Teoría estadística: Aplicaciones y métodos

Teoría estadística: Aplicaciones y métodos

Hanwen Zhang
Hugo Andrés Gutiérrez

Facultad de Estadística
Universidad Santo Tomás

ISBN: 978-958-631-675-0
Hecho el depósito legal que establece la ley

Corrección de estilo
Camilo Cuéllar
Diseño y diagramación
Hugo Andrés Gutiérrez
Universidad Santo Tomás

UNIVERSIDAD SANTO TOMÁS
Departamento de Publicaciones
Carrera 13 No. 54 - 39
Teléfonos: 2497121 - 2351975
http:// www.usta.edu.co
editorial@usantotomas.edu.co

Bogotá, D.C., 2010

A mis padres,
Nian Jiang y Desheng Zhang.

Hanwen Zhang

A mi abuela,
Lola Moreno de Gutiérrez,
y a los lectores del blog «Apuntes de Estadística».

Hugo Andrés Gutiérrez

Prólogo

Sobre teoría estadística se han escrito muchos libros, indudablemente más en el concierto internacional que en el nacional. Sin embargo, cada vez que un lector se enfrenta a una nueva publicación sobre el tema, quisiera detectar qué es lo nuevo, diferente o atractivo que se presenta o desarrolla en la obra que tiene en sus manos. Desde esta premisa, es muy agradable presentar este libro en el cual se marcan diferencias importantes con respecto a muchos otros escritos sobre la materia. En las líneas siguientes explicaré estas características significativas, para usar un término muy "estadístico".

En virtud de la gran experiencia y habilidad en el manejo del lenguaje R por parte de los autores, el libro incluye muchos ejemplos ilustrativos de los conceptos fundamentales de la inferencia estadística, los cuales se han desarrollado con dicho lenguaje. Esto permite al lector comprender, por ejemplo entre muchas otras, la noción intuitiva de distribución muestral (o de muestreo). Se incluye la teoría estadística básica de la inferencia multivariada, crucial en el entendimiento del comportamiento probabilístico de un vector de variables aleatorias y de las relaciones entre ellas. No es usual encontrar un trabajo en donde se incluyan conjuntamente los contextos univariado y multivariado de la inferencia estadística.

Este libro es un buen punto de partida para el conocimiento e interiorización de la teoría estadística por parte de estudiantes de una carrera de estadística, en el entendido de hacer de la práctica estadística una profesión. Además, podría ser un gran soporte para la realización de estudios de posgrado, bien sea a nivel de profundización de conocimientos o a nivel de investigación. En forma muy general, se puede afirmar que en la presente obra, la teoría y sus aplicaciones son presentadas de manera muy coherente y equilibrada; es decir, sin profundizar en lo teórico más allá de lo necesario y sin exagerar en la inclusión de las aplicaciones. Por esto y todo lo expresado anteriormente, me siento muy complacido de presentar este libro y de recomendarlo a un amplio conglomerado de lectores o usuarios de la estadística.

Fabio Nieto, Ph.D.
Profesor titular
Departamento de Estadística
Universidad Nacional de Colombia

Prefacio

La estadística es una herramienta poderosa en manos del investigador y del profesional. En la vida práctica del profesional que utiliza la estadística, es bien sabido que no es posible realizar un trabajo apropiado sin tener el conocimiento preciso que permita el entendimiento y comprensión de los fundamentos de esa herramienta. La teoría estadística da esos fundamentos y este libro pretende ser un camino que permita el entendimiento y comprensión de los mismos.

Con el pasar del tiempo, el pensamiento estadístico se está convirtiendo en una cultura. Es una forma de razonar que los profesionales de la mayoría de las áreas del saber deben tener para ejercer exitosamente sus trabajos aplicados concernientes al análisis de datos. Y precisamente, el comienzo de la formación de este pensamiento se da en un curso de inferencia. Por lo anterior, es muy importante que el estudiante termine el curso estando preparado de la mejor forma posible, y para esto es fundamental contar con materiales adecuados para el aprendizaje tanto de la teoría como de los métodos prácticos. En el ámbito estadístico existen abundantes textos sobre el tema de inferencia, algunos de los cuales desarrollan rigurosamente las teorías estadísticas formales. Se encuentran también textos de la teoría estadística para disciplinas como ingeniería, veterinaria o biología, entre otras. Estos textos van dirigidos directamente a los usuarios finales de la estadística, algunos de ellos con muy poca formación en matemáticas, razón por la cual los textos no son teóricos sino que se enfocan en la aplicación de la estadística en la vida real. Con el presente texto, quisimos situarnos en el punto de convergencia entre estas corrientes y hacer de la práctica un resultado directo del conocimiento teórico.

Este libro nació del curso de inferencia estadística que dictamos en la Facultad de Estadística de la Universidad Santo Tomas en Bogotá, Colombia. Allí surgió la necesidad de fusionar la teoría y la práctica de una manera óptima. Por esta razón, en este libro pretendemos abarcar ambos aspectos de la estadística: la teoría y la aplicación práctica. Por un lado, se desarrolla la teoría de la inferencia estadística con un alto grado de rigurosidad y por otro lado, se ilustra cómo es la aplicación de estas técnicas y métodos en la práctica. Hemos optado por el uso del programa estadístico R para ilustrar mediante gráficas y códigos las teorías expuestas; asimismo, la mayor parte de los cálculos en las aplicaciones también fueron realizados en dicho software. Algunos de estos códigos computacionales están disponibles en el libro; otros, por limitación de espacio, no fueron incluidos pero los proveeremos en caso de ser requeridos.

El contenido del libro se divide principalmente en dos partes y apéndices: en principio, se presenta la inferencia univariada, que tal como su nombre indica, estudia características de una variable aleatoria observada en una muestra; posteriormente, la inferencia multivariada, que estudia conjuntamente varias variables; por último, se encuentran los apéndices que complementan el proceso natural del análisis de datos. Aunque las partes tienen temas y técnicas similares, se pueden considerar como dos partes independientes.

En algunos apartados, hemos querido dar continuidad al nombre estándar (en inglés) de algunas técnicas estadísticas omitiendo la traducción al castellano puesto que se nos antoja apropiado, en términos de aprendizaje, que el estudiante se familiarice con estas definiciones y pueda fácilmente consultar otra bibliografía sobre el tema. Sin embargo, en otros apartados hemos querido modificar el lenguaje común de la práctica estadística puesto que algunas definiciones pueden inducir una malinterpretación del contexto estadístico. Por ejemplo, deliberadamente cambiamos la palabra ≪poblacional≫ por la palabra ≪teórica≫. Es así como conceptos básicos relacionados con las distribuciones de muestreo como ≪media poblacional≫ y ≪varianza poblacional≫, aparecerán en esta obra como ≪media teórica≫ y ≪varianza teórica≫, respectivamente. La razón de lo anterior se debe a que el estudiante puede interpretar el concepto de media poblacional como el promedio de la variable de interés en todos los individuos de una población finita, aunque, por el contrario, se debe interpretar como la esperanza de la distribución que define una población de naturaleza teórica e intangible. Además, este concepto de media poblacional es bien conocido y definido en otras áreas del saber estadístico como el muestreo y la inferencia en poblaciones finitas.

Este texto va dirigido a los profesionales y estudiantes que deban utilizar herramientas de inferencia estadística y puede servir como libro guía para un curso de cuatro meses con intensidad horaria de seis horas por semana, si se desea abarcar tanto la parte univariada como la parte multivariada. Es importante aclarar que, dependiendo del curso y el énfasis, el docente debe enfocarse en la parte relevante del curso realizando algunas demostraciones teóricas, sin pretender cubrir todos los temas del libro, pero siempre enfatizando la aplicabilidad de los resultados teóricos y la relación estrecha que existe entre el sentido común y los resultados encontrados.

También el texto puede servir para un curso introductorio de estadística en un programa de especialización o maestría donde la mayoría de los estudiantes no son de profesión estadística, y necesitan simultáneamente adquirir formas de pensamiento estadístico y técnicas de análisis de datos en la práctica.

Al final de cada capítulo, se provee de ejercicios sobre el tema desarrollado en las distintas secciones que lo conforman. Algunos de estos ejercicios son teóricos y exigen que los estudiantes estén familiarizados con las herramientas indicadas para derivar resultados. Otros ejercicios son de carácter práctico, en los cuales se describe un problema de la vida real que debe ser resuelto utilizando el sentido común, seguido del pensamiento estadístico y entendiendo el contexto del problema mediante el planteamiento de las preguntas que se deben resolver para, por último, aplicar las herramientas estadísticas apropiadas.

Agradecemos en primer lugar a Dios que nos dio la motivación y la perseverancia para escribir este libro, como también el apoyo que encontramos en la Universidad Santo Tomás por medio del Centro de Investigaciones y Estudios Estadísticos (CIEES) y mediante la siempre elegante gestión administrativa de Sander Rangel en la Decanatura de la Facultad de Estadística. Agradecemos también a los estudiantes del curso inferencia estadística de la Universidad Santo Tomás que colaboraron con la corrección del libro. Por último, agradecemos a los profesores Yesid Rodríguez de la Universidad Santo Tomás y Sergio Calderón Villanueva y Luis Guillermo Díaz de la Universidad Nacional por los valiosos comentarios sobre las notas.

Los autores aclaran que la responsabilidad por los errores que pueden haber en el libro es única y exclusivamente de ellos y agradecen los comentarios, las correcciones y las posibles críticas constructivas sobre la obra. Este es un producto del grupo de investigación en Muestreo y Marketing, adscrito al Centro de Investigaciones y Estudios Estadísticos (CIEES) de la Facultad de Estadística de la Universidad Santo Tomás.

Contenido

Parte I

Inferencia estadística univariada

Capítulo 1

Conceptos preliminares

Los modelos no son la realidad. Los datos no se ajustan a un modelo; por el contrario, los modelos se ajustan a la realidad de las observaciones. Por ejemplo, los modelos de mercadeo y, en general, de cualquier campo, son acepciones de la realidad que buscan describirla, mas no explicarla a cabalidad. Es así como el modelo astronómico de Tolomeo describía con gran precisión la posición de los planetas en la bóveda celeste, aunque como bien lo sabemos no era un modelo que explicara la realidad porque simplemente la tierra no es el centro del universo. Sin embargo, ¿era un mal modelo? Seguramente no, el modelo lograba su función y desde un punto de vista pragmático, era lo que se tenía en esa época y funcionaba bien.

Comparemos la noción general de un modelo cualquiera con un modelo estadístico y empecemos por considerar tres ejemplos concretos:

- Modelos arquitectónicos: planos o maquetas hechos a escala que son fundamentales en la etapa de diseño y el proceso de construcción de cualquier obra.
- Modelos de ingeniería: túneles de viento o simulación de corrientes fluviales.
- Modelos atómicos: teorías, visualizaciones acerca de movimientos, estructuras de un átomo y sus componentes.

Un modelo debe ser visto como un mapa. Incluso el mapa más barato de una ciudad puede responder a todas las preguntas razonables que uno pueda imaginar acerca del posicionamiento de la ciudad: ¿dónde queda el aeropuerto?, ¿qué tan lejos estoy de la alcaldía?, etc. Un buen mapa turístico es capaz de ubicar sitios históricos que ni siquiera, hoy en día, existen. Sin embargo, la construcción de un modelo estadístico requiere otro tipo de abstracciones. Los estadísticos usamos la palabra modelo de una forma bien diferente a los anteriores ejemplos, ya lo diría G.E.P. Box al afirmar que "Todos los modelos son errados, pero algunos son útiles".

Es común considerar la bondad del ajuste del modelo. Típicamente, un modelo estadístico se considera adecuado si, después de haber sido calibrado con los datos reales, cumple significativamente con los supuestos considerados en el diseño del estudio.

