INFERENCIA ESTADÍSTICA BÁSICA

INFERENCIA ESTADÍSTICA BÁSICA

Apoyo al estudio independiente

Víctor Miguel Ángel Burbano Pantoja
Margoth Adriana Valdivieso Miranda

Universidad Pedagógica y Tecnológica de Colombia
Tunja
2016

Inferencia Estadística Básica, apoyo al estudio independiente / Burbano Pantoja, Víctor Miguel Ángel; Valdivieso Miranda, Margoth Adriana. Tunja: Editorial UPTC, 2016. 222 p.
ISBN 978-958-660-240-2

1. Inferencia estadística 2. Muestra aleatoria 3. Estimación 4. Hipótesis

(Dewey 519.5/21).

Primera Edición, 2016
200 ejemplares (impresos)

Inferencia Estadística Básica, apoyo al estudio independiente
ISBN 978-958-660-240-2

Colección Académica UPTC

Libro financiado por la Vicerrectoría Académica y la Dirección de Investigaciones de la UPTC.

Citación: Burbano, V y Valdivieso, M. (2016). *Inferencia Estadística Básica, apoyo al estudio independiente.* Tunja: Editorial UPTC.

Editorial UPTC.
Edificio Administrativo – Piso 4
Avenida Central del Norte 39-115
comite.editorial@uptc.edu.co
www.uptc.edu.co

Impresión
Grupo Imprenta y Publicaciones
Coordiandor: Rafael Humberto Parra Niño
UPTC - Avenida Central del Norte
Tels.: (0*8) 740 5626 - Exts. 2366 - 2367 - Fax 2408
imprenta.publicaciones@uptc.edu.co
Tunja – Colombia

A nuestros estudiantes
y a nuestras amadas hijas:
Ángela Yereth y Ángela Saray

Contenido

Pag.

Introducción

La estadística tuvo su origen a la par con la génesis de las actividades del hombre antiguo; creció como técnica de registro de información en diferentes culturas, asociada al conteo del ganado, a la administración de la riqueza y a la producción de bienes de consumo; maduró como una tecnología del método científico, y hoy se consolida como "la ciencia de los datos" (Aliaga & Gunderson, 2005; Cabria, 1994; Carrasco, 2005). La palabra "estadística" procede del latín *status*, que significa "estado o situación", "estado y nación" u "hombre de estado". En este sentido, la estadística se sitúa en un contexto cuantitativo, relacionado con la administración en general (Rioboó, González y Tato, 1997).

Durante mucho tiempo, los principales usos de la estadística se orientaron hacia la recolección, la organización, la representación y el análisis descriptivo de información; pero, desde el siglo xv hasta el siglo xvii, en Europa se generaron múltiples cambios de orden social, filosófico y cultural que favorecieron el uso del pensamiento cuantitativo y la estadística para buscar soluciones a los problemas surgidos en esos contextos. Con la formulación de la teoría de las probabilidades (Laplace, 1820), el análisis de los datos empíricos toma nuevos rumbos hacia el desarrollo de procesos de estimación y predicción.

No obstante, es a inicios del siglo XX que se establece un puente entre la estadística descriptiva y la teoría de la probabilidad, con los trabajos de Pearson, Fisher y Neyman, y se originan procesos estadísticos tendientes a desarrollar una teoría para la prueba de hipótesis y la toma de decisiones en presencia de incertidumbre (García y Ríos-Insúa, 1998); desde entonces, la inferencia estadística ha crecido a pasos agigantados, constituyéndose en pilar fundamental para desarrollar investigación científica desde un enfoque cuantitativo.

Intuitivamente, la inferencia estadística se utiliza para obtener conclusiones válidas en una población objeto de estudio a partir de una muestra aleatoria, seleccionada apropiadamente, de dicha población. De manera formal, la inferencia estadística se centra en el desarrollo de procesos de estimación de ciertos parámetros desconocidos en una población determinada y de pruebas de hipótesis que involucran esos parámetros (Gutiérrez y De la Vara, 2008; Hurtado & Silvente, 2012); para ello, utiliza muestras y variables aleatorias con distribuciones de probabilidad específicas. Por su naturaleza, esta rama de la estadística presenta altos niveles de abstracción. Con el propósito de facilitar el acceso a ella y de contribuir con el aprendizaje inicial de este necesario y útil tema, se ha escrito el presente texto, en el que se conjuga lo intuitivo con lo formal.

Debido a que cada día la inferencia estadística influye con gran vigor de manera directa sobre los procesos de investigación desarrollados por estudiantes universitarios de diversas titulaciones, profesionales en distintas disciplinas del conocimiento, científicos e incluso estudiantes de la educación preuniversitaria, y puesto que la inferencia estadística se ha constituido en un elemento fundamental del método científico, resulta urgente y pertinente aprender contenidos de esta materia, de tal forma que se promueva la consolidación de una cultura estadística y la construcción del conocimiento científico.

En este sentido, desde mediados del siglo XX se intuían los alcances que tendría la estadística, cuando se afirmaba: "llegará el día en que pensar estadísticamente sea tan necesario para el ciudadano eficiente como leer y escribir" (H. G. Wells en Huff, 1954). En estas palabras proféticas, pensar estadísticamente ya insinuaba la necesidad de desarrollar una cultura estadística acorde con los desafíos de la sociedad actual globalizada, informatizada y afectada por el continuo cambio, el riesgo, la presencia de incertidumbre, lo aleatorio y lo no determinista.

Adicionalmente, la inferencia estadística es un tema de gran interés que se incluye en diversas áreas de gran parte de las carreras universitarias, y se recomienda su estudio en los grados superiores de la educación secundaria y media en Colombia, para potenciar el desarrollo del denominado pensamiento aleatorio y sistema de datos (Schmidt, 2006). El conocimiento estadístico-probabilístico es necesario para que los ciudadanos puedan desenvolverse en la sociedad actual; en este sentido, el propósito del presente texto es apoyar a un gran número de profesionales, docentes, estudiantes preuniversitarios y universitarios y demás personas que requieran de este tipo de conocimiento a fin de realizar un mejor análisis sobre la compleja realidad en que se vive, para así tomar decisiones apropiadas.

Este texto ha sido concebido, diseñado y elaborado teniendo presentes los elementos de la denominada Enseñanza estratégica (Alvermann, 1998; Fly Beau, 1987), el Conocimiento pedagógico del contenido (Shulman, 1987) y el Conocimiento del contenido para la enseñanza (Ball, Thames & Phelps, 2008). Estos enfoques, extrapolados al campo de la estadística, se constituyen en una alternativa para aprender estadística inferencial centrados en las actividades cognitivas acordadas entre los docentes y los estudiantes, siendo esta, a la vez, rol y proceso. Las etapas principales de la enseñanza estratégica son: preparación para el aprendizaje, presentación de los contenidos que se han de aprender y aplicación e integración de los nuevos conocimientos; las mencionadas etapas se hacen visibles en cada capítulo del texto. Se sugiere empezar con la lectura comprensiva de los contenidos expuestos en cada capítulo por medio de la revisión de los ejemplos y del desarrollo de los ejercicios incluidos.

Solamente cuando el lector tenga claros los conceptos, se recomienda abordar la sección que se encuentra al final de cada capítulo, titulada *Actividades para el estudio independiente,* cuyo propósito es aplicar e integrar los nuevos conocimientos. Al finalizar el texto se presenta la *información de retorno* de las actividades del estudio independiente, la cual se ha de usar para comparar los procedimientos realizados y los resultados obtenidos por el lector, en especial cuando haya tenido dificultades en la solución de las actividades propuestas; además, al finalizar cada capítulo se proponen ejercicios adicionales relacionados con las temáticas estudiadas. Con esta metodología también se espera obtener mejores resultados tanto en pruebas internas elaboradas en las instituciones educativas de nivel universitario y preuniversitario, como en pruebas externas (SABER), y contribuir en alguna medida con el mejoramiento de la enseñanza y el aprendizaje de los procesos de inferencia estadística, tan necesarios en diversos campos de la investigación científica.

El texto, que se ha titulado **Inferencia estadística básica**, ***Apoyo al estudio independiente,*** en principio trata de forma intuitiva algunos conceptos básicos, apoyados con ejemplos y ejercicios que de forma didáctica facilitan el estudio independiente de los temas propuestos; luego, aborda de manera formal ciertos tópicos de inferencia estadística, incluyendo un buen número de ejemplos y algunas demostraciones. El texto está dividido en cinco capítulos; los cuatro primeros están dirigidos a personas que se inicien en el tema de la inferencia estadística, tales como estudiantes de educación media (grados diez y once), profesores del nivel preuniversitario y universitario y estudiantes de las diversas carreras a nivel universitario, así como profesionales de distintas disciplinas que deseen utilizar la estadística para realizar investigación científica desde un enfoque cuantitativo.

En el primer capítulo se indican algunos conceptos asociados con la estadística descriptiva e inferencial, tales como: población, muestra, variable, variable aleatoria, muestra aleatoria, parámetros, estimadores y tipos de muestreo. En el segundo se presentan las distribuciones de probabilidad asociadas con los estimadores o estadísticas más frecuentes, entre ellas: la distribución normal, la *chi-cuadrado*, la *t-student* y la de *Fisher*. En el tercero se desarrollan procesos de inferencia estadística focalizados en la estimación de parámetros; se inicia con algunos conceptos básicos asociados a la estimación puntual y luego se determinan intervalos de confianza para estimar algunos de los parámetros usuales en estadística inferencial básica, entre ellos: el intervalo de confianza para estimar la media y la proporción poblacional, la diferencia de medias y de proporciones poblacionales, la varianza y el cociente de varianzas bajo el supuesto de normalidad. El cuarto capítulo se destina a la prueba de hipótesis; se inicia con algunos conceptos básicos asociados de este tema y luego se plantea un algoritmo que posibilita llevar a cabo diversos procedimientos inferenciales, entre ellos, la prueba de hipótesis para la media y la proporción poblacional, la diferencia de medias y de proporciones poblacionales, la varianza y el cociente de varianzas. En el quinto capítulo se proporciona la información de retorno a las actividades de estudio independiente propuestas en los cuatro primeros capítulos, y al finalizar se indica la forma de calcular algunos tamaños de muestra para casos específicos.

El presente libro se consolidó gracias al conocimiento estadístico-probabilístico logrado por sus autores en su actividad académica como profesores universitarios, y al acrecentamiento de este a través de su participación en procesos y proyectos de investigación científica. Para la comprensión de este texto, el lector ha de tener conocimientos básicos de aritmética, estadística descriptiva, probabilidades, cálculo diferencial e integral, entre otros. Es conveniente indicar que los gráficos presentados a lo largo del documento fueron elaborados por los autores usando el software libre R. Finalmente, es admisible que este texto pueda contener errores, por lo tanto, sería de gran ayuda que se nos den a conocer, dentro de un ambiente académico y crítico que permita corregirlos de forma oportuna y pertinente.

1

Conceptos asociados con la estadística

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

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

1.1 Algunos conceptos asociados con la estadística

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

1.1.1 Estadística descriptiva e inferencial

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

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

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

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

1.1.2 Población

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

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

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

1.1.3 Muestra

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

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

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

1.1.4 Variable

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.2 Variable aleatoria y muestra aleatoria

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

1.2.1 Variable aleatoria

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

$$X : \Omega \rightarrow R$$

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

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

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

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

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

la medida siguiente:

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

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

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

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

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

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

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

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

i) $f(x) \geq 0$

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

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

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

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

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

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

1.2.2 Muestra aleatoria

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

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

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

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

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

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

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

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

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

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

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

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

Así sucesivamente hasta obtener:

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

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

1.3 Parámetros y estimadores

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

1.3.1 Parámetros

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

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

denota con μ y se define mediante la siguiente expresión:

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

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

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

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

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

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

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

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

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

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

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

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

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

A continuación, se calcula la *varianza poblacional*:

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

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

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

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

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

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

Luego el *coeficiente de variación poblacional* es:

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

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

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

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

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

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

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

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

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

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

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

toda vez que la anterior integral exista.

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

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

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

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

Al número

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

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

1.3.2 Estimadores

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Usando las expresiones anteriores se puede obtener la siguiente relación:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Un valor del coeficiente de variación muestral es:

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

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

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

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

1.3.3 Teorema central del límite

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

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

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

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

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

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

1.4 Algunos tipos de muestreo

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

1.4.1 Muestreos aleatorios o probabilísticos

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

i) Poder definir el conjunto total de muestras posibles,

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

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

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

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

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

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

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

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

1.4.1.1 Muestreo aleatorio simple sin reemplazo

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

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

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

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

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

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

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

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

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

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

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

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

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

1.4.1.2 Muestreo aleatorio simple con reemplazo

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

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

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

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

$$10^4 = 10\ 000$$

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

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

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

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

1.4.1.3 Muestreo estratificado

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

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

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

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

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

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

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

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

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

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

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

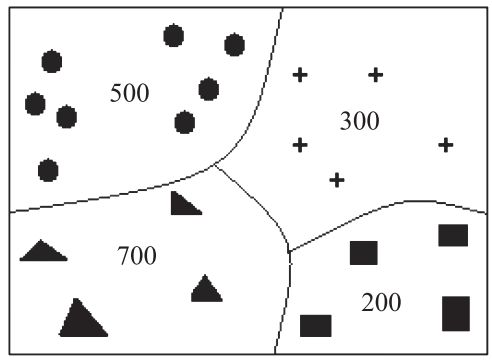

Figura 1.1 Población conformada por cuatro estratos

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

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

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

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

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

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

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

1.4.1.4 Muestreo sistemático o en serie

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

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

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

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

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

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

1.4.1.5 Muestreo por conglomerados

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

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

$N = 1000$
$n = 100$

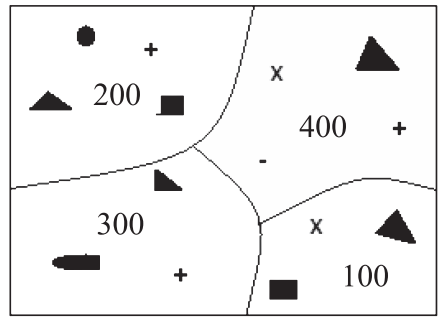

Figura 1.2. Población conformada por cuatro conglomerados

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

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

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

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

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

1.4.2 *Otros métodos de muestreo*

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

1.4.2.1 Muestreo por criterio

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

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

1.4.2.2 Muestreo por conveniencia

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

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

1.4.2.3 Muestreo teórico

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

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

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

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

Actividades para el estudio independiente Capítulo 1

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

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

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

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

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

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

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

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

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

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

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

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

3: poco, 5: mucho. _____

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

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

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

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

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

a) El tamaño de la muestra es_____

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

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

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

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

e) La desviación estándar _____

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

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

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

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

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

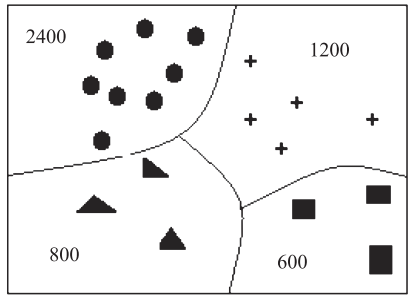

Figura 1.3 Población conformada por cuatro estratos

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

Ejercicios para el capítulo 1

1.1 Mencionar tres diferencias entre estadística descriptiva y estadística inferencial.

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

1.3 Proporcionar 2 ejemplos de variables:

- Cualitativas
- Discretas
- Continuas

1.4 Escribir 2 ejemplos de variables

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

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

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

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

2

Distribuciones usuales de muestreo

En este capítulo se presentan las distribuciones de probabilidad asociadas con los estimadores o estadísticas más frecuentes, entre ellas, la distribución normal, la *chi cuadrado*, la *t-student* y la de *Fisher;* estas distribuciones posibilitan el estudio de variables aleatorias como la media, la varianza y la proporción muestral, la diferencia de medias y de proporciones muestrales y el cociente de varianzas muestrales. Asimismo, se proporcionan ejemplos de aplicación alusivos.

2.1. Distribuciones de probabilidad de uso frecuente en inferencia estadística

Algunos de los aspectos teóricos considerados en esta sección son asumidos o adaptados de los conceptos expuestos al respecto por diversos autores: Alvarado & Obagi (2008), Bickel & Doksum (1977), Blanco (2004), Burbano y Valdivieso (2015), Canavos (1988), Freund y Miller (2000), Gutiérrez *et al.* (2008), Lindgren (1993), Mayorga (2003), Nieves, Sánchez & Clicero (2010) y Shao (1999), por citar algunos.

2.1.1 Distribución normal

Se dice que una variable aleatoria X sigue una distribución normal de parámetros μ y σ, donde μ es un número real y σ es un número real positivo, si su función de densidad es (Bickel & Doksum, 1977; Burbano *et al.*, 2015):

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

La representación gráfica de esta función de densidad se indica en la Figura 2.1; la curva de esta densidad suele denominarse campana de Gauss, y el área bajo la curva tiene un valor de 1.

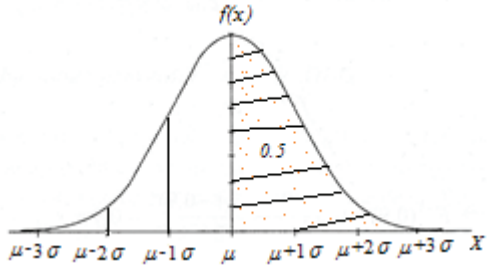

Figura 2.1 Función de densidad de la variable aleatoria X normal
Fuente: los autores apoyados en el software libre R.

Además, la función *f*, por tratarse de una función de densidad, cumple con las dos siguientes condiciones:

i) $f(x) \geq 0$

ii) $\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$

En cuanto al valor esperado y la varianza de la variable aleatoria X se cumple:

i) $E(X) = \mu$

ii) $Var(X) = \sigma^2$

En el caso particular de que $\mu = 0$ y $\sigma = 1$, se tiene la densidad de la distribución normal estándar; esta queda expresada de la siguiente manera:

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

La mencionada densidad se deduce realizando la siguiente estandarización de la variable involucrada:

$$z = \frac{x-\mu}{\sigma}$$

$$dz = \frac{1}{\sigma} dx$$

La función de densidad de la distribución normal estándar satisface la siguiente igualdad:

$$\int_{-\infty}^{\infty} f(z)dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz = 1$$

La densidad de la distribución normal estándar para la variable aleatoria Z es:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

Sin pérdida de generalidad, esta función se escribe de la siguiente forma:

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

Donde $f(x)$ es una función de densidad de probabilidad para la variable aleatoria X. La función de distribución de probabilidad para la variable aleatoria Z con distribución normal estándar es la siguiente:

$$F(z) = P(Z \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}t^2\right] dt$$

En la Figura 2.2 se observa la función de densidad de la distribución normal estándar; se trata de una curva simétrica con respecto al eje vertical, dado que esta función de densidad satisface que $f(-x) = f(x)$, hecho que también indica que *f* es una función par. El área bajo la curva *f* es igual a 1, en la parte izquierda se ubica un área igual a 0.5=50 % y en la parte derecha se tiene un área de 0.5 como resultado de la simetría. Los valores en el eje horizontal para la variable estandarizada están aproximadamente entre -3.6 y 3.6.

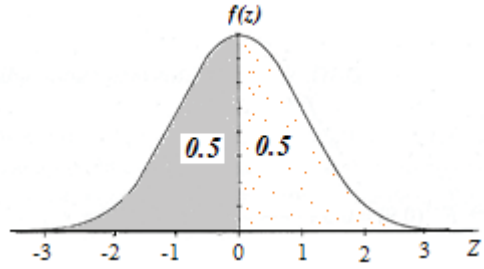

Figura 2.2. Función de densidad de la variable aleatoria Z normal estándar

Fuente: los autores apoyados en el *software* libre R.

Ejemplo 2.1. En la Figura 2.3 se indica el valor del área bajo la curva normal estándar equivalente al cálculo de la probabilidad $P(Z \leq 2.14)$; esta ha de leerse en una tabla para la distribución normal estándar.

$$F(2.14) = P(Z \leq 2.14) = \Phi(2.14) = 0.9838$$

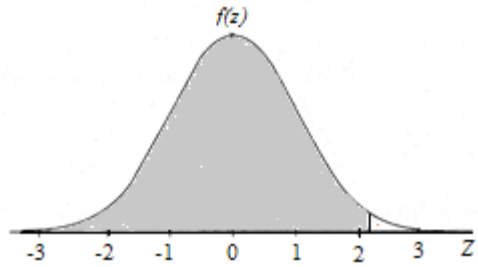

Figura 2.3 Área bajo la curva normal estándar $P(Z \leq 2.14)$

Fuente: los autores con la ayuda del *software* libre R.

Ejemplo 2.2. En la Figura 2.4 se ha sombreado el valor del área bajo la curva normal estándar equivalente al cálculo de la probabilidad,

$$F(-1.22) = P(Z \leq -1.22) = \Phi(-1.22) = 0.1112$$

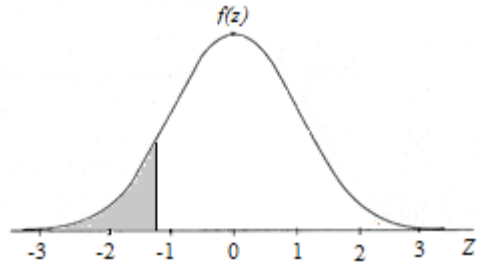

Figura 2.4. Área bajo la curva normal estándar $P(Z \leq -1.22)$

Fuente: los autores apoyados en el *software* libre R.

Ejemplo 2.3. En la Figura 2.5 se muestra el valor del área bajo la curva normal estándar equivalente al cálculo de la probabilidad

$$P(Z \leq 0.86) = 0.8051 = 80.51\ \%$$

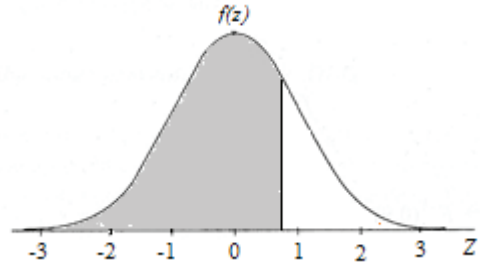

Figura 2.5 Área bajo la curva normal estándar $P(Z \leq 0.86)$

Fuente: los autores con la ayuda del *software* libre R.

Ejemplo 2.4. En la Figura 2.6 se indica el valor del área bajo la curva normal estándar equivalente al cálculo de la probabilidad

$$P(Z \geq 1.42) = 1 - P(Z \leq 1.42) = 1 - 0.9222 = 0.0778 = 7.78\ \%$$

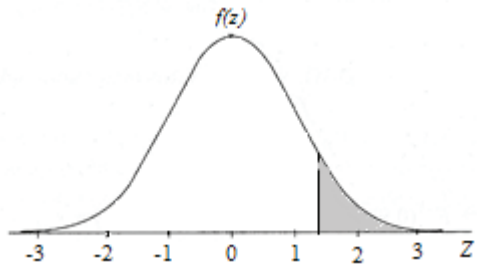

Figura 2.6. Área bajo la curva normal estándar $P(Z \geq 1.42)$

Fuente: los autores con la ayuda del *software* libre R.

Ejemplo 2.5. En la Figura 2.7 se presenta el valor del área bajo la curva normal estándar equivalente al cálculo de la probabilidad

$$P(-1.38 \leq Z \leq 2.47) = P(Z \leq 2.47 - P(Z \leq -1.38)$$

$$P(-1.38 \leq Z \leq 2.47) = 0.9932 - 0.0838 = 0.9094$$

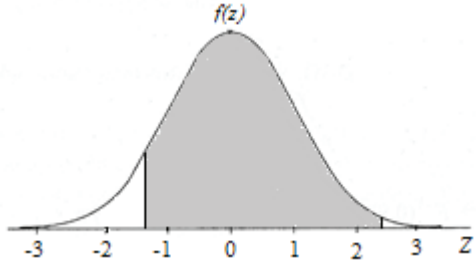

Figura 2.7. Área bajo la curva normal estándar $P(-1.38 \leq Z \leq 2.47)$

Fuente: los autores apoyados en el *software* libre R.

Ejemplo 2.6. En la fábrica A se ha establecido que la duración (en horas) de los bombillos que se producen se distribuye normalmente con media de 800 horas y desviación estándar de 50 horas. Si de forma aleatoria se toma un bombillo de la producción:

a) ¿Cuál es la probabilidad de que dure máximo 842 horas?
b) ¿Cuál es la probabilidad de que dure por lo menos 718 horas?
c) ¿Cuál es la probabilidad de que dure entre 782 y 917 horas?

$\mu = 800$ horas, $\sigma = 50$ horas

X : duración en horas de un bombillo cualquiera producido por la fábrica A.

a) $$P(X \leq 842) = P\left(\frac{X-\mu}{\sigma} \leq \frac{842-\mu}{\sigma}\right) = P\left(Z \leq \frac{842-800}{50}\right)$$

$$P(X \leq 840) = P\left(Z \leq \frac{42}{50}\right) = P(Z \leq 0.84) = 0.7995 = 79.95\ \%$$

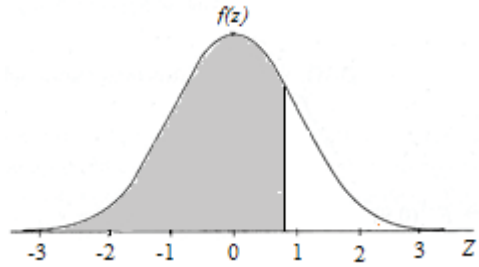

Figura 2.8. Área bajo la curva normal estándar $P(Z \leq 0.84)$

Fuente: los autores apoyados en el *software* libre R.

La probabilidad de que un bombillo escogido aleatoriamente de la producción presente una duración máxima de 842 horas es del 79.95 %. Una representación de este caso se indica en la Figura 2.8.

b) $$P(X \geq 718) = P\left(\frac{X-\mu}{\sigma} \geq \frac{718-\mu}{\sigma}\right) = P\left(Z \geq \frac{718-800}{50}\right)$$

$$P(X \geq 718) = P(Z \geq -1.64) = 1 - P(Z \leq -1.64) = 1 - 0.0505 = 0.9495 = 94.95\ \%$$

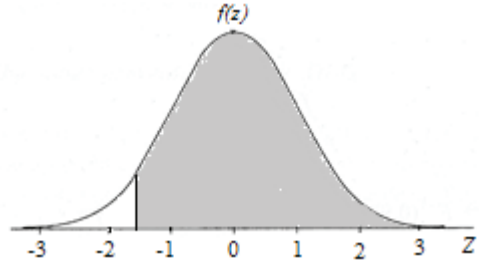

Figura 2.9. Área bajo la curva normal estándar $P(Z \geq -1.64)$

Fuente: los autores apoyados en el *software* libre R.

La probabilidad de que un bombillo escogido de forma aleatoria de esa producción dure por lo menos 718 horas es del 94.95 %. Una representación de la anterior situación se tiene en la Figura 2.9.

c) $$P(782 \leq X \leq 917) = P\left(\frac{782-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{917-\mu}{\sigma}\right)$$

$$P(782 \le X \le 903) = P\left(\frac{782-800}{50} \le Z \le \frac{917-800}{50}\right) = P(-0.36 \le Z \le 2.34)$$

$$P(782 \le X \le 917) = 0.9904 - 0.3594 = 0.631 = 63.1\ \%$$

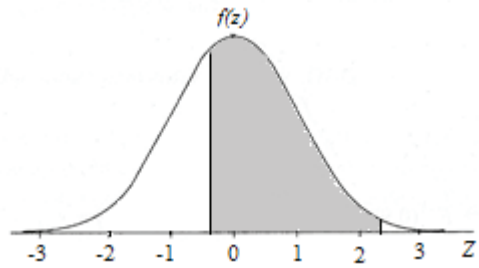

Figura 2.10. Área bajo la curva normal estándar $P(-0.36 \le Z \le 2.34)$

Fuente: los autores con la ayuda del *software* libre R.

La probabilidad de que un bombillo escogido aleatoriamente de la producción dure entre 782 y 917 horas es de 63.1 % (ver Figura 2.10).

2.1.2 Distribución chi-cuadrado

Si $X_1, X_2, ..., X_n$ son variables aleatorias independientes, cada una con distribución normal estándar, $\mu = 0$ y $\sigma = 1$, entonces la siguiente variable aleatoria tiene distribución *chi-cuadrado* con *n* grados de libertad:

$$\chi^2 = X_1^2 + X_2^2 +, ..., +X_n^2 = \sum_{i=1}^{n} X_i^2$$

Es decir, una variable aleatoria *chi-cuadrado* es la suma de *n* variables aleatorias en las condiciones mencionadas, cada una de ellas elevada al cuadrado.

Una variable aleatoria *X* tiene distribución *chi-cuadrado* con *n* grados de libertad si su función de densidad es:

$$f(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \exp\left(-\dfrac{x}{2}\right) & \text{si } x \ge 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Donde $\Gamma(.)$ es la función gamma dada por:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} \exp(-t) dt \ \text{con}\ \alpha > 0$$

La representación gráfica de la función de densidad para una variable aleatoria *X* que tiene distribución chi-cuadrado con *n* grados de libertad se indica en

la Figura 2.11; se trata de una curva asimétrica hacia la derecha; en esta, el área bajo la curva tiene valor de 1.

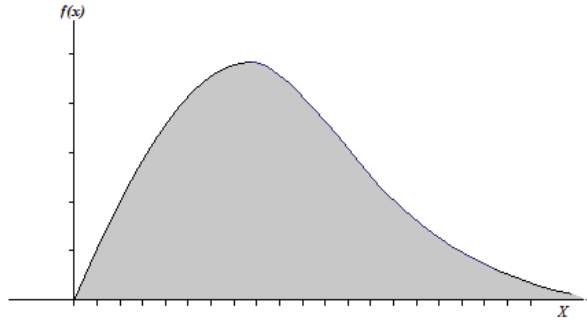

Figura 2.11. Área bajo la curva chi-cuadrado con *n* grados de libertad

Fuente: los autores apoyados en el *software* libre R.

Por supuesto, la función *f*, por tratarse de una función de densidad de probabilidad, cumple con las dos siguientes condiciones:

i) $f(x) \geq 0$

ii) $\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{\infty} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right) dx = 1$

De la función gamma se deducen los siguientes resultados:

$$\Gamma(n) = \int_{0}^{\infty} t^{n-1} \exp(-t)dt = (n-1)! \text{ siempre que } n \text{ sea entero positivo}$$

$$\Gamma(2) = \int_{0}^{\infty} t^{2-1} \exp(-t)dt = (2-1)! = 1! = 1$$

$$\Gamma(3) = \int_{0}^{\infty} t^{3-1} \exp(-t)dt = (3-1)! = 2! = 2$$

Asimismo, es posible demostrar que se cumple la siguiente igualdad:

$$\Gamma(\tfrac{1}{2}) = \int_{0}^{\infty} t^{\frac{1}{2}-1} \exp(-t)dt = \sqrt{\pi}$$

Si *X* es una variable aleatoria con distribución *chi-cuadrado* con *n* grados de libertad, entonces se deduce que:

i) $E(X) = n$

ii) $Var(X) = 2n$

La función de distribución de probabilidad para la variable aleatoria *X* con distribución *chi-cuadrado* con *n* grados de libertad es la siguiente:

$$P(X \le x) = \begin{cases} \int_0^x \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} t^{\frac{n}{2}-1} \exp\left(-\frac{t}{2}\right) dt & \text{si } x \ge 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Ejemplo 2.7. En la Figura 2.12 se indica el valor del área bajo la curva de la densidad *chi-cuadrado* equivalente al cálculo de la probabilidad $P(X \le 18.307)$ con n=10 grados de libertad; suele denotarse con $\chi^2_{0.95,10}$ la distancia sobre el eje horizontal cuya área bajo la curva es de 0.95; esta se lee en una tabla para la distribución *chi-cuadrado*.

$$P(X \le 18.307) = 0.95$$

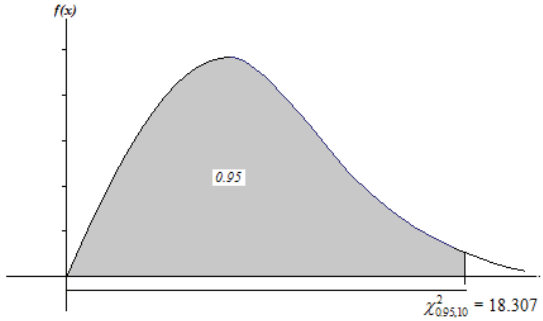

Figura 2.12 Área bajo la curva chi-cuadrado con n = 10 grados de libertad

Fuente: los autores apoyados en el *software* libre R.

Ejemplo 2.8. En la Figura 2.13 se indica el valor del área bajo la curva de la densidad chi-cuadrado, equivalente al cálculo de la probabilidad $P(X \le 5.226)$ con n = 12 grados de libertad; $\chi^2_{0.05,12}$ denota una distancia sobre el eje horizontal; esta se lee en una tabla para la distribución *chi-cuadrado*.

$$P(X \le 5.226) = 0.05$$

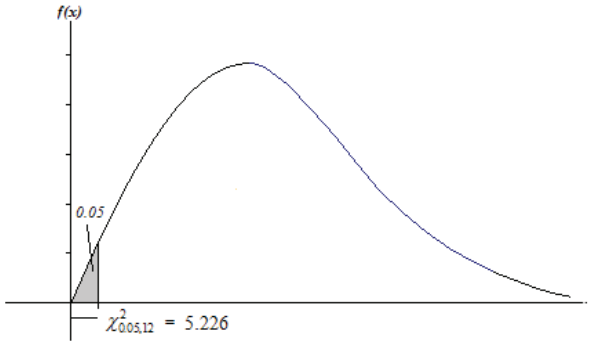

Figura 2.13 Área bajo la curva chi-cuadrado con n = 12 grados de libertad

Fuente: los autores apoyados en el *software* libre R.

Ejemplo 2.9. A continuación se presentan algunos valores de la distancia *chi-cuadrado* para valores específicos de probabilidad o del área bajo la curva de la densidad *chi-cuadrado;* estos son susceptibles de ser verificados en la tabla de esta distribución (ver Blanco, 2004):

$\chi^2_{0.99,15} = 30.578$

$\chi^2_{0.975,15} = 27.488$

$\chi^2_{0.90,8} = 13.36$

$\chi^2_{0.01,20} = 8.26$

$\chi^2_{0.025,14} = 5.629$

Ahora, si Z es una variable aleatoria con distribución normal estándar, $\mu = 0$ y $\sigma = 1$, entonces la variable aleatoria Z^2 tiene distribución *chi-cuadrado* con $n = 1$ grados de libertad. Lo anterior quiere decir que si X es una variable aleatoria con distribución normal con media μ y desviación estándar σ,

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1) \text{ implica que } Z^2 = \left(\frac{X - \mu}{\sigma}\right)^2$$

tiene distribución *chi-cuadrado* con $n = 1$ grados de libertad. Además, es posible mostrar que la variable aleatoria

$$\frac{(n-1)\hat{S}^2}{\sigma^2}$$

tiene distribución *chi-cuadrado* con $n - 1$ grados de libertad.

Si $X_1, X_2, ..., X_n$ es una muestra aleatoria para estudiar la variable aleatoria X, con distribución normal, entonces, las variables aleatorias $\bar{X}$ y $\hat{S}^2$ son independientes.

2.1.3 Distribución t-student

Si Z es una variable aleatoria con distribución normal estándar, $\mu = 0$ y $\sigma = 1$, y si Y es una variable aleatoria con distribución *chi-cuadrado* con n grados de libertad, pero Z y Y son independientes, entonces, la variable siguiente se denomina variable aleatoria *t-student* con n grados de libertad:

$$T = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

Es decir, una variable aleatoria t-student es el cociente entre una variable normal estándar y la raíz cuadrada de una variable chi-cuadrado dividida por sus grados de libertad.

Una variable aleatoria X tiene distribución *t-student* con n grados de libertad si su función de densidad de probabilidad es:

$$f(x)=\frac{1}{\sqrt{n\pi}}\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)} \quad \text{con} \quad \infty < x < \infty$$

Donde $\Gamma(.)$ es la función gamma definida en el apartado anterior.

La representación gráfica de la función de densidad para una variable aleatoria X que tiene distribución *t-student* con n grados de libertad se indica en la Figura 2.14; se trata de una curva simétrica similar a la curva normal estándar, pero más aplanada y con sus colas un poco más altas que aquella; el área bajo la curva también tiene un valor de 1.

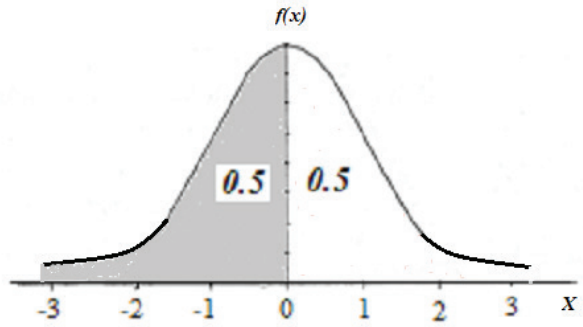

Figura 2.14. Área bajo la curva t-student con n grados de libertad
Fuente: los autores apoyados en el *software* libre R.

Por supuesto, la función f, por tratarse de una función de densidad de probabilidad, cumple con las dos siguientes condiciones:

i) $f(x) \geq 0$

ii) $\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{n\pi}}\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)} dx = 1$

Si X es una variable aleatoria con distribución *t-student* con n grados de libertad, entonces, se deduce que:

i) $E(X)=0$

ii) $Var(X)=\dfrac{n}{n-2}$ con $n\geq 3$

La función de distribución de probabilidad para la variable aleatoria X con distribución *t-student* con n grados de libertad es la siguiente:

$$F(x)=P(X\leq x)=\int_{-\infty}^{x}\frac{1}{\sqrt{n\pi}}\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}dt$$

Ejemplo 2.10. En la Figura 2.15 se indica el valor del área bajo la curva de la densidad t-student equivalente al cálculo de la probabilidad $P(t\leq 1.812)$ con n=10 grados de libertad; el valor denotado con $t_{0.95,10}$ corresponde a un valor sobre el eje horizontal; la probabilidad se ha de leer en una tabla para la distribución *t-student*.

$$P(t\leq 1.812)=0.95$$

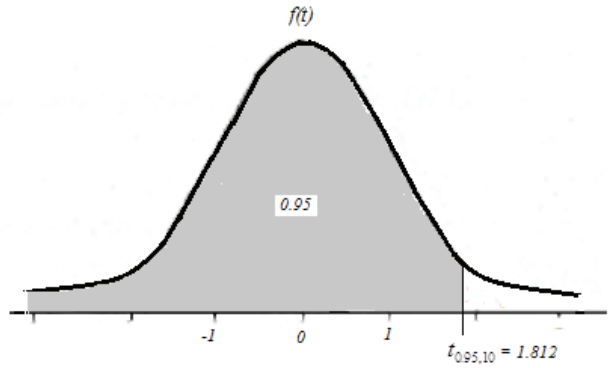

Figura 2.15. Área bajo la curva t-student con n = 10 grados de libertad

Fuente: los autores apoyados en el *software* libre R.

Ejemplo 2.11. En la Figura 2.16 se indica el valor del área bajo la curva de la densidad t-student, equivalente al cálculo de la probabilidad $P(t\leq -2.131)$ con n = 15 grados de libertad; el valor denotado con $t_{0.025,15}$ corresponde a un valor sobre el eje horizontal; la probabilidad se lee en una tabla para la distribución *t-student*.

$$P(t\leq -2.131)=0.025$$

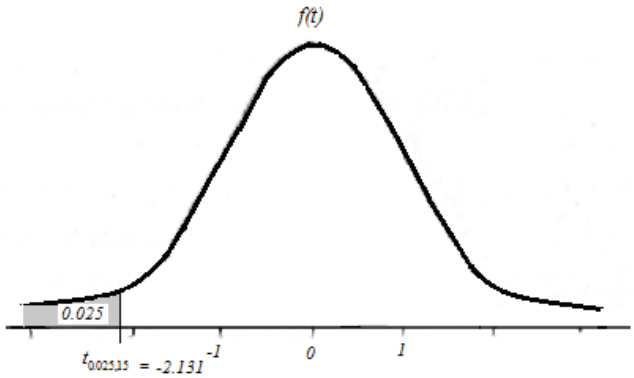

Figura 2.16. Área bajo la curva t-student con $n = 15$ grados de libertad

Fuente: los autores apoyados en el *software* libre R.

