮	
63988	0 131 9	···
67468	22553	···
⋮	⋮	
70321	26394	···
98710	5 063 9	···
⋮	⋮	
57652	46065	···
35933	3 120 3	···
⋮	⋮	
69865	39302	···

(a) Los siguientes números son 063 y 120.

⋮	⋮	
63988	01319	···
67468	22553	···
⋮	⋮	
70321	26394	···
98710	50639	···
⋮	⋮	
57652	46065	···
35933	31203	···
⋮	⋮	
69865	39 302	···

(b) Al final, se corre un dígito a la derecha.

⋮	⋮	
63988	01319	···
67468	22553	···
⋮	⋮	
70321	26394	···
98710	50639	···
⋮	⋮	
57652	46 065	···
35933	31203	···
⋮	⋮	
69865	39302	···

(c) El siguiente número utilizable es 065.

procedimiento seguido por el consejero arroja los siguientes números aleatorios:

$$131, \quad 063, \quad 120, \quad 065, \quad 154, \quad 117, \quad 002, \quad 166, \quad 031, \quad 101.$$

Por tanto, la muestra aleatoria simple consta de los 10 estudiantes identificados con estos números en la lista. ◀

Es imposible precisar por simple inspección si una muestra es aleatoria o no. Para determinarlo, debemos conocer el proceso de selección que se usó. Ilustremos esto a través del siguiente ejemplo:

Ejemplo 1.1.13 *Suponga que queremos elegir tres meses al año para estudiar cierto comportamiento ambiental y que hemos escogido enero, julio, octubre y noviembre. ¿Representan estos cuatros meses una muestra aleatoria?*

SOLUCIÓN:
A partir de la información dada, es imposible decir si esta muestra es aleatoria. Estos meses pueden haber sido escogidos porque están distribuidos a lo largo del año y, siendo así, la muestra no es aleatoria. Sin embargo, si se escogieron con la ayuda de una tabla de números aleatorios o de otros procedimientos aleatorios, entonces, sí representan una muestra aleatoria. ◀