Podríamos objetar esta definición. En particular, parece muy ingenuo ignorar que el comportamiento de las unidades seleccionadas en la muestra, en algunas ocasiones diverge radicalmente del comportamiento de las unidades que no están en la muestra, o que fueron seleccionadas en la muestra pero para las cuales existe ausencia de respuesta. Ahora, si el modelo falla en la incorporación de toda la información relevante, ¿debería ser considerado como un modelo no adecuado?

No se puede dejar de lado que el usuario de los modelos estadísticos (o de sus primos: los modelos estocásticos o econométricos) tiene unos objetivos claros y definidos al iniciar la investigación. El estadístico debe formular el modelo que mejor ajuste consiga de manera selectiva con los objetivos de la investigación, teniendo en cuenta los fundamentos teóricos y supuestos del modelo (tarea nada fácil). Ya lo diría Tukey cuando afirmaba: "mantén tu mirada en la ciencia y conserva tus herramientas estadísticas muy simples".

Con lo anterior, queremos enfatizar que también existen modelos para el ajuste de cierto tipo de datos, conocidos como distribuciones de probabilidad que son el soporte de la inferencia estadística. Lo importante es que el lector caiga en cuenta de que en la vida real y en la práctica profesional jamás va a encontrar datos que provengan de estas distribuciones; por el contrario, existen situaciones diversas enmarcadas en contextos especificos que permiten aseverar que los datos observados se ajustan a cierta distribución de probabilidad.

En esta parte del libro, se hace un breve repaso de las principales distribuciones de variables aleatorias. Para cada una de ellas, presentamos las principales características, tales como los diferentes momentos y relaciones entre distribuciones. Queremos hacer énfasis sobre las diferentes aplicaciones que pueden tener estas distribuciones, además de caracterizar los datos que provienen de las mismas. Para mayores detalles acerca de la teoría básica de probabilidad, consulte Blanco (2004).

1.1 Variables aleatorias y distribuciones de probabilidad

La teoría estadística estudia fenómenos cuyos comportamientos no pueden ser predeterminados. La vida práctica está llena de estos fenómenos, algunos de gran impacto socio-económico tales como la tasa de desempleo, precio del dólar, la inflación, etc.; otros asociados más a la vida cotidiana tales como el resultado de un juego de azar, de un partido de fútbol o el clima de mañana. Con las herramientas estadísticas apropiadas, se pueden conocer más a fondo estos fenómenos y así poder describirlos y/o predecirlos.

Estos fenómenos pueden ser descritos como un experimento aleatorio, esto es, un experimento cuyo resultado no se conoce de antemano. El conjunto que contiene todos los posibles resultados de un experimento aleatorio se denomina el espacio muestral, y en este libro será denotado por Ω. Así que para el experimento de observar la tasa de desempleo para el siguiente mes, el espacio muestral será $\Omega = [0, 1]$; mientras que mientras que para el resultado de un partido de fútbol, el espacio muestral puede

ser $\Omega = \{$gana el equipo A, pierde el equipo A, $textempatanlosdosequipos\}$, si en el experimento solo se observa si el equipo A gana o pierde, mas no la diferencia de goles.

Dado un experimento aleatorio con el espacio muestral Ω, una variable aleatoria X es una función definida sobre Ω que asigna a cada elemento de Ω un número real. Por ejemplo, en el ejemplo del partido de fútbol, podemos definir la variable X que vale 1 si el equipo A no pierde el partido y -1 si lo hace. De esta forma, X es una función que asigna el valor 1 a los resultados *gana el equipo A* y *empatan los dos equipos*, y asigna el valor -1 al resultado *pierde el equipo A*. En algunas situaciones, una variable aleatoria puede ser, simplemente, la función idéntica, como en el caso de observar la tasa de desempleo en el siguiente mes. La variable *tasa de desempleo en el siguiente mes* tomará valor en $[0, 1]$ y corresponde simplemente al resultado del experimento, esto es, una función idéntica.

Una forma de clasificar a las variables aleatorias es según los valores que toman y se tienen dos tipos de variables aleatorias: las variables discretas son aquellas que toman valores en un conjunto finito o enumerable[1], aunque en la teoría estadística, la mayoría de las variables discretas toman valores finitos o en el conjunto de los números naturales, mas no en conjuntos enumerables más extraños como los racionales. Por otro lado, tenemos las variables continuas que son aquellas que toman valores en un intervalo, entendiendo que el conjunto de los números reales $\mathbb{R}$ es un intervalo de la forma $(-\infty, \infty)$. En los ejemplos dados anteriormente, las variables *tasa de desempleo* y *precio de dólar* son continuas, mientras que el *resultado del partido de fútbol*, *clima de mañana*, se consideran discretas.

Dada una variable aleatoria X, estamos interesados en calcular probabilidades acerca de los valores que toma, por ejemplo, la probabilidad de que la tasa de empleo del siguiente mes sea inferior al 10 % o la probabilidad de que el equipo A no pierde un partido. Y estas probabilidades se resumen en la la función de distribución $F(x)$ o equivalentemente en la función de densidad, $f(x)$. Y para algunas funciones de densidad de alguna forma especial, se les dan algunos nombres específicos a la distribución de X. En los siguientes capítulos se repasan algunas de las distribuciones discretas y continuas. Como se mencionó antes, los distintos nombres de las distribuciones se dan cuando la función de densidad toma una forma especial. De esta forma, las definiciones de los siguientes capítulos se basan en la forma funcional de las funciones de densidad.

Adicionalmente, presentamos algunas instrucciones en el paquete R para generar números aleatorios de las distribuciones que presentaremos. Esto es importante, puesto que nos da una idea general sobre cómo es el comportamiento de un conjunto de valores provenientes de una distribución específica, lo cual nos ayuda a identificar distribuciones en contextos específicos. Por otro lado, la generación de números aleatorios también será útil cuando abordamos el tema de la inferencia estadística.

Antes de repasar las distribuciones de probabilidad, se definen los conceptos de parámetro de distribución y espacio paramétrico. Un parámetro de distribución es aquel valor fijo que define la forma funcional de una distribución de probabilidad, es decir, cuando el parámetro cambia de valor, la función de distribución y la función

[1] Un conjunto A es enumerable cuando existe una función inyectiva que tiene como dominio A y recorrido el conjunto de los números naturales.

de densidad cambian[2]. Las distribuciones de probabilidad pueden tener más de un parámetro. Cuando una distribución tiene solo un parámetro, éste se denota usualmente por θ; cuando se presenta más de un parámetro, la notación se cambia a $\boldsymbol{\theta}$, representando el vector de parámetros. El espacio paramétrico, $\boldsymbol{\Theta}$, es el conjunto que contiene todos los posibles valores del parámetro o el vector de parámetros. Para distribuciones con un solo parámetro, $\boldsymbol{\Theta}$ será un subconjunto de $\mathbb{R}$, mientras que para distribuciones con k parámetros, $\boldsymbol{\Theta}$ será un subconjunto de $\mathbb{R}^k$.

1.1.1 Distribuciones discretas

En esta parte, presentamos algunas de las distribuciones discretas más conocidas. En primer lugar, se tiene la distribución uniforme discreta que puede ser útil para describir algunos resultados en los juegos de azar.

Distribución uniforme discreta

Definición 1.1.1. *Una variable aleatoria X tiene distribución uniforme discreta sobre el conjunto $\{1, 2, \cdots, N\}$ si su función de densidad está dada por:*

$$f_X(x) = Pr(X = x) = \frac{1}{N} I_{\{1,2,\cdots,N\}}(x) \tag{1.1.1}$$

En la Figura 1.1, podemos visualizar la función de densidad de una distribución uniforme discreta sobre $\{1, \cdots, 5\}$.

Esta distribución describe situaciones donde el experimento aleatorio puede tener un finito de resultados, y la probabilidad de ocurrencia la misma para cada posible resultado. Entre los ejemplos de la distribución uniforme discreta en la vida práctica están el resultado (cara o sello) del lanzamiento de una moneda corriente, el resultado del lanzamiento de un dado corriente, resultado al extraer una bola al azar de una urna que contiene bolas enumeradas de 1 a N. También en las rifas, donde en una bolsa que contiene, digamos, 145 nombres de los empleados de una empresas, al seleccionar un nombre de la bolsa para ser ganador de un computador portátil, la probabilidad de que Juan Gómez sea el ganador es $1/145$, y es claro que entre menos empleados hayan en la empresa, más probable es que Juan Gómez sea el ganador. Ahora, suponga que de las 145 empleados, hay 60 mujeres y 85 hombres (donde las 60 mujeres se denotan por M1, M2, $\cdots$, M60), entonces la probabilidad de que el ganador sea mujer puede ser pensada como la probabilidad de que el ganador sea M1, o sea M2, o $\cdots$, o M60. Recurriendo a propiedades de la probabilidad, se tiene que la probabilidad requerida será $1/145 + 1/145 + \cdots + 1/145$, 60 veces, esto es, $60/145$. Más adelante, se verá que la anterior situación también puede ser descrita por una variable con distribución hipergeométrica.

En la vida práctica, para identificar variables con distribución uniforme discreta, en muchos casos basta con conocer el contexto del problema, es decir, con conocer

[2]En este enfoque clásico, los parámetros se consideran cantidades fijas. Existe otro enfoque en el cual se considera a los parámetros como variables aleatorias, dicho enfoque se denomina bayesiano.

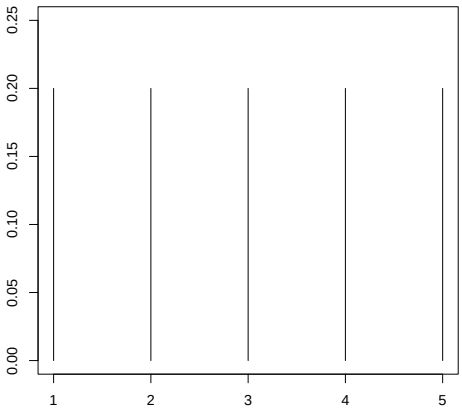

Figura 1.1: *Función de densidad de una distribución uniforme discreta sobre* $\{1, \cdots, 5\}$.

condiciones que garantizan que los valores ocurren con la misma probabilidad, como por ejemplo, en el lanzamiento de una moneda, saber que la moneda no está cargada garantiza que los resultados siguen esta distribución. Sin embargo, puede suceder que se desconoce si esta condición se tiene o no, y lo disponible es simplemente un conjunto de valores. Suponga que se tienen los valores, 3, 1, 3, 4, 2, 4, 2, 2, 1, 3 que denotan resultados de 10 selecciones de una bolsa con bolas enumeradas de 1 hasta 5. Dada la característica de la distribución uniforme discreta, afirmar que el resultado de selección tiene distribución uniforme discreta es equivalente a afirmar que el proceso de selección es completamente al azar, sin ninguna preferencia de números. Ahora, si los datos provinieran de una distribución uniforme sobre $\{1, \cdots, N\}$, entonces la probabilidad de ocurrencia de cualquier $n = 1, \cdots, N$ debe ser igual a $1/N$. Haciendo la analogía entre la probabilidad de ocurrencia con la frecuencia relativa que se puede ver en la Figura 1.2, podemos intuir que sí se presenta una variable uniforme discreta si las frecuencias relativas para 1, $\cdots$, 5 son todos cercanos a $1/5 = 0.2$. Por lo tanto, del histograma de los datos, se observa que la frecuencia relativa del valor 5 está muy alejada del valor 0.2, de donde se sospecha la afirmación de que la selección fue realizada completamente al azar, sin ninguna preferencia de números.

Para generar valores de una distribución uniforme discreta, se puede usar el comando `sample` con la opción `replace=TRUE`, el siguiente código simula dos conjuntos de valores a partir de una distribución uniforme discreta sobre $\{1, 2, 3\}$, con tamaño 500 y 1000, respectivamente, y grafica los dos histogramas. Estos dos histogramas se muestran en la Figura 1.3, donde podemos observar que las frecuencias de los valores parecen ser constantes. Especialmente cuando el tamaño es grande, no hay algún valor con una frecuencia muy grande o muy pequeña con respecto a otros valores.