Ejemplo 2.12. En la Figura 2.17 se indica el valor del área bajo la curva de la densidad t-student, equivalente al cálculo de la probabilidad $P(-1.697 \le t \le 2.042)$ con $n = 30$ grados de libertad; en este caso se aplica el siguiente criterio para leer la tabla de la distribución *t-student*.

$$P(-1.697 \le t \le 2.042) = P(t \le 2.042) - P(t \le -1.697) = 0.975 - 0.05 = 0.925$$

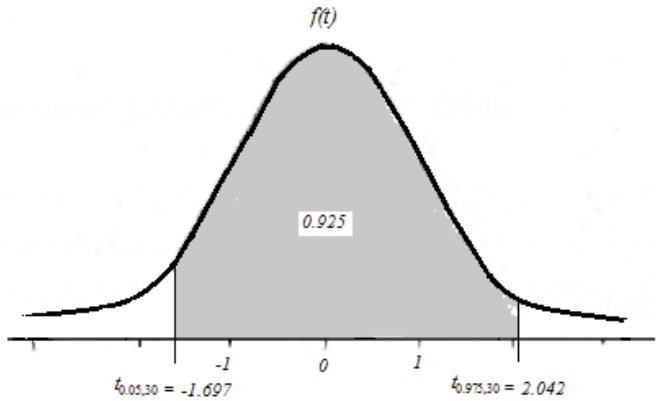

Figura 2.17. Área bajo la curva t-student con $n = 30$ grados de libertad

Fuente: los autores con la ayuda del *software* libre R.

Ejemplo 2.13. Enseguida, se presentan algunos valores sobre el eje horizontal, correspondientes a valores específicos de probabilidad o del área bajo la curva de la densidad *t-student;* estos son susceptibles de verificar en la tabla de esta distribución.

$t_{0.99,12} = 2.681$

$t_{0.01,14} = -2.624$

$t_{0.975,80} = 1.990$

$t_{0.95,24} = 1.71$

2.1.4 Distribución F de Fisher

Si X es una variable aleatoria con distribución *chi-cuadrado* con *m* grados de libertad, y si Y es otra variable aleatoria con distribución *chi-cuadrado* con *n* grados de libertad, pero *X* y *Y* son variables aleatorias independientes, entonces, la variable siguiente se denomina variable aleatoria de Fischer, con *m* y *n* grados de libertad, respectivamente:

$$F = \frac{\frac{X}{m}}{\frac{Y}{n}}$$

Es decir, una variable aleatoria de *Fisher* es el cociente entre dos variables chi-cuadrado, cada una dividida por sus grados de libertad.

Una variable aleatoria *X* tiene distribución de *Fisher* con *m* y *n* grados de libertad si su función de densidad de probabilidad es:

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{m+n}{2}\right)}{\Gamma\left(\dfrac{m}{2}\right)\Gamma\left(\dfrac{n}{2}\right)} m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\left(\frac{m}{2}-1\right)} (n+mx)^{-\left(\frac{m+n}{2}\right)} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Donde $\Gamma(.)$ es la función gamma definida en los apartados anteriores.

La representación gráfica de la función de densidad para una variable aleatoria *X* que tiene distribución de *Fisher* con *m* y *n* grados de libertad se indica en la Figura 2.18; se trata de una curva asimétrica hacia la derecha; en esta, el área bajo la curva tiene un valor de 1.

Figura 2.18. Área bajo la curva de Fisher con *m* **y** *n* grados de libertad
Fuente: los autores apoyados en el *software* libre R.

Por supuesto, la función *f,* por tratarse de una función de densidad de probabilidad, cumple con las siguientes dos condiciones:

i) $f(x) \geq 0$

ii) $\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{\infty} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\left(\frac{m}{2}-1\right)} (n+mx)^{-\left(\frac{m+n}{2}\right)} dx = 1$

Si *X* es una variable aleatoria con distribución de *Fisher* con *m* y *n* grados de libertad, entonces, se deduce que:

i) $E(X) = \frac{n}{n-2} \quad \text{con} \quad n > 2$

ii) $Var(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad \text{con} \quad n > 4$

La función de distribución de probabilidad para la variable aleatoria *X* con distribución de *Fisher* con *m* y *n* grados de libertad es la siguiente:

$$P(X \leq x) = \begin{cases} \int_{0}^{x} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} m^{\frac{m}{2}} n^{\frac{n}{2}} t^{\left(\frac{m}{2}-1\right)} (n+mt)^{-\left(\frac{m+n}{2}\right)} dt & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Ejemplo 2.14. En la Figura 2.19 se indica el valor del área bajo la curva de la densidad de *Fisher,* equivalente al cálculo de la probabilidad $P(X \leq 3.07)$ con

m = 8 y n = 10 grados de libertad; suele denotarse con $F_{0.95,8,10}$ la distancia sobre el eje horizontal, cuya área bajo la curva es de 0.95; esta ha de leerse en una tabla para la distribución de *Fisher*.

$$P(X \leq 3.07) = 0.95$$

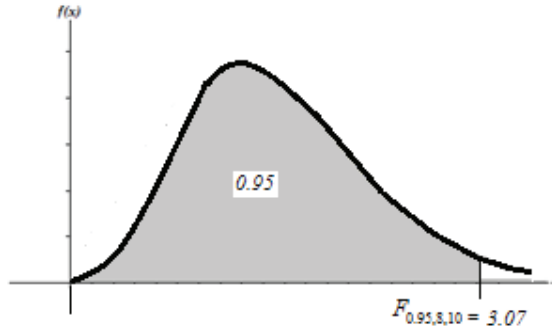

Figura 2.19 Área bajo la curva de Fisher con m = *8* **y** n = 10 grados de libertad

Fuente: los autores con la ayuda del *software* libre R.

Ejemplo 2.15. En la Figura 2.20 se observa el valor del área bajo la curva de la densidad de *Fisher,* equivalente al cálculo de la probabilidad $P(X \leq 0.2159)$ con m = 11 y n = 9 grados de libertad; se trata de calcular $F_{0.01,11,9}$, correspondiente a la distancia sobre el eje horizontal cuya área bajo la curva es de 0.01; esta no se puede leer directamente en una tabla de la distribución de *Fisher;* en este caso, se ha de recurrir a la siguiente fórmula de interpolación:

$$F_{0.01,11,9} = \frac{1}{F_{0.99,9,11}} = \frac{1}{4.63} \cong 0.2159$$

$$P(X \leq 0.2159) = 0.01$$

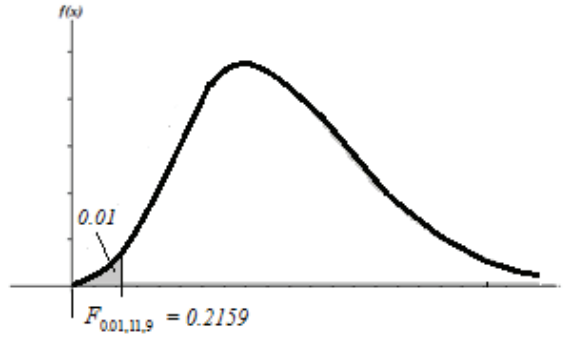

Figura 2.20 Área bajo la curva de Fisher con m = *11* **y** n = 9 grados de libertad

Fuente: los autores apoyados en el *software* libre R.

Ejemplo 2.16. Enseguida se presentan algunos valores de la distancia sobre el eje horizontal para valores específicos de probabilidad o del área bajo la curva

de la densidad Fisher; estos se determinan con la tabla de esta distribución (ver Blanco, 2004).

$$F_{0.99,7,4} = 14.98$$

$$F_{0.95,6,8} = 3.58$$

$$F_{0.01,5,9} = \frac{1}{F_{0.99,9,5}} = \frac{1}{10.16} \cong 0.0984$$

$$F_{0.05,6,12} = \frac{1}{F_{0.95,12,6}} = \frac{1}{4.00} = 0.25$$

2.2. Distribuciones muestrales

En esta sección se presentan las distribuciones correspondientes a la media, proporción y varianza muestral, además de las distribuciones de la diferencia de proporciones, medias y cociente de varianzas muestrales. Algunos de los aspectos teóricos considerados en esta sección son asumidos o adaptados de los conceptos expuestos al respecto por diversos autores, entre ellos: Bickel & Doksum (1977), Canavos (1988), Lindgren (1993), Shao (1999), Freund y Miller (2000), Mayorga (2003), Blanco (2004), Gutiérrez *et al.* (2008) y Burbano y Valdivieso (2015).

2.2.1 Distribución de la media muestral

Si $X_1, X_2, ..., X_n$ para $n > 1$ es una muestra aleatoria con $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$, entonces $\bar{X}$ es una variable aleatoria; esta tiene distribución de probabilidad y para determinarla se ha de tener presente que se está trabajando con una muestra de variables aleatorias independientes e idénticamente distribuidas, por consiguiente:

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{E(X_1) + E(X_2) + ... + E(X_n)}{n} = \frac{n\mu}{n} = \mu$$

$$Var(\bar{X}) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{Var(X_1) + Var(X_2) + ... + Var(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Luego

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

Usando el teorema central del límite, se tiene que la siguiente variable aleatoria Z tiene distribución normal estándar:

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{Var(\bar{X})}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Así, entonces,

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \qquad (2.1)$$

La expresión 2.1 se utiliza para una población infinita con desviación estándar conocida. Ahora, si el muestreo se realiza sin reemplazo en una población finita de tamaño N, entonces, en concordancia con Freund y Miller (2000), se deduce que:

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

Aplicando el teorema central del límite a la variable promedio muestral, se obtiene que la variable aleatoria Z siguiente tiene distribución normal estándar:

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{Var(\bar{X})}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

Luego

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)} \sim N(0,1) \qquad (2.2)$$

La expresión 2.2 se utiliza para una población finita de tamaño N con desviación estándar conocida.

Ahora, si la desviación estándar σ es desconocida, pero constante, entonces la variable aleatoria Z tiene distribución normal estándar:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Por otro lado, la siguiente variable aleatoria tiene distribución chi-cuadrado con *n-1* grados de libertad cuando se muestrea de una población normal:

$$\frac{(n-1)\hat{S}^2}{\sigma^2}$$

Luego el cociente entre la variable normal estándar anterior (expresión 2.1) y la raíz cuadrada de la variable aleatoria con distribución chi-cuadrado con *n-1* grados de libertad dividida por sus *n-1* grados de libertad tiene distribución t-student con *n-1* grados de libertad, así:

$$t = \frac{Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1)\hat{S}^2}{\sigma^2}}{(n-1)}}} = \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}}$$

En consecuencia,

$$t = \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \qquad (2.3)$$

tiene distribución t-student con n -1 grados de libertad; además, la expresión 2.3 se utiliza con frecuencia para *n* menor que 30.

Para una muestra aleatoria proveniente de una población finita de tamaño *n* con desviación estándar desconocida, entonces, se deduce que

$$t = \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} \qquad (2.4)$$

corresponde a una variable aleatoria con distribución *t-student* con *n-1* grados de libertad; esta se utiliza con frecuencia para *n* menor que 30 con poblaciones finitas y con desviación estándar poblacional desconocida.

Ejemplo 2.17. De los datos de la variable X (utilidades en millones de pesos por día de las cinco empresas más eficientes en la ciudad de Paipa, Boyacá-Colombia, en el año 2014), considerados en el *ejemplo 1.11,* a saber: 12, 11, 10, 9, 8, obtener todas las muestras de tamaño n=2 seleccionas por medio de un muestreo aleatorio simple sin reemplazo, determinar el promedio para cada muestra, construir la distribución de probabilidad para la variable aleatoria "media muestral", obtener su valor esperado y su varianza y verificar si se cumplen las igualdades para la esperanza matemática y la varianza presentadas por Freund y Miller (2000).

Como se tiene que $N = 5$ y $n = 2$, entonces, el número total de muestras distintas que son posibles de obtener es:

$$\binom{N}{n}=\binom{5}{2}=\frac{5!}{2!(5-2)!}=\frac{5!}{2!3!}=\frac{120}{12}=10$$

Estas muestras son las siguientes:

$$M_1=\{12,11\}, M_2=\{12,10\}, M_3=\{12,9\}, M_4=\{12,8\}, M_5=\{11,10\},$$

$$M_6=\{11,9\}, M_7=\{11,8\}, M_8=\{10,9\}, M_9=\{10,8\}, M_{10}=\{9,8\}$$

Sus correspondientes medias muestrales son:

$$\bar{X}_1=\frac{12+11}{2}=11.5, \bar{X}_2=\frac{12+10}{2}=11, \bar{X}_3=\frac{12+9}{2}=10.5, \bar{X}_4=\frac{12+8}{2}=10, \bar{X}_5=\frac{11+10}{2}=10.5$$

$$\bar{X}_6=\frac{11+9}{2}=10, \bar{X}_7=\frac{11+8}{2}=9.5, \bar{X}_8=\frac{10+9}{2}=9.5, \bar{X}_9=\frac{10+8}{2}=9, \bar{X}_{10}\frac{9+8}{2}=8.5$$

Los anteriores valores indican que la media muestral es una variable aleatoria que cambia de valor a medida que se cambia de muestra (con tamaño de muestra constante).

De acuerdo con lo establecido en el apartado 1.3 del primer capítulo, la función de probabilidad asociada a esta variable aleatoria es:

$$f\left(\bar{X}=11.5\right)=\frac{1}{10},\ f\left(\bar{X}=11\right)=\frac{1}{10},\ f\left(\bar{X}=10.5\right)=\frac{2}{10},\ f\left(\bar{X}=10\right)=\frac{2}{10},$$

$$f\left(\bar{X}=9.5\right)=\frac{2}{10},\ f\left(\bar{X}=9\right)=\frac{1}{10},\ f\left(\bar{X}=8.5\right)=\frac{1}{10}$$

El valor esperado o esperanza matemática de la variable aleatoria es:

$$E(\bar{X}) = 11.5\left(\frac{1}{10}\right)+11\left(\frac{1}{10}\right)+10.5\left(\frac{2}{10}\right)+10\left(\frac{2}{10}\right)+9.5\left(\frac{2}{10}\right)+9\left(\frac{1}{10}\right)+8.5\left(\frac{1}{10}\right)$$

$$E(\bar{X}) = \frac{11.5+11+21+20+19+9+8.5}{10} = \frac{100}{10} = 10$$

Ahora, para calcular la varianza, en primera instancia se calcula:

$$E(\bar{X}^2) = (11.5)^2\left(\frac{1}{10}\right)+11^2\left(\frac{1}{10}\right)+(10.5)^2\left(\frac{2}{10}\right)+10^2\left(\frac{2}{10}\right)+(9.5)^2\left(\frac{2}{10}\right)+9^2\left(\frac{1}{10}\right)+(8.5)^2\left(\frac{1}{10}\right)$$

$$E(\bar{X}^2) = \frac{(11.5)^2+11^2+2(10.5)^2+2(10)^2+2(9.5)^2+9^2+(8.5)^2}{10}$$

$$E(\bar{X}^2) = \frac{132.25+121+220.5+200+180.5+81+72.25}{10} = \frac{1007.5}{10} = 100.75$$

Con los resultados anteriores, se calcula así la varianza:

$$Var(\bar{X}) = E(\bar{X}^2) - (E(\bar{X}))^2 = 100.75 - (10)^2 = 100.75 - 100 = 0.75$$

En el ejemplo 1.11 ya se obtuvieron el valor esperado y la varianza de la variable utilidades; estos fueron:

$$\mu = 10$$

$$\sigma^2 = 2$$

Por otro lado, en concordancia con Freund y Miller (2000), se cumple que:

$$E(\bar{X}) = 10 = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{2}{2}\left(\frac{5-2}{5-1}\right) = \frac{3}{4} = 0.75$$

2.2.2 Distribución de la proporción muestral

Si *X* es una variable aleatoria con distribución binomial de parámetro *p,* donde este parámetro indica la probabilidad de éxito en una muestra aleatoria de tamaño *n* (de variables aleatorias Bernoulli), entonces, para la variable aleatoria

X: número de éxitos en la muestra de la característica A de interés,

$$E(X) = np$$

$$Var(X) = np(1-p)$$

Considerando la variable

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

su valor esperado y varianza son los siguientes:

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{np}{n} = p$$

$$Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{Var(X)}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}$$

Luego

$$E(\hat{p}) = p$$

$$Var(\hat{p}) = \frac{pq}{n}$$

Donde

$$q = 1 - p$$

Para un tamaño de muestra n "grande", se aplica el teorema central del límite, obteniéndose:

$$Z = \frac{X - E(X)}{\sqrt{Var(X)}} = \frac{X - np}{\sqrt{npq}}$$

Al realizar operaciones de tipo aritmético, la expresión anterior se transforma en:

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{\frac{X - np}{n}}{\frac{\sqrt{npq}}{n}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{pq}{n}}} = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Así, entonces,

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1) \qquad (2.5)$$

La expresión 2.5 corresponde a una variable aleatoria con distribución normal estándar para *n* grande; esta involucra a la variable aleatoria proporción muestral.

2.2.3 Distribución de la diferencia de proporciones muestrales

Si se tienen dos poblaciones binomiales con parámetros p_1 y p_2, respectivamente, y se ha seleccionado una muestra aleatoria de tamaño n_1 en la primera población, y una muestra aleatoria de tamaño n_2 en la segunda, de forma que estas sean independientes, entonces, para las variables aleatorias:

X_1: número de éxitos en la primera muestra de la característica A de interés y

X_2: número de éxitos en la segunda muestra de la característica A de interés

se definen las proporciones muestrales correspondientes de la siguiente manera:

$$\hat{p}_1 = \frac{X_1}{n_1} \quad \text{y} \quad \hat{p}_2 = \frac{X_2}{n_2}$$

Sus correspondientes valores esperados y sus varianzas son:

$$E(\hat{p}_1) = p_1$$

$$Var(\hat{p}_1) = \frac{p_1 q_1}{n_1}$$

Donde $q_1 = 1 - p_1$

$$E(\hat{p}_2) = p_2$$

$$Var(\hat{p}_2) = \frac{p_2 q_2}{n_2}$$

Donde $q_2 = 1 - p_2$

Se define la variable aleatoria denominada diferencia de proporciones muestrales; esta se denota con $\hat{p}_1 - \hat{p}_2$, y su valor esperado es:

$$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = p_1 - p_2$$

Si estas variables aleatorias $\hat{p}_1$ y $\hat{p}_2$ son independientes, entonces, su varianza equivale a:

$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

Para tamaños n_1 y n_2 "grandes", se aplica el teorema central del límite, obteniéndose,

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - E(\hat{p}_1 - \hat{p}_2)}{\sqrt{Var(\hat{p}_1 - \hat{p}_2)}} = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

Así, entonces, la siguiente variable aleatoria tiene distribución normal estándar:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1) \qquad (2.6)$$

La expresión 2.6 corresponde a una variable aleatoria con distribución normal estándar para n_1 y n_2 grandes; esta involucra a la variable aleatoria diferencia de proporciones muestrales.

2.2.4 Distribución de la diferencia de medias muestrales

Dadas dos poblaciones normales, se desea estudiar una determinada característica de tipo cuantitativo; para esto, se selecciona una muestra aleatoria de tamaño n_1 en la primera población y una muestra aleatoria de tamaño n_2 en la segunda, de forma que estas resulten independientes, y se obtienen sus respectivos promedios, para definir una nueva variable aleatoria llamada diferencia de medias muestrales. A continuación, se determina la distribución de probabilidad de esa nueva variable.

Si $X_{11}, X_{12}, ..., X_{1n_1}$ para $n_1 > 1$ es la muestra aleatoria extraída de la primera población con $E(X_{1i}) = \mu_1$ y $Var(X_{1i}) = \sigma_1^2$, entonces, $\overline{X}_1$ es una variable aleatoria; asimismo, si $X_{21}, X_{22}, ..., X_{2n_2}$, para $n_2 > 1$, es la muestra aleatoria extraída de la segunda población con $E(X_{2j}) = \mu_2$ y $Var(X_{2j}) = \sigma_2^2$, entonces, $\overline{X}_2$ es otra variable aleatoria; estas variables se expresan así:

$$\overline{X}_1 = \sum_{i=1}^{n_1} \frac{X_{1i}}{n_1} \qquad \overline{X}_2 = \sum_{j=1}^{n_2} \frac{X_{2j}}{n_2}$$

Luego la diferencia $\overline{X}_1 - \overline{X}_2$ es una nueva variable aleatoria. El valor esperado y la varianza de estas variables son, respectivamente:

$$E(\bar{X}_1)=\mu_1 \qquad Var(\bar{X}_1)=\frac{\sigma_1^2}{n_1}$$

$$E(\bar{X}_2)=\mu_2 \qquad Var(\bar{X}_2)=\frac{\sigma_2^2}{n_2}$$

$$E(\bar{X}_1-\bar{X}_2)=E(\bar{X}_1)-E(\bar{X}_2)=\mu_1-\mu_2$$

Por la independencia de las variables $\bar{X}_1$ y $\bar{X}_2$ resulta que:

$$Var(\bar{X}_1-\bar{X}_2)=Var(\bar{X}_1)+Var(\bar{X}_2)=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$$

Si las varianzas son conocidas, entonces, usando el teorema central del límite se tiene que la siguiente variable aleatoria Z tiene distribución normal estándar:

$$Z=\frac{\bar{X}_1-\bar{X}_2-E(\bar{X}_1-\bar{X}_2)}{\sqrt{Var(\bar{X}_1-\bar{X}_2)}}=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

Así, entonces,

$$Z=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1) \tag{2.7}$$

La expresión 2.7 se utiliza para trabajar con dos poblaciones infinitas que presenten desviaciones estándar conocidas.

Ahora, si para cada población la desviación estándar es desconocida y las muestras tienen tamaño inferior a 30 (muestras pequeñas), se maneja el supuesto de que las desviaciones son iguales, en este caso es posible probar que la siguiente variable aleatoria tiene distribución *t-student* con n_1+n_2-2 grados de libertad,

$$t=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{2.8}$$

En la expresión 2.8, el valor S_p se obtiene a través de la siguiente expresión:

$$S_p = \sqrt{\frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}}$$

Pero si en cada población la desviación estándar es desconocida y las muestras tienen tamaño inferior a 30 (muestras pequeñas), también es posible suponer que las desviaciones estándar son distintas; en este caso se deduce que la siguiente variable aleatoria tiene distribución *t-student* con *g* grados de libertad:

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \tag{2.9}$$

Para la expresión 2.9, los grados de libertad *g* se obtienen por medio de la siguiente expresión (Gutiérrez *et al.*, 2008):

$$g = \frac{\left(\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{S}_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{\hat{S}_2^2}{n_2}\right)^2}{n_2 + 1}} - 2$$

En cambio, si en cada población la desviación estándar es desconocida y las muestras tienen tamaño mayor o igual a 30 (muestras grandes), sea para varianzas iguales o desiguales, se deduce que la siguiente variable aleatoria (expresión 2.10) tiene distribución normal estándar:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \sim N(0,1) \tag{2.10}$$

2.2.5 Distribución del cociente de varianzas muestrales

Dadas dos poblaciones normales, si $X_{11}, X_{12}, ..., X_{1n_1}$ para $n_1 > 1$ es una muestra aleatoria extraída de la primera población con $E(X_{1i}) = \mu_1$ y $Var(X_{1i}) = \sigma_1^2$, entonces, $\hat{S}_1^2$ es una variable aleatoria; asimismo, si $X_{21}, X_{22}, ..., X_{2n_2}$ para $n_2 > 1$ es la muestra aleatoria extraída de la segunda población con $E(X_{2j}) = \mu_2$

y $Var(X_{2j}) = \sigma_2^2$, entonces, $\hat{S}_2^2$ es otra variable aleatoria; estas variables se expresan de la siguiente manera:

$$\hat{S}_1^2 = \sum_{i=1}^{n_1} \frac{(X_{1i} - \overline{X}_1)^2}{n_1 - 1} \qquad \hat{S}_2^2 = \sum_{j=1}^{n_2} \frac{(X_{2j} - \overline{X}_2)^2}{n_2 - 1}$$

Además, las siguientes variables aleatorias tienen distribución chi-cuadrado con n_1-1 y n_2-1 grados de libertad, respectivamente:

$$\frac{(n_1 - 1)\hat{S}_1^2}{\sigma_1^2} \qquad \text{y} \qquad \frac{(n_2 - 1)\hat{S}_2^2}{\sigma_2^2}$$

Por definición, al cociente entre las dos variables aleatorias anteriores, cada una dividida entre sus grados de libertad, le corresponde una distribución de Fisher, escrita como sigue:

$$F = \frac{\dfrac{\dfrac{(n_1 - 1)\hat{S}_1^2}{\sigma_1^2}}{n_1 - 1}}{\dfrac{\dfrac{(n_2 - 1)\hat{S}_2^2}{\sigma_2^2}}{n_2 - 1}} = \frac{\dfrac{\hat{S}_1^2}{\sigma_1^2}}{\dfrac{\hat{S}_2^2}{\sigma_2^2}} = \frac{\hat{S}_1^2 \sigma_2^2}{\hat{S}_2^2 \sigma_1^2}$$

Así, entonces, la variable de la expresión 2.11 tiene distribución de Fischer con n_1-1 grados de libertad para el numerador y n_2-1 grados de libertad para el denominador.

$$F = \frac{\hat{S}_1^2 \sigma_2^2}{\hat{S}_2^2 \sigma_1^2} \tag{2.11}$$

Para el caso particular en que las varianzas sean iguales, resulta que

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

con n_1-1 grados de libertad para el numerador y n_2-1 grados de libertad para el denominador.

Actividades para el estudio independiente Capítulo 2

2.1 Graficar y determinar las siguientes probabilidades utilizando la distribución normal estándar:

a) $P(Z \leq -1.25)$

b) $P(Z \geq 1.31)$

c) $P(0.29 \leq Z \leq 2.48)$

2.2 La empresa S&S produce lámparas cuya duración en horas se distribuye normalmente con media de 900 horas y desviación estándar de 50 horas. Si se toma al azar una lámpara de la producción,

a) ¿Cuál es la probabilidad de que dure máximo 940 horas?
b) ¿Cuál es la probabilidad de que dure por lo menos 822 horas?
c) ¿Cuál es la probabilidad de que dure entre 880 y 1020 horas?
d) ¿Cuál es la probabilidad de que dure más de 1200 horas?

2.3. El peso de los paquetes de arroz empacados por la máquina H&H es una variable aleatoria que se distribuye normalmente con media $\mu = 500$ g y una desviación estándar $\sigma = 20$ g. Si se escoge aleatoriamente un paquete de arroz empacado por la máquina H&H,

a) ¿Cuál es la probabilidad de que su peso sea por lo menos 486 g?
b) ¿Cuál es la probabilidad de que su peso sea máximo 480 g?
c) ¿Cuál es la probabilidad de que su peso esté entre 476 y 550 g?
d) ¿Cuál es la probabilidad de que su peso sea menos de 400 g?
e) ¿Cuál es la probabilidad de que su peso sea más de 440 g?

2.4 Determinar los siguientes valores correspondientes a las probabilidades indicadas y elaborar una representación gráfica:

a) $\chi^2_{0.99,17} = ?$

b) $\chi^2_{0.025,12} = ?$

2.5 Obtener las probabilidades y graficar:

a) $t_{0.99,16} = ?$

b) $t_{0.05,13} = ?$

2.6 Leer el valor de probabilidad en la distribución de Fisher y elaborar un gráfico:

a) $F_{0.99,12,10} = ?$

b) $F_{0.05,7,7} = ?$

2.7 Para la variable X (gastos en transporte por día [miles de pesos] de los cinco mejores estudiantes de la universidad B en la ciudad de Tunja, Boyacá, Colombia, en el año 2015), cuyos valores son: 8, 10, 14, 12, 6: *a)* determinar todas las muestras de tamaño $n = 2$ que son posibles de seleccionar por medio de un muestreo aleatorio simple sin reemplazo, *b)* calcular el promedio para cada muestra, *c)* construir la distribución de probabilidad para la variable aleatoria "media muestral", *d)* obtener su valor esperado y su varianza, *e)* verificar si se cumplen las igualdades para la esperanza matemática y la varianza presentadas por Freund y Miller (2000).

2.8 La duración promedio de cierta marca de teclados es de 900 días, con una desviación estándar de 70 días, siempre que se usen 8 horas por día. Determinar la probabilidad de que una muestra aleatoria de 36 teclados tenga una duración promedio: *a)* comprendida entre 870 y 925 días, *b)* menor o igual a 910 días.

2.9 El tiempo promedio que gasta el bus urbano en la ciudad de Cali es de 70 minutos. Si se toma una muestra aleatoria de 12 recorridos y con esos datos se obtiene una desviación estándar corregida de 8 minutos, ¿cuál es la probabilidad de que en esa muestra se tenga un tiempo promedio entre 64.94 y 76.84 minutos?

2.10 El 4 % de los artículos que produce una máquina son defectuosos; se toma una muestra aleatoria de 400 artículos. ¿Cuál es la probabilidad de que más del 5 % de los artículos de la muestra sean defectuosos?

2.11 En la ciudad A, para niños de grado quinto de educación básica primaria se tiene un peso promedio de 35 kg con varianza de 5 , mientras que en la ciudad B, para niños que cursan el mismo grado, se tiene un peso promedio de 45 kg con varianza de 8. En la ciudad A se toma una muestra aleatoria de 50, y en la ciudad B otra de 60. ¿Cuál es la probabilidad de que la media muestral del peso de los niños de la ciudad B difiera de la de los niños de la ciudad A en más de 11 kg?

2.12 Un candidato a la presidencia de la república tiene el 60 % de la intención de voto en el departamento de Nariño, y el 58 % en el Valle del Cauca, en

Colombia. Si se toma una muestra aleatoria de 400 votantes en Nariño y de 500 en el Valle del Cauca, ¿cuál es la probabilidad de que la diferencia entre las proporciones muestrales de los potenciales votantes en Nariño y el Valle no superen el 3 %?

Ejercicios para el capítulo 2

2.1. Leer en una tabla normal estándar los siguientes valores de probabilidad y elaborar la representación gráfica correspondiente:

a) $P(Z \leq 2.16)$

b) $P(Z \geq 1.62)$

c) $P(Z < -0.025)$

d) $P(Z > 2.33)$

e) $P(Z < -1.47)$

f) $P(Z = -0.94)$

g) $P(Z < 5.4).$

h) $P(Z > 3.8).$

i) $P(Z < -4.2).$

2.2. El salario de los trabajadores de la empresa J&T se distribuye normalmente con media de 800 (en miles de pesos) y desviación estándar de 15; si se escoge aleatoriamente a un trabajador de esta empresa, ¿cuál es la probabilidad de que su salario sea superior a 813, inferior a 819 o esté entre 812 y 832?

2.3. El peso de un producto de la marca H es una variable con distribución normal, con media de 1000 g y desviación estándar de 5 g; si se escoge aleatoriamente una unidad del producto de la marca H, ¿cuál es la probabilidad de que su peso sea superior a 1009 g?

2.4. En un proceso productivo se sabe que el 96 % de los artículos de la marca MM resultan de buena calidad; si se selecciona una muestra aleatoria de 100 unidades de estos, ¿cuál es la probabilidad de que más de 94 unidades de esta marca resulten de buena calidad?

2.5. Una fábrica produce unidades de medicamento del tipo A para vacunos; la probabilidad de que cualquier unidad producida por la fábrica no resulte efectiva es del 3 %; si se escoge una muestra aleatoria de 90 unidades, ¿cuál es la probabilidad de que, por lo menos, 5 no resulten efectivas?

2.6. El peso en un grupo de aves sigue una distribución normal con media 2000 g y desviación estándar de 100 g; si se escoge aleatoriamente un ave de ese

grupo, ¿cuál es la probabilidad de que su peso sea mayor que 1857 g?

2.7. Si X es una variable aleatoria con distribución normal estándar, mostrar que su función de densidad de probabilidad cumple la siguiente igualdad:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx = 1$$

2.8. Leer en una tabla los siguientes valores de probabilidad correspondientes a una distribución chi-cuadrado y elaborar la representación gráfica correspondiente:

a) $\chi^2_{0.95,16} = ?$

b) $\chi^2_{0.99,20} = ?$

c) $\chi^2_{0.975,6} = ?$

d) $\chi^2_{0.01,9} = ?$

e) $\chi^2_{0.025,8} = ?$

2.9. Si Z es una variable aleatoria con distribución normal estándar, $\mu = 0$ y $\sigma = 1$, demostrar que Z^2 es una variable aleatoria con distribución *chi-cuadrado* con $n=1$ grados de libertad.

2.10. Si $X_1, X_2, ..., X_n$ es una muestra aleatoria para estudiar la variable aleatoria X, con distribución normal con media μ y desviación estándar σ, deducir que la variable aleatoria siguiente:

$$\frac{(n-1)\hat{S}^2}{\sigma^2}$$

tiene distribución chi-cuadrado con $n - 1$ grados de libertad.

2.11. Si X es una variable aleatoria con distribución chi-cuadrado con n grados de libertad, usar la función generadora de momentos para deducir que:

i) $E(X) = n$

ii) $Var(X) = 2n$

2.12. Si X es una variable aleatoria con distribución chi-cuadrado con n grados de libertad, mostrar que su función de densidad de probabilidad cumple la siguiente igualdad:

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

2.13. Leer en una tabla los siguientes valores de probabilidad correspondientes a una distribución t-student y elaborar la representación gráfica correspondiente:

a) $t_{0.975,8} = ?$

b) $t_{0.01,14} = ?$

c) $t_{0.95,40} = ?$

d) $t_{0.025,14} = ?$

2.14. Leer en una tabla los siguientes valores de probabilidad correspondientes a una distribución de Fisher y elaborar la representación gráfica correspondiente:

a) $F_{0.99,6,13} = ?$

b) $F_{0.95,10,4} = ?$

c) $F_{0.01,12,12} = ?$

d) $F_{0.05,9,13} = ?$

2.15. Si se realiza un muestreo sin reemplazo de una población finita de tamaño *N*, entonces, deducir que

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

2.16. La duración promedio de 1000 teclados es de 40 meses, con una desviación estándar de 10 meses; determinar: *i)* la probabilidad de que en una muestra aleatoria de 64 teclados se tenga una duración promedio comprendida entre 38 y 42 meses, *ii)* hallar la probabilidad de que en una muestra aleatoria de 36 teclados la duración promedio sea inferior a 39 meses.

2.17. El 98 % de los artículos que produce una máquina son de buena calidad. Si se toma una muestra aleatoria de 800 artículos: *i)* ¿Cuál es la probabilidad de que más del 98 % de los artículos de la muestra sean de buena calidad? y *ii)* ¿Cuál es la probabilidad de que menos del 3 % de los artículos de la muestra sean defectuosos?

2.18. En la ciudad A los adultos mayores de 60 años tienen un peso promedio de 65 kg, con una varianza poblacional de 10, mientras que para los de la ciudad B el peso promedio es de 66 kg, con una varianza poblacional de 12; si se toma una muestra aleatoria de 40 individuos en la ciudad A y otra de 45 individuos en la ciudad B: *i)* ¿Cuál es la probabilidad de que la media muestral del peso de adultos de la ciudad A difiera de la de los adultos de la ciudad B en más de 13 kg? y *ii)* ¿Cuál es la probabilidad de que la diferencia absoluta de las medias muestrales no superen los 9 kg?

3

Estimación de parámetros

De manera general, un proceso de inferencia estadística hace referencia a la extracción de una muestra aleatoria de una población, la cual tiene una distribución de probabilidad que contiene uno o varios parámetros desconocidos, sobre los cuales es posible realizar dos tipos de inferencia: *i)* estimación puntual o por intervalo y *ii)* contrastes o pruebas de hipótesis sobre cada uno de los parámetros. En el presente capítulo se desarrollan procesos de inferencia estadística focalizados en la estimación de parámetros; se inicia con algunos conceptos básicos asociados a la estimación puntual y, luego, se determina un intervalo de confianza para estimar algunos de los parámetros usuales en estadística inferencial básica, entre ellos, el intervalo para estimar la media y proporción poblacionales, la diferencia de medias y de proporciones poblacionales, la varianza y el cociente de varianzas. El capítulo cuatro se dedica la prueba de hipótesis.

3.1. Estimación puntual

Varios de los aspectos teóricos considerados en esta sección son asumidos o adaptados de los conceptos expuestos al respecto por diversos autores, entre ellos: Bickel & Doksum (1977), Canavos (1988), Devore (2008), Freund y Miller (2000), Gutiérrez, *et al.* (2008), Lindgren (1993), Meeker & Escobar (1998), Mayorga (2003), Macchi *et al.* (2014) y Walpole, Myers, Myers & Ye (2007). A continuación se indican algunos conceptos y propiedades alusivos a los estimadores, y se describe el método de estimación de máxima verosimilitud.

3.1.1 Propiedades de los estimadores

Recuérdese que un estimador T es una función de las variables que conforman una muestra aleatoria, pero que no incluye ningún parámetro θ; de hecho, T es otra variable aleatoria y, por lo tanto, es posible determinar su valor esperado $E(T)$ y su varianza $Var(T)$. La varianza de T se expresa de la siguiente manera:

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

Se define el error cuadrático medio del estimador T de la siguiente manera:

$$e.c.m(T) = E\left(T - \theta\right)^2$$

Ahora, la anterior expresión se puede descomponer así:

$$e.c.m(T) = E\left(T - \theta\right)^2 = E\left(T - E(T) + (E(T) - \theta)\right)^2$$

$$e.c.m(T) = E\left(T - E(T)\right)^2 + 2E\left((T - E(T))(E(T) - \theta)\right) + \left(E(T) - \theta\right)^2$$

$$e.c.m(T) = E\left(T - E(T)\right)^2 + 2\left((E(T) - E(T))(E(T) - \theta)\right) + \left(E(T) - \theta\right)^2$$

Como el doble producto se anula, entonces, la anterior expresión toma la forma:

$$e.c.m(T) = E\left(T - E(T)\right)^2 + \left(E(T) - \theta\right)^2$$

Al valor

$$B(T) = E(T) - \theta$$

se le denomina el sesgo del estimador T.

Enseguida se indican algunas características susceptibles de presentarse en un determinado estimador.

3.1.1.1 Insesgamiento

Se dice que un estimador T para el parámetro θ es insesgado si su sesgo es cero, es decir, T es insegado si

$$B(T) = E(T) - \theta = 0$$

En otras palabras, T es insesgado si el valor esperado del estimador T es igual al parámetro θ; luego para verificar si T es insesgado se ha de verificar la siguiente igualdad:

$$E(T) = \theta$$

Ahora, si T es un estimador insesgado, entonces el error cuadrático medio de T es igual a la varianza de T, es decir:

$$e.c.m(T) = E\left(T - E(T)\right)^2 + \left(E(T) - \theta\right)^2 = E\left(T - E(T)\right)^2 = Var(T)$$

Ejemplo 3.1. Analizar si el promedio muestral $\overline{X}$ es insesgado.

Efectivamente, como se cumple que:

$$E(\overline{X}) = \mu$$

entonces, se concluye que $\overline{X}$ es un estimador insesgado para el parámetro media poblacional μ.

Ejemplo 3.2. Determinar si la proporción muestral $\hat{p}$ es un estimador insesgado.

En efecto,

$$E(\hat{p}) = p$$

Luego, entonces, se concluye que $\hat{p}$ es un estimador insesgado para el parámetro proporción poblacional p.

Ejemplo 3.3. Establecer si la varianza corregida $\hat{S}^2$ es un estimador insesgado.

Debido a que se satisface que

$$E(\hat{S}^2) = \sigma^2$$

se concluye que $\hat{S}^2$ es un estimador insesgado para el parámetro varianza poblacional σ^2.

Ejemplo 3.4. Analizar si la varianza muestral S^2 es un estimador insesgado.

Puesto que, por definición, la varianza muestral y cuasivarianza se relacionan mediante la siguiente igualdad:

$$\hat{S}^2 = \frac{n}{n-1} S^2 \qquad \text{o} \qquad S^2 = \frac{n-1}{n} \hat{S}^2$$

entonces, resulta que

$$E(S^2) = E\left(\frac{n-1}{n}\hat{S}^2\right) = \frac{n-1}{n} E\left(\hat{S}^2\right) = \frac{n-1}{n}\sigma^2$$

Lo anterior indica que S^2 es un estimador sesgado o no insesgado para el parámetro varianza poblacional σ^2.

3.1.1.2 Insesgamiento asintótico

Un estimador T del parámetro θ es asintóticamente insesgado si

$$\lim_{n\to\infty} E(T) = \theta$$

Ejemplo 3.5. Analizar si la varianza muestral S^2 es un estimador asintóticamente insesgado.

La varianza muestral y la cuasivarianza se relacionan mediante la siguiente igualdad:

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

Entonces,

$$\lim_{n\to\infty} E(S^2) = \lim_{n\to\infty} E\left(\frac{n-1}{n}\hat{S}^2\right) = \lim_{n\to\infty} \frac{n-1}{n} E\left(\hat{S}^2\right) = \lim_{n\to\infty} \frac{n-1}{n}\sigma^2 = \sigma^2$$

Por lo tanto,

$$\lim_{n\to\infty} E(S^2) = \sigma^2$$

Lo anterior indica que S^2 es un estimador asintóticamente insesgado para el parámetro varianza poblacional σ^2.

3.1.1.3 Eficiencia relativa

Si T_1 y T_2 son estimadores para el parámetro θ y se cumple la siguiente desigualdad:

$$Var(T_1) < Var(T_2)$$

entonces, el estimador T_1 es relativamente más eficiente que el estimador T_2. En consecuencia, el estimador con menor varianza es el más eficiente; además, de la definición se tiene que

$$\frac{Var(T_1)}{Var(T_2)} < 1$$

3.1.1.4 Consistencia

Antes de indicar la propiedad de consistencia, se presenta la denominada desigualdad de *Tchebychev*. Sea X una variable aleatoria con varianza finita,

entonces, para todo $\varepsilon > 0$ se satisface la siguiente desigualdad (Blanco, 2004):

$$P\left(\left|X - E(X)\right| \geq \varepsilon\right) \leq \frac{Var(X)}{\varepsilon^2}$$

La propiedad de consistencia se enuncia de la siguiente forma: si la sucesión de estimadores $T_1, T_2, \ldots, T_n$ del parámetro θ, entonces, T_n es consistente si

$$\lim_{n \to \infty} P\left(\left|T_n - \theta\right| \geq \varepsilon\right) = 0$$

Ejemplo 3.6. Analizar si el promedio muestral es un estimador consistente para el parámetro μ.

Se sabe que para cualquier muestra de tamaño *n*,

$$E(\bar{X}_n) = \mu$$

Aplicando la desigualdad de *Tchebychev* a la variable promedio muestral, resulta:

$$P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) \leq \frac{Var(\bar{X}_n)}{\varepsilon^2}$$

debido a que

$$Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

al tomar el límite en ambos miembros de la última desigualdad y reemplazar el valor de la varianza del promedio resulta:

$$\lim_{n \to \infty} P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) \leq \lim_{n \to \infty} \frac{\sigma^2}{n\varepsilon^2}$$

El límite de la parte derecha de la desigualdad es cero, y las probabilidades del lado izquierdo resultan mayores o iguales que cero, en consecuencia:

$$0 \leq \lim_{n \to \infty} P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) \leq 0$$

Por lo tanto,

$$\lim_{n \to \infty} P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) = 0$$

Lo anterior indica que $\bar{X}_n$ es un estimador consistente para el parámetro media poblacional μ.

3.1.1.5 Robustez

Intuitivamente, un estimador *T* es robusto si es insensible a datos extremos, es decir, no se deja afectar por la presencia de datos extremos.

Ejemplo 3.7. Ilustrar si el promedio muestral es un estimador robusto para el parámetro μ.

Sea *X:* el peso de unos vacunos en kilogramos

Los datos recolectados para una muestra de tamaño cinco fueron: 400, 399, 401, 395, 405. El promedio muestral es 400 y la mediana también es de 400 kg.

Ahora, se ha tomado una segunda muestra, resultando un dato extremo; esta muestra presenta los siguientes datos: 400, 399, 401, 405, 300.

En este caso, la muestra apenas ha cambiado un valor correspondiente al peso; sin embargo, el peso promedio en esta muestra es de 381, "ha disminuido de manera importante, se ha dejado afectar por el dato extremo 300"; en cambio, la mediana sigue siendo 400 kg; lo anterior indica que la mediana muestral es un estimador robusto, y el promedio muestral es un estimador no robusto.

3.1.2 Estimación de parámetros por el método de máxima verosimilitud

En el proceso de estimación puntual de parámetros es posible utilizar diversos métodos: el de *máxima verosimilitud*, el de los *momentos*, el del *pivote*, por *analogía*, la *estimación bayesiana*, entre otros (Aubone & Wöhler, 2000; Cao & Van Keilegom, 2006). Por ser el más usual, a continuación se trata el método de estimación máximo verosímil y se proporcionan algunos ejemplos alusivos.

Si $X_1, X_2, ..., X_n$ es una muestra aleatoria para estudiar la variable aleatoria *X* con función de densidad de probabilidad:

$$f_X(x_1, x_2, ..., x_n, \theta)$$

donde θ es un parámetro o un vector en el espacio de parámetros $\Theta \subseteq R^n$, entonces la función de verosimilitud se denota y se define de la siguiente manera:

$$L(\theta) = L(x_1, x_2, ..., x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

Se dice que un estimador $T = t(X_1, X_2, ..., X_n)$ es un estimador máximo verosímil del parámetro θ si el valor particular de $t = t(x_1, x_2, ..., x_n)$ es tal que el supremo de *L* siguiente:

$$Sup\{L(\theta) / \theta \in \Theta\}$$

se alcanza cuando $t = \theta$. En este caso *t* recibe el nombre de estimador máximo

verosímil de θ (Mayorga, 2003). Es de anotar que si $t=\theta$ maximiza a $L(\theta)$, entonces $t=\theta$ también maximiza a $Ln(L(\theta))$.

Ejemplo 3.8. Enseguida se obtiene el estimador para el parámetro *p* del modelo de probabilidad de *Bernoulli* (Burbano *et al.*, 2014), por el método de máxima verosimilitud.

Para una variable aleatoria *X*, cuya función de probabilidad está dada por:

$$f(x,p)=p^{x}(1-p)^{1-x}$$

donde p es el parámetro de la variable aleatoria X del modelo de *Bernoulli*, esta toma los valores 0, 1; si se considera una muestra aleatoria $X_1, X_2, ..., X_n$ de una población con $f(x,p)=p^{x}(1-p)^{1-x}$, la función de verosimilitud es:

$$L(p)=\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

Donde los x_i son observaciones correspondientes a las variables aleatorias $X_1, X_2, ..., X_n$.

Se trata de encontrar un valor del parámetro de tal manera que se maximice la función de verosimilitud:

$$L(p)=p^{x_1}(1-p)^{1-x_1}.p^{x_2}(1-p)^{1-x_2}....p^{x_n}(1-p)^{1-x_n}.$$

$$L(p)=p^{x_1+x_2+...+x_n}(1-p)^{1+1+...+1-(x_1+x_2+...+x_n)}$$

$$L(p)=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}$$

$$Ln\left(L(p)\right)=Ln\left(p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}\right)$$

$$Ln\left(L(p)\right)=\sum_{i=1}^{n}x_i Ln(p)+\left(n-\sum_{i=1}^{n}x_i\right)Ln(1-p)$$

Ahora, se hace uso del cálculo diferencial para derivar parcialmente con respecto al parámetro p:

$$\frac{\partial}{\partial p}\left(Ln\left(L(p)\right)\right)=\frac{\partial}{\partial p}\left(\sum_{i=1}^{n}x_i Ln(p)\right)+\frac{\partial}{\partial p}\left(\left(n-\sum_{i=1}^{n}x_i\right)Ln(1-p)\right)$$

$$\frac{\partial}{\partial p}\left(Ln\left(L(p)\right)\right)=\sum_{i=1}^{n}x_i\left(\frac{1}{p}\right)+\left(n-\sum_{i=1}^{n}x_i\right)\left(\frac{-1}{1-p}\right)$$

Igualando a cero se tiene:

$$\frac{\partial}{\partial p}\left(Ln\left(L(p)\right)\right)=\sum_{i=1}^{n}x_i\left(\frac{1}{p}\right)+\left(n-\sum_{i=1}^{n}x_i\right)\left(\frac{-1}{1-p}\right)=0$$

$$\sum_{i=1}^{n}x_i\left(\frac{1}{p}\right)=\left(n-\sum_{i=1}^{n}x_i\right)\left(\frac{1}{1-p}\right)$$

$$(1-p)\sum_{i=1}^{n}x_i=p\left(n-\sum_{i=1}^{n}x_i\right)$$

$$\sum_{i=1}^{n}x_i-p\sum_{i=1}^{n}x_i=np-p\sum_{i=1}^{n}x_i$$

$$np=\sum_{i=1}^{n}x_i$$

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

$$\hat{p}=\frac{\sum_{i=1}^{n}x_i}{n}$$

Esta última expresión corresponde al estimador de máxima verosimilitud del parámetro p de la distribución de Bernoulli.

Ejemplo 3.9. Para la variable aleatoria X con distribución de *Poisson*, determinar el estimador máximo verosímil. La función de probabilidad correspondiente está dada por:

$$f(x,\theta)=\begin{cases}\dfrac{\lambda^x e^{-\lambda}}{x!} & \text{si } x=0,1,2,3,....,...\\ 0 & \text{en otro caso.}\end{cases}$$

donde $\lambda=\theta$ es el parámetro de la variable aleatoria X con modelo de *Poisson;* si se toma una muestra aleatoria $X_1, X_2, ..., X_n$, la función de verosimilitud es:

$$L(\lambda)=\prod_{i=1}^{n}\frac{\lambda^{x_i}e^{-\lambda}}{x_i!}$$

donde los x_i son observaciones correspondientes a las variables aleatorias $X_1, X_2, ..., X_n$.

Se trata de encontrar un valor del parámetro de tal manera que se maximice la función de verosimilitud:

$$L(\lambda)=\left(\frac{\lambda^{x_1}e^{-\lambda}}{x_1!}\right).\left(\frac{\lambda^{x_2}e^{-\lambda}}{x_2!}\right)....\left(\frac{\lambda^{x_n}e^{-\lambda}}{x_n!}\right)$$

$$L(\lambda)=\frac{\lambda^{x_1+x_2+...+x_n}e^{-n\lambda}}{x_1!x_2!...x_n!}$$

Ahora, aplicando el logaritmo natural, resulta:

$$Ln\left(L(\lambda)\right)=Ln(\lambda)\sum_{i=1}^{n}x_i+Ln\left(e^{-n\lambda}\right)-Ln(x_1!x_2!...x_n!)$$

A continuación, se realiza el cálculo de la derivada parcial con respecto al parámetro λ:

$$\frac{\partial Ln\left(L(\lambda)\right)}{\partial\lambda}=\frac{1}{\lambda}\sum_{i=1}^{n}x_i-n$$

Igualando a cero, resulta:

$$\frac{1}{\lambda}\sum_{i=1}^{n}x_i-n=0$$

Por lo tanto,

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

$$\hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

Esta última expresión corresponde al estimador de máxima verosimilitud del parámetro λ de una variable aleatoria con distribución de *Poisson*.

Ejemplo 3.10. Para la variable aleatoria *X,* con posible función de densidad de probabilidad, definida por:

$$f(x,\theta)=\begin{cases}\theta x^{\theta-1} & \text{si } x\in(0,1)\\ 0 & \text{en otro caso.}\end{cases}$$

donde $\theta>0$ es el parámetro de la población.

Como se trata de un modelo no usual, inicialmente se ha de analizar si *f* corresponde a una función de densidad para la variable aleatoria *X*.

En efecto, si $\theta > 0$ y la variable aleatoria X toma valores x en el intervalo $(0,1)$, entonces,

$$f(x,\theta) = \theta x^{\theta-1} \geq 0$$

Enseguida se ha de probar que

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

Esta integral se descompone de la siguiente manera:

$$\int_{-\infty}^{\infty} f(x,\theta)dx = \int_{-\infty}^{0} f(x,\theta)dx + \int_{0}^{1} f(x,\theta)dx + \int_{1}^{\infty} f(x,\theta)dx$$

$$\int_{-\infty}^{\infty} f(x,\theta)dx = \int_{-\infty}^{0} 0dx + \int_{0}^{1} \theta x^{\theta-1}dx + \int_{1}^{\infty} 0dx = x^{\theta}\Big|_{0}^{1} = 1-0 = 1$$

En consecuencia, la función f sí es una densidad de probabilidad para la variable aleatoria X. Para la muestra aleatoria $X_1, X_2, ..., X_n$, la función de verosimilitud es:

$$L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \left(\theta x_1^{\theta-1}\right).\left(\theta x_2^{\theta-1}\right)...\left(\theta x_n^{\theta-1}\right)$$

Donde los x_i son observaciones correspondientes a las variables aleatorias $X_1, X_2, ..., X_n$.

Se trata de encontrar un valor del parámetro de tal manera que se maximice la función de verosimilitud,

$$L(\theta) = \theta^n (x_1.x_2...x_n)^{\theta-1}$$

Ahora, aplicando el logaritmo natural, resulta:

$$Ln\left(L(\theta)\right) = nLn\theta + (\theta - 1)Ln(x_1.x_2...x_n)$$

Luego, se realiza el cálculo de la derivada parcial con respecto al parámetro θ

$$\frac{\partial Ln\left(L(\theta)\right)}{\partial \theta} = \frac{n}{\theta} + Ln(x_1.x_2...x_n)$$

Igualando a cero, se tiene:

$$\frac{n}{\theta} + Ln(x_1.x_2...x_n) = 0$$

Por consiguiente,

$$\theta = \frac{-n}{Ln(x_1.x_2...x_n)}$$

$$\hat{\theta} = \frac{-n}{Ln(x_1.x_2...x_n)}$$

Esta última expresión corresponde al estimador de máxima verosimilitud del parámetro θ.

3.2 Estimación por intervalo

Se trata de determinar un intervalo de la forma *(a,b)* que contenga al parámetro θ de interés con una alta probabilidad $(1-\alpha)$; a esta se le denomina nivel de confianza o nivel de confiabilidad. En esta sección se obtienen los intervalos de confianza para estimar la media y la proporción poblacional, la diferencia de medias y la diferencia de proporciones, el intervalo para la varianza poblacional y para el cociente de varianzas poblacionales.

De manera sintética, se trata de determinar los valores *a* y *b* tales que

$$P(a \leq \theta \leq b) = 1-\alpha$$

Una representación gráfica de esta situación se observa en la Figura 3.1.

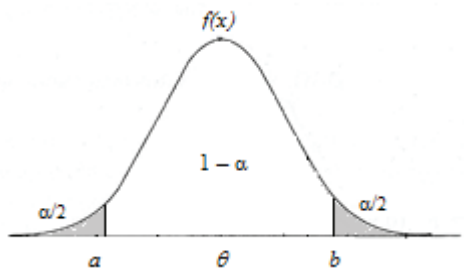

Figura 3.1 Intervalo de confianza

Fuente: los autores con la ayuda del *software* libre R.

3.2.1 Intervalos de confianza para estimar la media poblacional

Ahora, se especifica un intervalo de la forma *(a,b)* que contenga el parámetro μ con un nivel de confianza de $(1-\alpha)$. Se busca determinar los valores *a* y *b* tales que

$$P(a \leq \mu \leq b) = 1-\alpha$$

Al valor *a* se le denomina límite inferior del intervalo, y al valor *b* se le llama límite superior. Una representación gráfica de esta situación se visualiza en la Figura 3.2.

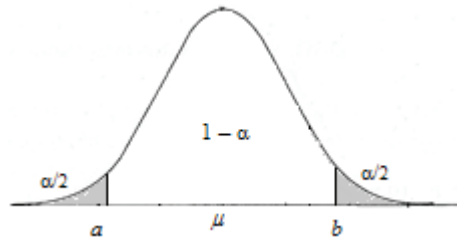

Figura 3.2 Intervalo de confianza para la media poblacional

Fuente: los autores con la ayuda del *software* libre R.

Caso 1. De la expresión 2.1 se tiene que

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Luego, en concordancia con la distribución normal estándar, se trata de determinar

$$P(Z_{\frac{\alpha}{2}} \leq Z \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(Z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

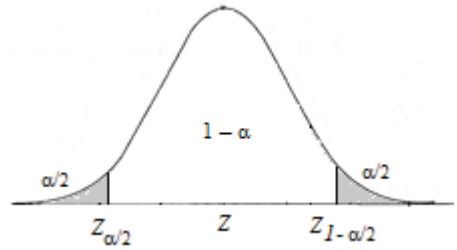

Figura 3.3 Intervalo de confianza sobre la curva normal

Fuente: los autores con la ayuda del *software* libre R.

En concordancia con la Figura 3.3, la función de densidad de una variable aleatoria *Z* con distribución normal estándar permite escribir:

$$P(Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

De aquí se deduce que

$$P(-\bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Multiplicando por -1 en cada uno de los miembros de la desigualdad del evento al que se le calcula la probabilidad, resulta:

$$P(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

La anterior expresión se escribe de la siguiente forma:

$$P(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Debido a que la curva normal es simétrica, resulta que:

$$Z_{\frac{\alpha}{2}} = -Z_{1-\frac{\alpha}{2}}$$

Usando la igualdad anterior en la expresión que le precede, se tiene que:

$$P(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Por lo tanto, el intervalo de confianza para estimar la media poblacional cuando se conoce la desviación estándar poblacional está dado por la expresión 3.1, para cuando se muestrea de una poblacional infinita.

$$\mu \in (\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) \qquad (3.1)$$

Al valor $\frac{\sigma}{\sqrt{n}}$ se le denomina error estándar, para estimar la media poblacional usando el promedio muestral (Gutiérrez *et al.*, 2008).

Al valor $Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ se le denomina error de estimación, porque hasta ese valor

puede diferir el estimador puntual denominado media muestral del parámetro media poblacional (Gutiérrez *et al.*, 2008).

Caso 2. De la expresión 2.3 se tiene que

$$t = \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}}$$

tiene distribución t-student con $n - 1$ grados de libertad, luego se trata de determinar

$$P(t_{\frac{\alpha}{2}} \leq t \leq t_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(t_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \leq t_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

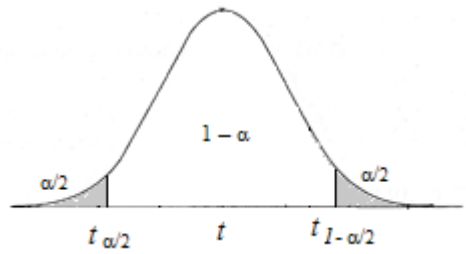

Figura 3.4 Intervalo de confianza sobre la curva t-student

Fuente: los autores con la ayuda del *software* libre R.

De acuerdo con la Figura 3.4, la función de densidad de una variable aleatoria t con distribución *t-student* permite escribir:

$$P(t_{\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \leq \bar{X} - \mu \leq t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}) = 1 - \alpha$$

De aquí se deduce que

$$P(-\bar{X} + t_{\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \leq -\mu \leq -\bar{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}) = 1 - \alpha$$

Multiplicando por -1 en cada uno de los miembros de la desigualdad del evento involucrado en el cálculo de la probabilidad, resulta

$$P(\bar{X} - t_{\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \geq \mu \geq \bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}) = 1 - \alpha$$

La anterior expresión se escribe de la siguiente manera:

$$P(\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \leq \mu \leq \bar{X} - t_{\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}) = 1 - \alpha$$

Puesto que la curva t-student es simétrica, se tiene que:

$$t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$$

Usando la igualdad anterior en la expresión que le precede, se deduce que:

$$P(\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}) = 1 - \alpha$$

Por lo tanto, el intervalo de confianza para estimar la media poblacional cuando se desconoce la desviación estándar poblacional y el tamaño de la muestra es inferior a 30; cuando se muestrea de una población infinita está dado por la expresión 3.2:

$$\mu \in (\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}) \tag{3.2}$$

Caso 3. Al proceder de manera similar a como se obtuvo el intervalo de confianza presentado por medio de la expresión 3.1, pero utilizando la expresión 2.2, que estableció que

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)} \sim N(0,1)$$

se deduce que el intervalo de confianza para estimar la media poblacional cuando se conoce la desviación estándar poblacional corresponde a la expresión 3.3 para cuando se muestrea de una poblacional finita de tamaño *N*:

$$\mu \in (\bar{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}) \tag{3.3}$$

Caso 4. Al realizar un proceso semejante al desarrollado para obtener el intervalo de confianza indicado a través de la expresión 3.2, pero usando la expresión 2.4, se estableció que:

$$t = \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)}$$

tenía una distribución *t-student* con $n-1$ grados de libertad; por consiguiente, se deduce que el intervalo de confianza para estimar la media poblacional cuando se desconoce la desviación estándar poblacional corresponde a la expresión 3.4 para cuando se muestrea de una poblacional finita de tamaño N.

$$\mu \in (\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \bar{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}) \quad (3.4)$$

Ejemplo 3.11. En la empresa de lácteos T&R, la cantidad de leche depositada por la máquina A en cada una de las bolsas se distribuye normalmente con desviación estándar de 5 mililitros; se toma una muestra aleatoria de 16 bolsas de tal máquina y se encuentra un contenido promedio de 900 mililitros. Construir un intervalo de confianza del 98 % para estimar la verdadera media de llenado en la producción que se haga a través de la máquina A.

En este caso, se tiene:

$$n = 16$$
$$\bar{X} = 900$$
$$\sigma = 5$$

Además,

$$1-\alpha = 0.98 \rightarrow \alpha = 0.02$$
$$\frac{\alpha}{2} = 0.01$$

En la Figura 3.5 se observa el valor de Z obtenido al leer una tabla normal estándar para un 99 % de probabilidad.

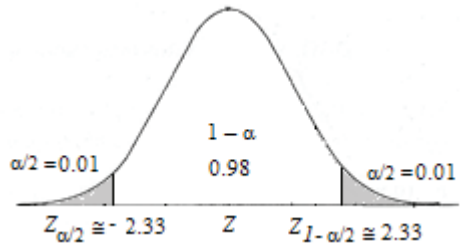

Figura 3.5 Valor de Z en una curva normal estándar
Fuente: los autores con la ayuda del *software* libre R.

$$Z_{1-\frac{\alpha}{2}} = Z_{0.99} \cong 2.33$$

Luego el intervalo de confianza es:

$$\mu \in (\bar{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

Reemplazando los valores, resulta:

$$\mu \in \left(900 - 2.33(\frac{5}{\sqrt{16}}), 900 + 2.33(\frac{5}{\sqrt{16}}) \right)$$

Realizando las operaciones de tipo aritmético se obtiene:

$$\mu \in (900 - 2.33(1.25), 900 + 2.33(1.25))$$

Luego,

$$\mu \in (900 - 2.9125, 900 + 2.9125)$$

Finalmente,

$$\mu \in (897.08, 902.91)$$

En conclusión, con un nivel de confianza del 98 % se puede afirmar que el promedio poblacional de llenado de leche de las bolsas producidas por la máquina A está ente 897.08 mililitros y 902.91 mililitros, aproximadamente. Lo anterior indica que de cada 100 muestras que se tomen, en 98 se encuentra el parámetro μ.

Ejemplo 3.12. En una muestra aleatoria de 10 latas de un producto se obtuvo un peso neto promedio de 184 g, con una desviación estándar corregida de 3 g; si se asume que los datos provienen de una distribución normal, determinar un intervalo de confianza del 95 % para estimar el verdadero peso promedio de las latas del producto en la población.

$$\hat{S} = 3 \qquad 1 - \alpha = 95\% = 0.95$$
$$n = 10 \qquad \alpha = 0.05$$
$$\bar{X} = 184 \qquad \frac{\alpha}{2} = 0.025$$

En la Figura 3.6 se observa el valor de *t* obtenido al leer una tabla de la distribución t-student con n – 1 = 10 – 1 = 9 grados de libertad para un 0.975 = 97.5 % de probabilidad.

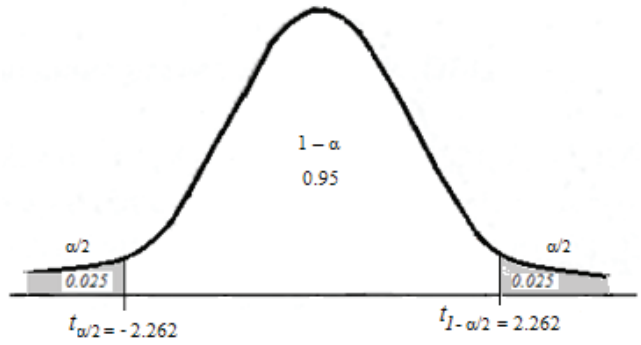

Figura 3.6 Valor de t en una curva t-student con $n = 9$ grados
Fuente: los autores con la ayuda del *software* libre R.

$$t_{1-\frac{\alpha}{2}} = t_{0.975,9} = 2.262$$

Así, el intervalo de confianza es:

$$\mu \in (\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}})$$

Al reemplazar los valores se tiene:

$$\mu \in \left(184 - 2.262(\frac{3}{\sqrt{10}}), 184 + 2.262(\frac{3}{\sqrt{10}}) \right)$$

Haciendo las operaciones de tipo aritmético se obtiene:

$$\mu \in (184 - 2.262(0.9487), 184 + 2.262(0.9487))$$

Entonces,

$$\mu \in (184 - 2.1459, 184 + 2.1459)$$

Por consiguiente,

$$\mu \in (181.85, 186.15)$$

En conclusión, con un nivel de confianza del 95 % se infiere que el peso promedio de las latas del producto en la poblacional está entre 181.85 g y 186.15 g, aproximadamente. Lo anterior indica que en 95 de cada 100 muestras tomadas se encuentra el parámetro μ.

Ejemplo 3.13. Este ejemplo es adaptado de Gutiérrez *et al.* (2008). En un proceso de inyección de plástico, una característica de calidad del producto

(disco) es su grosor, que ha de ser de 1.21 mm, con una tolerancia de ± 0.09 mm. En este contexto, el grosor del disco debe estar en un rango de 1.12 y 1.30 mm, para aceptar que el proceso de inyección resultó satisfactorio. Para evaluar esta característica de calidad, durante una semana se ha realizado un muestreo sistemático en una de las líneas de producción, seleccionando cuatro muestras de tamaño 20 y una de tamaño 21, bajo normalidad, para finalmente obtener una muestra de n = 101; de allí se obtienen la media muestral y la desviación estándar corregida, que tuvieron los siguientes valores: $\overline{X} = 1.18$ mm y $\hat{S} = 0.0259$ mm. Hallar el error estándar para estimar la media, construir un intervalo de confianza del 95 % para estimar la media poblacional y determinar el error de estimación.

En primera instancia, el error estándar para la media es:

$$\frac{\hat{S}}{\sqrt{n}} = \frac{0.0259}{\sqrt{101}} = \frac{0.0259}{10.0498} = 0.002577$$

Ahora, para construir el intervalo de confianza se tiene en cuenta que:

$$1 - \alpha = 95\% = 0.95$$

$$\alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

En la Figura 3.7 se observa el valor de *t* obtenido al leer una tabla de la distribución *t-student* con $n - 1 = 101 - 1 = 100$ grados de libertad para un 0.975 = 97.5 % de probabilidad.

$$t_{1-\frac{\alpha}{2}} = t_{0.975,100} = 1.984$$

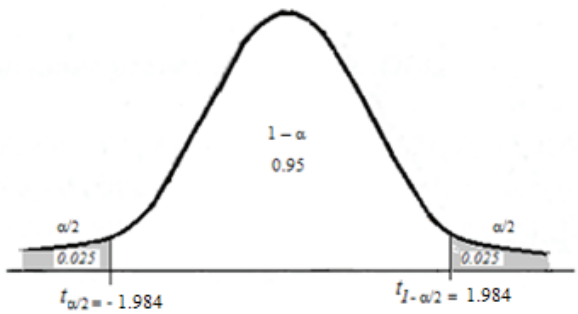

Figura 3.7 Valor de t en una curva t-student con *n* = 100 grados

Fuente: los autores con la ayuda del *software* libre R.

Así, el intervalo de confianza es:

$$\mu \in (\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}})$$

Al reemplazar los valores se tiene:

$$\mu \in \left(1.18 - 1.984(\frac{0.0259}{\sqrt{101}}), 1.18 + 1.984(\frac{0.0259}{\sqrt{101}}) \right)$$

Haciendo las operaciones de tipo aritmético se obtiene:

$$\mu \in (1.18 - 1.984(0.002577), 1.18 + 1.984(0.002577))$$

Entonces,

$$\mu \in (1.18 - 0.00511, 1.18 + 0.00511)$$

Por lo tanto,

$$\mu \in (1.17489, 1.18511)$$

En conclusión, con un nivel de confianza del 95 % se infiere que el grosor promedio de los discos en la poblacional está entre 1.17489 mm y 1.18511 mm, aproximadamente; estos resultados indican que tal grosor se encuentra dentro de los valores especificados de calidad. Lo anterior significa que en 95 de cada 100 muestras tomadas se encuentra el parámetro μ, correspondiente al grosor promedio en la población de discos.

Finalmente, el error de estimación es:

$$t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} = 1.984(\frac{0.0259}{\sqrt{101}}) = 0.00511$$

El valor 0.00511 indica que hasta ese valor puede diferir el estimador puntual $\bar{X}$ del parámetro μ.

3.2.2 Intervalos de confianza para estimar la proporción poblacional

Nuevamente se trata de construir un intervalo de la forma *(a,b)* que contenga el parámetro p con un nivel de confianza de $(1 - \alpha)$. Se busca determinar los valores *a* y *b* tales que:

$$P(a \leq p \leq b) = 1 - \alpha$$

Una representación gráfica de esta situación se observa en la Figura 3.8:

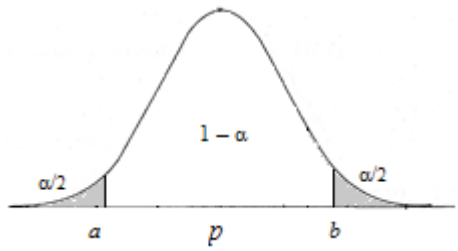

Figura 3.8 Intervalo de confianza para la proporción poblacional

Fuente: los autores con la ayuda del *software* libre R.

Caso 1. De la expresión 2.5 se tiene que

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

Para un tamaño de muestra n lo suficientemente grande, se asume que

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \sim N(0,1)$$

Así entonces, de acuerdo con la distribución normal estándar (ver Figura 3.3), se tiene:

$$P(Z_{\frac{\alpha}{2}} \leq Z \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(Z_{\frac{\alpha}{2}} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

Esta expresión se escribe de la siguiente forma:

$$P(Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq \hat{p} - p \leq Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}) = 1 - \alpha$$

De aquí se establece que

$$P(-\hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq -p \leq -\hat{p} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}) = 1 - \alpha$$

Multiplicando por -1 en cada uno de los miembros de la desigualdad del evento al que se le calcula la probabilidad, resulta:

$$P(\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \geq p \geq \hat{p} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}) = 1 - \alpha$$

Esta expresión se escribe de la siguiente manera:

$$P(\hat{p} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}) = 1 - \alpha$$

Puesto que la curva normal es simétrica, resulta que

$$Z_{\frac{\alpha}{2}} = -Z_{1-\frac{\alpha}{2}}$$

Usando la igualdad anterior en la expresión que le precede, se tiene que

$$P(\hat{p} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}) = 1 - \alpha$$

Por lo tanto, el intervalo de confianza para estimar la proporción poblacional cuando se muestrea de una poblacional infinita se especifica en la expresión 3.5.

$$p \in \left(\hat{p} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \ , \ \hat{p} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \right) \tag{3.5}$$

Al valor $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ se le denomina error estándar para estimar la proporción poblacional usando la proporción muestral, y al valor $Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$ se le llama error de estimación.

Caso 2. Ahora, si se muestrea de una población finita de tamaño N y el tamaño de la muestra es "grande", entonces se deduce que

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}} \sim N(0,1) \tag{3.6}$$

Para un tamaño de muestra n lo suficientemente grande, se asume que

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}}} \sim N(0,1)$$

Realizando un procedimiento similar al efectuado para el *caso 1*, se establece que el intervalo de confianza para estimar la proporción poblacional cuando

se muestrea de una población finita de tamaño N y el tamaño de la muestra es "grande" es aquel que se indica en la expresión 3.7.

$$p \in \left(\hat{p} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \sqrt{\frac{N-n}{N-1}} \ , \ \hat{p} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \sqrt{\frac{N-n}{N-1}} \right) \tag{3.7}$$

Ejemplo 3.14. Un jugador de baloncesto produce 140 aciertos en 400 lanzamientos al aro en entrenamientos seleccionados de manera aleatoria. Construir un intervalo de confianza del 99 % para estimar la verdadera proporción de la efectividad del jugador.

En primera instancia, se define así *X:* número de aciertos en la muestra de 400 lanzamientos, luego

$$n = 400$$

$$\hat{p} = \frac{x}{n} = \frac{140}{400} = 0.35$$

$$\hat{q} = 1 - 0.35 = 0.65$$

Adicionalmente,

$$1 - \alpha = 0.99 \rightarrow \alpha = 0.01$$

$$\frac{\alpha}{2} = 0.005$$

En la Figura 3.9 se observa el valor de Z obtenido al leer una tabla normal estándar para un 0.995 = 99.5 % de probabilidad.

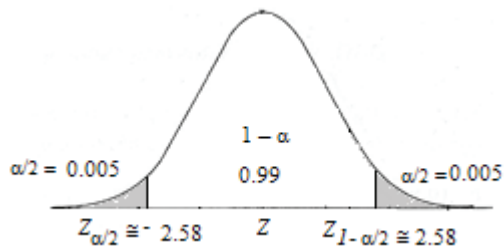

Figura 3.9 Valor de Z en una curva normal estándar
Fuente: los autores con la ayuda del *software* libre R.

$$Z_{1-\frac{\alpha}{2}} = Z_{0.995} \cong 2.58$$

Luego el intervalo de confianza es:

$$p \in \left(\hat{p} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \ , \ \hat{p} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$$

Reemplazando los valores, resulta:

$$p \in \left(0.35 - 2.58 \sqrt{\frac{0.35 * 0.65}{400}}, 0.35 + 2.58 \sqrt{\frac{0.35 * 0.65}{400}} \right)$$

Realizando las operaciones de tipo aritmético se obtiene:

$$p \in (0.35 - 2.58(0.0238) \ , \ 0.35 + 2.58(0.0238))$$

Luego

$$p \in (0.35 - 0.0614 \ , \ 0.35 + 0.0614)$$

En consecuencia,

$$p \in (0.2886 \ , \ 0.4114)$$

En conclusión, con un nivel de confianza del 99 % es posible afirmar que la verdadera proporción de la efectividad del jugador (en la población de entrenamientos) se encuentra entre el 28.86 % y 41.14 %. Lo anterior indica que en 99 de cada 100 muestras tomadas se encuentra el parámetro p .

3.2.3 Intervalos de confianza para estimar la diferencia de proporciones poblacionales

Ahora se pretende construir un intervalo de la forma *(a,b)* que contenga al parámetro $p_1 - p_2$ con un nivel de confianza de (1 – α). Se busca determinar los valores *a* y *b* tales que:

$$P(a \leq p_1 - p_2 \leq b) = 1 - \alpha$$

Una representación gráfica de esta situación se presenta en la Figura 3.10.

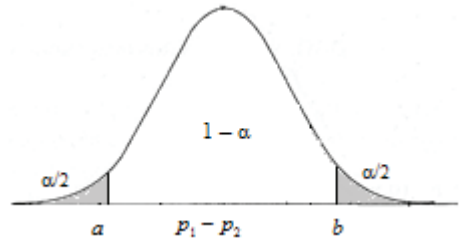

Figura 3.10 Intervalo de confianza para la diferencia de proporciones poblacionales

Fuente: los autores con la ayuda del *software* libre R.

De la expresión 2.6 se tiene que:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$

Para tamaños de las dos muestras n_1 y n_2 grandes, se asume que:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} \sim N(0,1)$$

En concordancia con la distribución normal estándar (ver Figura 3.3), resulta:

$$P(Z_{\frac{\alpha}{2}} \leq Z \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(Z_{\frac{\alpha}{2}} \leq \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

La anterior expresión se escribe de la siguiente manera:

$$P\left(Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq \hat{p}_1 - \hat{p}_2 - (p_1 - p_2) \leq Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right) = 1 - \alpha$$

De aquí se establece que:

$$P\left(-(\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq -(p_1 - p_2) \leq -(\hat{p}_1 - \hat{p}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right) = 1 - \alpha$$

Multiplicando por -1 en cada uno de los miembros de la desigualdad del evento involucrado en el cálculo de la probabilidad, resulta:

$$P\left((\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \geq (p_1 - p_2) \geq (\hat{p}_1 - \hat{p}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right) = 1 - \alpha$$

Esta expresión se escribe de la siguiente forma:

$$P\left((\hat{p}_1 - \hat{p}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right) = 1 - \alpha$$

Puesto que la curva normal es simétrica, resulta que:

$$Z_{\frac{\alpha}{2}} = -Z_{1-\frac{\alpha}{2}}$$

Usando la igualdad anterior en la expresión que le precede, se tiene que:

$$P\left((\hat{p}_1-\hat{p}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}} \le p_1-p_2 \le (\hat{p}_1-\hat{p}_2)+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}\right)=1-\alpha$$

Por lo tanto, el intervalo de confianza para estimar la diferencia de proporciones poblacionales se especifica en la expresión 3.8.

$$p_1-p_2 \in \left((\hat{p}_1-\hat{p}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}\ ,\ (\hat{p}_1-\hat{p}_2)+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}\right) \quad (3.8)$$

Ejemplo 3.15. Un candidato a la presidencia de la república de Colombia tiene, en una muestra aleatoria de 500 ciudadanos seleccionados en el departamento de Boyacá, el 60 % de la intención de voto, y en una muestra aleatoria de 450 ciudadanos tomada en el departamento de Cundinamarca logra un 56 % de la intención de voto; construir un intervalo de confianza del 98 % para estimar la verdadera diferencia de proporciones.
En este caso se tiene:

Boyacá	Cundinamarca
$n_1 = 500$	$n_2 = 450$
$\hat{p}_1 = 0.6$	$\hat{p}_2 = 0.56$
$\hat{q}_1 = 1-0.6 = 0.4$	$\hat{q}_2 = 1-0.56 = 0.44$

Además,

$$1-\alpha = 0.98 \rightarrow \alpha = 0.02$$

$$\frac{\alpha}{2} = 0.01$$

En la Figura 3.11 se observa el valor de Z obtenido al leer una tabla normal estándar para un 0.99 = 99 % de probabilidad.

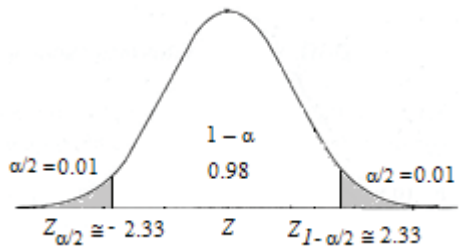

Figura 3.11 Valor de Z en una curva normal estándar
Fuente: los autores con la ayuda del *software* libre R.

$$Z_{1-\frac{\alpha}{2}} = Z_{0.99} \cong 2.33$$

Luego, en concordancia con la expresión 3.8, el intervalo de confianza es:

$$p_1 - p_2 \in \left((\hat{p}_1 - \hat{p}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \ , \ (\hat{p}_1 - \hat{p}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \right)$$

Reemplazando los valores, resulta:

$$\left((0.6-0.56) - 2.33\sqrt{\frac{(0.6)(0.4)}{500} + \frac{(0.56)(0.44)}{450}} \ , \ (0.6-0.56) + 2.33\sqrt{\frac{(0.6)(0.4)}{500} + \frac{(0.56)(0.44)}{450}} \right)$$

Realizando las operaciones de tipo aritmético se obtiene:

$$p_1 - p_2 \in \left((0.6-0.56) - 2.33\sqrt{0.000480 + 0.000547} \ , \ (0.6-0.56) + 2.33\sqrt{0.000480 + 0.000547} \right)$$

Luego:

$$p_1 - p_2 \in \left((0.6-0.56) - 2.33(0.0320) \ , \ (0.6-0.56) + 2.33(0.032) \right)$$

En consecuencia,

$$p_1 - p_2 \in \left(0.04 - 0.0746 \ , \ 0.04 + 0.0746 \right)$$

Finalmente,

$$p_1 - p_2 \in \left(-0.0346 \ , \ 0.1146 \right)$$

En conclusión, con un nivel de confianza del 98% se establece que la verdadera diferencia de proporciones poblacionales se encuentra entre -3.46% y 11.46%. Lo anterior indica que en 98 de cada 100 muestras tomadas se encuentra el parámetro $p_1 - p_2$.

Adicionalmente, en este ejemplo se infieren las tres situaciones siguientes:

i) $p_1 - p_2 = 0 \Rightarrow p_1 = p_2$;

ii) $p_1 - p_2 > 0 \Rightarrow p_1 > p_2$;

iii) $p_1 - p_2 < 0 \Rightarrow p_1 < p_2$

3.2.4 Intervalos de confianza para estimar la diferencia de medias poblacionales

En este apartado se determina un intervalo de la forma *(a,b),* de modo que este contenga el parámetro $\mu_1 - \mu_2$ con un nivel de confianza de $(1 - \alpha)$. Se busca especificar los valores *a* y *b* tales que:

$$P(a \leq \mu_1 - \mu_2 \leq b) = 1 - \alpha$$

Una representación gráfica de esta situación se observa en la Figura 3.12.

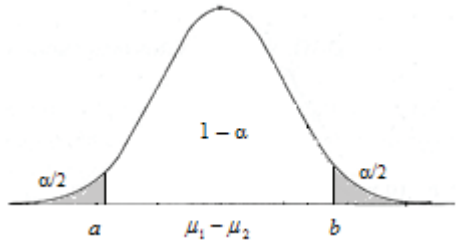

Figura 3.12 Intervalo de confianza para la diferencia de medias poblacionales
Fuente: los autores con la ayuda del *software* libre R.