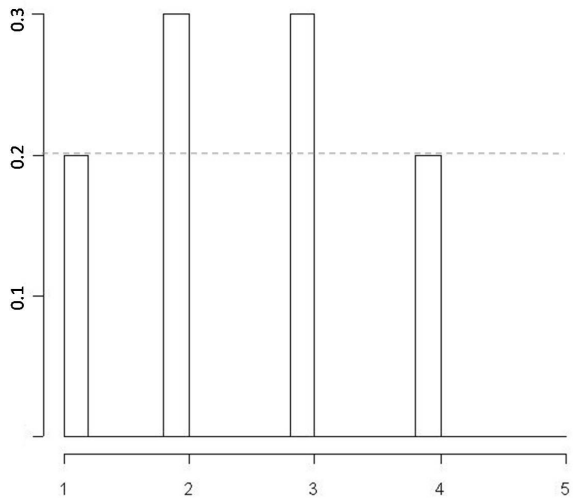

Figura 1.2: *Histograma de los datos 3, 1, 3, 4, 2, 4, 2, 2, 1, 3.*

```
> set.seed(123)
> n<-c(500,1000)
> theta<-3
> par(mfrow=c(1,2))
> for(i in 1:length(n)){
+ a<-n[i]
+ hist(sample(theta,n[i],replace=TRUE),main="",xlab=a,
+    ylab="Frecuencia")
+ }
```

Algunas propiedades básicas de una distribución uniforme se muestran a continuación.

Resultado 1.1.1. *Si X es una variable aleatoria con distribución uniforme discreta sobre el conjunto $\{1, 2, \cdots, N\}$, entonces*

1. $E(X) = \frac{N+1}{2}$.
2. $Var(X) = \frac{N^2-1}{12}$.
3. $m_X(t) = \sum_{i=1}^{N} \frac{e^{ti}}{N}$.

Distribución Bernoulli

La distribución Bernoulli debe su nombre al matemático suizo Jacob Bernoulli (1654-1705). Esta distribución es asociada con experimentos aleatorios que tienen solo dos

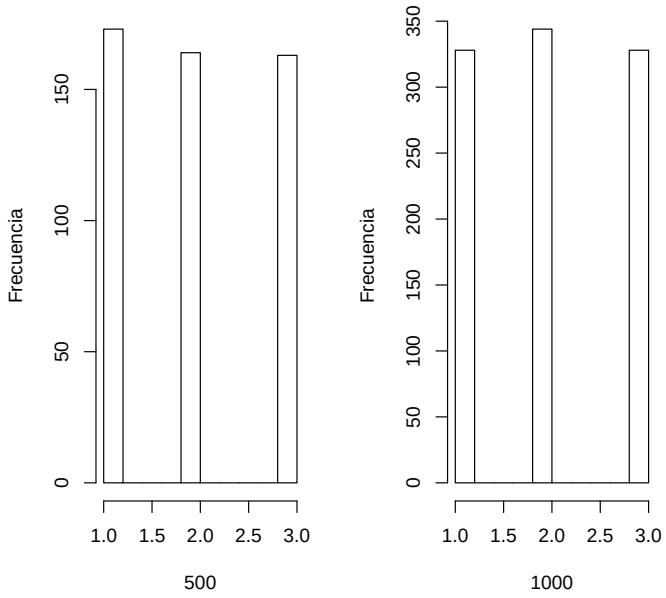

Figura 1.3: *Histograma de valores simulados de una distribución uniforme discreta sobre* $\{1, 2, 3\}$ *con tamaño de muestra 500 y 1000.*

posibles resultados, los cuales se etiquetan como *éxito* y *fracaso*, donde la probabilidad de obtener *éxito* es p, con $0 < p < 1$. De esta forma una variable aleatoria que toma valor 1 cuando se observa el *éxito* y 0 en el caso de *fracaso* tiene distribución Bernoulli con parámetro p. En la vida práctica, se presentan muchos ensayos del tipo Bernoulli; por ejemplo, el éxito o fracaso de un correo electrónico ofreciendo algún servicio o producto. En la teoría del muestreo, la pertenencia de un elemento de la población en la muestra también tiene distribución Bernoulli.

Figura 1.4: *Jacob Bernoulli (1654-1705)*

En términos de la función de densidad tenemos la siguiente definición de la distribución Bernoulli .

Definición 1.1.2. *Una variable aleatoria X tiene distribución Bernoulli con parámetro $p \in (0,1)$ si su función de densidad está dada por:*

$$f_X(x) = p^x(1-p)^{1-x} I_{\{0,1\}}(x), \tag{1.1.2}$$

y se nota como $X \sim Ber(p)$.

Nótese que si $X \sim Ber(p)$, entonces $Pr(X=1) = p$, y $Pr(X=0) = 1-p$. Y tenemos las siguientes propiedades para la distribución Bernoulli.

Resultado 1.1.2. *Si X es una variable aleatoria con distribución Bernoulli con parámetro p, entonces*

1. $E(X) = p$.
2. $Var(X) = p(1-p)$.
3. $m_X(t) = pe^t + 1 - p$.

Demostración. Las anteriores tres expresiones se pueden obtener fácilmente usando la definición de la esperanza para una variable discreta. En particular, $m_X(t) = E(e^{tX}) = e^t Pr(X=1) + e^0 Pr(X=0) = pe^t + 1 - p$. □

En muchos casos, no se observa un solo ensayo del tipo Bernoulli, sino una series de ensayos. Por ejemplo, una empresa que hace ventas virtuales, no manda el correo de promocionamiento a una sola persona, sino a muchos, y la empresa está interesada en cantidades como si se manda el mismo correo a 30 personas distintas, cuántas ventas exitosas obtendrá; es decir, estamos interesados en la variable definida como *el número de éxitos en n ensayos del tipo Bernoulli*. Este tipo de variables se describen con la distribución binomial que se estudiará a continuación.

Distribución binomial

Definición 1.1.3. *Una variable aleatoria X tiene distribución binomial con los parámetros $n \in \mathbb{N}$ y $p \in (0,1)$ si su función de densidad está dada por:*

$$f_X(x) = Pr(X=x) = \binom{n}{x} p^x (1-p)^{n-x} I_{\{0,1,\cdots,n\}}(x), \tag{1.1.3}$$

y se nota como $X \sim Bin(n,p)$.

De acuerdo a lo discutido al final de la sección anterior, una aplicación de la distribución binomial es cuando tenemos un número n de repeticiones independientes de un mismo experimento del tipo Bernoulli, donde la probabilidad del éxito en cada ensayo se denota por p, entonces la variable *número de éxitos obtenidos en las n*

repeticiones tiene distribución $Bin(n, p)$. Dada la anterior interpretación, podemos ver fácilmente que los valores que toma X son enteros entre 0 y n, denotando el valor de 0 la situación donde en todos los ensayos se obtuvo como resultado *fracaso*, y el valor de n cuando todos los ensayos tuvieron como resultado *éxito*. En la Figura 1.5, se muestra la función de densidad de una distribución $Bin(10, 0.35)$, donde se observa que a diferencia de la distribución uniforme discreta, la distribución binomial tiene un valor que tiene mayor probabilidad que otros, digamos x_0, y a medida que se aleja de x_0, la probabilidad disminuye, aunque no de la forma simétrica.

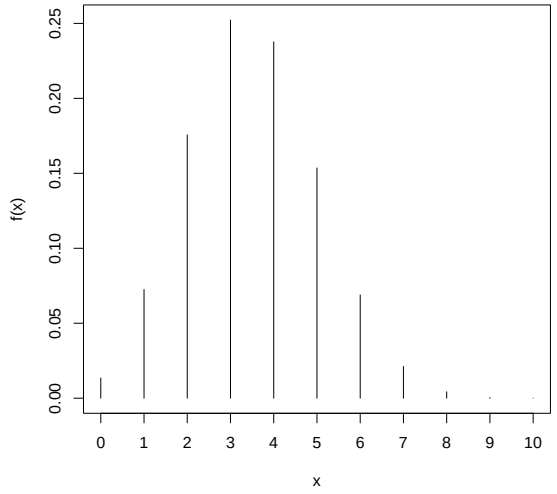

Figura 1.5: *Función de densidad de una distribución* $Bin(10, 0.35)$.

Esta distribución tiene dos parámetros: n y p. Sin embargo, en muchas aplicaciones en la vida práctica el número de repeticiones n es conocido, y la distribución dependerá sólo del valor p que sería el parámetro de la distribución con espacio paramétrico $\mathbf{\Theta} = (0, 1)$.

Algunas propiedades de la distribución binomial se enuncian a continuación.

Resultado 1.1.3. *Si X es una variable aleatoria con distribución binomial con parámetros n y p, entonces*

1. $E(X) = np$.
2. $Var(X) = np(1 - p)$.
3. $m_X(t) = (pe^t + 1 - p)^n$.

Demostración. Se deja como ejercicio (Ejercicio 1.4). □

Observación: el lector puede verificar fácilmente que la distribución Bernoulli es un caso particular de la distribución binomial cuando $n = 1$. Y el Resultado 1.1.2 también se puede obtener del anterior resultado con $n = 1$. También podemos ver que en una distribución binomial, el valor más probable está cercano de la esperanza de la distribución; por ejemplo, en la distribución $Bin(10, 0.35)$, la esperanza está dada por 3.5, mientras que en la Figura 1.5 se observa que el valor más probable es 3. De allí podemos sacar conclusiones muy sencillas sin mayores cálculos: por ejemplo, la empresa de ventas virtuales sabe por experiencia que la probabilidad de obtener una venta exitosa con un correo enviado es del 0.04, entonces si envía 200 correos, lo más probable es que obtenga aproximadamente $0.04 * 200 = 8$ ventas.

La generación de observaciones provenientes de una distribución binomial puede realizarse mediante el comando `rbinom`; de esta manera, si queremos simular 100 valores provenientes de una distribución $Bin(10, 0.35)$, podemos usar el siguiente código `rbinom(1000,10,0.35)`. El histograma de un conjunto de datos simulados con esta instrucción está dado en la Figura 1.6, donde se observa un comportamiento muy similar a la función de densidad teórica dada en la Figura 1.5.

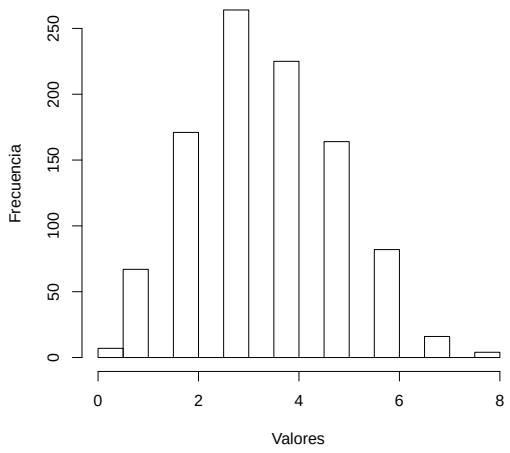

Figura 1.6: *Histograma de un conjunto de datos provenientes de* $Bin(10, 0.35)$.

Usando la función generadora de momentos del Resultado 1.1.3, podemos establecer el siguiente resultado que ilustra la relación entre las distribuciones Bernoulli y binomial.

Resultado 1.1.4. *Sea* $X_1, \cdots, X_n$ *variables aleatorias independientes e idénticamente distribuidas con distribución Bernoulli con parámetro* p*, entonces la variable* $\sum_{i=1}^{n} X_i$ *tiene distribución* $Bin(n, p)$.