Caso 1. De la expresión 2.7 se sigue que:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Mediante la distribución normal estándar, se trata de determinar:

$$P(Z_{\frac{\alpha}{2}} \leq Z \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(Z_{\frac{\alpha}{2}} \leq \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

La anterior expresión se escribe así:

$$P\left(Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \bar{X}_1-\bar{X}_2-(\mu_1-\mu_2) \le Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)=1-\alpha$$

De aquí se deduce que:

$$P\left(-(\bar{X}_1-\bar{X}_2)+Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le -(\mu_1-\mu_2) \le -(\bar{X}_1-\bar{X}_2)+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)=1-\alpha$$

Multiplicando por -1 en cada uno de los miembros de la desigualdad del evento para el cálculo de la probabilidad, resulta:

$$P\left((\bar{X}_1-\bar{X}_2)-Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \ge (\mu_1-\mu_2) \ge (\bar{X}_1-\bar{X}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)=1-\alpha$$

Esta expresión se escribe de la siguiente manera:

$$P\left((\bar{X}_1-\bar{X}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1-\mu_2 \le (\bar{X}_1-\bar{X}_2)-Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)=1-\alpha$$

Debido a que la curva normal es simétrica, se tiene que:

$$Z_{\frac{\alpha}{2}}=-Z_{1-\frac{\alpha}{2}}$$

Usando la igualdad anterior en la expresión que le precede, se deduce que:

$$P\left((\bar{X}_1-\bar{X}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1-\mu_2 \le (\bar{X}_1-\bar{X}_2)+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)=1-\alpha$$

Por lo tanto, el intervalo de confianza para estimar la diferencia de medias poblacionales cuando se conocen las respectivas desviaciones estándar poblacionales está dado por la expresión 3.9.

$$\mu_1-\mu_2 \in \left((\bar{X}_1-\bar{X}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\ ,\ (\bar{X}_1-\bar{X}_2)+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)=1-\alpha \qquad (3.9)$$

Al valor $\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ se le denomina error estándar para estimar la diferencia de medias poblacionales. Al valor $Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ se le denomina error de estimación.

Caso 2. De la expresión 2.8 se sigue que

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

tiene distribución t-student con $n_1 + n_2 - 2$ grados de libertad, donde de S_p se obtiene mediante la siguiente expresión, a partir de datos provenientes de muestras independientes de poblaciones normales:

$$S_p = \sqrt{\frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}}$$

Luego, se trata de determinar:

$$P(t_{\frac{\alpha}{2}} \leq t \leq t_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(t_{\frac{\alpha}{2}} \leq \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

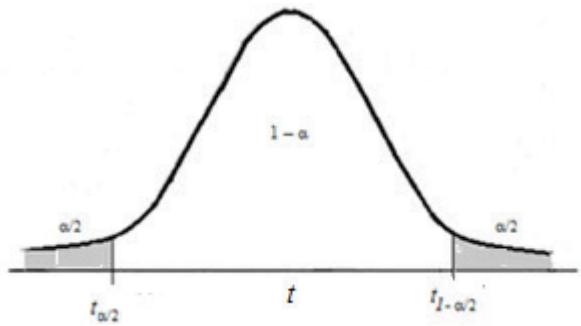

Figura 3.13 Intervalo de confianza sobre la curva t-student

Fuente: los autores con la ayuda del *software* libre R.

De acuerdo con la Figura 3.13, la función de densidad de una variable aleatoria *t* con distribución *t-student* permite escribir:

$$P\left(t_{\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2) \leq t_{1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) = 1 - \alpha$$

De aquí se deduce que:

$$P\left(-(\bar{X}_1-\bar{X}_2)+t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\leq-(\mu_1-\mu_2)\leq-(\bar{X}_1-\bar{X}_2)+t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)=1-\alpha$$

Multiplicando por -1 en cada uno de los miembros de la desigualdad del evento involucrado en el cálculo de la probabilidad, resulta:

$$P\left((\bar{X}_1-\bar{X}_2)-t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\geq(\mu_1-\mu_2)\geq(\bar{X}_1-\bar{X}_2)-t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)=1-\alpha$$

La anterior expresión se escribe de la siguiente manera:

$$P\left((\bar{X}_1-\bar{X}_2)-t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\leq\mu_1-\mu_2\leq(\bar{X}_1-\bar{X}_2)-t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)=1-\alpha$$

Puesto que la curva t-student es simétrica, se tiene que:

$$t_{\frac{\alpha}{2}}=-t_{1-\frac{\alpha}{2}}$$

Usando la igualdad anterior en la expresión que le precede, se establece que:

$$P\left((\bar{X}_1-\bar{X}_2)-t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\leq\mu_1-\mu_2\leq(\bar{X}_1-\bar{X}_2)+t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)=1-\alpha$$

Por lo tanto, el intervalo de confianza para estimar la diferencia de medias poblacionales, cuando se desconocen sus correspondientes desviaciones estándar poblacionales y el tamaño de las muestras, es inferior a 30; bajo el supuesto de varianzas poblacionales iguales (homocedasticidad) está dado por la expresión 3.10.

$$\mu_1-\mu_2\in\left((\bar{X}_1-\bar{X}_2)-t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\ ,\ (\bar{X}_1-\bar{X}_2)+t_{1-\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) \quad (3.10)$$

Caso 3. Al proceder de manera similar, pero utilizando la expresión 2.9, se obtiene que:

$$t=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}}}$$

con *g* grados de libertad. Se deduce que el intervalo de confianza para estimar la diferencia de medias poblacionales cuando se tienen muestras inferiores a

30 y se desconocen las desviaciones estándar poblacionales, pero considerando varianzas poblacionales desiguales (heterocedasticidad), corresponde a la expresión 3.11.

$$\mu_1-\mu_2 \in \left((\bar{X}_1-\bar{X}_2)-t_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}}\ ,\ (\bar{X}_1-\bar{X}_2)+t_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}} \right) \quad (3.11)$$

Caso 4. Al realizar un proceso semejante, pero usando la expresión 2.10, se estableció que:

$$Z=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}}} \sim N(0,1)$$

Luego se deduce que el intervalo de confianza para estimar la diferencia de medias poblacionales cuando se desconocen las desviaciones estándar poblacionales, pero las muestras tienen tamaños superiores a 30, está dado por la expresión 3.12.

$$\mu_1-\mu_2 \in \left((\bar{X}_1-\bar{X}_2)-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}}\ ,\ (\bar{X}_1-\bar{X}_2)+Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}} \right) \quad (3.12)$$

Ejemplo 3.16. En una muestra aleatoria conformada por 22 bolsas de jugo de la marca A se obtiene un contenido medio de 10 g de fibra con una desviación estándar corregida de 1.2 g; en otra, conformada por 25 bosas de esta clase de jugo, pero de la marca B, se obtuvo un contenido promedio de fibra de 9.8 g, con una desviación estándar de 0.95 g; se supone que las muestras provienen de poblaciones normales. Construir un intervalo de confianza del 95 % para estimar la diferencia de contenido medio de fibra entre las marcas A y B del mencionado jugo.

En este caso se tiene:

Marca A	Marca B
$n_1=22$	$n_2=25$
$\bar{X}_1=10$	$\bar{X}_2=9.8$
$\hat{S}_1=1.2$	$\hat{S}_2=0.9695$

El valor de la desviación estándar corregida para la marca B se obtiene de la siguiente manera:

$$\hat{S}_2^2 = \frac{n_2}{n_2 - 1} S_2^2 = \frac{25}{24}(0.95)^2 = 0.9401 \rightarrow \hat{S}_2 = 0.9695$$

Además,

$$1 - \alpha = 0.95 \rightarrow \alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

En la Figura 3.14 se observa el valor aproximado de *t* obtenido al leer una tabla *t-student* con n_1+n_2-2=22+25-2=45 grados de libertad, para un 0.975=97.5 % de probabilidad. En este caso se obtendrá un intervalo de confianza considerando que las varianzas poblacionales son iguales y se recurrirá a la expresión 3.10.

Figura 3.14 Intervalo de confianza sobre la curva t-student

Fuente: los autores con la ayuda del *software* libre R.

$$t_{1-\frac{\alpha}{2}} = t_{0.975,45} = 2.014$$

Ahora se calcula S_p así:

$$S_p = \sqrt{\frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{21(1.2)^2 + 24(0.9695)^2}{22 + 25 - 2}} = \sqrt{\frac{52.8024}{45}} = \sqrt{1.1733} = 1.083$$

Luego, usando:

$$\mu_1 - \mu_2 \in \left((\bar{X}_1 - \bar{X}_2) - t_{1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \ , \ (\bar{X}_1 - \bar{X}_2) + t_{1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

Y reemplazando los valores, resulta:

$$\mu_1 - \mu_2 \in \left((10 - 9.8) - 2.014(1.083)\sqrt{\frac{1}{22} + \frac{1}{25}} \ , \ (10 - 9.8) + 2.014(1.083)\sqrt{\frac{1}{22} + \frac{1}{25}} \right)$$

Efectuando las operaciones de tipo aritmético se obtiene:

$$\mu_1 - \mu_2 \in \left((10-9.8) - 2.014(1.083)(0.2923)\ ,\ (10-9.8) + 2.014(1.083)(0.2923)\right)$$

Luego,

$$\mu_1 - \mu_2 \in \left(0.2 - 0.6375\ ,\ 0.2 + 0.6375\right)$$

Finalmente,

$$\mu_1 - \mu_2 \in \left(-0.4375\ ,\ 0.8375\right)$$

En conclusión, con un nivel de confianza del 95 % se infiere que la diferencia de medias poblacionales referidas al contenido de fibra en el jugo de las marcas A y B está entre -0.4375 y 0.8375 g. Lo anterior indica que, de cada 100 muestras seleccionadas, en 95 se encuentra el parámetro $\mu_1 - \mu_2$.

También en este ejemplo se infieren las tres situaciones siguientes, considerando igualdad de varianzas poblacionales:

i) $\mu_1 - \mu_2 = 0 \ \Rightarrow\ \mu_1 = \mu_2$;

ii) $\mu_1 - \mu_2 > 0 \ \Rightarrow\ \mu_1 > \mu_2$;

iii) $\mu_1 - \mu_2 < 0 \ \Rightarrow\ \mu_1 < \mu_2$

Por otro lado, se supone que las varianzas poblacionales son distintas. En este caso se utiliza la expresión 3.11 para determinar el intervalo de confianza, pero inicialmente se calculan los grados de libertad, como se indica a continuación:

$$g = \frac{\left(\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{S}_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{\hat{S}_2^2}{n_2}\right)^2}{n_2+1}} - 2 = \frac{\left(\frac{(1.2)^2}{22} + \frac{(0.9695)^2}{25}\right)^2}{\frac{\left(\frac{(1.2)^2}{22}\right)^2}{22+1} + \frac{\left(\frac{(0.9695)^2}{25}\right)^2}{25+1}} - 2 = \frac{0.0106}{0.0001863 + 0.000054} - 2 \cong 42.11$$

En la Figura 3.15 se observa el valor aproximado de *t* obtenido al leer una tabla *t-student* con *g* = 42 grados de libertad, para un 0.975 = 97.5 % de probabilidad.

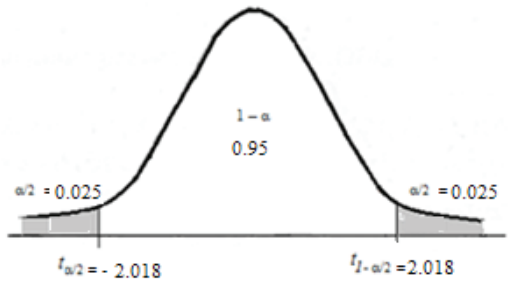

Figura 3.15 Intervalo de confianza sobre la curva t-student

Fuente: los autores con la ayuda del *software* libre R.

$$t_{1-\frac{\alpha}{2}} = t_{0.975,42} = 2.018$$

Luego, se usa la expresión 3.11

$$\mu_1 - \mu_2 \in \left((\bar{X}_1 - \bar{X}_2) - t_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}} \ , \ (\bar{X}_1 - \bar{X}_2) + t_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}} \right)$$

Al reemplazar los correspondientes valores resulta:

$$\mu_1 - \mu_2 \in \left((10-9.8) - 2.018 \sqrt{\frac{(1.2)^2}{22} + \frac{(0.9695)^2}{25}} \ , \ (10-9.8) + 2.018 \sqrt{\frac{(1.2)^2}{22} + \frac{(0.9695)^2}{25}} \right)$$

Efectuando las operaciones de tipo aritmético se obtiene:

$$\mu_1 - \mu_2 \in \left((10-9.8) - 2.018(0.3210) \ , \ (10-9.8) + 2.018(0.3210) \right)$$

Luego

$$\mu_1 - \mu_2 \in \left(0.2 - 0.6477 \ , \ 0.2 + 0.6477 \right)$$

Por lo tanto,

$$\mu_1 - \mu_2 \in \left(-0.4477 \ , \ 0.8477 \right)$$

En conclusión, con un nivel de confianza del 95 % se infiere que la diferencia de medias poblacionales referidas al contenido de fibra en el jugo de las marcas A y B está entre -0.4477 y 0.8477 g, considerando varianzas desiguales (heterocedasticidad). Lo anterior indica que, de cada 100 muestras seleccionadas, 95 tienen el parámetro $\mu_1 - \mu_2$.

Al considerar varianzas poblacionales diferentes, también se infieren las tres situaciones siguientes:

i) $\mu_1 - \mu_2 = 0 \Rightarrow \mu_1 = \mu_2;$

ii) $\mu_1 - \mu_2 > 0 \Rightarrow \mu_1 > \mu_2;$

iii) $\mu_1 - \mu_2 < 0 \Rightarrow \mu_1 < \mu_2$

3.2.5 Intervalos de confianza para estimar la varianza poblacional

En este apartado se busca obtener un intervalo de la forma *(a,b)* que contenga al parámetro σ^2 con un nivel de confianza de $(1 - \alpha)$. Se busca determinar los valores *a* y *b* tales que:

$$P(a \leq \sigma^2 \leq b) = 1 - \alpha$$

Una representación gráfica de esta situación se visualiza en la Figura 3.16. Es necesario señalar que se trata de una curva asimétrica.

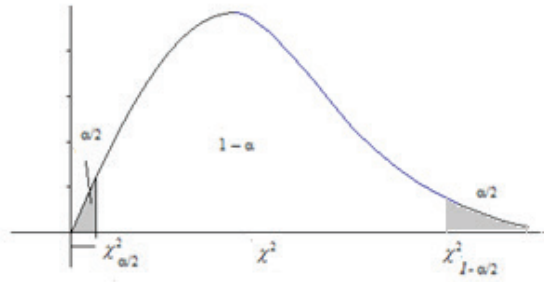

Figura 3.16 Intervalo de confianza para la varianza poblacional

Fuente: los autores con la ayuda del *software* libre R.

La siguiente variable tiene distribución chi-cuadrado con *n-1* grados de libertad:

$$\chi^2 = \frac{(n-1)\hat{S}^2}{\sigma^2}$$

Luego, en concordancia con su correspondiente distribución de probabilidad, se tiene:

$$P(\chi^2_{\frac{\alpha}{2}} \leq \chi^2 \leq \chi^2_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

En concordancia con la Figura 3.16, se escribe:

$$P(\chi^2_{\frac{\alpha}{2}} \leq \frac{(n-1)\hat{S}^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

Aplicando propiedades de las desigualdades de números reales positivos en cada uno de los miembros de la desigualdad del evento involucrado en el cálculo de la probabilidad, resulta:

$$P\left(\frac{1}{\chi^2_{\frac{\alpha}{2}}} \geq \frac{1}{\frac{(n-1)\hat{S}^2}{\sigma^2}} \geq \frac{1}{\chi^2_{1-\frac{\alpha}{2}}}\right) = 1-\alpha$$

De la anterior expresión se obtiene:

$$P\left(\frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}} \geq \sigma^2 \geq \frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}}\right) = 1-\alpha$$

La desigualdad anterior se expresa de la siguiente manera:

$$P\left(\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}}\right) = 1-\alpha$$

Por lo tanto, el intervalo de confianza para estimar la varianza poblacional está dado por la expresión 3.13.

$$\sigma^2 \in \left(\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}}, \frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}}\right) \tag{3.13}$$

Puesto que también es factible calcular la siguiente probabilidad:

$$P\left(\sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}}} \leq \sigma \leq \sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}}}\right) = 1-\alpha$$

entonces el intervalo de confianza para estimar la desviación estándar poblacional se obtiene mediante la expresión 3.14:

$$\sigma \in \left(\sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}}}, \sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}}}\right) \tag{3.14}$$

Ejemplo 3.17. El siguiente ejemplo ha sido adaptado de Freund *et al.* (2000). Una clase de motor experimental es sometida a 16 pruebas para evaluar su consumo

de combustible. En esa muestra aleatoria se obtiene una desviación estándar corregida de 2.2 litros. Bajo el supuesto de normalidad para los datos, construir un intervalo de confianza del 99 % para estimar la varianza poblacional como indicador de la verdadera variación del consumo de combustible; asimismo, determinar un intervalo de confianza para estimar la desviación estándar poblacional.

Los requerimientos son los siguientes:

$$\hat{S} = 2.2$$

$$n = 16$$

$$1 - \alpha = 0.99 \rightarrow \alpha = 0.01$$

$$\frac{\alpha}{2} = 0.005$$

$$1 - \frac{\alpha}{2} = 0.995$$

En la Figura 3.17 se observan los valores de los cuantiles obtenidos al leer una tabla *chi-cuadrado n -1*=16 – 1=15 grados de libertad, para un 0.005=0.5 % y un 0.995=99.5 % de probabilidad acumulada, respectivamente.

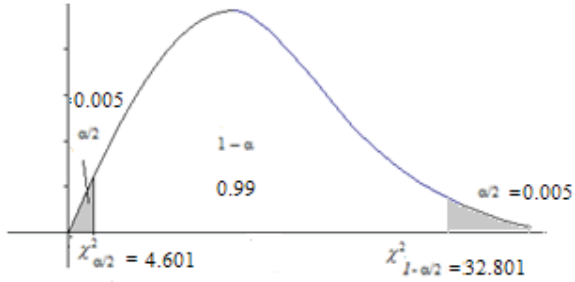

Figura 3.17 Intervalo de confianza sobre la curva chi-cuadrado

Fuente: los autores con la ayuda del *software* libre R.

$$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.995,15} = 32.801$$

$$\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.005,15} = 4.601$$

Luego, usando la expresión 3.13:

$$\sigma^2 \in \left(\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}} , \frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}} \right)$$

y reemplazando los valores, resulta:

$$\sigma^2 \in \left(\frac{(16-1)(2.2)^2}{32.801} \ , \ \frac{(16-1)(2.2)^2}{4.601} \right)$$

Efectuando las operaciones de tipo aritmético se obtiene:

$$\sigma^2 \in \left(\frac{72.6}{32.801} \ , \ \frac{72.6}{4.601} \right)$$

Por consiguiente,

$$\sigma^2 \in (2.213 \ , \ 15.779)$$

En conclusión, con un nivel de confianza del 99 % se infiere que la varianza poblacional en referencia al consumo de combustible está entre 2.213 y 15.779. Lo anterior indica que, de cada 100 muestras seleccionadas, 99 tienen el parámetro σ^2.

Ahora, para estimar la desviación estándar poblacional se usa la expresión 3.14:

$$\sigma \in \left(\sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}}} \ , \ \sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}}} \right)$$

y al sustituir los valores pertinentes se obtiene:

$$\sigma \in \left(\sqrt{\frac{(16-1)(2.2)^2}{32.801}} \ , \ \sqrt{\frac{(16-1)(2.2)^2}{4.601}} \right)$$

Así, entonces,

$$\sigma \in \left(\sqrt{2.213} \ , \ \sqrt{15.779} \right)$$

Por lo tanto,

$$\sigma \in (1.4876 \ , \ 3.9722)$$

En conclusión, con un nivel de confianza del 99 % se infiere que la desviación estándar poblacional en referencia al consumo de combustible está entre 1.4876 y 3.9722 litros. Lo anterior indica que, de cada 100 muestras seleccionadas, 99 tienen el parámetro σ.

3.2.6 Intervalos de confianza para estimar el cociente de varianzas poblacionales

En este apartado se desea obtener un intervalo de la forma *(a,b)* que contenga el parámetro $\frac{\sigma_1^2}{\sigma_2^2}$, con un nivel de confianza de $(1 - \alpha)$. Se busca determinar los valores *a* y *b* tales que:

$$P(a \leq \frac{\sigma_1^2}{\sigma_2^2} \leq b) = 1-\alpha$$

Una representación gráfica de esta situación se observa en la Figura 3.18. Es necesario señalar que se trata también de una curva asimétrica para la derecha.

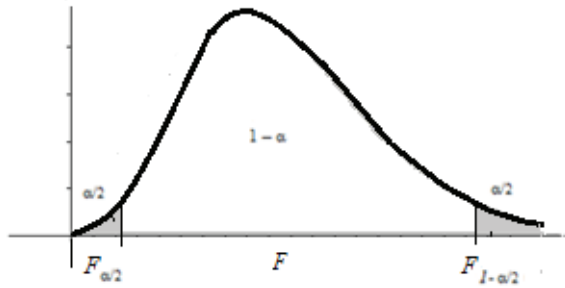

Figura 3.18 Intervalo de confianza para el cociente de varianzas poblacionales
Fuente: los autores con la ayuda del *software* libre R.

En consonancia con la expresión 2.11, la variable *F* tiene distribución de Fischer con n_1-1 grados de libertad para el numerador y n_2-1 grados de libertad para el denominador,

$$F = \frac{\hat{S}_1^2 \sigma_2^2}{\hat{S}_2^2 \sigma_1^2}$$

Luego, en concordancia con la distribución de probabilidad de *Fisher*, se tiene:

$$P(F_{\frac{\alpha}{2}} \leq F \leq F_{1-\frac{\alpha}{2}}) = 1-\alpha$$

De acuerdo con la Figura 3.18, se escribe:

$$P(F_{\frac{\alpha}{2}} \leq \frac{\hat{S}_1^2 \sigma_2^2}{\hat{S}_2^2 \sigma_1^2} \leq F_{1-\frac{\alpha}{2}}) = 1-\alpha$$

Aplicando propiedades de las desigualdades de números reales positivos en cada uno de los miembros de la desigualdad del evento al cual se le calcula la probabilidad, resulta:

$$P\left(\frac{1}{F_{\frac{\alpha}{2}}} \geq \frac{1}{\frac{\hat{S}_1^2\sigma_2^2}{\hat{S}_2^2\sigma_1^2}} \geq \frac{1}{F_{1-\frac{\alpha}{2}}}\right) = 1-\sigma$$

De la anterior expresión se obtiene:

$$P\left(\frac{\hat{S}_1^2}{\hat{S}_2^2 F_{\frac{\alpha}{2}}} \geq \frac{\sigma_1^2}{\sigma_2^2} \geq \frac{\hat{S}_1^2}{\hat{S}_2^2 F_{1-\frac{\alpha}{2}}}\right) = 1-\sigma$$

La desigualdad anterior se expresa de la siguiente forma:

$$P\left(\frac{\hat{S}_1^2}{\hat{S}_2^2 F_{1-\frac{\alpha}{2}}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{\hat{S}_1^2}{\hat{S}_2^2 F_{\frac{\alpha}{2}}}\right) = 1-\sigma$$

Por lo tanto, el intervalo de confianza para estimar el cociente de varianzas poblacionales está dado por la expresión 3.15.

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{\hat{S}_1^2}{\hat{S}_2^2 F_{1-\frac{\alpha}{2}}} \ , \ \frac{\hat{S}_1^2}{\hat{S}_2^2 F_{\frac{\alpha}{2}}}\right) \qquad (3.15)$$

Ejemplo 3.18. Se hizo un estudio para comparar los contenidos de nicotina de dos marcas de cigarrillos. En una muestra aleatoria de 10 cigarrillos de la marca A se encuentra un promedio de 3.5 mg de nicotina, con una desviación estándar corregida de 0.5 mg, mientras que en una muestra aleatoria de 8 cigarrillos de la marca B se obtiene un promedio de 2.9 mg de nicotina, con una desviación estándar de 0.7 mg; se supone que los dos conjuntos de datos provienen de muestras independientes seleccionadas de poblaciones normales. Construir un intervalo de confianza del 98 % para estimar el cociente de varianzas poblacionales.

En este caso se tiene:

Marca A	Marca B
$n_1 = 10$	$n_2 = 8$
$\overline{X}_1 = 3.5$	$\overline{X}_2 = 2.9$
$\hat{S}_1 = 0.5$	$\hat{S}_2 = 0.7483$

El valor de la desviación estándar corregida para la marca B se obtiene de la siguiente manera:

$$\hat{S}_2^2 = \frac{n_2}{n_2 - 1} S_2^2 = \frac{8}{7}(0.7)^2 = 0.56 \rightarrow \hat{S}_2 = 0.7483$$

Además,

$$1 - \alpha = 0.98 \rightarrow \alpha = 0.02$$

$$\frac{\alpha}{2} = 0.01$$

En la Figura 3.19 se observan los valores de los cuantiles obtenidos al leer una tabla *F* de Fisher con n_1-*1*=10-1=9 grados de libertad para el numerador, y n_2-*1*=8-1=7 grados de libertad para el denominador; estos corresponden a un 0.01=1 % y 0.99=99 % de probabilidad.

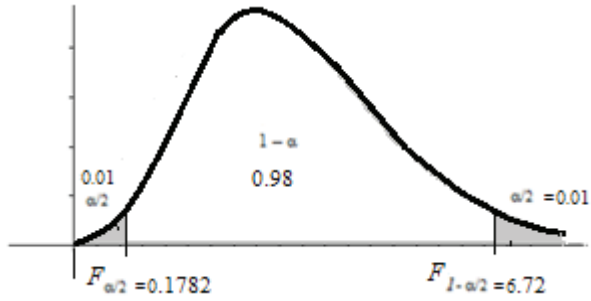

Figura 3.19 Intervalo de confianza sobre la curva de Fisher

Fuente: los autores con la ayuda del *software* libre R.

$$F_{1-\frac{\alpha}{2}} = F_{0.99,9,7} = 6.72$$

$$F_{\frac{\alpha}{2}} = F_{0.01,9,7} = \frac{1}{F_{0.99,7,9}} = \frac{1}{5.61} = 0.1782$$

Luego, usando la expresión 3.15:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{\hat{S}_1^2}{\hat{S}_2^2 F_{1-\frac{\alpha}{2}}} \ , \ \frac{\hat{S}_1^2}{\hat{S}_2^2 F_{\frac{\alpha}{2}}} \right)$$

al reemplazar los respectivos valores resulta:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{(0.5)^2}{(0.7483)^2(6.72)} \ , \ \frac{(0.5)^2}{(0.7483)^2(0.1782)} \right)$$

Efectuando las operaciones de tipo aritmético se obtiene:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{0.25}{3.7628} \ , \ \frac{0.25}{0.09978} \right)$$

Por consiguiente,

$$\frac{\sigma_1^2}{\sigma_2^2} \in (0.0664 \ , \ 2.5055)$$

En conclusión, con un nivel de confianza del 98 % se infiere que el cociente de las varianzas poblacionales en referencia a la cantidad de nicotina de las dos marcas de cigarrillos está entre 0.0664 y 2.5055. Lo anterior indica que, de cada 100 muestras seleccionadas, en 98 se encuentra el parámetro $\frac{\sigma_1^2}{\sigma_2^2}$.

Actividades para el estudio independiente Capítulo 3

3.1 Para la variable aleatoria *X,* con distribución exponencial, determinar el estimador máximo verosímil. Recuerde que la función de densidad de probabilidad está dada por:

$$f(x,\lambda)=\begin{cases}\lambda e^{-\lambda x} & \text{si } x\geq 0 \quad \text{con } \lambda>0\\ 0 & \text{si } x<0\end{cases}$$

3.2 En la empresa azucarera AA, la cantidad de azúcar depositada por la máquina M en cada uno de los paquetes se distribuye normalmente con desviación estándar de 2 g; de un lote de 500 paquetes se toma una muestra aleatoria de 25 paquetes (bolsas) empacados por tal máquina, y se encuentra un contenido promedio de 2500 g. Construir un intervalo de confianza del 98 % para estimar la verdadera media de empacado en las bolsas en el lote que se produzca a través de la máquina M.

3.3 De un lote de 200 unidades del producto LM se ha seleccionado una muestra aleatoria de 10 unidades, obteniéndose un peso neto promedio de 1000 g, con una desviación estándar corregida de 5 g; si se asume que los datos provienen de una distribución normal, determinar un intervalo de confianza del 95 % para estimar el verdadero peso promedio de las unidades del producto en la población.

3.4 En el departamento de Boyacá, Colombia, se tomó una muestra aleatoria de 500 ciudadanos y se les preguntó si pertenecen o no a la población económicamente activa de este departamento; 350 de los encuestados respondieron que sí pertenecen a esta población. Construir un intervalo de confianza del 99 % para estimar la verdadera proporción de ciudadanos que pertenecen a la población económicamente activa de este departamento.

3.5 En un centro de distribución de computadores se ofrecen computadores de dos marcas diferentes en un periodo de tiempo específico; se selecciona aleatoriamente un mes y se encuentra que se venden 350 computadores, de un total de 500, de la marca A, y 333, de un total de 450, de la marca B. Determinar un intervalo de confianza del 98 % para estimar la diferencia entre las verdaderas proporciones de las marcas A y B de computadores que se venden en todo el mercado en ese mes.

3.6 Se analiza el contenido de oro presente en una aleación; en una muestra especial de 40 circuitos integrados se encontró un contenido medio de 5.8 u.i

de oro, con una desviación estándar de 0.6 u.i de oro; asimismo, se inspecciona el contenido de oro en otra muestra aleatoria de 50 circuitos integrados corrientes, detectándose un contenido promedio de 5 u.i, con una desviación estándar de 0.8 u.i; se supone que las muestras provienen de poblaciones normales. Construir un intervalo de confianza del 95 % para estimar la diferencia de contenidos medios de oro de la primera clase de circuito con respecto a la segunda.

3.7 El siguiente ejemplo ha sido adaptado de Canavos (1988). Un determinado procedimiento produce cierta clase de cojinetes de bola cuyo diámetro interior es de 5 cm; se selecciona una muestra aleatoria de 12 de esos cojinetes, y al medir sus diámetros internos se obtiene una desviación estándar corregida 0.03 centímetros. Bajo el supuesto de normalidad para los datos, construir un intervalo de confianza del 99 % para estimar la varianza poblacional; asimismo, determinar un intervalo de confianza para estimar la desviación estándar poblacional.

3.8. Se tiene la creencia de que los egresados de la titulación de Administración de Empresas obtienen un salario promedio mayor que el de los egresados de la titulación de Economía; además, se quiere saber si la variación de sus correspondientes salarios difiere. Para comprobarlo se ha tomado una muestra aleatoria de 10 administradores, obteniéndose una media muestral de 2 600 000 pesos por mes, con una varianza corregida de 1 200 000 pesos, mientras que en una muestra aleatoria de 13 economistas se ha obtenido un promedio de 2 400 000 pesos por mes, con una varianza de 1 300 000; se supone que los dos conjuntos de datos provienen de muestras independientes seleccionadas de poblaciones normales. Construir un intervalo de confianza del 98 % para estimar el cociente de varianzas poblacionales.

Ejercicios para el capítulo 3

3.1 Indagar sobre la forma como se construye el intervalo de confianza para muestras pareadas, también denominadas muestras relacionadas o emparejadas; además, proporcionar un ejemplo de aplicación.

3.2 Se supone que la cantidad de cemento que una máquina empaca en cada bulto es una variable aleatoria con desviación típica poblacional de 1 kg. Se toma una muestra aleatoria de tamaño 20 y se obtiene una media de 50 kg. Obtener un intervalo de confianza del 95 % para estimar la media poblacional.

3.3 Un jugador de fútbol anota 120 goles en 500 lanzamientos (cobros) desde el punto penal, en los entrenamientos. Construir un intervalo de confianza del 99 % para estimar la verdadera proporción de la efectividad del jugador.

3.4 Una determinada clase de estufa de gas se somete a 11 pruebas para evaluar su consumo. En esa muestra se obtiene una desviación estándar corregida de 1.8 litros. Construir un intervalo de confianza del 95 % para estimar la varianza poblacional como indicador de la verdadera variación del consumo de gas; asimismo, construir un intervalo de confianza para la desviación estándar poblacional.

3.5 En un centro de distribución de monitores para computador de mesa se distribuyen monitores de dos marcas diferentes en un periodo de tiempo determinado; en una semana se venden 200 monitores, de un total de 350, de la marca A, y 180, de un total de 300, de la marca B. Hallar un intervalo de confianza del 96 % para estimar la verdadera diferencia entre las proporciones de las marcas A y B que se venden en todo el mercado.

3.6 En una muestra de 20 unidades de un producto se obtuvo un peso neto promedio de 250 g, con una desviación estándar corregida de 5. Encontrar un intervalo de confianza del 98 % para estimar el verdadero peso promedio de las unidades del producto.

4

Prueba de hipótesis

En el presente capítulo se desarrollan procesos de inferencia estadística centrados en la prueba de hipótesis, los cuales involucran uno o más parámetros desconocidos; se inicia con algunos conceptos básicos asociados al tema de hipótesis y luego se plantea un algoritmo que posibilita llevar a cabo diversos procedimientos de inferencia estadística básica, entre ellos, la prueba de hipótesis para la media y la proporción poblacionales, la diferencia de medias y de proporciones poblacionales, la varianza y el cociente de varianzas. Al final del capítulo se indica la forma de calcular algunos tamaños de la muestra para casos específicos.

4.1. Conceptos básicos sobre prueba de hipótesis

Algunos de los aspectos teóricos relacionados con el tema de prueba de hipótesis son asumidos o adaptados de los conceptos expuestos al respecto por diversos autores, entre ellos: Bickel & Doksum (1977), Canavos (1988), Devore (2008), Freund y Miller (2000), Gutiérrez *et al.* (2008), Hettmansperger (1984), Kandu *et al.* (2008), Lindgren (1993), Macchi *et al.* (2014), Mayorga (2003) y Walpole, Myers, Myers & Ye (2007). Enseguida se presentan algunos conceptos necesarios para desarrollar procesos de prueba de hipótesis.

4.1.1 Conceptos sobre hipótesis

Una hipótesis es una proposición, afirmación o conjetura sobre algo; puede ser verdadera o falsa, y se debe probar.

Una hipótesis estadística es una afirmación acerca de los parámetros poblacionales, para determinar si el parámetro ha cambiado o se mantiene en un valor fijo.

4.1.2 Clases de hipótesis

Los procesos de inferencia estadística suelen involucrar dos clases de hipótesis, denominadas hipótesis nula e hipótesis alternativa; generalmente, la primera se simboliza con H_0, y la segunda, con H_1.

La hipótesis nula (Ho) se formula como una afirmación que indica que un determinado parámetro se mantiene en un valor; esta hipótesis es aquella que se acepta o se rechaza (González, 2003; Krueger, 2001).

La hipótesis alternativa (H_1) se plantea en términos de que el parámetro ha cambiado (aumentado o disminuido); esta es la hipótesis de investigación, la cual se prueba utilizando los datos de una muestra aleatoria.

Ejemplo 4.1. Enseguida se plantean tres parejas de hipótesis relacionadas con el parámetro media poblacional asociado con el ingreso mensual (en pesos) de los trabajadores de la empresa M&T; sin embargo, en un proceso de inferencia estadística básica solamente se ha de plantear una pareja de hipótesis.

i) Ho: $\mu = 700\,000$

H_1: $\mu > 700\,000$

ii) Ho: $\mu = 700\,000$

H_1: $\mu < 700\,000$

iii) Ho: $\mu = 700\,000$

H_1: $\mu \neq 700\,000$

La primera pareja establece como hipótesis nula que el ingreso mensual promedio (en pesos) de los trabajadores de la empresa M&T es de 700 000; en cambio, la hipótesis alternativa indica que el ingreso mensual promedio (en pesos) de los trabajadores de la empresa M&T es mayor a 700 000. De modo similar, han de interpretarse la segunda y la tercera pareja de hipótesis, donde la alternativa indica que el promedio es inferior o diferente, respectivamente.

Ejemplo 4.2. A continuación se plantean tres parejas de hipótesis asociadas con el porcentaje poblacional de fumadores en la Universidad TRT; no obstante, en un proceso de inferencia estadística básica solo una de las tres parejas de hipótesis se ha de plantear.

i) Ho: $p = 24\ \%$
$H_1: p > 24\ \%$

ii) Ho: $p = 24\ \%$
$H_1: p < 24\ \%$

iii) Ho: $p = 24\ \%$
$H_1: p \neq 24\ \%$

La segunda pareja establece como hipótesis nula que el porcentaje poblacional de fumadores en la Universidad TRT es del 24 %, contra la hipótesis alternativa de que el porcentaje poblacional de fumadores en la Universidad TRT es inferior al 24 %. De forma similar, han de interpretarse la primera y la tercera pareja de hipótesis, considerando el porcentaje mayor o diferente al 24 % en la hipótesis alternativa.

4.1.3 Tipos de errores

En general, se suelen cometer dos tipos de errores al aceptar o rechazar la hipótesis nula: el error tipo I y el error tipo II.

El *error tipo I* consiste en rechazar la hipótesis nula Ho, dado que es verdadera (cierta). La probabilidad de rechazar la hipótesis nula, siendo verdadera, se denomina nivel de significancia de la prueba y se denota con α; de aquí se desprende que $1 - \alpha$ sea la probabilidad de no rechazar (aceptar) Ho, dado que esta es verdadera (Krueger, 2001).

El *error tipo II* consiste en aceptar la hipótesis nula Ho, dado que es falsa. La probabilidad de aceptar la hipótesis nula, siendo falsa, se denota con β; de aquí se sigue que $1 - \beta$ sea la probabilidad de rechazar Ho dado que esta es falsa (Krueger, 2001).

4.1.4 Algoritmo para desarrollar un proceso de prueba de hipótesis

1. Plantear las hipótesis Ho y H_1
2. Establecer el nivel de significación $\alpha \leq 0.05$
3. Determinar la dirección de la prueba: unilateral izquierda, unilateral derecha o bilateral
4. Determinar la estadística de prueba: "Z", "t", "X^2", "F" y calcularla

5. Comparar el valor de la estadística de prueba con el valor en la distribución teórica
6. Tomar una decisión: aceptar Ho o, en caso contrario, rechazar Ho
7. Escribir una conclusión

El anterior algoritmo guarda una estrecha relación con los pasos del método científico: observación (tomar muestras), planteamiento de hipótesis, comprobación (prueba) y conclusión.

Ejemplo 4.3. De forma general, si θ es el parámetro involucrado en una prueba de hipótesis simples, y θ_0 es un valor particular de este parámetro, entonces se ha de plantear una de las tres parejas de hipótesis:

i) Ho: $\theta = \theta_0$
H_1: $\theta > \theta_0$

ii) Ho: $\theta = \theta_0$
H_1: $\theta < \theta_0$

iii) Ho: $\theta = \theta_0$
H_1: $\theta \neq \theta_0$

En este contexto, las hipótesis contempladas en *i)* generan una prueba unilateral derecha; en esta, el nivel de significancia α ha de ubicarse en la cola derecha de la curva correspondiente a la función de densidad de probabilidad (modelo teórico para los datos) y genera una región de rechazo de la hipótesis nula, simbolizada con RH_0 (Ares, 1999), y una región de aceptación de la hipótesis nula, denotada con AH_0 (ver Figura 4.1).

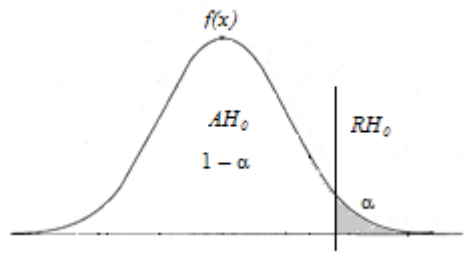

Figura 4.1 Prueba unilateral derecha
Fuente: los autores con la ayuda del *software* libre R.

Las hipótesis consideradas en *ii)* generan una prueba unilateral izquierda; ahora, el nivel de significancia α ha de ubicarse en la cola izquierda y genera una región de rechazo de la hipótesis nula (RH_0) y una región de aceptación de la hipótesis nula (AH_0) (ver Figura 4.2).

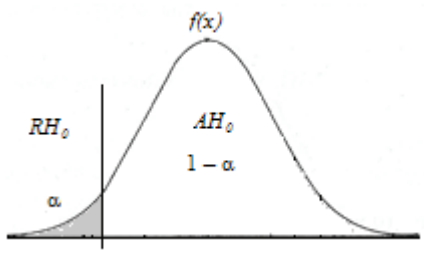

Figura 4.2 Prueba unilateral izquierda

Fuente: los autores con la ayuda del *software* libre R.

Las hipótesis consideradas en *iii)* generan una prueba bilateral; ahora, la mitad del nivel de significancia α se ubicará en la cola izquierda, y la otra mitad, en la cola derecha; esto genera dos regiones de rechazo de la hipótesis nula (RH_0) y una región de aceptación de la hipótesis nula (AH_0) (ver Figura 4.3).