Demostración. La demostración radica en el hecho de que la función generadora de momentos caracteriza la distribución probabilística, entonces basta demostrar que la función generadora de momentos de $\sum_{i=1}^{n} X_i$ es la de una distribución $Bin(n,p)$. Tenemos lo siguiente:

$$\begin{aligned} m_{\sum X_i}(t) = E(e^{\sum tX_i}) &= E\left(\prod_{i=1}^{n} e^{tX_i}\right) \\ &= \prod_{i=1}^{n} E(e^{tX_i}) \quad \text{(por independencia)} \\ &= \prod_{i=1}^{n} (pe^t + 1 - p) \quad \text{(definición de } m_{X_i}(t)) \\ &= (pe^t + 1 - p)^n. \end{aligned}$$

Y el resultado queda demostrado. □

Distribución hipergeométrica

Definición 1.1.4. *Una variable aleatoria X tiene distribución hipergeométrica con parámetros n, R y N si su función de densidad está dada por:*

$$f_X(x) = Pr(X = x) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}, \tag{1.1.4}$$

para x entero entre $\max(n-N+R, 0)$ y $\min(R, n)$, y se nota como $X \sim Hg(n, R, N)$.

Una variable con distribución hipergeométrica se da cuando se desea extraer n unidades de un conjunto de N objetos en total, que se pueden dividir en dos grupos: el primero de R unidades y el segundo de $N - R$. Entonces la variable definida como el número de unidades extraídas del primer grupo tiene distribución hipergeométrica con parámetros n, R y N. Suponga que se desea extraer 6 estudiantes de 15 en total, donde 10 son hombres y 5 son mujeres; la variable X definida como *el número de estudiantes hombres seleccionados* tiene distribución $Hg(6, 10, 15)$. Nótese que X solo toma valores entre 1 y 6, puesto que al seleccionar 6 estudiantes, a lo más 5 mujeres pueden quedar en la muestra, es decir, por lo menos estarán en la muestra. En la Figura 1.7, se muestra la función de densidad para esta distribución.

Algunas propiedades de la distribución hipergeométrica se enuncian a continuación.

Resultado 1.1.5. *Si X es una variable aleatoria con distribución hipergeométrica con parámetros n, R y N, entonces*

1. $E(X) = \frac{nR}{N}$.
2. $Var(X) = \frac{nR(N-R)(N-n)}{N^2(N-1)}$.

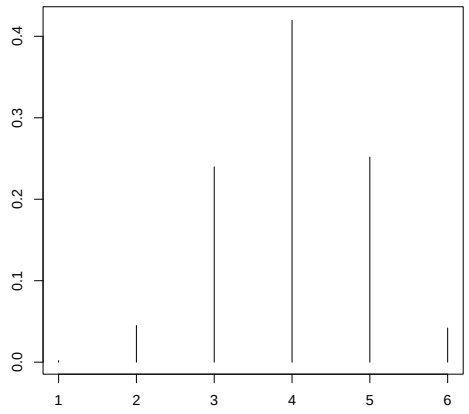

Figura 1.7: *Función de densidad de una distribución* $Hg(6, 10, 15)$.

El anterior resultado no incluye la función generadora de momentos, pues éste no ha resultado ser útil en la teoría relacionada con la distribución hipergeométrica.

La distribución hipergeométrica también es útil en la teoría de muestreo, tal como ilustran Ardilly & Tillé (2006), al considerar una región agricultural que consiste en $N = 2010$ fincas de donde se desea extraer una muestra aleatoria simple sin reemplazo de tamaño $n = 100$. Nótese que cada elemento de la población tiene la misma probabilidad de pertenecer a la muestra y puede ser seleccionado a lo más una vez. Suponga que adicionalmente se dispone la información sobre el área total cultivada de cada finca, de donde se sabe que en la población total existen 1580 fincas de menos de 160 hectáreas (subgrupo 1) y 430 de más de 160 hectáreas (subgrupo 2). Aunque el tamaño muestral n está fijo, el número de fincas del subgrupo 1 seleccionadas denotado por n_1 es aleatorio y sigue una distribución $Hg(100, 1580, 2010)$; también el número de fincas del subgrupo 2 seleccionadas n_2 sigue una distribución hipergeométrica $Hg(100, 430, 2010)$.

Usando el resultado anterior, podemos obtener que $E(n_1) = 100 \times 1580/2010 = 78.6$, esto es, se espera seleccionar 78 o 79 fincas del subgrupo 1.

Otro uso de la distribución hipergeométrica es el problema de captura-recaptura, donde se necesita estimar el tamaño de una población de interés. Para ello, se identifica un número R menor que N de individuos, luego se deja que estos R individuos se mezclen bien con el resto de la población. Después de esto, se selecciona n individuos de la mezcla homogénea, y se cuenta el número, x, de individuos marcados que quedaron seleccionados. Dado que la población estaba homogénea al momento de la selección, podemos pensar que las proporciones de objetos marcados en la muestra y en la población deben ser similares, esto es,

$$\frac{x}{n} \approx \frac{R}{N}, \tag{1.1.5}$$

de donde se tiene que $N \approx nR/x$. Para una ilustración de este escenario, ver Figura 1.8. Otra aplicación de la distribución hipergeométrica es cuando se desea estimar el tamaño de un subgrupo de una población conocida, el procedimiento es el mismo de la captura recaptura, y se tiene la relación de (1.1.5), de donde se tiene que $R \approx xN/n$. Suponga que en una ciudad existen 2396 empresas que pueden clasificar en empresas grandes, medianas o pequeñas según el número de empleados. Si en una muestra aleatoria simple sin reemplazos de tamaño 200 se encuentran 28 empresas grandes, podemos estimar el número total de empresas grandes en la población total como $28 * 2396/200 \approx 335$ empresas grandes. Más detalles sobre la estimación en las anteriores situaciones se describen en el siguiente capítulo.

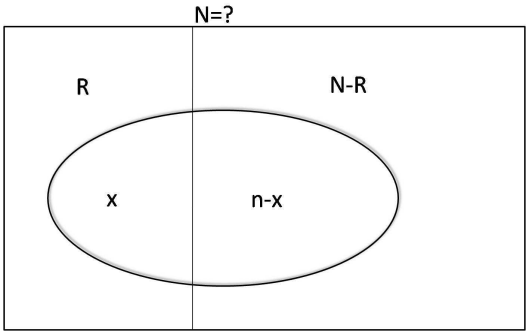

Figura 1.8: *Ilustración del problema de captura recaptura.*

Nótese que un experimento del tipo hipergeométrico está muy relacionado con un experimento Bernoulli, puesto que en la i-ésima extracción para $i = 1, \cdots, n$, podemos definir X_i como 1 si el resultado es uno de los R objetos y 0 si no. De esta forma, tenemos n variables con distribución Bernoulli, y la variable $Hg(n, R, N)$ viene siendo la variable $X = X_1 + \cdots + X_n$. Aunque las variables X_1, $\cdots$, X_n son del tipo Bernoulli, no podemos afirmar que X tiene distribución binomial, puesto que dado el mecanismo de selección, estas n variables no son independientes. Sin embargo, bajo algunas condiciones, sí podemos afirmar que una distribución hipergeométrica puede ser aproximada como una distribución binomial, tal como lo afirma el siguiente resultado.

Resultado 1.1.6. *Dada una variable aleatoria X con distribución $Hg(n, R, N)$, si se tiene que $R/N \to p$, con $0 < p < 1$ cuando $R, N \to \infty$, entonces, la función de densidad de X tiende a la función de densidad de una distribución $Bin(n, p)$.*

Para estudiar la convergencia enunciada en el anterior resultado, se calculó la función de densidad de cuatro distribuciones hipergeométricas con diferentes valores de R y N, y las respectivas distribución $Bin(n, R/N)$. El código de R es como sigue y la gráfica arrojada se muestra en la Figura 1.9. Podemos observar de la gráfica

resultante que la aproximación por medio de la distribución binomial puede ser muy adecuada para valores grandes de R y N.

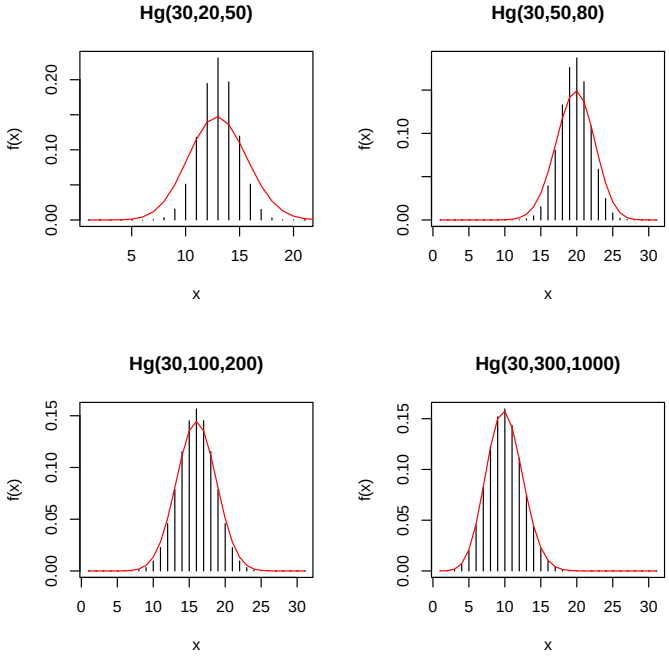

Figura 1.9: *Ilustración de la aproximación de la distribución hipergeométrica mediante la distribución binomial. (Línea roja indica la correspondiente distribución binomial)*

```
> Hg<-function(x,n,R,N){
+ x<-c(max(0,n-N+R):min(R,n))
+ res<-rePr(NA,length(x))
+ for(i in 1:length(x)){
+ res[i]<-choose(R,x[i])*choose(N-R,n-x[i])/choose(N,n)}
+ return(res)
+ }
>
> par(mfrow=c(2,2))
>
> plot(Hg(x,30,20,50),type="h",main="Hg(30,20,50)",xlab="x",
+  ylab="f(x)")
> lines(dbinom(x,n,20/50),col=2)
>
> plot(Hg(x,30,50,80),type="h",main="Hg(30,50,80)",xlab="x",
+  ylab="f(x)")
> lines(dbinom(x,n,50/80),col=2)
```

```
> plot(Hg(x,30,100,200),type="h",main="Hg(30,100,200)",xlab="x",
+  ylab="f(x)")
> lines(dbinom(x,n,100/200),col=2)
>
> plot(Hg(x,30,300,600),type="h",main="Hg(30,300,600)",xlab="x",
+  ylab="f(x)")
> lines(dbinom(x,n,300/600),col=2)
```

Volviendo al problema de la selección de una muestra con $n = 100$ de fincas que se dividen en dos grupos, podemos aproximar la variable aleatoria n_1 con una distribución $Bin(100, 1580/2010)$, y de esta forma calcular probabilidades acerca de los posibles valores de n_1.

Distribución Poisson

La distribución Poisson debe su nombre al francés Siméon-Denis Poisson (1781-1840), quien descubrió esta distribución en el año 1838, cuando la usó para describir el número de ocurrencias de algún evento durante un intervalo de tiempo de longitud dada. Como de costumbre, damos la definición de esta distribución en términos de la función de densidad.

Figura 1.10: *Siméon-Denis Poisson (1781-1840)*

Definición 1.1.5. *Una variable aleatoria X tiene distribución Poisson con parámetros $\lambda > 0$ si su función de densidad está dada por:*

$$f_X(x) = Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} I_{\{0,1,\cdots\}}(x) \tag{1.1.6}$$

y se nota como $X \sim Pois(\lambda)$.

La función de densidad de una distribución Poisson presenta un pico en valores cercanos al parámetro λ. En la Figura 1.11, se ilustra la densidad de una distribución $Pois(5.5)$, donde se observa que el valor con mayor probabilidad corresponde al valor 5, y a medida que el valor de x se aleja de 5, las probabilidades disminuyen.