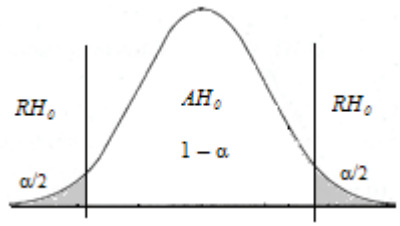

Figura 4.3 Prueba bilateral

Fuente: los autores con la ayuda del *software* libre R.

4.2 Prueba de hipótesis sobre la media poblacional

El proceso de inferencia estadística para la prueba de hipótesis asociadas con la media poblacional involucra los siguientes pasos (algoritmo):

1. Planteamiento de hipótesis

i) Ho: $\mu = \mu_0$

H_1: $\mu > \mu_0$

ii) Ho: $\mu = \mu_0$

H_1: $\mu < \mu_0$

iii) Ho: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$

2. Fijación del nivel de significancia α con un valor inferior o igual al 5 %

3. Dirección de la prueba en concordancia con una de las parejas de hipótesis: unilateral derecha, unilateral izquierda o bilateral

4. Estadística de prueba. Se presentan cuatro casos, a saber:

Caso 1. Si se conoce la desviación estándar poblacional, bajo la hipótesis nula, se usa:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Caso 2. Si no se conoce la desviación estándar poblacional, para muestras inferiores a 30 se utiliza una estadística de prueba asociada con la distribución *t-student* con *n -1* grados de libertad, bajo la hipótesis nula, resulta:

$$t = \frac{\bar{X} - \mu_0}{\frac{\hat{S}}{\sqrt{n}}}$$

Caso 3. Si se conoce la desviación estándar poblacional y se muestrea de una población finita de tamaño *N*, entonces, bajo la hipótesis nula, se usa la estadística de prueba siguiente:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)}$$

Caso 4. Si no se conoce la desviación estándar poblacional y se muestrea de una población normal finita de tamaño *N*, entonces, bajo la hipótesis nula, se usa la estadística de prueba siguiente:

$$t = \frac{\bar{X} - \mu_0}{\frac{\hat{S}}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)}$$

5. Comparar el valor de la estadística de prueba con el valor en la distribución teórica
6. Tomar una decisión: aceptar Ho si la estadística de prueba cae en la región *AHo* o, en caso contrario, rechazar Ho
7. Conclusión

Ejemplo 4.4. Un productor de computadores afirma que sus teclados duran 20 000 horas en promedio. Un distribuidor potencial de teclados quiere verificar la afirmación del productor; para esto, somete 100 teclados a prueba y encuentra una duración media de 19 320 horas; si la experiencia pasada indica que los teclados producidos tienen una desviación estándar poblacional de 2000 horas, probar la afirmación del productor usando un nivel de significación del 2.5 %.

1. Planteamiento de hipótesis

Ho: $\mu = 20\,000$
H_1: $\mu < 20\,000$

2. Nivel de significación $\alpha = 0.025$
3. La dirección de la prueba y la región crítica se observan en la Figura 4.4.

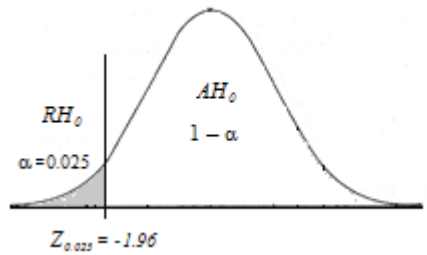

Figura 4.4 Prueba unilateral izquierda para la media poblacional
Fuente: los autores con la ayuda del *software* libre R.

4. Estadística de prueba, se utiliza la siguiente información:

$$\bar{X} = 19320 \qquad \sigma = 2000 \qquad n = 100$$

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{19320 - 20000}{\frac{2000}{\sqrt{100}}} = \frac{-680}{200} = -3.4$$

5. Como la estadística de prueba se ubica en la región *Rho,* dado que el valor -3.4 es inferior al valor teórico -1.96, entonces se rechaza la hipótesis Ho: μ=20 000; es decir, no hay evidencias suficientes para aceptarla. De otro modo, el valor-P correspondiente a la estadística Z=-3.4 de prueba es 0.0003; este valor es menor que el nivel de significancia del 2.5 % = 0.025, esto también indica que se rechaza la hipótesis nula.

6. Decisión: rechazar Ho: μ= 20 000

7. Finalmente, con un nivel de significancia del 2.5 % (confiabilidad del 97.5%) se concluye que la duración promedio de los teclados no es 20 000 horas, es inferior a este valor; luego hay evidencias suficientes para rechazar la afirmación del productor. En consecuencia, el distribuidor tiene la razón.

Ejemplo 4.5. Un distribuidor de artículos eléctricos afirma que las lámparas que expende duran, en promedio, 15 000 horas. Un usuario difiere de esa afirmación y decide seleccionar una muestra aleatoria de 21 lámparas, y encuentra una duración media de 14 800 horas, con una desviación estándar de 1500 horas; probar la afirmación del distribuidor usando un nivel de significación del 5 %.

1. Planteamiento de hipótesis

Ho: μ = 15 000

H_1: $\mu \neq$ 15 000

2. Nivel de significación α = 0.05, $\frac{\alpha}{2} = 0.025$

3. Se trata de una prueba bilateral, la región crítica se observa en la Figura 4.5. En esa se han de ubicar los valores:

$$t_{\frac{\alpha}{2}} = t_{0.025,20} = -2.086 \qquad t_{1-\frac{\alpha}{2}} = t_{0.975,20} = 2.086$$

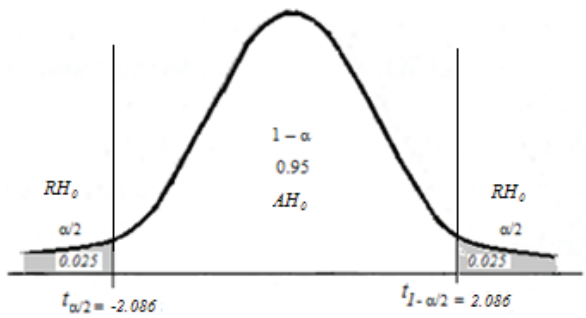

Figura 4.5 Prueba bilateral para la media poblacional

Fuente: los autores con la ayuda del *software* libre R.

4. Estadística de prueba, se usa la siguiente información:

$$\bar{X} = 14\,800 \qquad S = 1500 \qquad n = 21$$

Además, se requiere determinar la desviación estándar corregida a fin de aplicar una estadística asociada con una *t-student*:

$$\hat{S}^2 = \frac{n}{n-1} S^2 = \frac{21}{20}(1500)^2) = 2362500 \;\rightarrow\; \hat{S} = \sqrt{2362500} = 1537.04$$

Luego

$$t = \frac{\bar{X} - \mu_0}{\frac{\hat{S}}{\sqrt{n}}} = \frac{14800 - 15000}{\frac{1537.04}{\sqrt{21}}} = \frac{-200}{335.41} = -0.5962$$

5. Como la estadística de prueba se ubica en la región *AHo,* dado que el valor -0.5962 se encuentra entre los valores -2.086 y 2.086 de la distribución teórica, entonces se acepta la hipótesis Ho:μ=15 000; es decir, no hay evidencias suficientes para rechazarla.

6. Decisión: aceptar Ho: μ= 15 000

7. Finalmente, con un nivel de significancia del 5 % se concluye que la duración promedio de las lámparas es igual a 15 000 horas, luego hay evidencias suficientes para aceptar la afirmación hecha por el productor, quien sí tiene la razón.

Ejemplo 4.6. Tomando como referencia un lote conformado por 1000 latas de atún que presenta una desviación estándar poblacional de 5 g, un funcionario de control de pesas y medidas afirma que esas latas no presentan un peso promedio de 184 g, como se anuncia en la etiqueta del producto; por el contrario, el productor sostiene que efectivamente las latas que produce sí tienen ese peso promedio. El funcionario toma una muestra aleatoria de 25 latas y encuentra un peso promedio de 182 g. Usar un nivel de significancia del 4 % para aceptar o rechazar la afirmación del productor.

1. Planteamiento de hipótesis

Ho:μ= 184

H_1: μ ≠ 184

2. Nivel de significación α = 0.04; $\frac{\alpha}{2} = 0.02$

3. En este ejemplo se tiene una prueba bilateral y la región crítica se observa en la Figura 4.6.

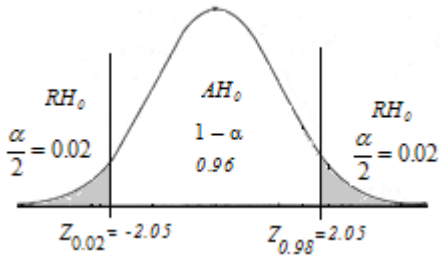

Figura 4.6 Prueba bilateral para la media poblacional

Fuente: los autores con la ayuda del *software* libre R.

4. Para la estadística de prueba, se utiliza la siguiente información:

$$\bar{X} = 182 \quad \sigma = 5 \quad n = 25 \quad N = 1000$$

La estadística de prueba es:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} = \frac{182 - 184}{\frac{5}{\sqrt{25}}\sqrt{\frac{1000-25}{1000-1}}} = \frac{-2}{\sqrt{\frac{975}{999}}} = \frac{-2}{\sqrt{0.9759}} = \frac{-2}{0.9878} = -2.024$$

5. Como la estadística de prueba se ubica en la región AH_0, entonces, se acepta la hipótesis Ho:μ=184; es decir, no hay evidencias suficientes para rechazarla.

6. Decisión: aceptar Ho: μ=184

7. Finalmente, con un nivel de significancia del 4 % se concluye que el peso promedio de las latas de atún en la población (lote) es de 184 g; luego el productor tiene la razón.

4.3 Prueba de hipótesis para la proporción poblacional

El proceso de inferencia estadística para realizar una prueba de hipótesis asociadas con la proporción o porcentaje poblacional incluye un algoritmo con los siguientes pasos:

1. Planteamiento de hipótesis

i) Ho: $p = p_0$

H_1: $p > p_0$

ii) Ho: $p = p_0$

H_1: $p < p_0$

iii) Ho: $p = p_0$

H_1: $p \neq p_0$

2. Fijación del nivel de significancia α

3. Dirección de la prueba. Se especifica en correspondencia con cada una de las parejas de hipótesis: para la pareja *i)* se usa una prueba unilateral derecha, como en la Figura 4.1; el caso *ii)* corresponde a una prueba unilateral izquierda, como se observa en la Figura 4.2, y para la pareja *iii)* se utiliza una prueba bilateral, como se ilustra en la Figura 4.3.

4. Estadística de prueba

Se presentan dos casos, a saber:

Caso 1. Si se tiene una poblacional infinita y un tamaño de muestra *n* grande, entonces, bajo la hipótesis nula, se usa la siguiente estadística de prueba:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

En este caso, se ha de tener presente que $q_0 = 1 - p_0$.

Caso 2. Si se tiene una poblacional finita de tamaño *N* y un tamaño de muestra *n* grande, entonces, bajo la hipótesis nula, se utiliza la estadística de prueba que se indica a continuación:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}\sqrt{\frac{N-n}{N-1}}}$$

5. Comparación del valor de la estadística de prueba con el valor en la distribución teórica

6. Decisión: aceptar Ho, en caso contrario, rechazar Ho.

7. Conclusión

Ejemplo 4.7. Un distribuidor de gas en el departamento de Boyacá, Colombia, afirma que el 40 % de los hogares boyacenses adquieren el gas domiciliario en su compañía; un competidor duda de esa afirmación y selecciona una muestra aleatoria de 400 hogares boyacenses, y encuentra que 180, efectivamente, le compran el gas al citado distribuidor; probar la afirmación hecha por el distribuidor de gas usando un nivel de significancia del 3 %.

1. Planteamiento de hipótesis

Ho: $p = 0.4$

H_1: $p < 0.4$

2. Nivel de significancia $\alpha = 0.03$

3. Dirección de la prueba. Se trata de una prueba unilateral izquierda, la región crítica se observa en la Figura 4.7

4. Estadística de prueba
Se tienen los siguientes datos provenientes de la muestra:

$$n = 400 \quad x = 180 \quad \hat{p} = \frac{x}{n} = \frac{180}{400} = 0.45$$

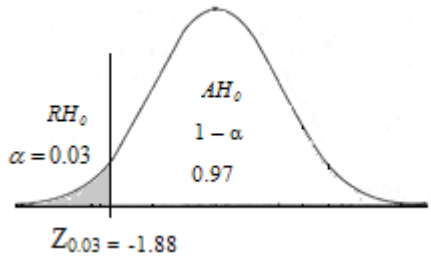

Figura 4.7 Prueba unilateral izquierda para la proporción poblacional
Fuente: los autores con la ayuda del *software* libre R.

La estadística de prueba por utilizar es:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.45 - 0.4}{\sqrt{\frac{(0.4)(0.6)}{400}}} = \frac{0.05}{\sqrt{0.0006}} = \frac{0.05}{0.024494} \cong 2.041$$

En este caso, se ha tenido en cuenta que $q_0 = 1 - p_0 = 1 - 0.4 = 0.6$.

5. El valor de la estadística de prueba, 2.041, cae en la región AH_0, de aceptación de la hipótesis nula H_0, dado que es mayor que el valor de Z=-1.88 en la distribución teórica normal estándar.

6. Decisión: aceptar Ho: $p = 0.4$

7. Finalmente, con un nivel de significancia del 3 % se concluye que la afirmación hecha por el distribuidor, de que el 40 % de los hogares boyacenses le compra el gas domiciliario a su compañía es cierta; no hay evidencias suficientes para rechazar tal afirmación.

Ejemplo 4.8. Un importador de peras considera que en un lote conformado por 5000, que acaba de adquirir, solamente el 90 % ha llegado en condiciones óptimas y la fruta no se perderá; el vendedor quiere probarle al importador que un porcentaje mayor de peras ha llegado en óptimas condiciones; para esto selecciona una muestra aleatoria de 200 peras y encuentra que 184 están en óptimas condiciones. Probar la hipótesis del importador usando un nivel de significancia del 5 %.

1. Planteamiento de hipótesis

Ho: $p = 0.9$

H_1: $p > 0.9$

2. Nivel de significancia $\alpha = 0.05$

3. Dirección de la prueba. Se trata de una prueba unilateral derecha; la región crítica se observa en la Figura 4.8.

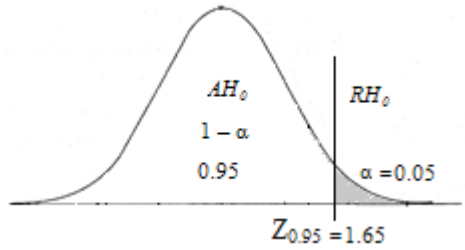

Figura 4.8 Prueba unilateral derecha para la proporción poblacional

Fuente: los autores con la ayuda del *software* libre R.

4. Estadística de prueba

Se tienen los siguientes datos provenientes de la muestra:

$$n = 200 \quad x = 184 \quad \hat{p} = \frac{x}{n} = \frac{184}{200} = 0.92$$

La estadística de prueba por utilizar para un tamaño de la población N=5000 es:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}\sqrt{\frac{N-n}{N-1}}} = \frac{0.92 - 0.9}{\sqrt{\frac{(0.9)(0.1)}{200}}\sqrt{\frac{5000-200}{5000-1}}} = \frac{0.02}{(0,02121)(0.97989)} = \frac{0.02}{0.02078} \cong 0.9624$$

En este caso, se ha tenido en cuenta que $q_0 = 1 - p_0 = 1 - 0.9 = 0.1$.

5. El valor de la estadística de prueba, 0.9624, cae en la región AH_0 de aceptación de la hipótesis nula H_0, puesto que es menor que el valor de Z=1.65 en la distribución teórica normal estándar.

6. Decisión: aceptar Ho: $p = 0.9$

7. En consecuencia, con un nivel de significancia del 5 % se concluye que la afirmación hecha por el importador de que solamente el 90 % de las peras del lote que acababa de adquirir han llegado en condiciones óptimas es cierta; no hay evidencias suficientes para rechazar tal afirmación. El importador tiene la razón.

4.4 Prueba de hipótesis para la diferencia de proporciones poblacionales

Un proceso de inferencia estadística para efectuar una prueba de hipótesis relacionada con la diferencia de proporciones poblacional se desarrolla a través de un algoritmo con los siguientes pasos:

1. Planteamiento de hipótesis

i) Ho: $p_1 - p_2 = 0$

H_1: $p_1 - p_2 > 0$

ii) Ho: $p_1 - p_2 = 0$

H_1: $p_1 - p_2 < 0$

iii) Ho: $p_1 - p_2 = 0$

H_1: $p_1 - p_2 \neq 0$

Estas tres parejas de hipótesis también se escriben de la siguiente manera:

i) Ho: $p_1 = p_2$

H_1: $p_1 > p_2$

ii) Ho: $p_1 = p_2$

H_1: $p_1 < p_2$

iii) Ho: $p_1 = p_2$

H_1: $p_1 \neq p_2$

2. Fijación del nivel de significancia α

3. Dirección de la prueba. Se determina en concordancia con cada una de las parejas de hipótesis: para la pareja *i)* se usa una prueba unilateral derecha, como en la Figura 4.1; el caso *ii)* corresponde a una prueba unilateral izquierda, como se observa en la Figura 4.2, y para la pareja *iii)* se utiliza una prueba bilateral, como se ilustra en la Figura 4.3.

4. Estadística de prueba

De la expresión 2.6 se tiene que:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

La anterior expresión, bajo la hipótesis nula H_0: $p_1 = p_2$, se simplifica y escribe de la siguiente forma:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_1 q_1 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Sin embargo, es imposible calcular la cantidad $p_1 q_1$, debido a que los parámetros p_1 y q_1 son desconocidos; con el propósito de obtener un valor para la estadística de prueba Z se utiliza la siguiente estimación:

$$p = p_1 = p_2$$

Donde el valor de *p* se obtiene de la siguiente manera:

$$p = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

En este contexto, la estadística de prueba por utilizar es:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Con $q = 1 - p$

5. Comparación del valor de la estadística de prueba con el valor de la distribución teórica; esta corresponde a una distribución normal estándar.

6. Decisión: aceptar Ho o, en caso contrario, rechazar Ho.

7. Conclusión

Ejemplo 4.9. El gobierno nacional afirma que el porcentaje de desempleo en la ciudad de Medellín (Antioquia) y en la ciudad de Bogotá es igual; un periodista difiere de esa afirmación y decide tomar una muestra aleatoria de 1500 ciudadanos en Medellín, encontrando que el 11 % de ellos está desempleado, y selecciona otra muestra de 2000 ciudadanos en Bogotá, que revela un 9 % de desempleo; utilizar un nivel de significancia del 3 % para aceptar o rechazar la afirmación del gobierno.

1. Planteamiento de hipótesis

Ho: $p_1 = p_2$

H_1: $p_1 \neq p_2$

2. Nivel de significancia $\alpha = 0.03$; $\frac{\alpha}{2} = 0.015$

3. Dirección de la prueba. Se trata de una prueba bilateral; la región crítica se observa en la Figura 4.9, donde se presenta una curva normal estándar.

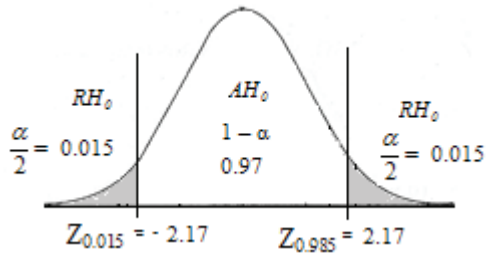

Figura 4.9 Prueba bilateral para la diferencia de proporciones poblacionales

Fuente: los autores con la ayuda del *software* libre R.

4. Estadística de prueba

En este ejemplo se tiene que:

Medellín	Bogotá
$n_1 = 1500$	$n_2 = 2000$
$\hat{p}_1 = 0.11$	$\hat{p}_2 = 0.09$
$\hat{q}_1 = 1 - 0.11 = 0.89$	$\hat{q}_2 = 1 - 0.09 = 0.91$

La estadística de prueba por utilizar es:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

En primera instancia, se realiza el cálculo siguiente:

$$p = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{1500(0.11) + 2000(0.09)}{1500 + 2000} = \frac{345}{3500} = 0.09857$$

Luego, se determina el valor $q = 1 - p = 1 - 0.09857 = 0.90143$

Reemplazando los valores en la estadística de prueba, resulta:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.11 - 0.09}{\sqrt{(0.09857)(0.90143)\left(\frac{1}{1500} + \frac{1}{2000}\right)}} = \frac{0.02}{\sqrt{0.0001036}} = \frac{0.02}{0.01017} = 1.966$$

5. El valor de la estadística de prueba, 1.966, cae en la región AH_0, de aceptación de la hipótesis nula H_0, puesto que se encuentra entre Z=-2.17 y Z=2.17 de la distribución teórica normal estándar.

6. Decisión: aceptar Ho: $p_1 = p_2$

7. Luego, con un nivel de significancia del 3 %, se concluye que la afirmación hecha por el gobierno de que el porcentaje de desempleo es igual en las ciudades de Medellín y Bogotá es cierta; no hay evidencias suficientes para rechazar tal afirmación. En consecuencia, el gobierno tiene la razón.

4.5 Prueba de hipótesis para la diferencia de medias poblacionales

En esta sección se indica el proceso de inferencia estadística para la prueba de hipótesis asociadas con la diferencia de medias poblacionales, el cual ha de desarrollarse a través de los siguientes pasos:

1. Planteamiento de hipótesis

i) Ho: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 > 0$

ii) Ho: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 < 0$

iii) Ho: $\mu_1 - \mu_2 = 0$
H_1: $\mu_1 - \mu_2 \neq 0$

Estas tres parejas de hipótesis se expresan de forma equivalente, así:

i) Ho: $\mu_1 = \mu_2$
H_1: $\mu_1 > \mu_2$

ii) Ho: $\mu_1 = \mu_2$
H_1: $\mu_1 < \mu_2$

iii) Ho: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$

2. Fijación del nivel de significancia α

3. Dirección de la prueba. Se determina en concordancia con cada una de las parejas de hipótesis: para la pareja *i)* se usa una prueba unilateral derecha, como en la Figura 4.1; el caso *ii)* corresponde a una prueba unilateral izquierda, como

se puede observar en la Figura 4.2, y para la pareja *iii)* se utiliza una prueba bilateral, como se ilustra en la Figura 4.3. Ahora, la curva trazada corresponde a una distribución *t-student* o a una distribución normal estándar.

4. Estadística de prueba. Se presentan cuatro casos, a saber:

Caso 1. Si se conocen las desviaciones estándar poblacionales, en concordancia con la expresión 2.7 se ha de utilizar:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Sin embargo, bajo la hipótesis nula H_0: $\mu_1 = \mu_2$ la anterior expresión se simplifica y reduce a la siguiente estadística de prueba:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Caso 2. Si las desviaciones estándar poblacionales son desconocidas, pero se suponen iguales, para muestras inferiores a 30 se utiliza una estadística de prueba asociada con la distribución t-student presentada a través de la expresión 2.8:

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Nuevamente, bajo la hipótesis nula H_0: $\mu_1 = \mu_2$ la anterior expresión se simplifica y reduce a la siguiente estadística de prueba:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Donde S_p se obtiene por medio de la expresión:

$$S_p = \sqrt{\frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}}$$

Caso 3. Si las desviaciones estándar poblacionales son desconocidas, pero se suponen distintas, para muestras inferiores a 30 se utiliza una estadística de prueba asociada con la distribución t-student con *g* grados de libertad presentada a través de la expresión 2.9.

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}$$

Otra vez, bajo la hipótesis nula H_0: $\mu_1 = \mu_2$ la anterior expresión se simplifica y reduce a la siguiente estadística de prueba:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}$$

Caso 4. Si las desviaciones estándar poblacionales son desconocidas, pero las muestras tienen tamaños superiores a 30, se utiliza una estadística de prueba asociada con la distribución normal estándar dada en la expresión 2.10.

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}$$

Bajo la hipótesis nula H_0: $\mu_1 = \mu_2$ la anterior expresión se simplifica; luego se usa la siguiente estadística de prueba:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}$$

5. Comparar el valor de la estadística de prueba con el valor en la distribución teórica.

6. Tomar una decisión: aceptar Ho, en caso contrario, rechazar Ho.

7. Conclusión

Ejemplo 4.10. El gerente de la empresa S&T quiere determinar si los salarios por hora para los trabajadores de esa empresa en la ciudad A y la ciudad B son iguales; para ello toma una muestra aleatoria de 40 trabajadores en la primera

ciudad y encuentra que el salario medio es de 6000 pesos por hora, con una desviación estándar de 200 pesos; luego, selecciona una muestra aleatoria de 54 trabajadores en la segunda ciudad y obtiene un salario promedio de 5940 por hora, con una desviación estándar de 180 pesos.

1. Planteamiento de hipótesis:

Ho: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

2. Nivel de significancia $\alpha = 0.05$; $\frac{\alpha}{2} = 0.025$

3. Dirección de la prueba. Se trata de una prueba bilateral; la región crítica se observa en la Figura 4.10, donde se presenta una curva normal estándar.

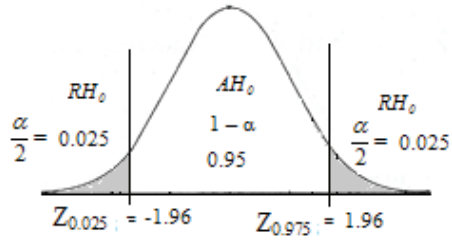

Figura 4.10 Prueba bilateral para la diferencia de medias poblacionales

Fuente: los autores con la ayuda del *software* libre R.

4. Estadística de prueba

En este ejemplo los datos muestrales son:

Ciudad A	Ciudad B
$n_1 = 40$	$n_2 = 54$
$\bar{X}_1 = 6000$	$\bar{X}_2 = 5940$
$S_1 = 200$	$S_2 = 180$

Por tratarse de muestras mayores a 30 y desviaciones estándar desconocidas, la estadística de prueba que se ha de usar en la prueba de hipótesis es:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}$$

No obstante, para calcular el valor de la estadística se requiere determinar las varianzas corregidas o cuasivarianzas, tarea que se desarrolla a continuación:

$$\hat{S}_1^2 = \frac{n_1}{n_1 - 1} S_1^2 = \frac{40}{39}(200)^2 = 41025.64$$

$$\hat{S}_2^2 = \frac{n_2}{n_2 - 1} S_2^2 = \frac{54}{53}(180)^2 = 33011.32$$

Reemplazando los valores en la estadística de prueba, resulta:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} = \frac{6000 - 5940}{\sqrt{\frac{41025.64}{40} + \frac{33011.32}{54}}} = \frac{60}{\sqrt{1025.64 + 611.32}} = \frac{60}{40.459} = 1.4829$$

5. El valor de la estadística de prueba, 1.4829, cae en la región AH_0, de aceptación de la hipótesis nula H_0, puesto que se encuentra en Z=-1.96 y Z=1.96 en la distribución teórica normal estándar.

6. Decisión: aceptar Ho: $\mu_1 = \mu_2$

7. Conclusión, con un nivel de significancia del 5 % se concluye que los salarios promedio por hora para los trabajadores de la empresa S&T no difieren significativamente en las ciudades A y B; es decir, los salarios promedio por hora para tales trabajadores en las dos ciudades son iguales.

4.6 Prueba de hipótesis para la varianza poblacional

El proceso de inferencia estadística para realizar una prueba de hipótesis asociadas con la varianza poblacional incluye el siguiente algoritmo:
1. Planteamiento de hipótesis

i) Ho: $\sigma^2 = \sigma_0^2$
H_1: $\sigma^2 > \sigma_0^2$

ii) Ho: $\sigma^2 = \sigma_0^2$
H_1: $\sigma^2 < \sigma_0^2$

iii) Ho: $\sigma^2 = \sigma_0^2$
H_1: $\sigma^2 \neq \sigma_0^2$

2. Fijación del nivel de significancia α

3. Dirección de la prueba. Se especifica en concordancia con cada pareja de hipótesis; en el primer caso, la región de rechazo se ubica en la cola derecha de una curva asimétrica correspondiente a la distribución *chi-cuadrado*, en el segundo, sobre la cola izquierda, y en el tercero sobre las dos colas; es decir, las zonas de rechazo se establecen de manera similar a las indicadas en las Figuras 4.1, 4.2 y 4.3, solamente que ahora se trabaja sobre una curva asimétrica.

4. Estadística de prueba

La estadística de prueba está asociada con una distribución *chi-cuadrado* con *n -1* grados de libertad; esta, bajo la hipótesis nula, es:

$$\chi^2 = \frac{(n-1)\hat{S}^2}{\sigma_0^2}$$

5. Comparación del valor de la estadística de prueba con el valor en la distribución teórica

6. Decisión: aceptar Ho o, en caso contrario, rechazar Ho.

7. Conclusión

Ejemplo 4.11. El espesor de una muestra aleatoria de 21 barras de chocolate es una variable aleatoria con desviación estándar corregida de 0.8 milésimas de pulgada; el proceso de fabricado de las barras de chocolate se encuentra bajo absoluto control si la varianza es de 0.38, y fuera de control si es mayor; un grupo de inspectores afirma que el proceso actualmente está fuera de control; probar esta hipótesis con un nivel de significancia del 5 %.

1. Planteamiento de hipótesis

$\text{Ho}: \sigma^2 = 0.38$

$\text{H}_1: \sigma^2 > 0.38$

2. Nivel de significancia $\alpha = 0.05$

3. Dirección de la prueba. Se trata de una prueba unilateral derecha; la región crítica se observa en la Figura 4.11, donde se presenta una curva asimétrica correspondiente a una distribución chi-cuadrado.

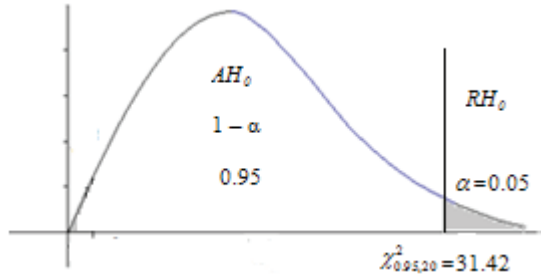

Figura 4.11 Prueba unilateral derecha para la varianza poblacional

Fuente: los autores con la ayuda del *software* libre R.

4. Estadística de prueba

En este ejemplo los datos muestrales son:

$$n = 21$$

$$\hat{S} = 0.8$$

Reemplazando los valores en la estadística de prueba, resulta:

$$\chi^2 = \frac{(n-1)\hat{S}^2}{\sigma_0^2} = \frac{(21-1)(0.8)^2}{0.38} = \frac{(20)(0.64)}{0.38} = \frac{12.8}{0.38} = 33.68$$

5. El valor de la estadística de prueba, 33.68, cae en la región RH_0, de rechazo de la hipótesis nula H_0, puesto que es mayor al valor teórico, 31.42, en la distribución chi- cuadrado con $n-1$ = 21 – 1 = 20 grados de libertad.

6. Decisión: rechazar Ho: $\sigma^2 = 0.38$; en consecuencia, se acepta que la varianza es mayor que 0.38.

7. Así, entonces, con un nivel de significancia del 5 % se concluye que el proceso de fabricación de las barras de chocolate está fuera de control; es decir, la afirmación hecha por el grupo de inspectores es cierta; no hay suficiente evidencia para rechazarla.

Ejemplo 4.12. En una muestra aleatoria proveniente de una población normal se obtuvo una desviación estándar corregida de 3 minutos para la cantidad de tiempo que tardaron 16 mujeres en terminar una prueba escrita para acceder a un puesto de trabajo nocturno; usar un nivel de significancia del 5 % para probar la hipótesis de que la desviación estándar poblacional es de 2.8 minutos contra la alternativa de que es inferior a este valor.

1. Planteamiento de hipótesis

Ho: $\sigma^2 = (2.8)^2$

H_1: $\sigma^2 < (2.8)^2$

2. Nivel de significancia $\alpha = 0.05$

3. Dirección de la prueba. Se trata de una prueba unilateral izquierda; la región crítica se observa en la Figura 4.12, donde se presenta una curva asimétrica correspondiente a una distribución chi-cuadrado.

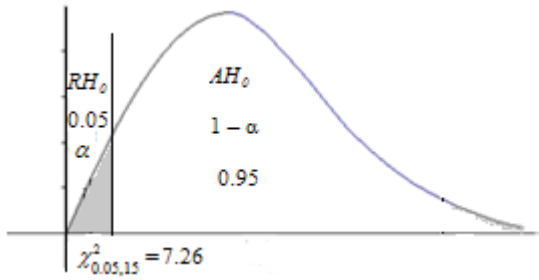

Figura 4.12 Prueba unilateral izquierda para la varianza poblacional
Fuente: los autores con la ayuda del *software* libre R.

4. Estadística de prueba

En este ejemplo los datos muestrales son:

$$n = 16$$

$$\hat{S} = 3$$

Reemplazando los valores en la estadística de prueba, resulta:

$$\chi^2 = \frac{(n-1)\hat{S}^2}{\sigma_0^2} = \frac{(16-1)(3)^2}{(2.8)^2} = \frac{(15)(9)}{7.84} = \frac{135}{7.8} = 17.30$$

5. El valor de la estadística de prueba, 17.30, cae en la región AH_0, de aceptación de la hipótesis nula H_0, puesto que es mayor al valor teórico, 7.26, en la distribución *chi- cuadrado* con *n -1* = 16 – 1 = 15 grados de libertad.

6. Decisión: aceptar Ho: $\sigma^2 = (2.8)^2$, por lo tanto, se acepta que la desviación estándar poblacional es de 2.8 minutos.

7. Por consiguiente, con un nivel de significancia del 5 % se concluye que la deviación estándar poblacional es de 2.8 minutos; no hay evidencias suficientes para rechazarla.

Por otro lado, es conveniente indicar que la prueba *chi-cuadrado* también se utiliza para probar hipótesis referidas a la asociación de variables cualitativas (Fernández & Díaz, 2004).

4.7 Prueba de hipótesis para el cociente de varianzas poblacionales

En esta sección se indica un proceso de inferencia estadística para realizar una prueba de hipótesis asociadas con el cociente de varianzas poblacionales; este involucra los siguientes pasos:

1. Planteamiento de hipótesis

i) Ho: $\frac{\sigma_1^2}{\sigma_2^2} = 1$

H_1: $\frac{\sigma_1^2}{\sigma_2^2} > 1$

ii) Ho: $\frac{\sigma_1^2}{\sigma_2^2} = 1$

H_1: $\frac{\sigma_1^2}{\sigma_2^2} < 1$

iii) Ho: $\frac{\sigma_1^2}{\sigma_2^2} = 1$

H_1: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$

Estas tres parejas de hipótesis también se escriben de la siguiente manera:

i) Ho: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 > \sigma_2^2$

ii) Ho: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 < \sigma_2^2$

iii) Ho: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

2. Fijación del nivel de significancia α

3. Dirección de la prueba. Se especifica en correspondencia con cada una de las parejas de hipótesis por medio de curvas asimétricas asociadas con una distribución *F* de *Fisher;* se establecen zonas de rechazo de manera similar a las indicadas en las Figuras 4.1, 4.2 y 4.3.

4. Estadística de prueba

La estadística de prueba está asociada con una distribución *F* de Fisher con n_1-1 grados de libertad para el numerador y n_2-1 grados de libertad para el denominador, como se indica enseguida:

$$F = \frac{\hat{S}_1^2 \sigma_2^2}{\hat{S}_2^2 \sigma_1^2}$$

Bajo la hipótesis nula H_0: $\sigma_1^2 = \sigma_2^2$, la anterior expresión se simplifica al cancelar las varianzas poblacionales; luego se usa la siguiente estadística de prueba:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

5. Comparación del valor de la estadística de prueba con el valor en la distribución teórica

6. Decisión: aceptar o rechazar la hipótesis nula Ho

7. Conclusión

Ejemplo 4.13. Al comparar la resistencia a la tracción de las clases A y B de acero estructural, el diseño experimental permitió establecer los siguientes resultados: una varianza corregida de 19.7 en la primera muestra aleatoria de 11 unidades y una varianza corregida de 3.8 en la segunda muestra aleatoria de 9 unidades, donde las unidades de medición están dadas en miles de libras por pulgada cuadrada. Se supone que los datos de las muestras provienen de poblaciones normales independientes; probar la hipótesis de que las varianzas de tal resistencia difieren usando un nivel de significancia del 2 %.

1. Planteamiento de hipótesis

$$\text{Ho: } \sigma_1^2 = \sigma_2^2$$

$$\text{H}_1\text{: } \sigma_1^2 \neq \sigma_2^2$$

2. Nivel de significancia $\alpha = 0.02 \rightarrow \frac{\alpha}{2} = 0.01$

3. Dirección de la prueba. Se trata de una prueba bilateral; la región crítica se observa en la Figura 4.13, donde se presenta una curva asimétrica correspondiente a una distribución *F* de Fisher.

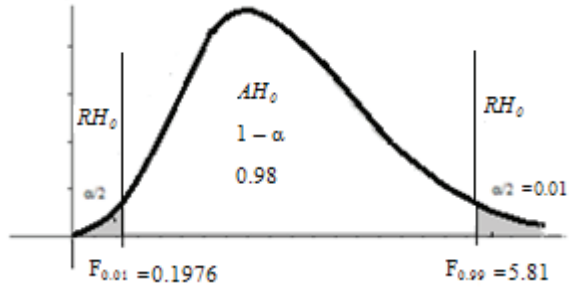

Figura 4.13 Prueba bilateral para la igualdad de varianzas poblacionales
Fuente: los autores con la ayuda del *software* libre R.

4. Estadística de prueba

En este caso los datos muestrales son:

Acero clase A	Acero clase B
$n_1 = 11$	$n_2 = 9$
$\hat{S}_1^2 = 19.7$	$\hat{S}_2^2 = 3.8$

Reemplazando los valores en la estadística de prueba, resulta:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2} = \frac{19.7}{3.8} = 5.1842$$

5. El valor de la estadística de prueba, 5.1842, cae en la región AH_0, de aceptación de la hipótesis nula H_0, puesto que este valor está entre los valores teóricos 0.1976 y 5.81 en la distribución *F* de Fisher n_1-1=11-1=10 grados de libertad para el numerador y n_2 -1 = 9 -1 = 8 grados de libertad para el denominador.

6. Decisión: aceptar Ho: $\sigma_1^2 = \sigma_2^2$, implica que se acepta que las varianzas poblacionales son iguales.

7. Por lo tanto, con un nivel de significancia del 2 % se concluye que las varianzas poblacionales correspondientes a las dos clases de acero estructural son iguales; no hay evidencias suficientes para afirmar que sean distintas.

Ejemplo 4.14. Para comparar la variabilidad en la duración de dos clases de bombillos etiquetados como W y T se tomaron dos muestras aleatorias; la primera, de 8 unidades, proporcionó una desviación estándar de 36 horas, y la segunda, de 7 unidades, generó una desviación estándar corregida de 26.7 horas; bajo el supuesto de que las muestras provienen de poblaciones normales independientes, probar la hipótesis de que la varianza de la clase de bombillos W es superior a la varianza de los tipo T usando un nivel de significancia del 5 %.

1. Planteamiento de hipótesis

$\text{Ho}: \sigma_1^2 = \sigma_2^2$

$\text{H}_1: \sigma_1^2 > \sigma_2^2$

2. Nivel de significancia $\alpha = 0.05$

3. Dirección de la prueba. Se trata de una prueba unilateral derecha; la región crítica se observa en la Figura 4.14, donde se presenta una curva asimétrica correspondiente a una distribución *F* de Fisher.

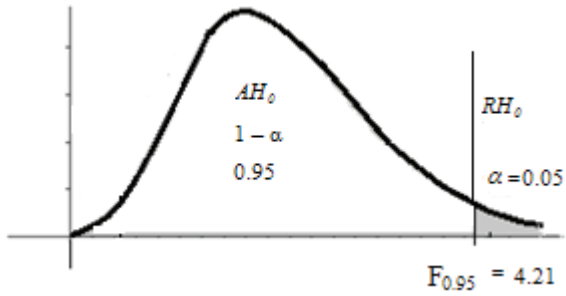

Figura 4.14 Prueba unilateral derecha para la igualdad de varianzas poblacionales

Fuente: los autores con la ayuda del *software* libre R.

4. Estadística de prueba

En este otro caso los datos de la muestra son:

Bombillos W	Bombillos T
$n_1 = 8$	$n_2 = 7$
$S_1 = 36$	$\hat{S}_2 = 26.7$

En primera instancia, conviene calcular la desviación estándar corregida correspondiente a los datos de la primera muestra:

$$\hat{S}_1^2 = \frac{n_1}{n_1 - 1} S_1^2 = \frac{8}{7}(36)^2 = 1481.14$$

Reemplazando estos resultados en la estadística de prueba, se tiene:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2} = \frac{1481.14}{(26.7)^2} = \frac{1481.14}{712.89} = 2.0776$$

5. El valor de la estadística de prueba, 2.0776, cae en la región AH_0, de aceptación de la hipótesis nula H_0, puesto que es menor que el valor teórico 4.21 en la distribución *F* de Fisher n_1 -1=8-1=7 grados de libertad para el numerador, y n_2-1=7-1=6 grados de libertad para el denominador.