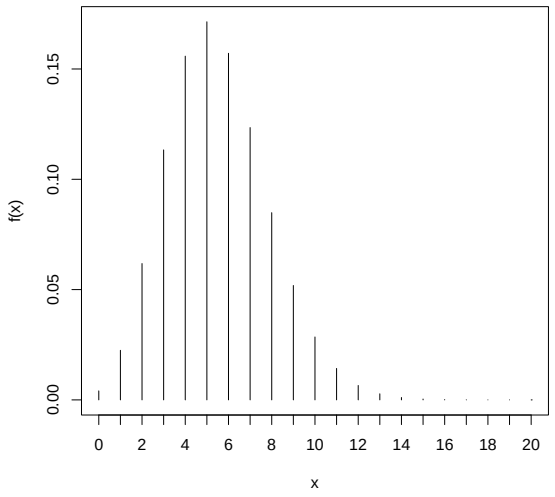

Figura 1.11: *Función de densidad de una distribución* $Pois(5)$.

Nótese que una variable con distribución Poisson puede tomar cualquier valor entero no negativo, y por esta razón, es usada frecuentemente para describir datos de conteo. Cuando en la práctica se presentan un conjunto de valores que son conteos, debe tener en cuenta la diferencia entre las distribuciones binomial, hipergeométrica y Poisson. En primer lugar, los valores que toma una variable con distribución Poisson no debe tener un límite superior, es decir, puede ser cualquier entero positivo. Por lo tanto, contextos donde la variable en cuestión no puede ser más grande que algún valor no deben ser considerados como una variable Poisson. En segundo lugar, tanto la distribución binomial como la hipergeométrica, la variable puede ser vista como un número de éxito obtenido en una sucesión de ensayos, mientras que la distribución Poisson carece de esta interpretación. De esta forma, una variable que describe, por ejemplo, número de accidentes automovilísticos en una determinada localidad, número de transacciones en una entidad durante diez minutos, o en general, número de eventos ocurridos en un punto geográfico y/o en un determinado rango del tiempo puede ser vista como una variable con distribución Poisson.

Por otro lado, aunque una variable con distribución Poisson solo toma valores enteros, el parámetro de la distribución puede ser cualquier número real positivo, esto es, el espacio paramétrico de la distribución es $\mathbf{\Theta} = (0, \infty)$.

Ahora bien cuando no se puede conocer la procedencia de un conjunto de los datos, podemos utilizar el histograma de éstos para identificar la distribución de donde provienen, puesto que el histograma debe ser similar a la densidad teórica de la distribución. Considere el siguiente ejercicio: en R, la generación de números aleatorios puede ser llevada a cabo usando el comando `rpois`. En la Figura 1.12, se muestra el histograma de 300 datos provenientes de la distribución $Pois(5.5)$ generados usando la instrucción `rpois(300,5.5)`. Podemos observar que el histograma tiene un comportamiento muy similar a la función de densidad teórica de la distribución, presentando mayor frecuencia en el valor 5 y comportamiento decreciente para valores alejados de 5.

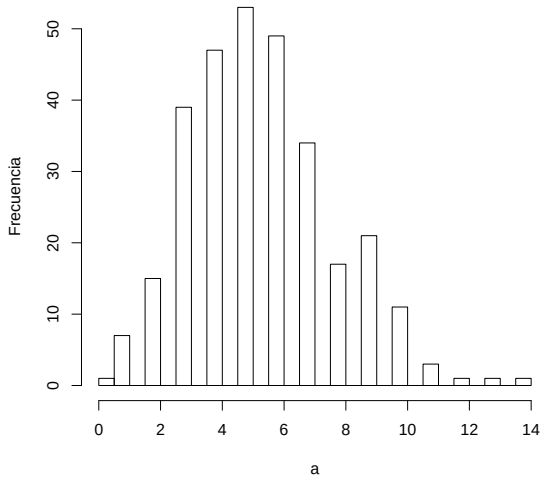

Figura 1.12: *Histograma de un conjunto de datos provenientes de* $Pois(5.5)$.

Algunas propiedades de la distribución Poisson se enuncian a continuación:

Resultado 1.1.7. *Si* X *es una variable aleatoria con distribución Poisson con parámetro* λ*, entonces*

1. $E(X) = \lambda$.
2. $Var(X) = \lambda$.
3. $m_X(t) = \exp\{\lambda(e^t - 1)\}$.

Demostración. La demostración del anterior resultado se deja como ejercicio (Ejercicio 1.6) y consiste en encontrar la función generadora de momentos, $m_X(t)$, usando directamente su definición y la expansión de Tayler de e^u dada por $e^u = \sum_{i=0}^{\infty} \frac{u^i}{i!}$. Una vez encontrada $m_X(t)$, se encuentra la esperanza y la varianza de manera habitual. □

El anterior resultado, a parte de proveernos propiedades de la distribución Poisson, también nos brinda una herramienta a la hora de identificar la distribución de datos cuya procedencia no se conoce, puesto que de acuerdo al resultado anterior, si un conjunto de datos proviene de la distribución Poisson, entonces el promedio debe ser cercano a la varianza; más aún, el promedio y la varianza deben ser cercanos al parámetro de la distribución. En la práctica, una distribución que puede ser confundida con la distribución Poisson es la distribución binomial, puesto que ambas toman valores enteros positivos, pero en la distribución Binomial, la varianza teórica está dada por $np(1-p)$, la cual es siempre menor que la esperanza np, situación que no ocurre si se tratara de una distribución Poisson.

Para corroborar la anterior afirmación en la práctica, se simularon muestras de tamaño 10, 30, 50, 100, 300, 500 y 1000 que provienen de la distribución $Pois(5)$ y $Bin(20, 0.25)$, y en cada muestra se calculó el promedio y la varianza, el código en R es como sigue, y tiene como resultado la Figura 1.13, donde podemos observar que en las muestras con distribución Poisson, la varianza muestral se asemeja al promedio muestral, mientras que en las muestras con distribución binomial, la varianza siempre estuvo por debajo del promedio con una diferencia considerable, corroborando las propiedades teóricas. De lo anterior, podemos concluir que en un conjunto de datos, si la varianza es muy similar al promedio, hay más evidencia a favor de la distribución Poisson que la binomial; mientras que si la varianza es considerablemente menor que el promedio, se puede decir que la distribución binomial ajusta mejor a los datos[3].

```
> set.seed(1234)
>
> n<-c(10,30,50,100,300,500,1000)
> mp<-matrix(NA)
> vp<-matrix(NA)
>
> for(i in 1:length(n)){
+ a<-rpois(n[i],5)
+ mp[i]<-mean(a)
+ vp[i]<-var(a)
+ }
>
> ####################
> mb<-matrix(NA)
> vb<-matrix(NA)
>
> for(i in 1:length(n)){
+ b<-rbinom(n[i],20,5/20)
+ mb[i]<-mean(b)
+ vb[i]<-var(b)
+ }
>
```

[3]En otras distribuciones como la distribución binomial negativa la varianza es más grande que la esperanza, y se denomina este fenómeno como la sobredispersión.

```
> par(mfrow=c(2,1))
>
> plot(mp,type="b",ylim=c(0,7),ylab="",xaxt="n",xlab="n",
+  main="Poisson(5)")
> lines(vp,type="b", pch=4)
> axis(1,1:length(n),n)
> legend("bottomright",c("Promedio","Varianza"), pch=c(1,4),bty="n")
>
> plot(mb,type="b",ylim=c(0,7),ylab="",xaxt="n",xlab="n",
+  main="Binomial(20,0.25)")
> lines(vb,type="b", pch=4)
> axis(1,1:length(n),n)
> legend("bottomright",c("Promedio","Varianza"), pch=c(1,4),bty="n")
```

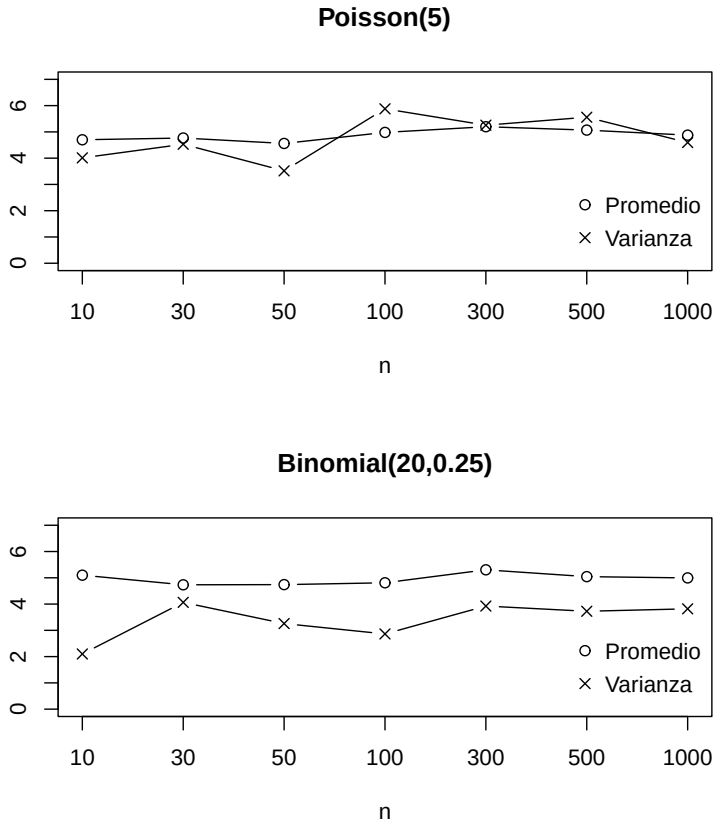

Figura 1.13: *Promedio y varianza de muestras provenientes de distribuciones* $Pois(5.5)$ *y* $Bin(20, 0.25)$.

Como se mencionaba anteriormente, en la práctica, si no se conoce la procedencia de los datos, sino solo los valores, unos datos provenientes de la distribución Poisson podrían confundirse con la distribución binomial. El siguiente resultado nos plantea una relación entre estas dos distribuciones bajo algunas circunstancias especiales.

Resultado 1.1.8. *Considera n eventos del tipo Bernoulli, donde p denota la probabilidad del éxito de cada uno de los n eventos. Si el valor de p es pequeño y $np \to \lambda > 0$ cuando $n \to \infty$, entonces la variable X con distribución $Bin(n, p)$ se distribuye aproximadamente con distribución $Pois(\lambda)$.*

Demostración. Blanco (2004, p. 114). □

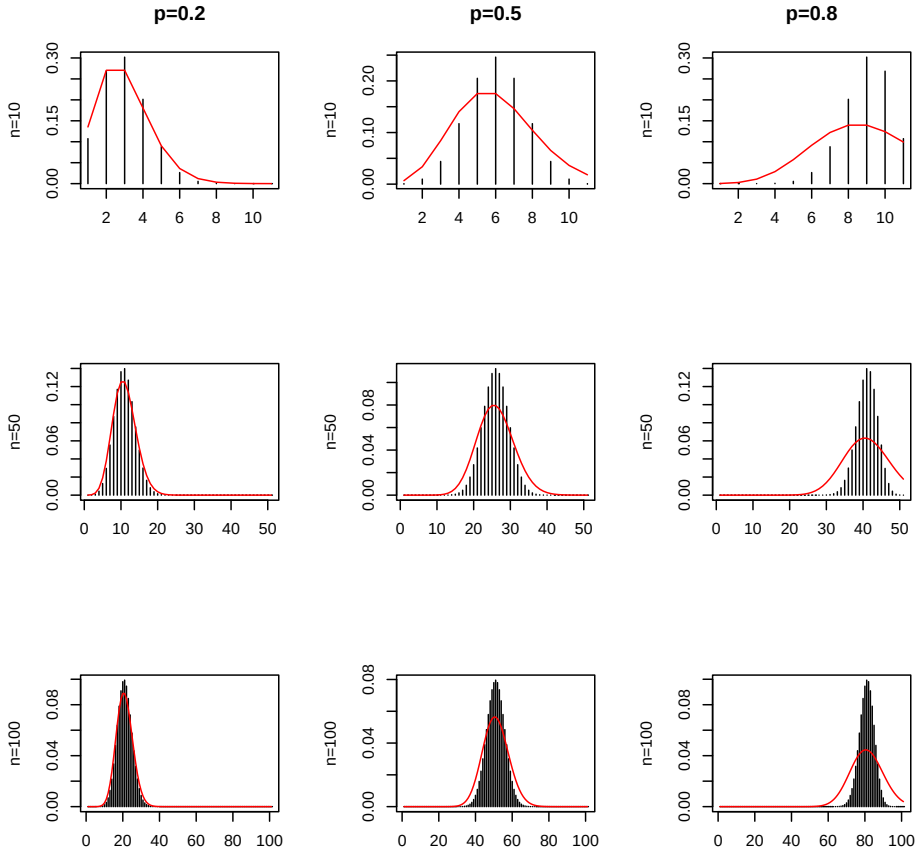

Figura 1.14: *Ilustración de la aproximación de la distribución Poisson mediante la distribución binomial del Resultado 1.1.8. (La línea gris indica la correspondiente distribución Poisson)*

En el anterior resultado, existen dos condiciones para garantizar la convergencia, la probabilidad de éxito en cada ensayo p debe ser pequeña y el número de ensayos n debe ser grande. Para hacerse una idea sobre qué tan importantes son estas dos condiciones, se elaboró la Figura 1.14, la cual ilustra funciones de densidad de distribución binomial con diferentes valores de n y de p, y también la correspondiente distribución Poisson. Se observa que efectivamente, a medida que aumenta el valor de p, la aproximación se torna cada vez más mala, sin importar el tamaño muestral n. Por otro lado, se observa que la condición de que p sea pequeña es más importante que la condición de que n sea grande, puesto que para la distribución $Bin(10, 0.2)$, aunque n sea pequeña, la aproximación sigue siendo buena.

Ahora, aunque en la Figura 1.14 se observó que cuando p es pequeña, la aproximación por la distribución Poisson resulta no adecuada, podemos transformar a una variable $X \sim Bin(n, p)$ con p grande para que siga siendo la válida la aproximación. En este caso, es fácil ver que la variable $Y = n - X \sim Bin(n, 1 - p)$, si p es grande, $1 - p$ es pequeño, entonces para calcular $f_X(x) = Pr(X = x)$, tenemos que ésta es igual a $Pr(Y = n - x)$, y utilizando la distribución Poisson para aproximar la distribución de Y tenemos que

$$Pr(X = x) = Pr(Y = n - x) \approx \frac{e^{-\lambda}\lambda^{n-x}}{(n-x)!},$$

con $\lambda = n(1-p)$. En la Figura 1.15, se ilustra la bondad de la anterior aproximación para densidades de la distribución binomial con distintos valores de p y n, se observa que la aproximación es bastante buena para valores grandes de p, mientras que cuando p toma un valor cercano al 0.5, la anterior aproximación es muy similar a la presentada en el Resultado 1.1.8.

Una propiedad interesante de la distribución Poisson es el hecho de que la suma de variables independientes con distribución Poisson sigue teniendo la distribución Poisson. Lo anterior lo afirma el siguiente resultado.

Resultado 1.1.9. *Sea $X_1, \cdots, X_n$ variables aleatorias independientes con distribución $Pois(\lambda_i)$ para $i = 1, \cdots, n$, entonces la variable $\sum_{i=1}^{n} X_i$ tiene distribución $Pois(\sum_{i=1}^{n} \lambda_i)$.*

Demostración. La demostración es análoga a la demostración del Resultado 1.1.4 y se deja como ejercicio (Ejercicio 1.7). □

Suponga que una central telefónica de atención de clientes cuenta con 5 operadores en cada turno, la variable de interés es el número de llamadas que atiende la central durante 5 minutos. Un estudio acerca del rendimiento de los 5 operadores de turno revela que el número de llamadas que atiende en 5 minutos sigue una distribución Poisson con parámetros 2, 3, 2, 4 y 3, respectivamente. Si además los operadores trabajan de forma independiente, entonces el anterior resultado garantiza que el número total de llamadas atendidas en 5 minutos puede ser modelado como una distribución $Pois(14)$, y podemos calcular probabilidades acerca de esta variable de la manera habitual.

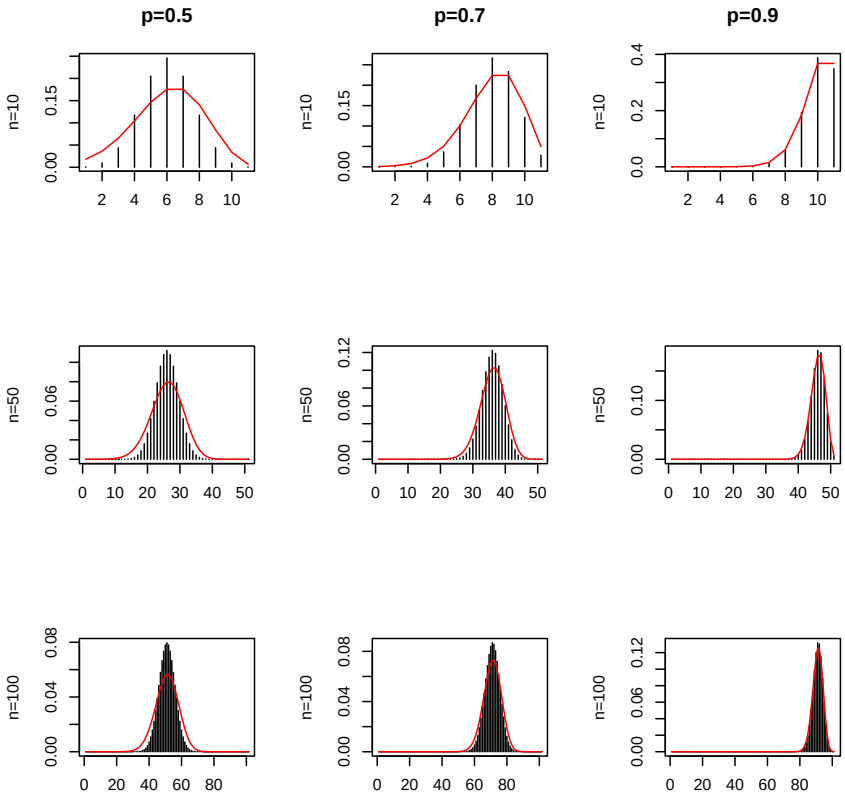

Figura 1.15: *Ilustración de la aproximación de la distribución Poisson mediante la distribución binomial cuando p es grande. (La línea gris indica la correspondiente distribución Poisson)*

1.1.2 Distribuciones continuas

En esta parte del libro consideraremos las distribuciones continuas, esto es, algunas distribuciones comunes para variables aleatorias continuas.

Distribución Uniforme Continua

Una de las distribuciones continuas más simples es la distribución uniforme continua sobre un intervalo $[a, b]$, la cual se caracteriza en que para un subintervalo de $[a, b]$ de longitud fija, una variable con esta distribución tiene la misma probabilidad de ubicarse en cualquiera de estos subintervalos. Por consiguiente, esta distribución es apropiada para situaciones donde para un experimento no hay resultados que son más probables

que otros, un aspecto similar a la distribución uniforme discreta. De esta forma, si suponemos que el primer bus puede demorar a lo más 15 minutos para llegar al portal de transporte, y puede llegar en cualquier momento en ese rango, en este caso, la variable definida como el tiempo de llegada del bus tiene una distribución uniforme $[0, 15]$. Claramente el límite inferior 0 está dado por el contexto del problema y la naturaleza de la variable.

La definición de esta distribución en términos de la función de densidad está dada a continuación.

Definición 1.1.6. *Una variable aleatoria X tiene distribución uniforme continua sobre el intervalo $[a, b]$ con $a < b$ si su función de densidad está dada por:*

$$f_X(x) = \frac{1}{b-a} I_{[a,b]}(x), \tag{1.1.7}$$

y se denotará por $X \sim U[a, b]$.

Análogo a lo discutido en la parte de la distribución uniforme discreta, cuando los datos provienen de la distribución $U[a, b]$, entonces el histograma debe ser plano, similar a la función de densidad teórica. Para observar lo anterior, simulamos dos muestras provenientes de la distribución $U[1, 3]$ del tamaño 500 y 1000 usando la instrucción `runif`, y graficamos los correspondientes histogramas. El código usado es

```
set.seed(123)
n<-c(500,1000)
par(mfrow=c(1,2))
for(i in 1:length(n)){
a<-n[i]
hist(runif(a,1,3),main="",xlab=a,ylab="Frecuencia")
}
```

Y la resultante gráfica está dada en la Figura 1.16, donde se observa que efectivamente no hay algún patrón reconocible en estos histogramas, sino que cada clase tiene aproximadamente la misma frecuencia, características propias de una distribución uniforme continua.

La generación de números aleatorios de una distribución $U[0, 1]$ es particularmente importante, puesto que son de utilidad para simular otras distribuciones más complicadas que la distribución uniforme. El procedimiento viene dado por el siguiente resultado tomado de Robert & Casella (1999).

Resultado 1.1.10. *Si $U \sim U(0, 1)$ y $F(x)$ es una función de distribución, entonces la función de distribución de la variable $F^-(U)$ está dada por F, donde F^- denota la función inversa generalizada de F dada por $F^-(u) = \text{ínf}\{x : \ F(x) \geq u\}$.*

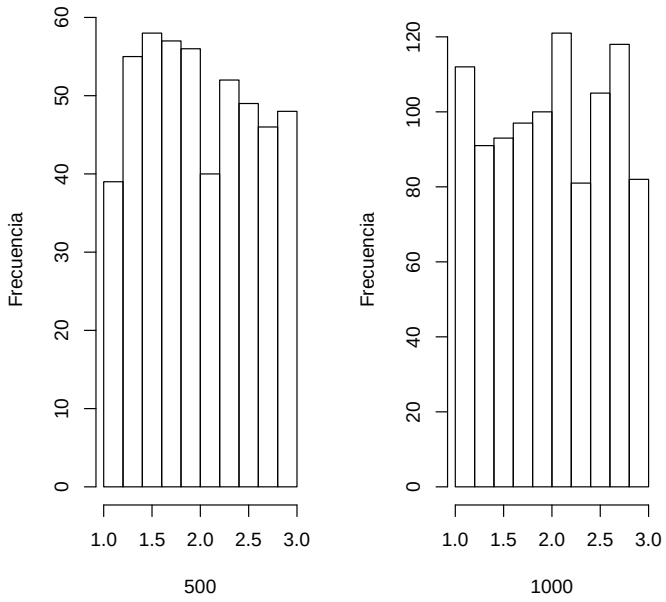

Figura 1.16: *Histograma de valores simulados de una distribución* $U[1,3]$ *con tamaño de muestra 500 y 1000.*

El anterior resultado nos indica que si queremos simular n valores a partir de una cierta distribución F, se debe, en primer lugar, hallar la función de distribución inversa generalizada[4] F^-, y en segundo lugar, simular n observaciones de la distribución $U(0,1)$ que denotamos por $u_1, \cdots, u_n$. Finalmente, el anterior resultado garantiza que los valores $F^-(u_1), \cdots, F^-(u_n)$ provienen de la distribución F.

Por ejemplo, si queremos simular observaciones de la función de densidad $f(x) = e^{-x}I_{0,\infty}(x)$, esto es, la función de densidad de una distribución $Exp(1)$ que se describirá con mayor detalle más adelante, la función de distribución está dada por $F(x) = 1 - e^{-x}$, de donde la inversa de esta función está dada por $F^-(x) = -\ln(1-x)$, así que podemos simular observaciones usando esta función inversa.