6. Decisión: aceptar Ho: $\sigma_1^2 = \sigma_2^2$, lo cual implica que se acepta que las varianzas poblacionales son iguales.

7. Finalmente, con un nivel de significancia del 5 % se concluye que la varianza de la duración de la clase de bombillos W es igual a la varianza (poblacional) de la duración de los bombillos tipo T; no hay evidencias suficientes para afirmar que las varianzas poblacionales sean distintas estadísticamente.

4.8 Algunos tamaños de muestra

En esta sección se aborda la manera de obtener algunos tamaños de la muestra; la primera, asociada con una variable de tipo cualitativo, y la segunda, cuando el interés principal recae en una variable cuantitativa.

Si el interés principal recae en una característica de tipo cualitativo, entonces, la proporción muestral es una variable aleatoria pertinente para estudiar tal característica, y su distribución de probabilidad, de acuerdo con la expresión 2.5, es:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

Si la diferencia entre el estimador y el parámetro se denomina "*e*: error absoluto", para un determinado nivel de confianza 1- α, se tiene la siguiente igualdad:

$$Z = \frac{e}{\sqrt{\frac{pq}{n}}} \quad \rightarrow \quad Z\sqrt{\frac{pq}{n}} = e$$

De la anterior expresión se despeja *n*, que representa el tamaño de la muestra:

$$Z^2 \frac{pq}{n} = e^2 \quad \rightarrow \quad Z^2 pq = ne^2$$

De aquí se deduce que el tamaño de la muestra proveniente de una población infinta es:

$$n = \frac{Z^2 pq}{e^2} \tag{4.1}$$

El error absoluto y el nivel de confianza son aportados por el investigador y determinados con base en su experiencia. Si de estudios anteriores se conocen los parámetros *p* y *q,* entonces se usa directamente la expresión 4.1; en caso contrario, se trabaja bajo total incertidumbre, es decir, asignando el valor 0.5 tanto a *p* como a *q*, de la forma como se indica en la expresión 4.2; otra posibilidad consiste en seleccionar una muestra piloto y de allí obtener los estimadores de *p* y *q*, como se indica en la expresión 4.3.

$$n = \frac{Z^2(0.5)(0.5)}{e^2} \quad \rightarrow \quad n = \frac{Z^2}{4e^2} \tag{4.2}$$

$$n = \frac{Z^2 \hat{p}\hat{q}}{e^2} \tag{4.3}$$

Ahora, si se muestrea de una población finita de tamaño *N,* entonces, la distribución para la proporción muestral es una normal estándar, dada por:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}}$$

Nuevamente, haciendo *e* igual a la diferencia entre el estimador y el parámetro, se realizan los procesos algebraicos pertinentes y se despeja *n* para obtener la expresión 4.4.

$$n = \frac{NZ^2 pq}{(N-1)e^2 + Z^2 pq} \tag{4.4}$$

Si de estudios anteriores se conocen los parámetros *p* y *q,* entonces se usa directamente la expresión 4.4; en caso contrario se trabaja bajo total incertidumbre y se asigna el valor 0.5 tanto a *p* como a *q,* de la manera como se indica en la expresión 4.5; otra posibilidad consiste en seleccionar una muestra piloto y, de allí, obtener los estimadores de *p* y *q,* como se indica en la expresión 4.6.

$$n = \frac{NZ^2 (0.5)(0.5)}{(N-1)e^2 + Z^2 (0.5)(0.5)} \tag{4.5}$$

$$n = \frac{NZ^2 \hat{p}\hat{q}}{(N-1)e^2 + Z^2 \hat{p}\hat{q}} \tag{4.6}$$

Por otro lado, si el interés principal recae en una característica de tipo cuantitativo, entonces el promedio muestral es una variable aleatoria adecuada para estudiar tal característica; cuando los datos sean relativamente homogéneos, su distribución de probabilidad, de acuerdo con la expresión 2.1, es la siguiente, cuando la desviación estándar es conocida:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Nuevamente, si llamamos error absoluto (*e*) a la diferencia entre el estimador y el parámetro, para un determinado nivel de confianza 1- α, se tiene la siguiente igualdad:

$$Z = \frac{e}{\frac{\sigma}{\sqrt{n}}} \quad \rightarrow \quad Z\frac{\sigma}{\sqrt{n}} = e$$

De la anterior expresión se despeja *n*,

$$\frac{Z^2\sigma^2}{n} = e^2 \quad \rightarrow \quad Z^2\sigma^2 = ne^2$$

De aquí se deduce que el tamaño de la muestra proveniente de una población infinta es:

$$n = \frac{Z^2\sigma^2}{e^2} \tag{4.7}$$

El error absoluto y el nivel de confianza son aportados por el investigador y determinados con base en su experiencia. Si de estudios anteriores se conoce la varianza poblacional, entonces se usa directamente la expresión 4.7; en caso contrario se selecciona una muestra piloto y de allí se obtiene el estimador insesgado de la varianza y se conforma la expresión 4.8.

$$n = \frac{Z^2\hat{S}^2}{e^2} \tag{4.8}$$

Ahora, si se muestrea de una población finita de tamaño *N,* entonces, la distribución para la media muestral es una normal estándar, dada por:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

Otra vez, haciendo *e* igual a la diferencia entre el estimador y el parámetro, se realizan los procesos algebraicos pertinentes y se despeja *n* para obtener la expresión 4.9,

$$n = \frac{NZ^2\sigma^2}{(N-1)e^2 + Z^2\sigma^2} \tag{4.9}$$

Si de estudios anteriores se conoce el parámetro varianza poblacional, entonces se usa directamente la expresión 4.4; en caso contrario, se selecciona una muestra piloto y de allí se obtiene el estimador insesgado para la varianza, con esto resulta la expresión 4.10.

$$n = \frac{NZ^2\hat{S}^2}{(N-1)e^2 + Z^2\hat{S}^2} \tag{4.10}$$

Ejemplo 4.15. Se requiere estudiar el grado de escolaridad (años promedio de estudio) de los individuos cabeza de familia residentes en el casco urbano de la ciudad M; los archivos del DANE en Colombia registran 51 000 núcleos familiares en el citado casco urbano; se necesita calcular el tamaño de la muestra

para estimar el promedio de los años de estudio, admitiendo un error absoluto de medio año y utilizando un nivel de confianza del 95 %. De una muestra piloto se ha obtenido una desviación estándar corregida de 5 años.

En el ejemplo que nos ocupa, los datos son:

$$1-\alpha = 0.95$$

$$\alpha = 0.05 \rightarrow \frac{\alpha}{2} = 0.025$$

$$N = 51000$$

$$\hat{S} = 5$$

$$e = 0.5$$

En este caso conviene utilizar la expresión 4.10; sin embargo, hace falta el valor de Z, que se obtiene de una distribución normal estándar para una probabilidad acumulada del 0.975, obteniéndose un valor de 1.96, como se observa en la Figura 4.15

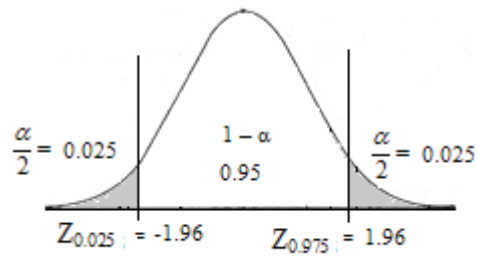

Figura 4.15 Valor de Z para un nivel de confianza del 95%

Fuente: los autores con la ayuda del *software* libre R.

Reemplazando los valores, resulta:

$$n = \frac{NZ^2\hat{S}^2}{(N-1)e^2 + Z^2\hat{S}^2} = \frac{51000(1.96)^2(5)^2}{50999(0.5)^2 + (1.96)^2(5)^2} = \frac{51000(3.8416)(25)}{50999(0.25) + (3.8416)(25)}$$

$$n = \frac{4898040}{12749.75 + 96.04} = \frac{4898040}{12845.79} \cong 381.29$$

En el contexto anterior, el tamaño de la muestra para estimar el promedio de los años de estudio, admitiendo un error absoluto de medio año y utilizando un nivel de confianza del 95 %, es de 381 individuos que sean cabeza de familia.

Esta cantidad de individuos se han de seleccionar apropiadamente, usando algunos de los métodos de muestreo descritos en el primer capítulo de este texto u otros.

Ejemplo 4.16. Se pretende investigar el mercado para el producto A; se sabe que el 30 % de los hogares de la ciudad H poseen este producto (han manifestado en algún momento que les ha gustado tal producto); calcular el tamaño de la muestra si se desea que el error máximo absoluto sea del 5 % con un nivel de confianza del 96 %.

En este caso, se tiene que:

$$1-\alpha = 0.96$$

$$\alpha = 0.04 \rightarrow \frac{\alpha}{2} = 0.02$$

$$e = 0.05$$

$$p = 0.3$$

En estas circunstancias, se recomienda utilizar la expresión 4.1; el valor de Z=2.055 se obtiene de una distribución normal estándar para una probabilidad acumulada del 0.98, como se indica en la Figura 4.16.

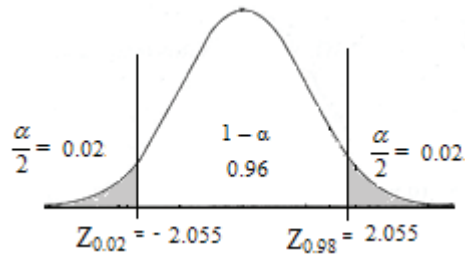

Figura 4.16 Valor de Z para un nivel de confianza del 96%

Fuente: los autores con la ayuda del *software* libre R.

Reemplazando los valores, resulta:

$$n = \frac{Z^2 pq}{e^2} = \frac{(2.055)^2(0.3)(0.7)}{(0.05)^2} = \frac{(4.223)(0.21)}{0.0025} = \frac{0.88683}{0.0025} \cong 354.73$$

En estas circunstancias, el tamaño de la muestra para investigar el mercado del producto A, en la ciudad H, es de 354 individuos, admitiendo un error del 5 % y un nivel de confianza del 96 %. Este número de individuos se han de seleccionar de forma pertinente, utilizando un método de muestreo apropiado.

Ejemplo 4.17. Una población de 20 000 estudiantes de la universidad B está interesada en elegir al rector de esa institución; el candidato JA aspira a esa rectoría porque siente el apoyo del sector estudiantil; se quiere establecer el tamaño de la muestra para estimar el porcentaje de estudiantes que potencialmente apoyan a este candidato; usar un nivel de confianza del 99 % y asumir un error máximo del 2 %.

En esta situación se ha de trabajar bajo total incertidumbre, los datos asociados son:

$$1-\alpha = 0.99$$

$$\alpha = 0.01 \rightarrow \frac{\alpha}{2} = 0.005$$

$$e = 0.02$$

$$p = 0.5 = q$$

En este caso se recomienda utilizar la expresión 4.5; el valor de Z=2.58 se obtiene de una distribución normal estándar para una probabilidad acumulada del 0.995, como se indica en la Figura 4.17.

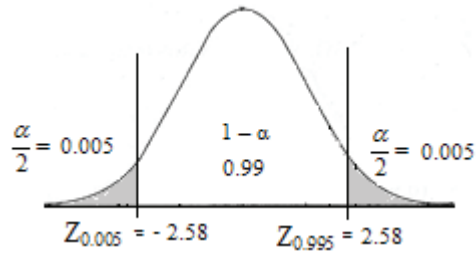

Figura 4.17 Valor de Z para un nivel de confianza del 99%
Fuente: los autores con la ayuda del *software* libre R.

Reemplazando los valores, resulta:

$$n = \frac{NZ^2(0.5)(0.5)}{(N-1)e^2 + Z^2(0.5)(0.5)} = \frac{20000(2.58)^2(0.5)(0.5)}{19999(0.02)^2 + (2.58)^2(0.5)(0.5)}$$

$$n = \frac{20000(6.6564)(0.25)}{19999(0.0004) + (6.6564)(0.25)} = \frac{33282}{7.9996 + 1.6641} = \frac{33282}{9.6637} = 3444.02$$

En este contexto, el tamaño de la muestra para para estimar el porcentaje de estudiantes que potencialmente apoyan al candidato JA es de 3444, admitiendo un error del 2 % y un nivel de confianza del 99 %.

Ejemplo 4.18. Se desea estudiar el salario promedio en miles de pesos por mes de una población de trabajadores del suroccidente colombiano; calcular el tamaño de la muestra con un nivel de confianza del 95 % y admitiendo un error de 600 pesos. De una muestra piloto de 20 trabajadores de esa zona del país se ha obtenido una desviación estándar de 10 000 pesos.

Los datos de que se disponen son los siguientes:

$$1-\alpha = 0.95$$

$$\alpha = 0.05 \rightarrow \frac{\alpha}{2} = 0.025$$

$$S = 10$$

$$e = 0.6$$

Se ha de tener presente que los datos queden expresados en miles de pesos, tanto la desviación estándar como el error máximo. En este caso conviene utilizar la expresión 4.8; no obstante, hace falta calcular la desviación estándar corregida y el valor de Z; este valor se obtiene de una distribución normal estándar para una probabilidad acumulada del 0.975, el cual es 1.96, como se observa en la Figura 4.18.

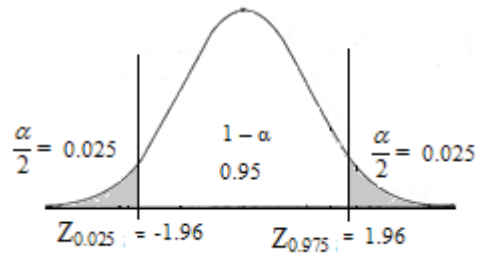

Figura 4.18 Valor de Z para un nivel de confianza del 95 %
Fuente: los autores con la ayuda del *software* libre R.

El valor de la desviación estándar corregida se obtiene así:

$$\hat{S}^2 = \frac{n}{n-1} S^2 = \frac{20}{19}(10)^2 = 1.0526(100) = 105.26$$

Reemplazando los valores, resulta:

$$n = \frac{Z^2 \hat{S}^2}{e^2} = \frac{(1.96)^2(105.26)}{(0.6)^2} = \frac{(3.8416)(105.26)}{(0.36)} = \frac{404.3668}{0.36} = 1123.24$$

Finalmente, el tamaño para estudiar el salario promedio en miles de pesos por mes de una población de trabajadores del suroccidente colombiano es de 1123 trabajadores, con un nivel de confianza del 95 % y un error máximo de 600 pesos. Esta cantidad de trabajadores se han de seleccionar usando métodos de muestreo apropiados.

Actividades para el estudio independiente Capítulo 4

4.1 Un fabricante de llantas afirma que sus cauchos duran, en promedio, 21 000 km. Un distribuidor potencial de llantas considera que duran menos; por eso quiere verificar la afirmación del fabricante, y de un lote de 200 llantas selecciona 11 de manera aleatoria y las pone a rodar, encontrando un promedio de 20 500 kilómetros; si el fabricante informa que las llantas que ha producido tienen una desviación estándar poblacional de 1000 kilómetros, probar la afirmación del fabricante usando un nivel de significación del 2.5 %.

4.2 Un representante estudiantil afirma que al 10 % de los estudiantes de la universidad B les sirven el almuerzo casi frio; los administradores del restaurante en esta universidad reconocen que algunas veces eso ha sucedido, sin embargo, consideran que el porcentaje es muy inferior al 10 %. En el intento de desvirtuar la afirmación del representante se toma una muestra aleatoria de 150 estudiantes de esta universidad, y el 12 de ellos responden que alguna vez han recibido el almuerzo casi frio; probar la afirmación hecha por el representante, usando un nivel de significancia del 3 %.

4.3 Un estudio sobre el gusto por la práctica deportiva en hombres y mujeres reveló que en una muestra aleatoria de 380 hombres, a 171 de ellos les gusta esta práctica, y que en una muestra aleatoria de 350 mujeres, a 168 de ellas les gusta la práctica deportiva; estos resultados se constituyen en suficiente evidencia para afirmar que el porcentaje de gusto por la práctica deportiva de los hombres difiere del de las mujeres; usar un nivel de significancia del 3 % para realizar la prueba de hipótesis.

4.4 El Director general de Cámaras de Comercio en Colombia desea determinar al 5 % de significancia si las utilidades en millones por mes de las pequeñas empresas en Cali y Bucaramanga son iguales. Para esto toma en Cali una muestra aleatoria de 20 empresas pequeñas, y en Bucaramanga, una de 17 empresas, y encuentra que la utilidad promedio en la primera es de 8 millones de pesos por mes, con una desviación estándar de 500 mil pesos, y en la segunda, de 7.5 millones de pesos por mes, con una desviación estándar de 400 mil pesos.

4.5 El diámetro en una muestra aleatoria de 10 varillas de hierro de media pulgada es una variable aleatoria con desviación estándar corregida de 0.2 milésimas de pulgada; el proceso de fabricado de las varillas de hierro se encuentra bajo absoluto control si la varianza es de 0.09, y fuera de control si es mayor; un empleado que lleva el control de calidad sobre el diámetro de las

varillas producidas considera que el proceso actualmente está fuera de control; probar esta hipótesis con un nivel de significancia del 5 %.

4.6 Para comparar la variabilidad en la duración de dos clases de calzado, A y B, se tomaron dos muestras aleatorias; la primera, de 6 pares, proporcionó una desviación estándar de 3 meses, y la segunda, de 5 pares, generó una desviación estándar corregida de 2.5 meses; bajo el supuesto de que las muestras provienen de poblaciones normales independientes, probar la hipótesis de que la varianza de la clase de calzado A es superior a la varianza de la clase de calzado B, usando un nivel de significancia del 5 %.

4.7 Se tiene una población finita de 600 personas; se seleccionó una muestra piloto del 2 % de la población total y se obtuvo una media muestral de 300 dólares para el salario promedio mensual; también se obtuvo de la muestra piloto una desviación estándar corregida de 50 dólares; si se admite un error del 3 % de la media muestral, calcular el tamaño muestral al nivel de confianza del 95 %.

4.8 Se requiere indagar el mercado para el producto H; se sabe que el 60 % de los individuos de la ciudad A utiliza diariamente este producto; calcular el tamaño de la muestra si se desea que el error máximo absoluto sea del 4 %, con un nivel de confianza del 96 %.

Ejercicios para el capítulo 4

4.1 Consultar sobre la forma como se realiza una prueba de hipótesis para muestras pareadas, también llamadas muestras relacionadas; asimismo, proporcionar un ejemplo de prueba de hipótesis.

4.2 Los siguientes datos corresponden a los ingresos por día de unos trabajadores independientes, en miles de pesos: 50, 30, 15, 20, 55, 35. Si se asume un error del 8 % de la media muestral, determinar el tamaño de una muestra con un $\alpha = 5\ \%$.

4.3 La muestra piloto para estimar la proporción y establecer el porcentaje de personas que tienen ingresos mensuales inferiores a 800 000 pesos mensuales en una población de tamaño $N = 200$ está dada por: 2 500 000, 2 090 000, 2 190 000, 750 000, 1 250 000, 700 000, 790 000 y 790 000. Asumir un error del 10 % y un $\alpha = 5\ \%$ para determinar el tamaño de la muestra.

4.4 Un estudio sobre el gusto por el producto A en hombres y mujeres reveló que, en una muestra aleatoria de 600 hombres, a 480 de ellos les gusta este producto, y en una muestra aleatoria de 1000 mujeres, a 800 de ellas les gusta este producto; probar la hipótesis de que el porcentaje de gusto por el producto A de los hombres difiere del de las mujeres; usar un nivel de significancia del 5 % para realizar la prueba de hipótesis.

4.5 El Director general de Cámaras de Comercio en Colombia quiere determinar, al 5 % de significancia, si las utilidades en millones por mes de las pequeñas empresas en Medellín y Barranquilla son iguales. Para esto toma una muestra aleatoria de 24 pequeñas empresas en Medellín, y encuentra que la utilidad promedio es de 10 millones de pesos por mes, con una desviación estándar de 600 mil pesos, y selecciona una muestra aleatoria de 18 pequeñas empresas en Barranquilla, y obtiene una utilidad promedio de 9.5 millones de pesos por mes, con una desviación estándar de 550 mil pesos. Realizar la prueba de hipótesis considerando varianzas poblacionales diferentes.

4.6 Los siguientes datos corresponden a la variable calificación en estadística inferencial, en la universidad B, en una muestra piloto: 3.5, 4.5, 3.5, 4.0, 5.0, 2.0, 1.0, 4.5. Si el error es del 10 % de la calificación promedio y se trabaja con un nivel de confianza del 99 %, calcular el tamaño muestral si en toda la universidad han cursado 800 estudiantes la mencionada asignatura.

4.7 En un centro de distribución de computadores portátiles se venden dos marcas diferentes en un periodo de tiempo específico; en una semana se venden

300 portátiles, de un total de 450, de la marca A, y 280, de un total de 400, de la marca B. El administrador considera que los porcentajes poblacionales de venta de las dos marcas son diferentes; probar la hipótesis del administrador usando un nivel de significancia del 4 %.

5

Información de retorno sobre las actividades de estudio independiente

En este capítulo se proporciona la información de retorno referente a las actividades de estudio independiente; se desarrollan diversos procesos de cálculo y se generan las respuestas a las preguntas formuladas en las mencionadas actividades; aquellas que el estudiante o el lector debería haber resuelto una vez hubiese realizado una lectura comprensiva de los temas y revisado los ejemplos indicados a través de cada capítulo. Con el propósito de aclarar algunas dudas y preguntas de investigación, se desarrollan completamente los ejercicios que se propusieron en esas actividades por capítulos.

Capítulo 1

1.1

a) La estadística *descriptiva*

b) La estadística *inferencial*

c) Una *muestra*

d) Se denominan *variables*

e) Variables *cualitativas*

f) Variables *cuantitativas.*

g) Una variable *continua*

h) Una variable *discreta*

1.2

a) cualitativa en escala nominal

b) cualitativa en escala ordinal

c) cuantitativa en escala ordinal

d) cuantitativa en escala de intervalo

e) cuantitativa en escala de razón

1.3

a) n = 30

b) 3/30 = 0.3 = 30 %

c) No es adecuado calcular el promedio, porque son datos correspondientes a una variable cualitativa; para esta se recomienda calcular un porcentaje o una proporción.

d) 63.8914

e) 3.14225

f) 0.0491 = 4.91 %, este es un valor inferior al 8 % e indica que los datos de la variable estatura en esta muestra de estudiantes son homogéneos.

1.4 Del conjunto de datos del ejemplo 1.11 es posible seleccionar 10 muestras aleatorias de tamaño $n = 3$ mediante un muestreo aleatorio sin reemplazo, una de ellas está constituida por los datos 9, 10, 8.

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

$$\overline{x} = \frac{\sum_{i=1}^{3} x_i}{3} = \frac{9+10+8}{3} = \frac{27}{3} = 9$$

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

$$s^2 = \frac{\sum_{i=1}^{3} (x_i - 9)^2}{3} = \frac{(9-9)^2 + (10-9)^2 + (8-9)^2}{3}$$

$$s^2 = \frac{(0)^2 + (1)^2 + (-1)^2}{3} = \frac{0+1+1}{3} = \frac{2}{3} \cong 0.6666$$

Otra estimación de la varianza poblacional es $\hat{s}^2$, valor particular de la variable

$\hat{S}^2$; esta se obtiene así:

$$\hat{s}^2 = \frac{\sum_{i=1}^{3}(x_i - 9)^2}{3-1} = \frac{2}{2} = 1$$

Se observa que el valor $\hat{s}^2$ se obtiene usando la siguiente expresión:

$$\hat{s}^2 = \frac{n}{n-1}s^2 = \frac{3}{3-1}\left(\frac{2}{3}\right) = \frac{6}{6} = 1$$

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

$$s = \sqrt{\frac{\sum_{i=1}^{3}(x_i - 9)^2}{3}} = \sqrt{0.6666} \cong 0.8164$$

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

$$\hat{s} = \sqrt{\frac{\sum_{i=1}^{3}(x_i - 9)^2}{3-1}} = \sqrt{1} = 1$$

Un valor del coeficiente de variación muestral es:

$$cv = \frac{\hat{s}}{\overline{x}} = \frac{1}{9} \cong 0.1111 = 11.11\ \%$$

El anterior valor se encuentra entre el 8 % y el 18 %, por lo tanto, los datos de esta muestra son casi homogéneos.

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

$$\hat{p} = \frac{x}{n} = \frac{1}{3} \cong 0.3333 = 33.33\ \%$$

1.5 Se tiene una población conformada por 100 individuos; en ellos interesa estudiar sus gastos semanales en miles de pesos. Se quiere seleccionar una muestra aleatoria de tamaño 5 usando muestreo aleatorio simple sin reemplazo. El número total de muestras distintas que se obtiene es:

$$\binom{100}{5} = \frac{100!}{6!(100-5)!} = \frac{100!}{5!95!} = 75\ 287\ 520$$

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

$$\frac{1}{75287520} = 0.0000000132$$

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

$$\frac{5}{100} = 0.05$$

1.6 Se tiene una población conformada por 20 individuos, de los cuales interesa estudiar la relación peso-talla, a fin de implementar una dieta para disminuir su peso. Se quiere seleccionar una muestra aleatoria de tamaño 3 usando muestreo aleatorio simple con reemplazo. El número total de muestras de tamaño 3 con repetición que se pueden obtener es:

$$20^3 = 8\ 000$$

La probabilidad de que una sucesión cualquiera conformada por 3 individuos sea seleccionada es:

$$\frac{1}{20^3} = \frac{1}{8000} = 0.000125$$

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

$$1-\left(\frac{20-1}{20}\right)^3 = 1-\left(\frac{19}{20}\right)^3 = 1-\left(\frac{6859}{8000}\right) = 1-0.857375 = 0.142625$$

1.7 Se tiene una población de 5000 individuos dividida en 4 estratos, como se indica en la Figura 5.1; se quiere seleccionar una muestra aleatoria de tamaño 100.

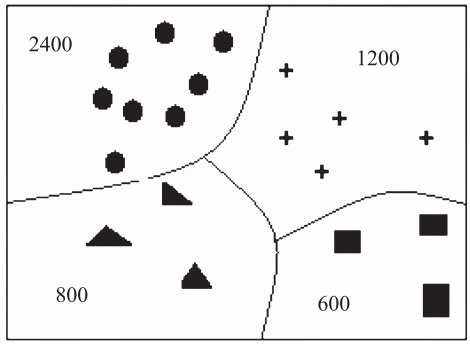

Figura 5.1 Población conformada por cuatro estratos

El tamaño de la población es N =5 000, el tamaño de la muestra es n =100

$$n_1 = \frac{2400}{5000} \star 100 = 48$$

$$n_2 = \frac{1200}{5000} \star 100 = 24$$

$$n_3 = \frac{600}{5000} \star 100 = 12$$

$$n_4 = \frac{800}{5000} \star 100 = 16$$

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

$$n = n_1 + n_2 + n_3 + n_4 = 48+24+12+16=100$$

En este contexto, 48 individuos se han de seleccionar mediante muestreo aleatorio simple del estrato 1; además, 24 individuos del estrato 2, 12 individuos del estrato 3 y 16 individuos del estrato 4.

ii) Ahora, usando muestreo aleatorio no proporcional, o por cuotas iguales, se toma igual número de elementos en cada estrato, usando la expresión:

$$n_i = \frac{n}{k} = \frac{100}{4} = 25$$

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

En este ejemplo se han de seleccionar 25 individuos de cada uno de los cuatro estratos por medio de muestreo aleatorio simple.

1.8 Si se tiene una población con N=2 500 individuos y se desea seleccionar una muestra de n=90 individuos, entonces se ordenan los datos correspondientes a los individuos y se calcula:

$$c = \frac{N}{n} = \frac{2500}{90} = 27.77$$

Ahora el número λ se toma de forma aleatoria como un número menor que 27, y se usará como punto de partida. Si después de realizar un procedimiento aleatorio para elegir el valor de λ, el valor resultante fuera 20, entonces el segundo individuo es aquel en la posición 57, el tercero ocupa la posición 84 y así sucesivamente; el individuo 90 se encuentra en la posición:
20 + (90-1) 27 = 2 423.

Capítulo 2

2.1 Graficar y determinar las siguientes probabilidades utilizando la distribución normal estándar a) $P(Z \leq -1.25)$, b) $P(Z \geq 1.31)$ c) $P(0.29 \leq Z \leq 2.48)$

$$a)\ P(Z \leq -1.25) = \Phi(-1.25) = 0.1056 \cong 10.56\ \%$$

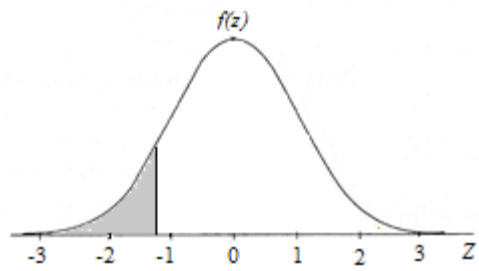

Figura 5.2 Área bajo la curva normal estándar $P(Z \leq -1.25)$
Fuente: los autores con la ayuda del *software* libre R.

La Figura 5.2 muestra el valor del área bajo la curva normal estándar equivalente al cálculo de la probabilidad.

$$b)\ P(Z \geq 1.31) = 1 - P(Z \leq 1.31) = 1 - 0.9049 = 0.0951 \cong 9.51\ \%$$

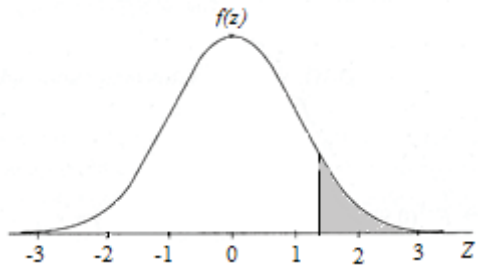

Figura 5.3. Área bajo la curva normal estándar $P(Z \geq 1.31)$

Fuente: los autores con la ayuda del *software* libre R.

La Figura 5.3 muestra el valor del área bajo la curva normal estándar, equivalente al cálculo de la probabilidad.

$$c)\ P(0.29 \leq Z \leq 2.48) = P(Z \leq 2.48) - P(Z \leq 0.29)$$

$$P(0.29 \leq Z \leq 2.48) = 0.9934 - 0.6141 = 0.3793 \cong 37.93\ \%$$

La Figura 5.4 muestra el valor del área bajo la curva normal estándar, equivalente al cálculo de la probabilidad.

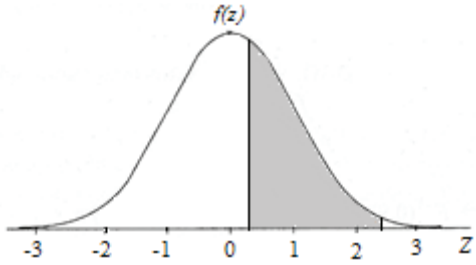

Figura 5.4 Área bajo la curva normal estándar $P(0.29 \leq Z \leq 2.48$

Fuente: los autores con la ayuda del *software* libre R.

2.2 La empresa S&S produce lámparas cuya duración en horas se distribuye normalmente con media de 900 horas y desviación estándar de 50 horas. Si se toma al azar una lámpara de la producción,

a) ¿Cuál es la probabilidad de que dure máximo 940 horas?
b) ¿Cuál es la probabilidad de que dure por lo menos 822 horas?
c) ¿Cuál es la probabilidad de que dure entre 880 y 1020 horas?
d) ¿Cuál es la probabilidad de que dure más de 1200 horas?

$\mu =$ 900 horas, $\sigma =$50 horas

X : duración en horas de una lámpara tomada al azar producida por la empresa S&S.

$$A)\ P(X \le 940) = P\left(\frac{X-\mu}{\sigma} \le \frac{940-\mu}{\sigma}\right) = P\left(Z \le \frac{940-900}{50}\right)$$

$$P(X \le 940) = P\left(Z \le \frac{40}{50}\right) = P(Z \le 0.8) = 0.7881 \cong 78.81\ \%$$

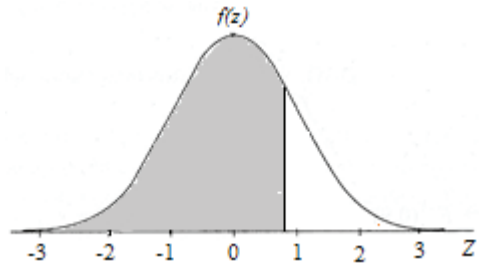

Figura 5.5. Área bajo la curva normal estándar $P(Z \le 0.8)$

Fuente: los autores con la ayuda del *software* libre R.

La probabilidad de que una lámpara escogida al azar de la producción dure máximo 940 horas es del 78.81 %. Una representación de la situación se tiene en la Figura 5.5.

$$b)\ P(X \ge 822) = P\left(\frac{X-\mu}{\sigma} \ge \frac{822-\mu}{\sigma}\right) = P\left(Z \ge \frac{822-900}{50}\right)$$

$$P(X \ge 822) = P(Z \ge -1.56) = 1 - P(Z \le -1.56) = 1 - 0.0594 = 0.9406$$

$$P(X \ge 822) = 0.9406 \cong 94.06\ \%$$

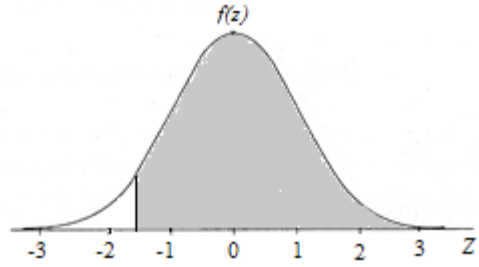

Figura 5.6. Área bajo la curva normal estándar $P(Z \ge -1.56)$

Fuente: los autores con la ayuda del *software* libre R.

La probabilidad de que una lámpara escogida al azar de la producción dure por lo menos 822 horas es del 94.06 %. Una representación de la anterior situación se tiene en la Figura 5.6.

$$c)\ P(880 \le X \le 1020) = P\left(\frac{880-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{1020-\mu}{\sigma}\right)$$

$$P(880 \le X \le 1020) = P\left(\frac{880-900}{50} \le Z \le \frac{1020-900}{50}\right) = P(-0.4 \le Z \le 2.4)$$

$$P(880 \le X \le 1020) = 0.9918 - 0.3446 = 0.6472 \cong 64.72\ \%$$

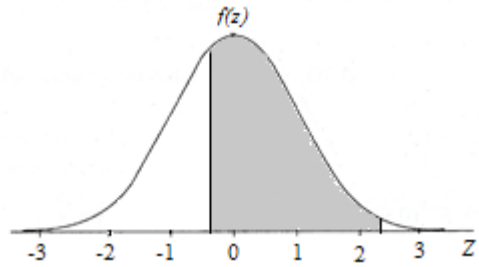

Figura 5.7. Área bajo la curva normal estándar $P(-0.4 \le Z \le 2.4)$

Fuente: los autores con la ayuda del *software* libre R.

La probabilidad de que una lámpara escogida al azar de la producción dure entre 880 y 1020 horas es de 64.72 %.Ver Figura 5.7.

$$d)\ P(X \ge 1200) = P\left(\frac{X-\mu}{\sigma} \ge \frac{1200-900}{50}\right) = P(Z \ge 6) \cong 0$$

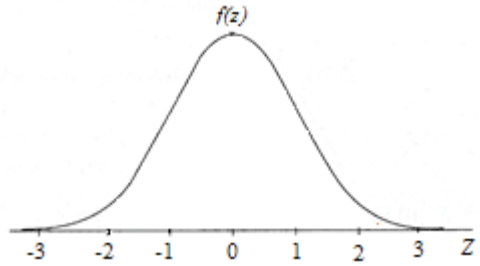

Figura 5.8. Área bajo la curva normal estándar $P(Z \ge 6)$

Fuente: los autores con la ayuda del *software* libre R.

La probabilidad de que una lámpara escogida al azar de la producción dure más de 1200 horas es del 0 % aproximadamente.Ver Figura 5.8.

2.3 El peso de los paquetes de arroz empacados por la máquina H&H es una variable aleatoria que se distribuye normalmente con media $\mu = 500$ g, y una desviación estándar $\sigma = 20$ g. Si se escoge aleatoriamente un paquete de arroz empacado por la máquina H&H,

a) ¿Cuál es la probabilidad de que su peso sea de por lo menos 486 gramos?

b) ¿Cuál es la probabilidad de que su peso sea máximo de 480 gramos?

c) ¿Cuál es la probabilidad de que su peso esté entre 476 y 550 gramos?

d) ¿Cuál es la probabilidad de que su peso sea de menos de 400 gramos?

e) ¿Cuál es la probabilidad de que su peso sea de más de 440 gramos?

$\mu = 500$ gramos, $\sigma = 20$ gramos

X : peso del paquete de arroz escogido aleatoriamente.

$$a)\ P(X \geq 486) = 1 - P\left(\frac{X-\mu}{\sigma} \leq \frac{486-500}{20}\right)$$

$$P(X \geq 486) = P(Z \geq -0.7) = 1 - P(Z \leq -0.7) = 1 - 0.2420 = 0.758 \cong 75.8\ \%$$

La probabilidad de que el peso del paquete de arroz escogido aleatoriamente sea de por lo menos 486 g es de 75.8 %. Obsérvese la Figura 5.9.

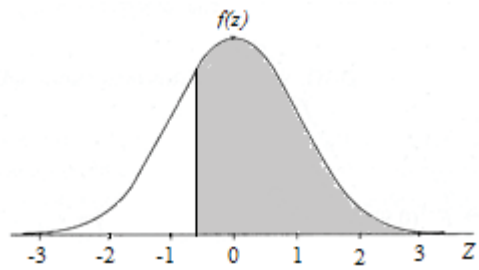

Figura 5.9. Área bajo la curva normal estándar $P(Z \geq -0.7)$

Fuente: los autores con la ayuda del *software* libre R.

$$b)P(X \leq 480) = P\left(\frac{X-\mu}{\sigma} \leq \frac{480-500}{20}\right) = P(Z \leq -1) = 0.1587 \cong 15.87\ \%.$$

La probabilidad de que el peso del paquete de arroz escogido aleatoriamente sea de máximo 480 g es de 15.87 %.Ver Figura 5.10.

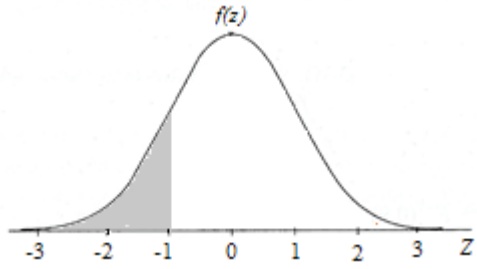

Figura 5.10. Área bajo la curva normal estándar $P(Z \leq -1)$

Fuente: los autores con la ayuda del *software* libre R.

c) $P(476 \leq X \leq 550) = P\left(\dfrac{476-500}{20} \leq Z \leq \dfrac{550-500}{20}\right)$

$$P(476 \leq X \leq 550) = P\left(-1.2 \leq Z \leq 2.5\right) = P(Z \leq 2.5) - P(Z \leq -1.2)$$

$$P(476 \leq X \leq 550) = 0.9938 - 0.1151 = 0.8787 \cong 87.87\ \%.$$

La situación anterior se presenta en la Figura 5.11.

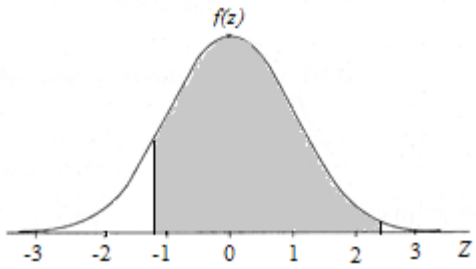

Figura 5.11. Área bajo la curva normal estándar $P(-1.2 \leq Z \leq 2.5)$

Fuente: los autores con la ayuda del *software* libre R.

La probabilidad de que el peso del paquete de arroz escogido aleatoriamente esté entre 476 y 550 g es del 87.87 %.

d) $P(X \leq 400) = P\left(\dfrac{X-\mu}{\sigma} \leq \dfrac{400-500}{20}\right) = P(Z \leq -5) \cong 0\ \%$

Esta situación se observa en la Figura 5.12.

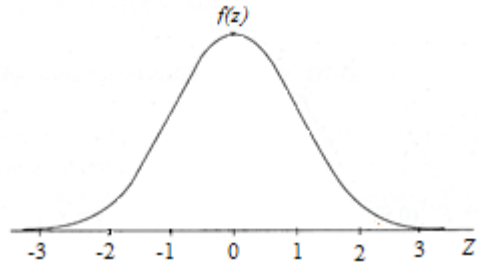

Figura 5.12. Área bajo la curva normal estándar $P(Z \leq -5)$

Fuente: los autores con la ayuda del *software* libre R.

La probabilidad de que el peso del paquete de arroz escogido aleatoriamente sea de menos de 400 g es de 0 %, aproximadamente.

$$e)\ P(X \geq 440) = P\left(\frac{X-\mu}{\sigma} \geq \frac{440-500}{20}\right) = P(Z \geq -3) = 1 - P(Z \leq -3)$$

$$P(X \geq 440) = 1 - 0.0013 = 0.9987 \cong 99.87\ \%$$

La probabilidad de que el peso del paquete de arroz escogido aleatoriamente sea de más de 440 g es de 99.87 %.Ver Figura 5.13.

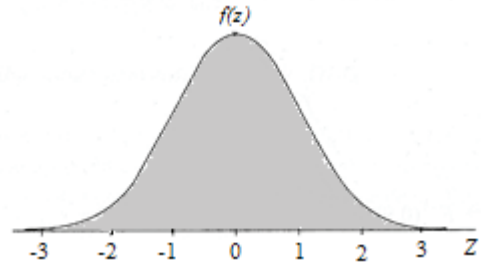

Figura 5.13. Área bajo la curva normal estándar $P(Z \geq -3)$

Fuente: los autores con la ayuda del *software* libre R.

2.4 Determinar los siguientes valores correspondientes a las probabilidades indicadas y elaborar una representación gráfica:

a) $\chi^2_{0.99,17} = ?$

b) $\chi^2_{0.025,12} = ?$

Para la situación *a)* se debe determinar la distancia *chi-cuadrado* correspondiente a una probabilidad de 0.99 con n=17 grados de libertad; esta corresponde al

valor 33.4087, es decir, $\chi^2_{0.99,17} = 33.4087$; una representación gráfica se indica en la Figura 5.14.

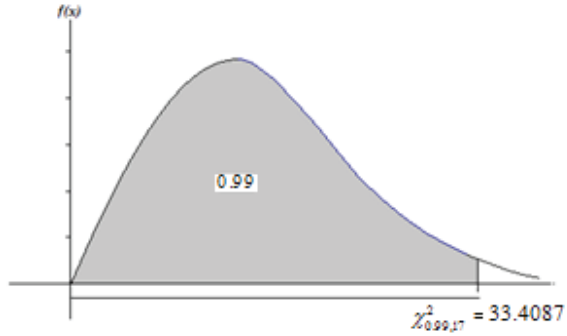

Figura 5.14 Área bajo la curva chi-cuadrado con *n*=17 grados de libertad
Fuente: los autores con la ayuda del *software* libre R

Para la situación *b)* se debe determinar la distancia *chi-cuadrado* correspondiente a una probabilidad de 0.025 con *n*=12 grados de libertad; esta corresponde al valor 4.404; es decir,

$\chi^2_{0.025,12} = 4.404$; la representación gráfica se observa en la Figura 5.15.

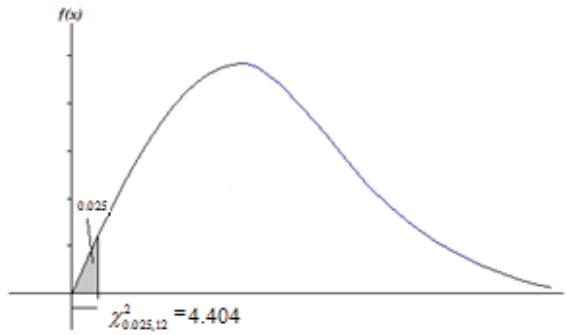

Figura 5.15 Área bajo la curva chi-cuadrado con *n* = 12 grados de libertad
Fuente: los autores con la ayuda del *software* libre R

2.5 Obtener las probabilidades y graficar:

a) $t_{0.99,16} = ?$

b) $t_{0.05,13} = ?$

Para el caso *a)* se debe determinar un valor en eje horizontal correspondiente a una probabilidad de 0.99 con *n*=16 grados de libertad, usando la distribución

t-student; este corresponde a 2.583, es decir, $t_{0.99,16} = 2.583$; una representación gráfica se observa en la Figura 5.16.

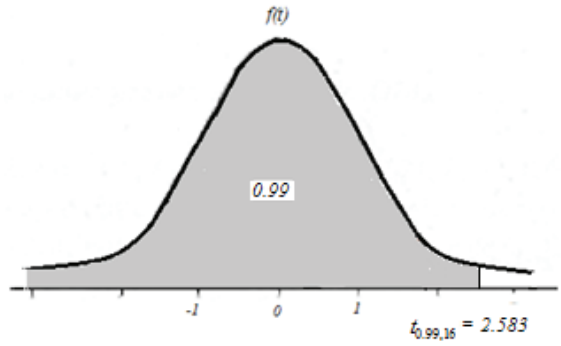

Figura 5.16 Área bajo la curva t-student con n = 16 grados de libertad
Fuente: los autores con la ayuda del software libre R

Para el caso *b)* se debe determinar un valor en eje horizontal correspondiente a una probabilidad de 0.05 con n=13 grados de libertad, usando la distribución *t-student;* este corresponde a 1.771; es decir, $t_{0.05,13} = -1.771$; una representación gráfica se observa en la Figura 5.17.

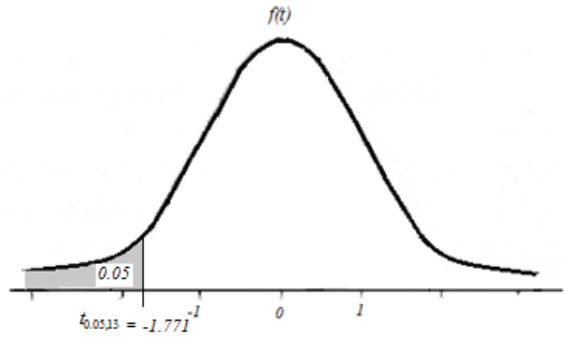

Figura 5.17 Área bajo la curva t-student con n=13 grados de libertad
Fuente: los autores con la ayuda del software libre R

2.6 Leer el valor de probabilidad en la distribución de Fisher y elaborar un gráfico:

a) $F_{0.99,12,10} = ?$

b) $F_{0.05,7,7} = ?$

Para el caso *a)* se debe determinar un valor en eje horizontal correspondiente a una probabilidad de 0.99 con m=12 y n=10 grados de libertad, usando la

distribución de *Fisher;* este corresponde a 4.71, es decir, $F_{0.99,12,10} = 4.71$; la representación gráfica se indica en la Figura 5.18.

Figura 5.18 Área bajo la curva de Fisher con *m*= 12 y *n*=10 grados de libertad

Fuente: los autores con la ayuda del software libre R.

Para la situación *b)* se debe determinar un valor en eje horizontal correspondiente a una probabilidad de 0.05 con *m*=7 y *n*=7 grados de libertad, usando la distribución de *Fisher;* este corresponde a 0.2638 y se obtiene así:

$F_{0.05,7,7} = \dfrac{1}{F_{0.95,7,7}} = \dfrac{1}{3.79} \cong 0.2638$; la representación gráfica se presenta en la Figura 5.19.

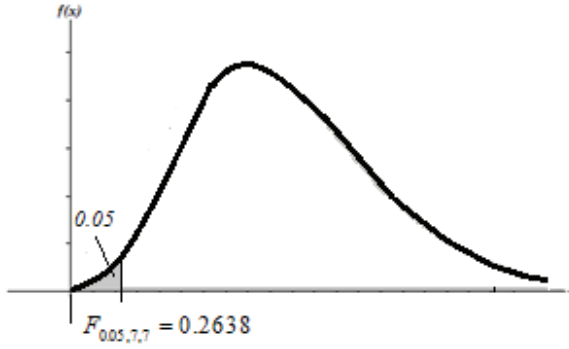

Figura 5.19 Área bajo la curva de Fisher con *m* =7 y *n* = 7 grados de libertad

Fuente: los autores con la ayuda del software libre R.

2.7 Para la variable *X: g*astos en transporte por día (miles de pesos) de los cinco mejores estudiantes de la universidad B, en la ciudad de Tunja en Boyacá-Colombia, en el año 2015, cuyos valores son: 8, 10, 14, 12, 6; *a)* determinar todas las muestras de tamaño *n*=2 que son posibles de seleccionar por medio de un muestreo aleatorio simple sin reemplazo, *b)* calcular el promedio para cada muestra, *c)* construir la distribución de probabilidad para la variable aleatoria

"media muestral", *d)* obtener su valor esperado y su varianza, *e)* verificar si se cumplen las igualdades para la esperanza matemática y la varianza presentadas por Freund y Miller (2000).

a) Como se tiene que N=5 y n=2, entonces, en este caso también el número total de muestras distintas por obtener es:

$$\binom{N}{n}=\binom{5}{2}=\frac{5!}{2!(5-2)!}=\frac{5!}{2!3!}=\frac{120}{12}=10$$

Estas muestras son las siguientes:

$$M_1=\{14,12\}, M_2=\{14,10\}, M_3=\{14,8\}, M_4=\{14,6\}, M_5=\{12,10\},$$

$$M_6=\{12,8\}, M_7=\{12,6\}, M_8=\{10,8\}, M_9=\{10,6\}, M_{10}=\{8,6\}$$

b) Las correspondientes medias muestrales son:

$$\bar{X}_1=\frac{14+12}{2}=13, \bar{X}_2=\frac{14+10}{2}=12, \bar{X}_3=\frac{14+8}{2}=11, \bar{X}_4=\frac{14+6}{2}=10, \bar{X}_5=\frac{12+10}{2}=11$$

$$\bar{X}_6=\frac{12+8}{2}=10, \bar{X}_7=\frac{12+6}{2}=9, \bar{X}_8=\frac{10+8}{2}=9, \bar{X}_9=\frac{10+6}{2}=8, \bar{X}_{10}\frac{8+6}{2}=7$$

Los valores anteriores indican que la media muestral es una variable aleatoria; esta cambia de valor a medida que se cambia de muestra.

c) De acuerdo con lo establecido en el apartado 1.3 del primer capítulo, la función de probabilidad asociada a esta variable aleatoria es:

$$f\left(\bar{X}=13\right)=\frac{1}{10},\ f\left(\bar{X}=12\right)=\frac{1}{10},\ f\left(\bar{X}=11\right)=\frac{2}{10},\ f\left(\bar{X}=10\right)=\frac{2}{10},$$

$$f\left(\bar{X}=9\right)=\frac{2}{10},\ f\left(\bar{X}=8\right)=\frac{1}{10},\ f\left(\bar{X}=7\right)=\frac{1}{10}$$

d) El valor esperado o esperanza matemática de la variable aleatoria es:

$$E\left(\bar{X}\right)=13\left(\frac{1}{10}\right)+12\left(\frac{1}{10}\right)+11\left(\frac{2}{10}\right)+10\left(\frac{2}{10}\right)+9\left(\frac{2}{10}\right)+8\left(\frac{1}{10}\right)+7\left(\frac{1}{10}\right)$$

$$E(\bar{X}) = \frac{13+12+22+20+18+8+7}{10} = \frac{100}{10} = 10$$

Ahora, para calcular la varianza, en primera instancia se calcula:

$$E(\bar{X}^2) = 13^2\left(\frac{1}{10}\right) + 12^2\left(\frac{1}{10}\right) + 11^2\left(\frac{2}{10}\right) + 10^2\left(\frac{2}{10}\right) + 9^2\left(\frac{2}{10}\right) + 8^2\left(\frac{1}{10}\right) + 7^2\left(\frac{1}{10}\right)$$

$$E(\bar{X}^2) = \frac{169+144+2(121)+2(100)+2(81)+64+49}{10}$$

$$E(\bar{X}^2) = \frac{169+144+242+200+162+64+49}{10} = \frac{1030}{10} = 103$$

Con los resultados anteriores se calcula la varianza así:

$$Var(\bar{X}) = E(\bar{X}^2) - (E(\bar{X}))^2 = 103 - (10)^2 = 103 - 100 = 3$$

Usando los cinco datos de la población se establecen los parámetros:

$$\mu = \frac{14+12+10+8+6}{5} = \frac{50}{5} = 10$$

$$\sigma^2 = \frac{(14-10)^2 + (12-10)^2 + (10-10)^2 + (8-10)^2 + (6-10)}{5}$$

$$\sigma^2 = \frac{4^2 + 2^2 + 0^2 + (-2)^2 + (-4)^2}{5} = \frac{16+4+0+4+16}{5} = \frac{40}{5} = 8$$

e) Por otro lado, en concordancia con Freund y Miller (2000), se cumple que:

$$E(\bar{X}) = 10 = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{8}{2}\left(\frac{5-2}{5-1}\right) = \frac{24}{8} = 3$$

2.8 La duración promedio de cierta marca de teclados es de 900 días, con una desviación estándar de 70 días, siempre que se usen 8 horas por día. Determinar la probabilidad de que una muestra aleatoria de 36 teclados tenga una duración promedio: *a)* comprendida entre 870 y 925 días, *b)* menor o igual a 910 días. En este caso se tiene:

$$\mu = 900,\ \sigma = 70,\ \ n = 36$$

Se define la variable:

$\bar{X}$: duración promedio en días de los teclados en la muestra

Así, entonces, para el caso *a)* se ha de usar la distribución de la media muestral, expresión 2.1, porque se conoce la desviación estándar poblacional:

$$P(870 \le \bar{X} \le 925) = P\left(\frac{870-\mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{925-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(\frac{870-900}{\frac{70}{\sqrt{36}}} \le Z \le \frac{925-900}{\frac{70}{\sqrt{36}}}\right)$$

$$P(870 \le \bar{X} \le 925) = P(-2.57 \le Z \le 2.14) = P(Z \le 2.14) - (Z \le -2.57)$$

$$P(870 \le \bar{X} \le 925) = 0.9838 - 0.0051 = 0.9787 \cong 97.87\ \%$$

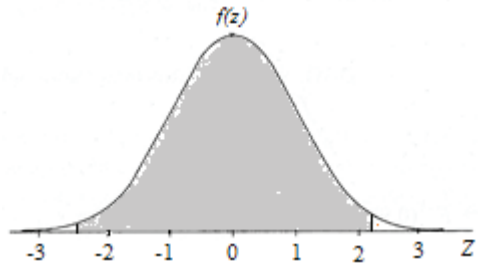

Figura 5.20. Área bajo la curva normal estándar $P(-2.57 \le Z \le 2.14)$

Fuente: los autores con la ayuda del software libre R.

La probabilidad de que en una muestra aleatoria de 36 teclados se obtenga una duración promedio entre 870 y 925 días es el 97.87 %, aproximadamente. Ver Figura 5.20.

Para el caso *b)* se tiene,

$$P(\bar{X} \le 910) = P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{910-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z \le \frac{910-900}{\frac{70}{\sqrt{36}}}\right)$$

$$P(\bar{X} \le 910) = P(Z \le 0.857) \cong P(Z \le 0.86) = 0.8051 \cong 80.51\ \%$$

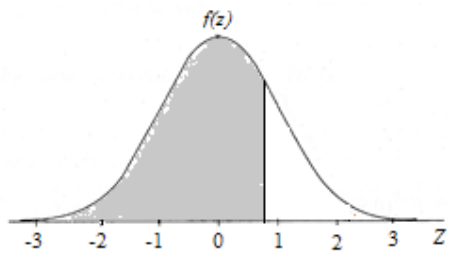

Figura 5.21. Área bajo la curva normal estándar $P(Z \leq 0.857)$

Fuente: los autores con la ayuda del software libre R.

La probabilidad de que una muestra aleatoria de 36 teclados tenga una duración promedio menor o igual a 910 días es del 80.51 %, aproximadamente. Ver Figura 5.21.

2.9 El tiempo promedio que gasta el bus urbano en la ciudad de Cali es de 70 minutos. Se toma una muestra aleatoria de 12 recorridos; con esos datos se obtuvo una desviación estándar corregida de 8 minutos. ¿Cuál es la probabilidad de que en esa muestra se tenga un tiempo promedio entre 64.94 y 76.84 minutos?

En este ejemplo se tiene:

$$\mu = 70,\ \hat{S} = 8,\ \ n = 12$$

Se define la variable:

$\overline{X}$: tiempo promedio en minutos que gasta el bus urbano en la muestra.

Así, entonces, se ha de usar una distribución de la media muestral correspondiente a la expresión 2.3, porque se desconoce la desviación estándar poblacional,

$$t = \frac{\overline{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \text{ con } n\text{ -1 } = 12 - 1 = 11 \text{ grados de libertad}$$

En efecto,

$$P(64.94 \leq \overline{X} \leq 96.84) = P\left(\frac{64.94 - 70}{\frac{8}{\sqrt{12}}} \leq \frac{\overline{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \leq \frac{76.84 - 70}{\frac{8}{\sqrt{12}}} \right)$$

$$P(64.94 \leq \overline{X} \leq 96.84) = P(-2.19 \leq t \leq 2.96)$$

$$P(64.94 \leq \overline{X} \leq 96.84) = P(t \leq 2.96) - P(t \leq -2.19) = 0.9935 - 0.0255 = 0.968 \cong 96.8\ \%$$

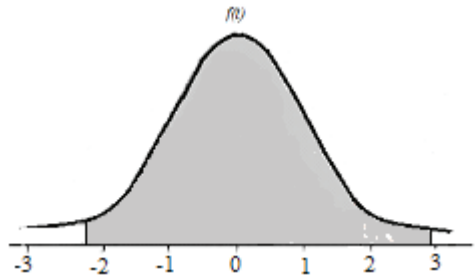

Figura 2.22. Área bajo la curva t-student con n = 11 grados de libertad
Fuente: los autores con la ayuda del software libre R.

La probabilidad de que en esa muestra se obtenga un tiempo promedio de recorrido para el bus urbano entre 64.94 y 76.84 minutos es del 96.8 %. Ver Figura 2.22. En este caso, los valores de probabilidad fueron obtenidos por medio del software libre R.

2.10 El 4 % de los artículos que produce una máquina son defectuosos; se toma una muestra aleatoria de 400 artículos. ¿Cuál es la probabilidad de que más del 5 % de los artículos de la muestra sean defectuosos?

Los datos asociados al problema planteado son:

n=400
p=4 % = 0.04
q=1 – p = 1 – 0.04 = 0.96

Se define la variable:

$\hat{p}$: porcentaje de artículos defectuosos producidos por la máquina en la muestra.

En este problema se ha de utilizar la distribución de la proporción muestral, que sigue una distribución normal estándar cuando el tamaño de la muestra es grande; esta se indicó en la expresión 2.5.

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

En efecto,

$$P(\hat{p} \geq 0.05) = P\left(\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \geq \frac{0.05 - p}{\sqrt{\frac{pq}{n}}} \right) = P\left(Z \geq \frac{0.05 - 0.04}{\sqrt{\frac{0.04 * 0.96}{400}}} \right)$$

$$P(\hat{p} \geq 0.05) = P(Z \geq 1.02) = 1 - p(Z \leq 1.02) = 1 - 0.8461 = 0.1539 \cong 15.93\ \%$$

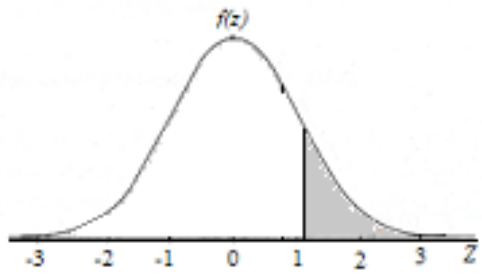

Figura 5.23. Área bajo la curva normal estándar $P(Z \geq 1.02)$

Fuente: los autores con la ayuda del software libre R.

La probabilidad de que el porcentaje de artículos defectuosos encontrados en la muestra supere el 5 % es del 15.39 %.Ver Figura 5.23.

2.11 En la ciudad A para niños de grado quinto de educación básica primaria se tiene un peso promedio $\mu_2 = 35\ kg$, con $\sigma_2^2 = 5$; mientras que para la ciudad B para niños que cursan este grado se tiene un peso promedio $\mu_1 = 45\ kg$, con $\sigma_1^2 = 8$; se toma en la ciudad A una muestra aleatoria de $n_2 = 50$, y otra en la ciudad B de $n_1 = 60$. ¿Cuál es la probabilidad de que la media muestral del peso de los niños de la ciudad B difiera de la de los niños de la ciudad A en más de 11 kg?

Los datos relacionados con la anterior situación son:

Ciudad A	Ciudad B
$\mu_2 = 35\ kg$	$\mu_1 = 45\ kg$
$\sigma_2^2 = 5$	$\sigma_1^2 = 8$
$n_2 = 50$	$n_1 = 60$

Se define la variable.

$\bar{X}_1 - \bar{X}_2$: diferencia de medias muestrales asociadas con el peso promedio de los

niños en las ciudades B y A, respectivamente, en sus correspondientes muestras.

En este caso se ha de utilizar la distribución de la diferencia de medias muestrales establecida en la expresión 2.7.

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Luego, entonces,

$$P(\bar{X}_1 - \bar{X}_2 \geq 11) = P\left(\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq \frac{11 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right)$$

$$P(\bar{X}_1 - \bar{X}_2 \geq 11) = P\left(Z \geq \frac{11 - (45 - 35)}{\sqrt{\frac{8}{60} + \frac{5}{50}}} \right) = P\left(Z \geq \frac{1}{\sqrt{0.233}} \right)$$

$$P(\bar{X}_1 - \bar{X}_2 \geq 11) = P(Z \geq 2.07) = 1 - P(Z \leq 2.07) = 1 - 0.9808 = 0.0192 \cong 1.92\ \%$$

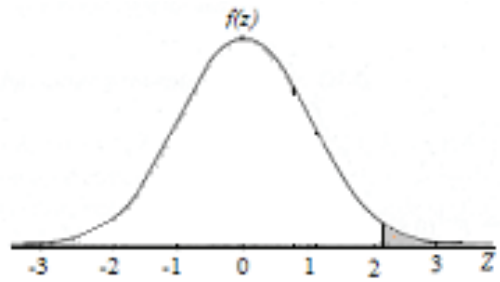

Figura 5.24. Área bajo la curva normal estándar $P(Z \geq 2.07)$

Fuente: los autores con la ayuda del software libre R.

La probabilidad de que la media muestral del peso de los niños de la ciudad B difiera en más de 11 kg de la media muestral del peso de los niños de la ciudad A es de 1.92 %. Ver Figura 5.24.

2.12 Un candidato a la presidencia de la república tiene el 60 % de la intención de voto en los potenciales votantes en el departamento de Nariño, y el 58 % en los del Valle del Cauca; se toma una muestra aleatoria de 400 votantes en

Nariño y de 500 en el Valle del Cauca. ¿Cuál es la probabilidad que la diferencia entre las proporciones muestrales de los potenciales votantes en Nariño y el Valle no superen el 3 %?

Los datos asociados con el anterior problema son:

Población 1: Nariño	Población 2: Valle
$p_1 = 60\ \% = 0.6$	$p_2 = 58\ \% = 0.58$
$n_1 = 400$	$n_2 = 500$

Se define la variable,

$\hat{p}_1 - \hat{p}_2$: diferencia de proporciones muestrales asociadas con la intención de voto de los potenciales votantes en las muestras de los departamentos de Nariño y Valle del Cauca, respectivamente.

En este problema se ha de usar la distribución de la diferencia de proporciones muestrales indicada en la expresión 2.6.

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

Por lo tanto,

$$P(\hat{p}_1 - \hat{p}_2 \leq 0.03) = P\left(\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \leq \frac{0.03 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \right)$$

$$P(\hat{p}_1 - \hat{p}_2 \leq 0.03) = P\left(Z \leq \frac{0.03 - (0.6 - 0.58)}{\sqrt{\frac{0.6 * 0.4}{400} + \frac{0.58 * 0.42}{500}}} \right)$$

$$P(\hat{p}_1 - \hat{p}_2 \leq 0.03) = P(Z \leq 0.3032) = 0.6179 \cong 61.79\%$$

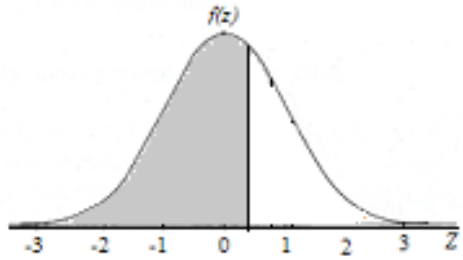

Figura 5.25. Área bajo la curva normal estándar $P(Z \leq 0.3032)$

Fuente: los autores con la ayuda del software libre R.

La probabilidad de que la diferencia entre las proporciones muestrales de los potenciales votantes por el candidato a la presidencia de la república en Nariño y Valle del Cauca no supere el 3 % es del 61.89 %, aproximadamente. Ver Figura 5.25.

Capítulo 3

3.1 Para la variable aleatoria X con distribución exponencial, determinar el estimador máximo verosímil. La función de densidad de probabilidad está dada por:

$$f(x,\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \quad \text{con } \lambda > 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Donde $\lambda = \theta$ es el parámetro de la variable aleatoria X con modelo exponencial.

Se toma una muestra aleatoria $X_1, X_2, ..., X_n$; la función de verosimilitud es:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

Aquí los x_i son observaciones correspondientes a las variables aleatorias $X_1, X_2, ..., X_n$.

Se trata de encontrar un valor del parámetro de tal manera que se maximice la función de verosimilitud,

$$L(\lambda) = \left(\lambda e^{-\lambda x_1}\right).\left(\lambda e^{-\lambda x_2}\right)....\left(\lambda e^{-\lambda x_n}\right)$$

$$L(\lambda) = \lambda^n e^{-\lambda(x_1 + x_2 + .. + x_n)}$$

Ahora, aplicando el logaritmo natural resulta:

$$Ln\left(L(\lambda)\right) = nLn(\lambda) - \lambda(x_1 + x_2 + ... + x_n)$$

A continuación, se realiza el cálculo de la derivada parcial con respecto al parámetro λ.

$$\frac{\partial Ln\left(L(\lambda)\right)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

Igualando a cero, resulta:

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

Por lo tanto,

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

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

Esta última expresión corresponde al estimador de máxima verosimilitud del parámetro λ de una variable aleatoria con distribución exponencial.

3.2 En la empresa azucarera AA, la cantidad de azúcar depositada por la máquina M en cada uno de los paquetes se distribuye normalmente, con desviación estándar de 2 g; de un lote de 500 paquetes se toma una muestra aleatoria de 25 paquetes (bolsas) empacados por tal máquina, y se encuentra un contenido promedio de 2500 g. Construir un intervalo de confianza del 98 % para estimar la verdadera media de empacado en las bolsas en el lote que se haga a través de la máquina M.

En este caso, se tiene:

$$N = 500$$
$$n = 25$$
$$\bar{X} = 2500$$
$$\sigma = 2$$

Además,

$$1-\alpha = 0.98 \rightarrow \alpha = 0.02$$

$$\frac{\alpha}{2} = 0.01$$

En la Figura 5.26 se observa el valor de Z obtenido al leer una tabla normal estándar para un 99 % de probabilidad.

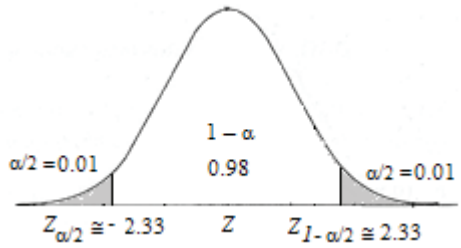

Figura 5.26 Valor de Z en una curva normal estándar
Fuente: los autores con la ayuda del software libre R.

$$Z_{1-\frac{\alpha}{2}} = Z_{0.99} \cong 2.33$$

Luego el intervalo de confianza es:

$$\mu \in \left(\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right)$$

Reemplazando los valores, resulta:

$$\mu \in \left(2500 - 2.33\left(\frac{2}{\sqrt{25}}\right)\sqrt{\frac{500-25}{500-1}}, 2500 + 2.33\left(\frac{2}{25}\right)\sqrt{\frac{500-25}{500-1}} \right)$$

Realizando las operaciones de tipo aritmético se obtiene:

$$\mu \in \left(2500 - 2.33(0.4)(0.97565), 2500 + 2.33(0.4)(0.97565) \right)$$

Luego

$$\mu \in \left(2500 - 0.9093, 2500 + 0.9093 \right)$$

Finalmente,

$$\mu \in \left(2499.09, 2500.91 \right)$$

En conclusión, con un nivel de confianza del 98 % se infiere que el promedio poblacional de empacado de las bolsas de azúcar por la máquina M está ente 2499.09 g y 2500.91 g, aproximadamente. Lo anterior indica que en 98 de cada 100 muestras tomadas se encuentra el parámetro media poblacional.

3.3 De un lote de 200 unidades del producto LM se ha seleccionado una muestra aleatoria de 10 unidades, obteniéndose un peso neto promedio de 1000 g, con una desviación estándar corregida de 5 g; si se asume que los datos provienen de una distribución normal, determinar un intervalo de confianza del 95 % para estimar el verdadero peso promedio de las unidades del producto en la población.

$$\hat{S} = 5 \qquad 1-\alpha = 95\ \% = 0.95$$
$$n = 10 \qquad \alpha = 0.05$$
$$\bar{X} = 1000 \qquad \frac{\alpha}{2} = 0.025$$

En la Figura 5.27 se observa el valor de *t* obtenido al leer una tabla de la distribución *t-student* con $n - 1 = 10 - 1 = 9$ grados de libertad para un 0.975=97.5 % de probabilidad.

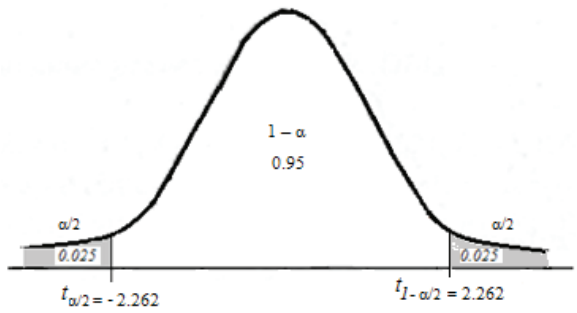

Figura 5.27 Valor de t en una curva t-student con n=9 grados

Fuente: los autores con la ayuda del software libre R.

$$t_{1-\frac{\alpha}{2}} = t_{0.975,9} = 2.262$$

Así, el intervalo de confianza es:

$$\mu \in \left(\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \bar{X} + t_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right)$$

Al reemplazar los valores se tiene:

$$\mu \in \left(1000 - 2.262(\frac{5}{\sqrt{10}})\sqrt{\frac{200-10}{200-1}}, 1000 + 2.262(\frac{5}{\sqrt{10}})\sqrt{\frac{200-10}{200-1}} \right)$$

Haciendo las operaciones de tipo aritmético se obtiene:

$$\mu \in (1000 - 2.262(1.5811)(0.9771), 1000 + 2.262(1.5811)(0.9771))$$

Entonces,

$$\mu \in (1000 - 3.4945, 1000 + 3.4945)$$

Por consiguiente,

$$\mu \in (996.50, 1003.49)$$

En conclusión, con un nivel de confianza del 95 % se infiere que el peso promedio poblacional de las unidades del producto LM está entre 996.50 g y 1003.49 g, aproximadamente. Lo anterior indica que en 95 muestras, de cada 100 tomadas, se encuentra el parámetro μ.

3.4 En el departamento de Boyacá, Colombia, se tomó una muestra aleatoria de 500 ciudadanos y se les preguntó si pertenecen o no a la población económicamente activa de este departamento; 350 de los encuestados respondieron que sí pertenecen a esta población. Construir un intervalo de confianza del 99 % para estimar la verdadera proporción de ciudadanos que pertenecen a la población económicamente activa de este departamento.

En primera instancia, se define así *X:* número de ciudadanos en la muestra de 500 que sí pertenecen a la población económicamente activa, luego:

$$n = 500$$
$$\hat{p} = \frac{x}{n} = \frac{350}{500} = 0.7$$
$$\hat{q} = 1 - 0.7 = 0.3$$

Adicionalmente,

$$1 - \alpha = 0.99 \rightarrow \alpha = 0.01$$
$$\frac{\alpha}{2} = 0.005$$

En la Figura 5.28 se observa el valor de Z obtenido al leer una tabla normal estándar para un 0.995 = 99.5 % de probabilidad.

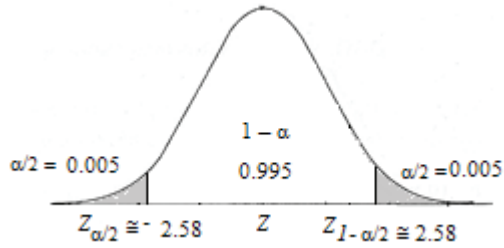

Figura 5.28 Valor de Z en una curva normal estándar
Fuente: los autores con la ayuda del software libre R.

$$Z_{1-\frac{\alpha}{2}} = Z_{0.995} \cong 2.58$$

Luego el intervalo de confianza es:

$$p \in \left(\hat{p} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \ , \ \hat{p} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$$

Reemplazando los valores resulta:

$$p \in \left(0.7 - 2.58\sqrt{\frac{0.7*0.3}{500}}, 0.7 + 2.58\sqrt{\frac{0.7*0.3}{500}} \right)$$

Realizando las operaciones de tipo aritmético se obtiene:

$$p \in \left(0.7 - 2.58(0.02049), 0.7 + 2.58(0.02049)\right)$$

Luego

$$p \in \left(0.7 - 0.05286, 0.7 + 0.05286\right)$$

En consecuencia,

$$p \in \left(0.6471, 0.7529\right)$$

En conclusión, con un nivel de confianza del 99 % se establece que la verdadera proporción de ciudadanos del departamento de Boyacá se encuentra, aproximadamente, entre el 64.71 % y el 75.29 %. Lo anterior indica que en 99 de cada 100 muestras tomadas se encuentra el parámetro p.

3.5 En un centro de distribución de computadores se ofrecen computadores de dos marcas diferentes en un periodo de tiempo específico; se selecciona

aleatoriamente un mes y se encuentra que se venden 350 computadores, de un total de 500, de la marca A, y 333, de un total de 450, de la marca B. Determinar un intervalo de confianza del 98 % para estimar la diferencia entre las verdaderas proporciones de las marcas A y B de computadores que se venden en todo el mercado en ese mes.

En este caso se tiene:

Marca A	Marca B
$n_1 = 500, \;\; x_1 = 350$	$n_2 = 450, x_2 = 333$
$\hat{p}_1 = \frac{x_1}{n_1} = \frac{350}{500} = 0.7$	$\hat{p}_2 = \frac{x_2}{n_2} = \frac{333}{450} = 0.74$
$\hat{q}_1 = 1 - 0.7 = 0.3$	$\hat{q}_2 = 1 - 0.74 = 0.26$

Además,

$$1-\alpha = 0.98 \rightarrow \alpha = 0.02$$

$$\frac{\alpha}{2} = 0.01$$

En la Figura 5.29 se observa el valor de Z obtenido al leer una tabla normal estándar para un 0.99 = 99 % de probabilidad.

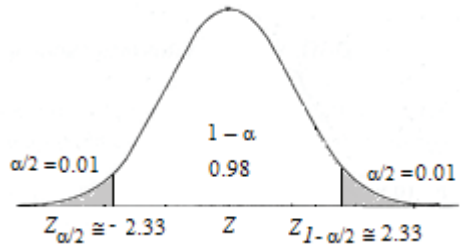

Figura 5.29 Valor de Z en una curva normal estándar
Fuente: los autores con la ayuda del software libre R.

$$Z_{1-\frac{\alpha}{2}} = Z_{0.99} \cong 2.33$$

Luego, en concordancia con la expresión 3.8, el intervalo de confianza es:

$$p_1 - p_2 \in \left((\hat{p}_1 - \hat{p}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \;,\; (\hat{p}_1 - \hat{p}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \right)$$

Reemplazando los valores, resulta:

$$\left((0.7-0.74) - 2.33\sqrt{\frac{(0.7)(0.3)}{500} + \frac{(0.74)(0.26)}{450}} \ , \ (0.7-0.74) + 2.33\sqrt{\frac{(0.7)(0.3)}{500} + \frac{(0.74)(0.26)}{450}} \right)$$

Realizando las operaciones de tipo aritmético se obtiene:

$$p_1 - p_2 \in \left((0.7-0.74) - 2.33\sqrt{0.00042 + 0.000427} \ , \ (0.7-0.74) + 2.33\sqrt{0.00042 + 0.000427} \ \right)$$

Luego

$$p_1 - p_2 \in \left((0.7-0.74) - 2.33(0.0291) \ , \ (0.7-0.74) + 2.33(0.0291) \right)$$

En consecuencia,

$$p_1 - p_2 \in \left(-0.04 - 0.0678 \ , \ -0.04 + 0.0678 \right)$$

Finalmente,

$$p_1 - p_2 \in \left(-0.1078, \ 0.0278 \right)$$

En conclusión, con un nivel de confianza del 98 % se infiere que la verdadera diferencia de proporciones poblacionales se encuentra entre el -10.78 % y el 2.78 %, aproximadamente. Lo anterior indica que en 98 de cada 100 muestras tomadas se encuentra el parámetro $p_1 - p_2$. La verdadera diferencia de proporciones de las marcas A y B de computadores que se venden en todo el mercado en ese mes está entre los porcentajes indicados.

3.6 Se analiza el contenido de oro presente en una aleación. En una muestra de 40 circuitos integrados especiales se encontró un contenido medio de 5.8 u.i de oro, con una desviación estándar de 0.6 u.i; asimismo, se inspecciona el contenido de oro en otra muestra aleatoria de 50 circuitos integrados corrientes, detectándose un contenido promedio de 5 u.i, con una desviación estándar de 0.8 u.i; se supone que las muestra provienen de poblaciones normales. Construir un intervalo de confianza del 95 % para estimar la diferencia de contenidos medios de oro de la primera clase de circuito con respecto a la segunda.

En este caso se tiene:

Clase 1	Clase 2
$n_1 = 40$	$n_2 = 50$
$\overline{X}_1 = 5.8$	$\overline{X}_2 = 5$
$S_1 = 0.6$	$S_2 = 0.8$

El valor de la desviación estándar corregida para la clase 1 "circuitos integrados especiales" se obtiene de la siguiente manera:

$$\hat{S}_1^2 = \frac{n_1}{n_1 - 1} S_1^2 = \frac{40}{39}(0.6)^2 = 0.3692 \rightarrow \hat{S}_1 = 0.6076$$

De modo similar,

$$\hat{S}_2^2 = \frac{n_2}{n_2 - 1} S_2^2 = \frac{50}{49}(0.8)^2 = 0.6530 \rightarrow \hat{S}_2 = 0.8080$$

Además,

$$1 - \alpha = 0.95 \rightarrow \alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

En la Figura 5.30 se observa el valor aproximado de Z, obtenido al leer una tabla normal estándar para una probabilidad acumulada de 0.975=97.5 %. En este caso, para obtener el intervalo de confianza se recurre a la expresión 3.12.

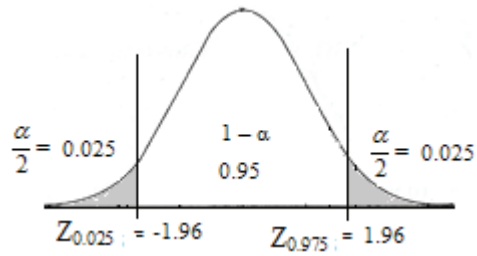

Figura 5.30 Intervalo de confianza sobre la curva normal estándar

Fuente: los autores con la ayuda del *software* libre R.

$$Z_{0.975} = 1.96$$

Luego, usando:

$$\mu_1 - \mu_2 \in \left((\bar{X}_1 - \bar{X}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}} \ , \ (\bar{X}_1 - \bar{X}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}} \right)$$

Y reemplazando los valores, resulta:

$$\mu_1 - \mu_2 \in \left((5.8 - 5) - 1.96 \sqrt{\frac{0.3692}{40} + \frac{0.6530}{50}} \ , \ (5.8 - 5) + 1.96 \sqrt{\frac{0.3692}{40} + \frac{0.6530}{50}} \right)$$

Efectuando las operaciones de tipo aritmético se obtiene:

$$\mu_1 - \mu_2 \in \left((5.8-5) - 1.96(0.149298), (5.8-5) + 1.96(0.149298)\right)$$

Luego

$$\mu_1 - \mu_2 \in (0.8 - 0.292624, 0.8 + 0.292624)$$

Finalmente,

$$\mu_1 - \mu_2 \in (0.5073\ ,\ 1.0926)$$

En conclusión, con un nivel de confianza del 95 % se infiere que la diferencia de medias poblacionales referidas al contenido de oro presente en cada una de las aleaciones de la primera clase de circuitos integrados con respecto a la segunda clase está entre 0.5073 y 1.0926 u.i. Lo anterior indica que en 95 de cada 100 muestras seleccionadas se encuentra el parámetro $\mu_1 - \mu_2$. Se deduce que el contenido promedio de oro en la segunda clase de circuitos es mayor que el contenido promedio de oro en la primera clase de circuitos.