Asimismo, en el software R se disponen las instrucciones para simular observaciones de la mayoría de las distribuciones de probabilidades, y podemos utilizarlos directamente sin tener que recurrir al anterior resultado manualmente. El siguiente comando simula 100 observaciones de la distribución $Exp(1)$ con el Resultado 1.1.10 y usando la instrucción `rexp` incorporado en R. En la Figura 1.17 se observan los histogramas de los valores obtenidos, donde se puede observar la similitud en las estructuras de los datos; por facilidad, usaremos en este libro las instrucciones de R.

[4] Cuando la inversa de F existe, ésta coincide con la inversa generalizada.

```
> set.seed(1234)
> n<-100
> u<-runif(100)
> e<--log(1-u)
> e1<-rexPr(100,1)
> par(mfrow=c(1,2))
> hist(e,ylab="Frecuencia",main="(a)")
> hist(e1,ylab="Frecuencia",main="(b)")
```

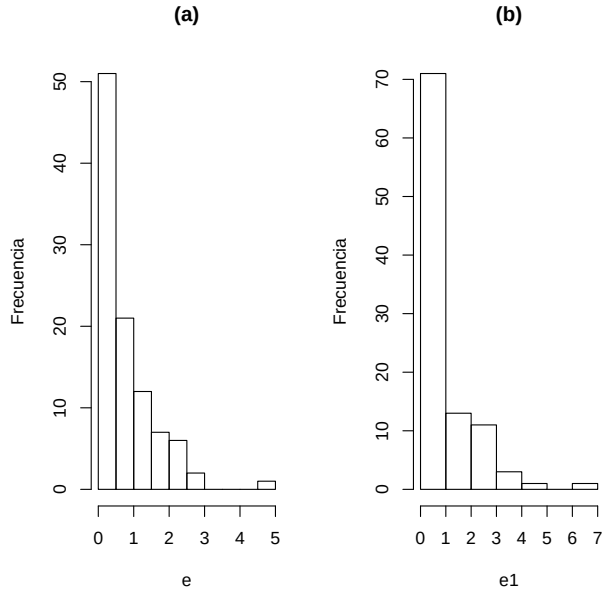

Figura 1.17: *Histogramas de 100 observaciones simulados de una distribución* $Exp(1)$*, (a) con el Resultado 1.1.10, (b) con la instrucción rexp.*

Las siguientes propiedades de una distribución uniforme continua se pueden comprobar fácilmente.

Resultado 1.1.11. *Si* X *es una variable aleatoria con distribución uniforme continua sobre* $[a, b]$*, entonces*

1. $E(X) = \frac{a+b}{2}$.
2. $Var(X) = \frac{(b-a)^2}{12}$.
3. $m_X(t) = \frac{e^{bt}-e^{at}}{(b-a)t}$.

Demostración. Se deja como ejercicio (Ejercicio 1.8). □

Distribución Gamma

La distribución Gamma es una distribución muy importante, puesto que muchas distribuciones de uso común, como la distribución exponencial y la distribución Ji-cuadrado, son casos particulares de esta distribución. La definición de esta distribución en término de la función de densidad de probabilidad está dada por

Definición 1.1.7. *Una variable aleatoria X tiene distribución Gamma con parámetro de forma $k > 0$ y parámetro de escala $\theta > 0$ si su función de densidad está dada por:*

$$f_X(x) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k \Gamma(k)} I_{(0,\infty)}(x), \tag{1.1.8}$$

donde $\Gamma(k)$ es la función Gamma dada por

$$\Gamma(k) = \int_0^\infty u^{k-1}e^{-u}du. \tag{1.1.9}$$

En este libro, se usará la notación $X \sim Gamma(k, \theta)$.

La distribución Gamma tiene dos parámetros: k que se denomina el parámetro de forma y θ el de escala. En este caso, el vector de parámetros es $\boldsymbol{\theta} = (k, \theta)'$ donde el espacio paramétrico está dado por $\boldsymbol{\Theta} = (0, \infty) \times (0, \infty)$. Pero cuando uno de los dos parámetros es fijo por ejemplo, si θ es fijo, entonces la distribución tendría un solo parámetro: k.

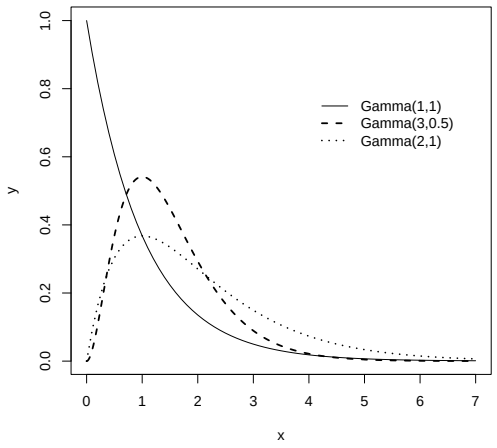

Figura 1.18: *Funciones de densidad de la distribución Gamma.*

Nótese que, en primer lugar, una variable con distribución sólo puede tomar valores positivos, y en segundo lugar, la función de densidad no es simétrica puesto que el

coeficiente de asimetría es positivo. En la Figura 1.18, se muestran algunas funciones de densidad de la distribución Gamma en las cuales se observa claramente la característica no simétrica.

Los datos del ingreso salarial cuentan con la estructura de la distribución Gamma, puesto que la mayoría de la población tiene ingreso inferior a, por ejemplo, los 500 mil pesos colombianos, y a medida que aumenta el salario, menor número de individuos puede obtener este ingreso. Observe la Figura 1.19 donde se dispone el histograma de datos que denotan el ingreso: nótese que la clase dominante se encuentra alrededor de los 500 mil pesos, y a medida que incrementa el ingreso, menos datos se ubican en ese rango, y además se presenta una cola larga hacia la derecha, las cuales son características propias de una distribución Gamma.

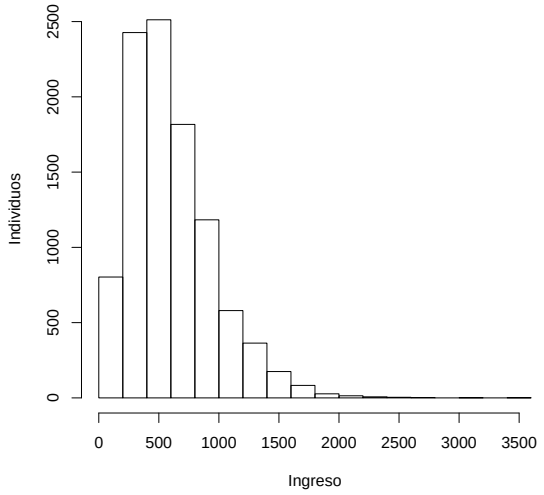

Figura 1.19: *Histograma de un conjunto de 10 mil datos con características de una distribución Gamma.*

Algunas propiedades de la distribución Gamma se enuncian a continuación.

Resultado 1.1.12. *Si X es una variable aleatoria con distribución Gamma con parámetro de forma k y parámetro de escala θ, entonces*

1. $E(X) = k\theta$.
2. $Var(X) = k\theta^2$.
3. $m_X(t) = \left(\frac{1}{1-\theta t}\right)^k$ *para* $t < 1/\theta$, *y no existe para otros valores de* t.

Del anterior resultado, podemos ver que el parámetro de escala θ se puede escribir como función de la esperanza y la varianza de la distribución como $\theta = Var(X)/E(X)$,

y por consiguiente, se tiene que el parámetro de forma se puede escribir como $k = E(X)/\theta = (E(X))^2/Var(X)$. De esta forma, para un conjunto de datos, una vez haya identificado que siguen una distribución Gamma, podemos calcular el promedio y la varianza de los datos y usarlos para tener un acercamiento a los dos parámetros de la distribución como $\theta' = s^2/\bar{x}$ y $k' = \bar{x}^2/s^2$.[5] El siguiente programa en R simula muestras de diferentes tamaños provenientes de una distribución $Gamma(3, 2)$, y en cada una de estas muestras calculan θ' y k'.

```
> set.seed(1234)
> n<-c(10,30,50,100,300,500,1000)
> tg<-matrix(NA)
> kg<-matrix(NA)
>
> for(i in 1:length(n)){
+ d<-rgamma(n[i],shape=3,scale=2)
+ tg[i]<-var(d)/mean(d)
+ kg[i]<-mean(d)/tg[i]
+ }
>
> par(mfrow=c(2,1))
> plot(tg,type="b",ylab="",xaxt="n",xlab="n",
+ main="Parámetro de escala")
> axis(1,1:length(n),n)
> abline(h=2)
>
> plot(kg,type="b",ylab="",xaxt="n",xlab="n",
+ main="Parámetro de forma")
> axis(1,1:length(n),n)
> abline(h=3)
```

Como resultado del anterior programa, se tiene la Figura 1.20, donde las dos líneas horizontales representan los valores verdaderos de θ y k. Podemos observar que los valores de θ' y k' se encuentran aproximadamente alrededor de los valores verdaderos, pero a medida que la muestra crece, no se observa mejora alguna.

La distribución Gamma tiene una buena propiedad que establece que la suma de variables con distribución Gamma puede seguir teniendo la distribución Gamma bajo algunos supuestos. Esta propiedad será útil para algunos desarrollos en los siguiente capítulos, y lo enunciamos en el siguiente resultado.

Resultado 1.1.13. *Sea $X_1, \cdots, X_n$ variables aleatorias independientes con distribución Gamma con parámetro de forma k_i y parámetro de escala θ para $i = 1, \cdots, n$, entonces la variable $\sum_{i=1}^n X_i$ tiene distribución Gamma con parámetro de forma $\sum_{i=1}^n k_i$ y parámetro de escala θ.*

[5] Este concepto se conoce como la estimación de los parámetros que se discutirá en el siguiente capítulo.

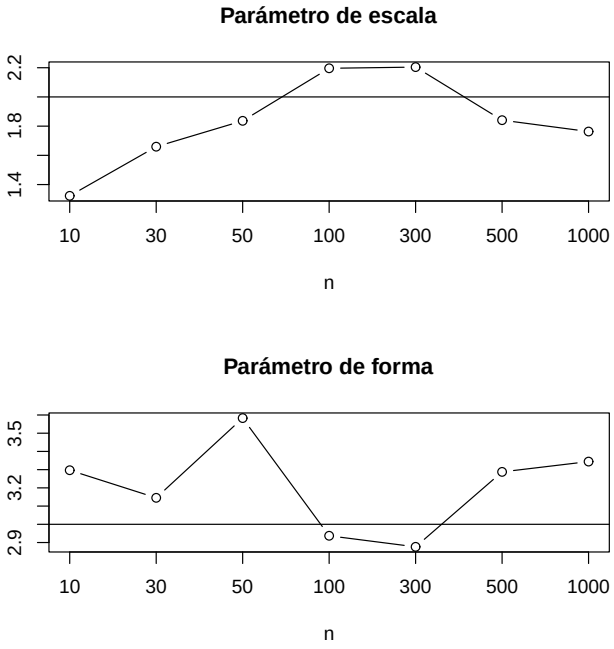

Figura 1.20: *Estimación de k y θ en muestras provenientes de una distribución Gamma de diferentes tamaños.*

Demostración. Análogo a la demostración del Resultado 1.1.4 y se deja como ejercicio (Ejercicio 1.9). □

Existe otro resultado interesante que liga la distribución Gamma con la distribución Poisson, y nos será de utilidad más adelante.