3.7 El siguiente ejemplo ha sido adaptado de Canavos (1988). Un determinado procedimiento produce cierta clase de cojinetes de bola, cuyo diámetro interior es de 5 cm; se selecciona una muestra aleatoria de 12 de esos cojinetes; al medir sus diámetros internos se obtiene una desviación estándar corregida 0.03 cm. Bajo el supuesto de normalidad para los datos, construir un intervalo de confianza del 99 % para estimar la varianza poblacional; asimismo, determinar un intervalo de confianza para estimar la desviación estándar poblacional.

Los requerimientos son los siguientes:

$$\hat{S} = 0.03$$
$$n = 12$$

$$1 - \alpha = 0.99 \rightarrow \alpha = 0.01$$

$$\frac{\alpha}{2} = 0.005$$

$$1 - \frac{\alpha}{2} = 0.995$$

En la Figura 5.31 se observan los valores de los cuantiles obtenidos al leer una tabla *chi-cuadrado* $n-1$ =12 – 1= 11 grados de libertad para un 0.005 =0.5 % y un 0.995 = 99.5 % de probabilidad acumulada, respectivamente.

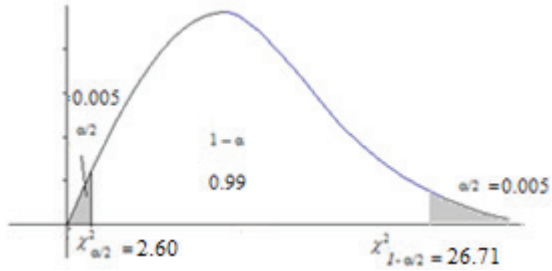

Figura 5.31 Intervalo de confianza sobre la curva chi-cuadrado

Fuente: los autores con la ayuda del *software* libre R.

$$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.995,11} = 26.757$$

$$\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.005,11} = 2.603$$

Luego, usando la expresión 3.13:

$$\sigma^2 \in \left(\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}} \ , \ \frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}} \right)$$

Y reemplazando los valores, resulta:

$$\sigma^2 \in \left(\frac{(12-1)(0.03)^2}{26.757} \ , \ \frac{(12-1)(0.03)^2}{2.603} \right)$$

Efectuando las operaciones de tipo aritmético se obtiene:

$$\sigma^2 \in \left(\frac{0.0099}{26.757} \ , \ \frac{0.0099}{2.603} \right)$$

Por consiguiente,

$$\sigma^2 \in (0.00037 \ , \ 0.0038)$$

En conclusión, con un nivel de confianza del 99 % se infiere que la varianza poblacional está entre 0.00037 y 0.0038. Lo anterior indica que en 99 de cada 100 muestras seleccionadas se encuentra el parámetro σ^2.

Ahora, para estimar la desviación estándar poblacional se usa la expresión 3.14:

$$\sigma \in \left(\sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{1-\frac{\alpha}{2}}}} \ , \ \sqrt{\frac{(n-1)\hat{S}^2}{\chi^2_{\frac{\alpha}{2}}}} \right)$$

Al sustituir los valores pertinentes se obtiene:

$$\sigma \in \left(\sqrt{\frac{(12-1)(0.03)^2}{26.757}} \ , \ \sqrt{\frac{(12-1)(0.03)^2}{2.603}} \right)$$

Así, entonces,

$$\sigma \in \left(\sqrt{0.00037} \ , \sqrt{0.0038} \right)$$

Por lo tanto,

$$\sigma \in (0.01923 \ , 0.0616)$$

En conclusión, con un nivel de confianza del 99 % se infiere que la desviación estándar poblacional está entre 0.01923 y 0.0616 cm. Lo anterior indica que en 99 de cada 100 muestras seleccionadas se encuentra el parámetro σ.

3.8 Se tiene la creencia de que los egresados de la titulación de Administración de Empresas obtienen un salario promedio mayor que el de los egresados de la titulación de Economía; además, se quiere saber si la variación de sus correspondientes salarios difiere. Para comprobarlo se ha tomado una muestra aleatoria de 10 administradores, obteniéndose una media muestral de 2 600 000 pesos por mes, con una varianza corregida de 1 200 000, y una muestra aleatoria de 13 economistas, obteniéndose un promedio de 2 400 000 pesos por mes, con una varianza de 1 300 000; se supone que los dos conjuntos de datos provienen de muestras independientes seleccionadas de poblaciones normales. Construir un intervalo de confianza del 98 % para estimar el cociente de varianzas poblacionales.

En este caso se tiene:

Administradores	Economistas
$n_1 = 10$	$n_2 = 13$
$\bar{X}_1 = 2\ 600\ 000$	$\bar{X}_2 = 2\ 400\ 000$
$\hat{S}_1^2 = 1\ 200\ 000$	$\hat{S}_2^2 = 1\ 300\ 000$

Además,

$$1-\alpha = 0.98 \rightarrow \alpha = 0.02$$

$$\frac{\alpha}{2} = 0.01$$

En la Figura 5.32 se presentan los valores de los cuantiles obtenidos al leer una tabla *F* de Fisher con n_1-1=10-1=9 grados de libertad para el numerador, y n_2-1=13-1=12 grados de libertad para el denominador; estos corresponden a un 0.01=1 % y 0.99=99 % de probabilidad.

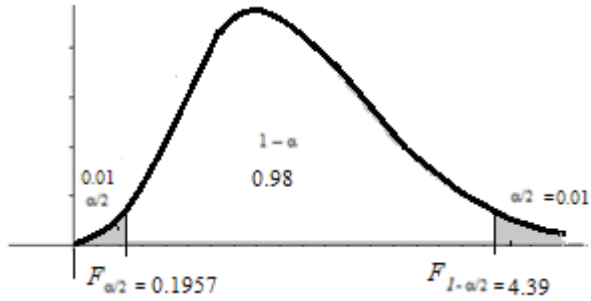

Figura 5.32 Intervalo de confianza sobre la curva de Fisher

Fuente: los autores con la ayuda del *software* libre R.

$$F_{1-\frac{\alpha}{2}} = F_{0.99,9,12} = 4.39$$

$$F_{\frac{\alpha}{2}} = F_{0.01,9,12} = \frac{1}{F_{0.99,12,9}} = \frac{1}{5.11} = 0.1957$$

Luego, usando la expresión 3.15:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{\hat{S}_1^2}{\hat{S}_2^2 F_{1-\frac{\alpha}{2}}} , \frac{\hat{S}_1^2}{\hat{S}_2^2 F_{\frac{\alpha}{2}}} \right)$$

Al reemplazar los respectivos valores resulta:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{1200000}{1300000(4.39)} , \frac{1200000}{1300000(0.1957)} \right)$$

Efectuando las operaciones de tipo aritmético se obtiene,

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left(\frac{1\ 200\ 000}{5\ 707\ 000} , \frac{1\ 200\ 000}{254\ 410} \right)$$

Por consiguiente,

$$\frac{\sigma_1^2}{\sigma_2^2} \in (0.21 , 4.7167)$$

En conclusión, con un nivel de confianza del 98 % se infiere que el cociente de las varianzas poblacionales está entre 0.21 y 4.7167. Lo anterior indica que en 98 de cada 100 muestras seleccionadas se encuentra el parámetro σ_1^2 / σ_2^2. Como el intervalo de confianza contiene el valor 1, entonces cabe la posibilidad de que las varianzas poblacionales resulten iguales; asimismo, se evidencia que el cociente de varianzas también es inferior a 1; esto indica que la varianza del salario de los administradores es inferior a la de los economistas, y como además el cociente de varianzas es mayor que 1, se tiene que la varianza del salario de los administradores resulta mayor que la de los economistas.

Capítulo 4

4.1 Un fabricante de llantas afirma que sus cauchos duran hasta que, en promedio, rueden 21 000 km, pero un distribuidor potencial de estos considera que duran menos; por eso quiere verificar la afirmación del fabricante, y de un lote de 200 llantas selecciona 11 de manera aleatoria y las pone a rodar, encontrando un promedio de 20 500 kilómetros. Si el fabricante informa que las llantas que ha producido tienen una desviación estándar poblacional de 1000 kilómetros, probar la afirmación del fabricante usando un nivel de significación del 2.5 %.

1. Planteamiento de hipótesis

Ho: $\mu = 21\ 000$

H_1: $\mu < 21\ 000$

2. Nivel de significación $\alpha = 0.025$

3. La dirección de la prueba y la región crítica se presentan en la Figura 5.33.

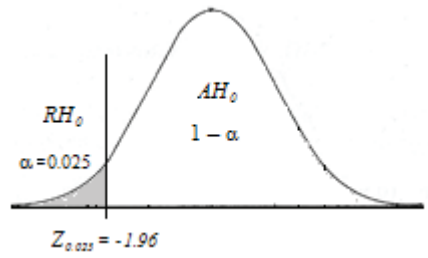

Figura 5.33 Prueba unilateral izquierda para la media poblacional

Fuente: los autores con la ayuda del software libre R.

4. Para la estadística de prueba se usa la siguiente información:

$$\bar{X}=20\ 500 \qquad \sigma=1000 \qquad n=11$$

$$Z=\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}=\frac{20\ 500-21\ 000}{\frac{1000}{\sqrt{11}}\sqrt{\frac{200-11}{200-1}}}=\frac{-500}{315.51(0.9745)}=\frac{-500}{307.4644}=-1.6262$$

5. La estadística de prueba tiene un valor de -1.6262, que es mayor que el valor -1.96 en el modelo teórico correspondiente a una distribución normal estándar; por lo tanto, se ubica en la región AH_0 de aceptación de la hipótesis nula.

6. Decisión: se acepta la hipótesis Ho: $\mu=21\ 000$, es decir, no hay evidencias suficientes para rechazarla.

7. Por lo tanto, con un nivel de significancia del 2.5 % (confiabilidad del 97.5 %) se concluye que, en promedio, las llantas duran hasta rodar 21 000 kilómetros; en consecuencia, el fabricante tiene la razón.

4.2 Un representante estudiantil afirma que al 10 % de los estudiantes de la universidad B les sirven el almuerzo casi frío; los administradores del restaurante de esta universidad reconocen que algunas veces eso ha sucedido, sin embargo, consideran que el porcentaje es muy inferior al 10 %. En el intento de desvirtuar la afirmación del representante se toma una muestra aleatoria de 150 estudiantes de esta universidad, 12 de los cuales responden que alguna vez han recibido el almuerzo casi frío; probar la afirmación hecha por el representante, usando un nivel de significancia del 3 %.

1. Planteamiento de hipótesis

Ho: $p = 0.1$

H_1: $p < 0.1$

2. Nivel de significancia $\alpha = 0.03$

3. Dirección de la prueba: se trata de una prueba unilateral izquierda; la región crítica se observa en la Figura 5.34.

4. Estadística de prueba

Se tienen los siguientes datos provenientes de la muestra:

$$n = 150 \quad x = 12 \quad \hat{p} = \frac{x}{n} = \frac{12}{150} = 0.08$$

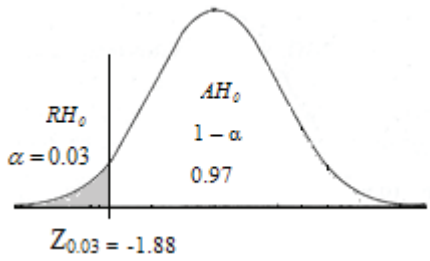

Figura 5.34 Prueba unilateral izquierda para la proporción poblacional

Fuente: los autores con la ayuda del software libre R.

La estadística de prueba por utilizar es:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.08 - 0.1}{\sqrt{\frac{(0.1)(0.9)}{150}}} = \frac{-0.02}{\sqrt{0.0006}} = \frac{-0.02}{0.024494} \cong -0.8165$$

En este caso, se ha de tenido en cuenta que $q_0 = 1 - p_0 = 1 - 0.1 = 0.9$.

5. El valor de la estadística de prueba, -0.8165, cae en la región AH_0, de aceptación de la hipótesis nula H_0, dado que es mayor que el valor de Z=-1.88 en la distribución teórica normal estándar.

6. Decisión: aceptar H_0: $p = 0.1$

7. Finalmente, con un nivel de significancia del 3 % se concluye que la afirmación hecha por el representante estudiantil de que al 10 % de los estudiantes de

la universidad B les sirven el almuerzo casi frío es cierta; no hay evidencias suficientes para rechazar tal afirmación.

4.3 Un estudio sobre el gusto por la práctica deportiva en hombres y mujeres reveló que a 171 de ellos, en una muestra aleatoria de 380, y a 168 de ellas, en una muestra aleatoria de 350, les gusta esta práctica; estos resultados se constituyen en suficiente evidencia para afirmar que el porcentaje de gusto por la práctica deportiva de los hombres difiere del de las mujeres; usar un nivel de significancia del 3 % para realizar la prueba de hipótesis.

1. Planteamiento de hipótesis

Ho: $p_1 = p_2$

H_1: $p_1 \neq p_2$

2. Nivel de significancia $\alpha = 0.03; \frac{\alpha}{2} = 0.015$

3. Dirección de la prueba: se trata de una prueba bilateral; la región crítica se indica en la Figura 5.35, donde se presenta una curva normal estándar.

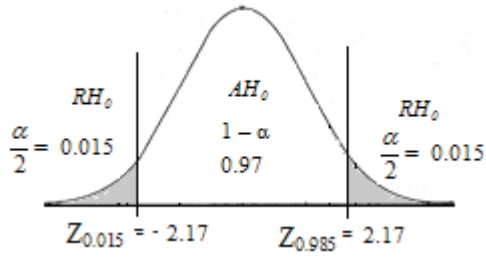

Figura 5.35 Prueba bilateral para la diferencia de proporciones poblacionales

Fuente: los autores con la ayuda del software libre R.

4. Estadística de prueba

En este ejemplo se tiene que:

Hombres	Mujeres
$n_1 = 380, \quad x_1 = 171$	$n_2 = 350, \quad x_2 = 168$
$\hat{p}_1 = \frac{x_1}{n_1} = \frac{171}{380} = 0.45$	$\hat{p}_2 = \frac{x_2}{n_2} = \frac{168}{350} = 0.48$
$\hat{q}_1 = 1 - 0.45 = 0.55$	$\hat{q}_2 = 1 - 0.48 = 0.52$

La estadística de prueba por utilizar es:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

En primera instancia, se realiza el cálculo siguiente:

$$p = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{380(0.45) + 350(0.48)}{380 + 350} = \frac{171 + 168}{730} = \frac{339}{730} \cong 0.4644$$

Luego, se determina el valor $q = 1 - p = 1 - 0.4644 = 0.5356$

Reemplazando los valores en la estadística de prueba, resulta:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.45 - 0.48}{\sqrt{(0.4644)(0.5356)\left(\frac{1}{380} + \frac{1}{350}\right)}} = \frac{-0.03}{\sqrt{0.001364}} = \frac{-0.03}{0.036932} = -0.8123$$

5. El valor de la estadística de prueba, -0.8123, cae en la región AH_0, de aceptación de la hipótesis nula H_0, puesto que se encuentra entre Z = -2.17 y Z = 2.17 en la distribución teórica normal estándar.

6. Decisión: aceptar H_0: $p_1 = p_2$

7. En consecuencia, con un nivel de significancia del 3 % se concluye que el porcentaje de gusto por la práctica deportiva de los hombres es igual al de las mujeres; no hay evidencias suficientes para aceptar que haya diferencias significativas en tales porcentajes.

4.4 El Director general de Cámaras de Comercio en Colombia desea determinar, al 5 % de significancia, si las utilidades en millones por mes de las pequeñas empresas en Cali y Bucaramanga son iguales; para esto toma una muestra aleatoria de 20 empresas pequeñas en Cali, y otra de 17 empresas pequeñas en Bucaramanga, y encuentra que la utilidad promedio en la primera es de 8 millones de pesos por mes, con una desviación estándar de 500 mil pesos, y en la segunda, de 7.5 millones de pesos por mes, con una desviación estándar de 400 mil pesos.

Esta prueba de hipótesis se ha de abordar de dos formas: *i)* considerando igualdad de varianzas poblacionales y *ii)* considerando varianzas poblacionales diferentes. A continuación, se desarrolla la prueba de hipótesis para el primer caso; el segundo se deja para que el lector se ejercite resolviéndolo.

1. Planteamiento de hipótesis

Ho: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

2. Nivel de significancia $\alpha = 0.05$; $\frac{\alpha}{2} = 0.025$

3. Se trata de una prueba bilateral; la región crítica se observa en la Figura 5.36. En esa figura se han de ubicar los valores del modelo teórico correspondiente a una distribución *t-student* con n_1+n_2-2=20+17-2= 35 grados de libertad, considerando igualdad de varianzas:

$$t_{\frac{\alpha}{2}} = t_{0.025,35} = -2.030 \qquad t_{1-\frac{\alpha}{2}} = t_{0.975,35} = 2.030$$

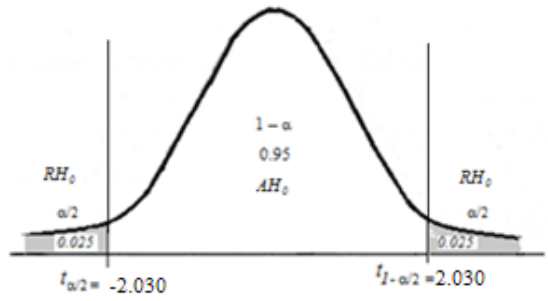

Figura 5.36 Prueba bilateral para la diferencia de medias poblacionales

Fuente: los autores con la ayuda del software libre R.

4. Estadística de prueba

En este ejemplo los datos muestrales son:

Cali	Bucaramanga
$n_1 = 20$	$n_2 = 17$
$\bar{X}_1 = 8$	$\bar{X}_2 = 7.5$
$S_1 = 0.5$	$S_2 = 0.4$

Por tratarse de muestras inferiores a 30 y por ser desconocidas las desviaciones estándar, la estadística de prueba por usar en la prueba de hipótesis es:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

No obstante, para calcular el valor de la estadística se requiere determinar las varianzas corregidas o cuasivarianzas y calcular el valor de S_p; tarea que se desarrolla a continuación:

$$\hat{S}_1^2 = \frac{n_1}{n_1 - 1} S_1^2 = \frac{20}{19}(0.5)^2 = 0.2631$$

$$\hat{S}_2^2 = \frac{n_2}{n_2 - 1} S_2^2 = \frac{17}{16}(0.4)^2 = 0.17$$

$$S_p = \sqrt{\frac{(20-1)(0.2631) + (17-1)(0.17)}{20 + 17 - 2}} = \sqrt{\frac{7.7189}{35}} = 0.4696$$

Reemplazando los valores en la estadística de prueba, resulta:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{8 - 7.5}{0.4696\sqrt{\frac{1}{20} + \frac{1}{17}}} = \frac{0.5}{0.4696(0.3298)} = \frac{0.5}{0.154874} = 3.2284$$

5. El valor de la estadística de prueba, 3.2282, cae en la región RH_0, de rechazo de la hipótesis nula H_0, puesto que es mayor que 2.030 en la distribución teórica *t-student,* con 35 grados de libertad.

6. Decisión: rechazar H_0: $\mu_1 = \mu_2$

7. Luego, con un nivel de significancia del 5 %, se concluye que las utilidades promedio en millones de pesos por mes de las pequeñas empresas en Cali y Bucaramanga difieren.

4.5 El diámetro en una muestra aleatoria de 10 varillas de hierro de media pulgada es una variable aleatoria con desviación estándar corregida de 0.2 milésimas de pulgada; el proceso de fabricado de las varillas de hierro se

encuentra bajo absoluto control si la varianza es de 0.09, y fuera de control si es mayor; un empleado que lleva el control de calidad sobre el diámetro de las varillas producidas considera que el proceso actualmente está fuera de control; probar esta hipótesis con un nivel de significancia del 5 %.

1. Planteamiento de hipótesis

$$\text{Ho}: \sigma^2 = 0.09$$
$$\text{H}_1: \sigma^2 > 0.09$$

2. Nivel de significancia $\alpha = 0.05$

3. Dirección de la prueba: se trata de una prueba unilateral derecha; la región crítica se observa en la Figura 5.37, donde se presenta una curva asimétrica correspondiente a una distribución *chi-cuadrado* para una probabilidad acumulada de 0.95 con 9 grados de libertad.

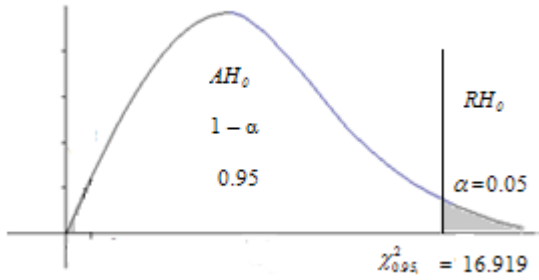

Figura 5.37 Prueba unilateral derecha para la varianza poblacional
Fuente: los autores con la ayuda del software libre R.

4. Estadística de prueba

En este ejemplo los datos muestrales son:

$$n = 10$$
$$\hat{S} = 0.2$$

Reemplazando los valores en la estadística de prueba, resulta:

$$\chi^2 = \frac{(n-1)\hat{S}^2}{\sigma_0^2} = \frac{(10-1)(0.2)^2}{0.09} = \frac{9(0.04)}{0.09} = \frac{0.36}{0.09} = 4$$

5. El valor de la estadística de prueba, 4, cae en la región AH_0, de aceptación de la hipótesis nula H_0, puesto que es menor que el valor teórico, 16.919, en la distribución *chi- cuadrado* con *n*-1=10-1=9 grados de libertad.

6. Decisión: aceptar H_0: $\sigma^2 = 0.09$; en consecuencia, se acepta que la varianza es igual a 0.09.

7. Por consiguiente, con un nivel de significancia del 5 % se concluye que el proceso de fabricación de las varillas en referencia al diámetro está bajo absoluto control; es decir, la consideración hecha por el empleado no tiene fundamento.

4.6 Para comparar la variabilidad en la duración de dos clases de calzado, A y B, se tomaron dos muestras aleatorias; la primera, de 6 pares, proporcionó una desviación estándar de 3 meses, y la segunda, de 5 pares, generó una desviación estándar corregida de 2.5 meses; bajo el supuesto de que las muestras provienen de poblaciones normales independientes, probar la hipótesis de que la varianza de la clase de calzado A es superior a la varianza de la clase de calzado B, usando un nivel de significancia del 5 %.

1. Planteamiento de hipótesis

Ho: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 > \sigma_2^2$

2. Nivel de significancia $\alpha = 0.05$

3. Dirección de la prueba: se trata de una prueba unilateral derecha; la región crítica se presenta en la Figura 5.38, donde se presenta una curva asimétrica correspondiente a una distribución *F* de *Fisher,* con 5 grados de libertad para el numerador y 4 grados para el denominador.

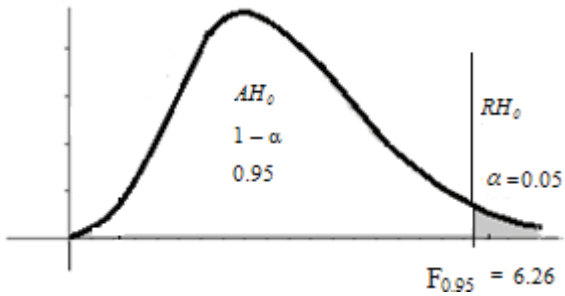

Figura 5.38 Prueba unilateral derecha para la igualdad de varianzas poblacionales

Fuente: los autores con la ayuda del software libre R.

4. Estadística de prueba

En este caso los datos de la muestra son:

Calzado A	Calzado B
$n_1 = 6$	$n_2 = 5$
$S_1 = 3$	$\hat{S}_2 = 2.5$

En primera instancia, conviene calcular la desviación estándar corregida correspondiente a los datos de la primera muestra:

$$\hat{S}_1^2 = \frac{n_1}{n_1 - 1} S_1^2 = \frac{6}{5}(3)^2 = 10.8$$

Reemplazando estos resultados en la estadística de prueba, se tiene:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2} = \frac{10.8}{(2.5)^2} = \frac{10.8}{6.25} = 1.728$$

5. El valor de la estadística de prueba, 1.728, cae en la región AH_0, de aceptación de la hipótesis nula H_0, puesto que es menor que el valor teórico, 6.26, en la distribución teórica *F* de *Fisher* con n_1-1=6-1=5 grados de libertad para el numerador y n_2 -1 = 5 -1 = 4 grados de libertad para el denominador.

6. Decisión: aceptar H_0: $\sigma_1^2 = \sigma_2^2$; implica que se acepta que las varianzas poblacionales son iguales.

7. Así, entonces, con un nivel de significancia del 5 % se concluye que la varianza de la duración de la clase de calzado A es igual a la varianza (poblacional) de la duración de la clase de calzado B; no hay evidencias suficientes para afirmar que las varianzas poblacionales sean distintas.

4.7 Se tiene una población finita de 600 personas; se seleccionó una muestra piloto del 2 % de la población total y se obtuvo una media muestral de 300 dólares para el salario promedio mensual; también se obtuvo de la muestra piloto una desviación estándar corregida de 50 dólares. Si se admite un error del 3 % de la media muestral, calcular el tamaño muestral al nivel de confianza del 95 %.

En este caso, los datos son:

$$1-\alpha = 0.95 \qquad N = 600$$

$$\alpha = 0.05 \rightarrow \frac{\alpha}{2} = 0.025 \qquad \hat{S} = 50$$

$$e = 9$$

Conviene en este caso utilizar la expresión 4.10; sin embargo, hace falta el valor de Z, que se obtiene de una distribución normal estándar para una probabilidad acumulada del 0.975, obteniéndose un valor de 1.96, como se observa en la Figura 5.39.

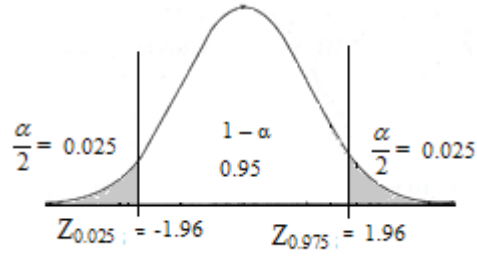

Figura 5.39 Valor de Z para un nivel de confianza del 95 %
Fuente: los autores con la ayuda del software libre R.

Reemplazando los valores, resulta:

$$n = \frac{NZ^2\hat{S}^2}{(N-1)e^2 + Z^2\hat{S}^2} = \frac{600(1.96)^2(50)^2}{599(9)^2 + (1.96)^2(50)^2} = \frac{600(3.8416)(2500)}{599(81) + (3.8416)(2500)}$$

$$n = \frac{5762400}{48519 + 9604} = \frac{5762400}{58123} \cong 99.14$$

En el contexto anterior, el tamaño de la muestra para estimar el salario promedio mensual, usando un nivel de confianza del 95 % es de 99 individuos. Esta cantidad de individuos se han de seleccionar apropiadamente usando métodos aleatorios de muestreo.

4.8 Se requiere indagar el mercado para el producto H; se sabe que el 60 % de los individuos de la ciudad A utilizan diariamente este producto; calcular el tamaño de la muestra si se desea que el error máximo absoluto sea del 4 % con un nivel de confianza del 96 %.

En este caso, se tiene que:

$$1-\alpha = 0.96$$

$$\alpha = 0.04 \rightarrow \frac{\alpha}{2} = 0.02$$

$$e = 0.04$$

$$p = 0.6$$

En estas circunstancias se recomienda utilizar la expresión 4.1; el valor de Z=2.055 se obtiene de una distribución normal estándar para una probabilidad acumulada del 0.98, como se indica en la Figura 5.40.

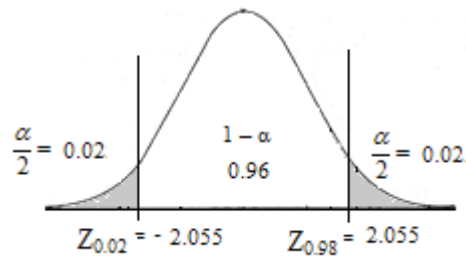

Figura 5.40 Valor de Z para un nivel de confianza del 96 %
Fuente: los autores con la ayuda del software libre R.

Reemplazando los valores, resulta:

$$n = \frac{Z^2 pq}{e^2} = \frac{(2.055)^2(0.6)(0.4)}{(0.04)^2} = \frac{(4.223)(0.24)}{0.0016} = \frac{1.01352}{0.0016} \cong 633.45$$

En estas circunstancias, el tamaño de la muestra para indagar el mercado del producto H en la ciudad A es de 633 individuos, admitiendo un error del 4 % y un nivel de confianza del 96 %. Este número de individuos se han de seleccionar de forma pertinente, utilizando un método de muestreo apropiado.

GLOSARIO DE SÍMBOLOS

N: tamaño de la población o número de individuos que la conforman

n: tamaño de la muestra o número de individuos que la constituyen

E: conjunto

X, Y, Z: variables numéricas o aleatorias

Ω : espacio muestral o conjunto de todos los resultados posibles de un experimento aleatorio

$\emptyset$: conjunto vacío

$\leq, \geq$: menor o igual, mayor o igual, respectivamente

$\neq$: diferente

$\cong$: aproximadamente igual

$\Im$: sigma álgebra o familia de subconjuntos del espacio muestral

P: medida de probabilidad

$(\Omega, \Im, P)$: indica un espacio de probabilidad.

R: conjunto de los números reales

β: sigma álgebra de *Borel* sobre el conjunto de los números reales

(R, β): espacio medible sobre los números reales

$X^{-1}(E)$: denota la imagen inversa del conjunto E a través de la variable aleatoria X

R_X : simboliza el rango de la variable aleatoria X

$P_X(B)$: probabilidad inducida por la variable aleatoria X

$f_X(x)$: función de probabilidad o de densidad de probabilidad

$\sum_{x_i \in R_X} f(x_i) = 1$: suma de los valores de la función f sobre el rango de la variable aleatoria X

$F_X(x) = P(X \leq x)$: función de distribución de probabilidad de la variable aleatoria X

$X: \ N(0,1)$: la variable aleatoria X tiene distribución normal estándar

μ : media o promedio poblacional

σ^2 : varianza poblacional

σ : desviación estándar poblacional

CV: coeficiente de variación poblacional

p : proporción poblacional

$\overline{X}$: media o promedio muestral

S^2 : varianza muestral

S: desviación estándar muestral

$\hat{S}^2$: varianza corregida o cuasivarianza

$\hat{S}$: desviación estándar corregida o cuasidesviación estándar

cv : coeficiente de variación muestral

$\hat{p}$: proporción o porcentaje muestral

$\binom{N}{n}$: combinatorio de los N elementos tomados de n en n.

$n!$: factorial de n

$x \in R$: x pertenece al conjunto de los números reales

$E(X)$: valor esperado o media de la variable aleatoria X

$Var(X)$: varianza de la variable aleatoria X

$F(z) = \Phi(z)$: distribución normal estándar

$P(Z \leq 1.64)$: probabilidad de Z menor o igual al valor 1.64 en la normal estándar

$\chi^2_{0.99,15} = 30.578$: valor de la *chi-cuadrado* para una probabilidad del 99 % y 15 grados

$t_{0.95,24} = 1.71$: valor de la *t-student* para una probabilidad del 95 % y 24 grados de libertad

$F_{0.99,7,4} = 14.98$: valor de la F de Fisher para una probabilidad del 99 % con 7 y 4 grados

θ : parámetro en el espacio de parámetros $\Theta \subseteq R^n$ como subconjunto de R a la n

T, $\hat{\theta}$: estimador del parámetro θ

$1 - \alpha$: nivel de confianza

$\hat{p}_1 - \hat{p}_2$: diferencia de proporciones o de porcentajes muestrales

$p_1 - p_2$: diferencia de proporciones o de porcentajes poblacionales.

$\overline{X}_1 - \overline{X}_2$: diferencia de promedios o de medias muestrales

$\mu_1 - \mu_2$: diferencia de promedios o de medias poblacionales

$\prod_{i=1}^{n} f(x_i)$: productoria desde i=1 hasta *n* de los valores de la función f en x_i

$Z = \dfrac{\bar{X} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$: estadística de prueba para el test de la media con σ conocida

$t = \dfrac{\bar{X} - \mu_0}{\dfrac{\hat{S}}{\sqrt{n}}}$: estadística de prueba para el test de la media con σ desconocida

$Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}$: estadística de prueba para el test de la proporción

$Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$: estadística de prueba para el test de la diferencia de proporciones

$Z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$: estadística de prueba para el test de la diferencia de medias

$\chi^2 = \dfrac{(n-1)\hat{S}^2}{\sigma_0^2}$: estadística de prueba para el test de la varianza

$F = \dfrac{\hat{S}_1^2}{\hat{S}_2^2}$: estadística de prueba para el cociente de varianzas

$e = \hat{p} - p$: error máximo absoluto para la proporción

$n = \dfrac{Z^2 pq}{e^2}$: tamaño de la muestra para la proporción en población infinita

$n = \dfrac{NZ^2 pq}{(N-1)e^2 + Z^2 pq}$: tamaño de la muestra para la proporción en población finita

$e = \bar{X} - \mu$: error máximo absoluto para la media

$n = \frac{Z^2\sigma^2}{e^2}$: tamaño de la muestra para la media en población infinita con σ conocida

$n = \frac{Z^2\hat{S}^2}{e^2}$: tamaño de la muestra para la media en población infinita, con σ desconocida

$n = \frac{NZ^2\sigma^2}{(N-1)e^2 + Z^2\sigma^2}$: tamaño de la muestra de la media en población finita, σ conocida

$n = \frac{NZ^2\hat{S}^2}{(N-1)e^2 + Z^2\hat{S}^2}$: tamaño de la muestra, en población finita, σ desconocida

REFERENCIAS

Aliaga, M., & Gunderson, B. (2005). *Interactive Statistics*. (3rd Edition). Prentice Hall.

Alvarado, J., & Obagi, J. (2008). *Fundamentos de inferencia estadística*. Bogotá: Pontificia Universidad Javeriana.

Alvermann, D. (1998). *Estrategias para enseñar a aprender: un enfoque cognitivo para todas las áreas y niveles*. Buenos Aires: Aique.

Ares, V. M. (1999). La prueba de significación de la «hipótesis cero» en las investigaciones por encuesta. *Metodología de encuestas*, *1*(1), 47-68.

Aubone, A., & Wöhler, O. C. (2000). *Aplicación del método de máxima verosimilitud a la estimación de parámetros y comparación de curvas de crecimiento de von Bertalanffy*.

Ball, D. L., Thames, M. H, & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389-407.

Batanero, C. (2001). *Didáctica de la estadística*. Granada: Grupo de Investigación en Educación Estadística. Universidad de Granada, España.

Bickel, P., & Doksum, K. (1977). *Mathematical Statistics: Basic ideas and select topics*. San Francisco: Holden-Day.

Blanco, L. (2004). *Probabilidad*, Colección texto. Bogotá: Facultad de Ciencias, Universidad Nacional de Colombia.

Botero, D. O. (2001). *Introducción al muestreo*. Bogotá: Editorial Unibiblos, Universidad Nacional de Colombia.

Burbano, V. M., Valdivieso, M. A. y Salcedo, L. A. (2014). *Simulación con modelos aleatorios: conocimiento estadístico-probabilístico y simulación*. Tunja: Uptc.

Burbano, V. M. y Valdivieso, M. A. (2015). *Elementos de probabilidad. Apoyo al estudio independiente*. Tunja: Uptc.

Cabria, S. G. (1994). *Filosofía de la estadística* (Vol. 26). Universitat de València.

Campos, A. A. (2009). *Métodos mixtos de investigación*. Bogotá: Magisterio.

Canavos, G. (1988). *Probabilidad y estadística: Aplicaciones y métodos.* México: McGraw-Hill.

Cao, R., & Van Keilegom, I. (2006). Empirical Likelihood Tests for Two-Sample Problems via Nonparametric Density Estimation. *Canad. J. Statist*. 34, 61-77.

Carrasco, A. S. (2005). *Aproximación a la Estadística desde las Ciencias Sociales.* Universidad de Valencia. Recuperado de http://www.uv.es/~carrascs/PDF/aproxi macion%20 estadistica.pdf

Denzin & Lincoln, Y. S. (Eds.), *Handbook of qualitative research* (105-117). Thousand Oaks, California: Sage Publications, Inc.

Devore, J. L. (2008). *Probabilidad y estadística para ingenierías y ciencias*. Cengage Learning Editores.

Fernández, S. P., & Díaz, S. P. (2004). Asociación de variables cualitativas: test de Chi-cuadrado. *Metodología de la Investigación*, *1*, 5.

Flores, J. G., Gómez, G. R., y Jiménez, E. G. (1999). *Metodología de la investigación cualitativa*. Málaga: Aljibe.

Fly, B. (1987). *Estrategia para enseñar a aprender*. Buenos Aires: Aique.

Freund, J. y Miller, I. (2000). *Estadística matemática con aplicaciones*. México: Prentice Hall.

García, S. R., y Ríos-Insúa, S. (1998). La teoría de la decisión, de Pascal a Von Neumann. *Historia de la Matemática*, 11-42.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. *Handbook of qualitative research*, *2*(163-194), 105.

Gutiérrez, H. (2005). *Calidad total y productividad*. México D.F: McGraw-Hill.

Gutiérrez, H. y De la Vara, R. (2008). *Análisis y diseño de experimentos*. México D.F.: McGraw-Hill.

Hettmansperger, T. (1984). *Statistical Inference Based on Ranks*. USA: John Wiley & Sons.

Hair, A. Taham, B. (2008). *Análisis Multivariante*. 5.ta ed. España: Pearson Prentice Hall.

Huff, D. (1954). *How to Lie with Statistics*. New York: W.W. Norton.

Hurtado, M. J. R., & Silvente, V. B. (2012). Cómo aplicar las pruebas paramétricas bivariadas t de Student y ANOVA en SPSS. Caso práctico. *REIRE*, *5*(2).

Kandu, D., Kannan, N., & Balakrishnan, N. (2008). On the hazard function of the Birnbaum-Saunders distribution and associated inference. *Computational Statistics & Data Analysis 52*, 2692-2702.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*(1), 16.

Lindgren, B. (1993). *Statistical Theory,* Fourth Edition, USA: Chapman & Hall.

Macchi, R. L. L. M., Martínez Damian, M. A., Diaz Carreno, M. A., Guerra, D. D., León, E. E.,

Barron, L. J. R., & Stephens, L. J. (2014). *Introducción a la estadística en ciencias de la salud* (No. 338.43: 338.5 (72)). e-libro, Corp.

Marquis de Laplace, P. S. (1820). *Théorie analytique des probabilités*. V. Courcier.

Mayorga, H. (2003). *Inferencia estadística*. Bogotá: Unibiblos, Universidad Nacional de Colombia.

Meeker, W. Q., & Escobar, L. A. (1998). *Statistical methods for reliability data*. USA: John Wiley & Sons.

Mejía, M. R. (2011). Pensar la Educación y la Pedagogía en el siglo XXI. *Colección*

Seminario Permanente de Pedagogía - SPP; N.° 1, Tunja: Universidad Pedagógica y Tecnológica de Colombia.

Nieves, H. A., Sánchez, D., & CLicero, F. (2010). *Probabilidad y Estadística para Ingeniería, un enfoque moderno*. México D.F.: McGraw-Hill.

Papoulis, A. (1991). *Probability, random variables and stochastic processes*. New York: McGraw-Hill.

Peña, D. y Romo, J. (1997). *Introducción a la Estadística para las Ciencias Sociales*. Madrid: McGraw-Hill.

Rioboó, J., González, P., y Tato, M. (1997). Resumen Histórico de la Evolución de la Estadística. *Estudios de Economía Aplicada, 8,* 141-162.

Särndal, C. E., Swenson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlang.

Schmidt, Q. (2006). Estándares básicos de competencias en lenguaje, matemáticas, ciencias y ciudadanas: guía sobre lo que los estudiantes deben sabe y saber hacer con lo que aprenden. *Ministerio*, 47-95. Recuperado de: http://www.mineducación.gov.co/1621/articles-340021_recurso_1.pdf

Siegel, S. (1970). *Diseño experimental no paramétrico.* México D.F.: Trillas.

Simons, H. (2011). *El estudio de caso: Teoría y práctica*. Ediciones Morata.

Shao, J. (1999). *Mathematical Statistics*. New York: Springer-Verlang.

Shulman, L. (1987). Knowledge and teaching: foundations of the new reforms. *Harvard Educational Review, 57*(1), 1-23.

Stake, R. (1998). *Investigación con estudio de casos*. (2a. ed.). Madrid: Morata.

Strauss, A. & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. USA: Sage Publications.

Straus, A. y Corbin, J. (2002). *Bases de investigación cualitativa*. Edición en español, Traducción de Eva Zimmerman. Medellín: Universidad de Antioquia.

Valdivieso, M. A. (2011). *Estadística descriptiva: Apoyo al estudio independiente*. Tunja: Uptc.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2007). *Probabilidad y estadística para ingeniería y ciencias* (No. TA430. P76 2012.). Pearson Educación.

Yin, R. K. (2009). *Case Study Research*. London: Sage.

Yin, R. K. (2014). *Case Study Research: Design and Methods* (5 ed.). USA: Sage.