Resultado 1.1.14. *Si X es una variable aleatoria con distribución $Gamma(k, \theta)$ donde k es un número entero positivo, entonces para todo x, se tiene que $P(X \leq x) = P(Y \geq k)$ donde Y es una variable aleatoria con distribución $Pois(x/\theta)$.*

Podemos verificar de forma numérica la validez del anterior resultado. Sea $X \sim Gamma(2, 3.5)$, y $x = 4.2$ entonces $Pr(X \leq x) = Pr(X \leq 4.2)$ y se puede calcular con el comando `pgamma(4.2,shape=2,scale=3.5)` y arroja como resultado 0.337. Por otro lado, de acuerdo al resultado anterior, $Y \sim Pois(4.2/3.5) = Pois(1.2)$ y $Pr(Y \geq k) = Pr(Y \geq 2)$, el cual se calcula con el comando `ppois(k,4.2/3.5,lower.tail=F)+dpois(k,4.2/3.5)`, y arroja el mismo resultado 0.337.

De lo anterior observamos que una probabilidad con respecto a una variable con distribución Gamma se puede calcular en términos de una variable Poisson; el recíproco también es cierto. Suponga que $Y \sim Pois(5.3)$, entonces $Pr(Y \geq 4)$ se puede calcular con `ppois(4,5.3,lower.tail=F)+dpois(4,5.3)` y da como resultado 0.774. Pero

de acuerdo al anterior resultado, esta probabilidad también se puede calcular como $Pr(X \leq x)$ donde X tiene distribución Gamma como parámetro de escala $k = 4$, y el parámetro de escala θ debe satisfacer $x/\theta = 5.3$, y de esta ecuación se pueden hallar infinitas soluciones para x y θ; por ejemplo, al tomar $x = 1$, tenemos que $\theta = 1/5.3$, y así tenemos $Pr(X \leq 1)$ con $X \sim Gamma(4, 1/5.3)$. Al calcular esta probabilidad con `pgamma(1,shape=4,scale=1/5.3)` tenemos la misma probabilidad 0.774. El lector puede comprobar que si se hubiera escogido otro valor para x, por ejemplo $x = 10$, y se halla el valor $\theta = x/5.3$, la probabilidad de $Pr(X \leq 10)$ también es 0.774.

Con lo anterior, vemos que la probabilidad concerniente a una distribución Poisson puede ser calculada en términos de infinitas distribuciones Gamma. En particular si escogemos $\theta = 2$, la distribución Gamma se reduce a una distribución χ^2, y tenemos el siguiente resultado que es una consecuencia inmediata del Resultado 1.1.14.

Resultado 1.1.15. *Si Y es una variable aleatoria con distribución $Pois(\lambda)$, entonces para cualesquieras enteros y_1 y y_2, se tiene que*

$$P(Y \geq y_1) = P(X \leq 2\lambda)$$

donde $X \sim \chi^2_{2y_1}$. Y

$$P(Y \leq y_2) = 1 - P(Y \geq y_2 + 1) = 1 - P(X \leq 2\lambda)$$

donde $X \sim \chi^2_{2(y_2+1)}$.

Distribución exponencial

La distribución exponencial es un caso particular de la distribución Gamma cuando el parámetro de forma k toma el valor 1, y por consiguiente se puede obtener fácilmente la función de densidad dada a continuación.

Definición 1.1.8. *Una variable aleatoria X tiene distribución exponencial con parámetro de escala $\theta > 0$ si su función de densidad está dada por:*

$$f_X(x) = \frac{1}{\theta} e^{-x/\theta} I_{(0,\infty)}(x), \tag{1.1.10}$$

y en este libro, se usará la notación $X \sim Exp(\theta)$.

Una variable exponencial toma valores en el intervalo $(0, \infty)$, y puede ser utilizada para describir el tiempo necesario para la ocurrencia de algún evento o la vida útil de un componente eléctrico. En la Figura 1.21 se muestra la función de densidad de la distribución exponencial con diferentes valores de θ. En primer lugar, se observa que la función de densidad es siempre decreciente, y por consiguiente, para una variable $X \sim Exp(\theta)$, se tiene que $Pr(t_1 < X < t_1 + \delta) < Pr(t_2 < X < t_2 + \delta)$ si $t_1 > t_2$. Para ver la interpretación de eso, suponga que X denota la vida útil (en años) de una referencia de lavadora, entonces, como es natural, se afirma que es más probable que la lavadora funcione entre 2 y 3 años que entre 6 y 7 años, esto es,

$Pr(6 < X < 7) < Pr(2 < X < 3)$, que es una característica reflejada en la función de densidad de una distribución exponencial.

Más aún, suponga que dos tipos de lavadoras, A y B, tienen la vida útil (en años) que puede ser descrita por la distribución $Exp(2)$ y $Exp(5)$, respectivamente. Entonces, por el Resultado 1.1.12, donde provee propiedades de una distribución Gamma, podemos afirmar que las vidas útiles promedio de A y B son 2 y 5 respectivamente, es decir, las lavadoras del tipo B pueden funcionar por más años que los del tipo A. Por tanto, intuitivamente podemos afirmar que la probabilidad de que una lavadora funcione más de 6 años debe ser mayor en las del tipo B que en el tipo A, y recordando que esta probabilidad corresponde al área bajo la función de densidad en el intervalo $(6, \infty)$, podemos observar que la anterior afirmación sí se refleja en la Figura 1.21. Dadas las anteriores observaciones, podemos ver por qué la distribución exponencial es usada para describir este tipo de variables.

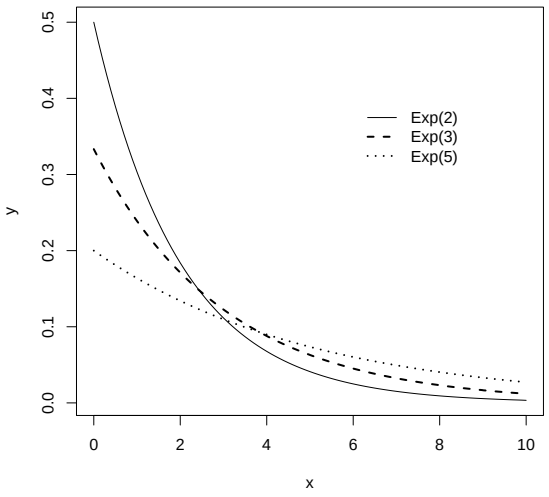

Figura 1.21: *Función de densidad de distribuciones* $Exp(2)$*,* $Exp(3)$ *y* $Exp(5)$.

Ahora bien aunque muchas veces se ha dicho en la literatura estadística que una variable aleatoria que denota el tiempo puede ser descrita por la distribución exponencial dado que ésta siempre toma valores positivos, característica propia de la variable tiempo, la función de densidad de la distribución exponencial es siempre decreciente, y puede no ser apta para algunas situaciones; por ejemplo, considere una central telefónica que atiende quejas de consumidores. Si se desea estudiar la variable X definida como el tiempo de duración de una llamada hecha por un consumidor, no es natural pensar que la probabilidad de que la llamada dura menos de un minuto sea mayor a la probabilidad de que dura menos de 5 minutos, esto es, puede no suceder que $Pr(X < 1) > Pr(X < 5)$; en otras palabras, no necesariamente, entre más corta sea la llamada, mayor probabilidad tiene asociada. Y en este caso, la distribución

exponencial no resulta adecuada, sino posiblemente una distribución Gamma. Nótese que la distribución exponencial es un caso particular de la distribución Gamma, por consiguiente se pueden obtener fácilmente sus propiedades usando el Resultado 1.1.12.

Resultado 1.1.16. *Si X es una variable aleatoria con distribución exponencial con parámetro θ, entonces*

1. $E(X) = \theta$.
2. $Var(X) = \theta^2$.
3. $m_X(t) = \frac{1}{1-\theta t}$ *para* $t < 1/\theta$, *y no existe para otros valores de* t.

Nótese que la varianza teórica de la distribución es el cuadrado de la esperanza. De esta forma, si un conjunto de datos continuos positivo tiene la varianza aproximadamente igual al promedio al cuadrado, podríamos afirmar que los datos provienen de una distribución exponencial.

El siguiente resultado es un caso particular del Resultado 1.1.13 y será de utilidad en los siguientes capítulos.

Resultado 1.1.17. *Sea $X_1, \cdots, X_n$ variables aleatorias independientes e idénticamente distribuidas con distribución exponencial con parámetro de escala θ, entonces la variable $\sum_{i=1}^{n} X_i$ tiene distribución Gamma con parámetro de forma n y parámetro de escala θ.*

Distribución Weibull

La distribución Weibull debe su nombre al sueco Ernst Hjalmar Waloddi Weibull (1887-1979) y es útil en la rama de la estadística denominada análisis de sobrevivencia donde se estudian variables que denotan el tiempo transcurrido hasta que suceda un evento como fallecimiento de un paciente o falla de algún componente eléctrico.

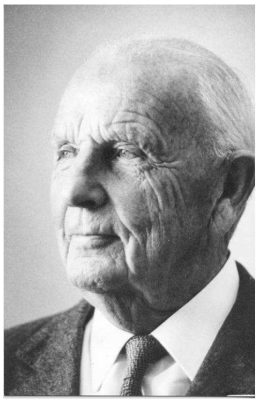

Figura 1.22: *Ernst Hjalmar Waloddi Weibull (1887-1979).*

La función de densidad de esta distribución se da a continuación.

Definición 1.1.9. *Una variable X tiene distribución Weibull con parámetro de forma $k > 0$ y parámetro de escala $\theta > 0$ si la función de densidad de X está dada por*

$$f_X(x) = \frac{k}{\theta^k} x^{k-1} \exp\left\{-\frac{x^k}{\theta^k}\right\} I_{(0,\infty)}(x). \tag{1.1.11}$$

Denotaremos esta distribución con $X \sim Weibull(k, \theta)$.

Nótese que la función de densidad de una distribución Weibull es similar en algunos términos a la de la distribución Gamma. De hecho, cuando el parámetro k toma valor 1, la distribución Weibull se reduce a una distribución Exponencial de media θ. En la Figura 1.23 se muestran algunas funciones de densidad de la distribución Weibull con diferentes valores de k y θ. Podemos observar que la función de densidad de la distribución Weibull(1,1) tiene la misma función de densidad que una distribución exponencial.

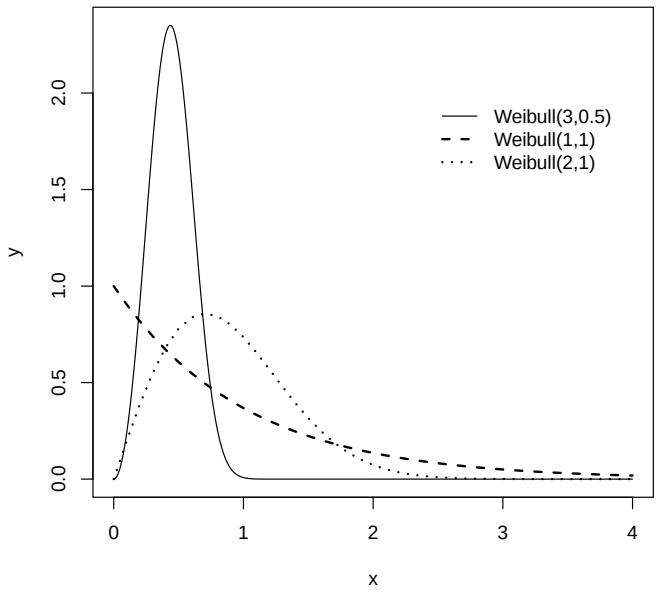

Figura 1.23: *Función de densidad de una distribución Weibull con diferentes parámetros.*

Algunas propiedades de la distribución Weibull se dan a continuación.

Resultado 1.1.18. *Si X es una variable aleatoria con distribución $Weibull(k, \theta)$, entonces*

1. $E(X) = \theta\Gamma\left(1+\frac{1}{k}\right)$,
2. $Var(X) = \theta^2\left[\Gamma\left(1+\frac{2}{k}\right) - \left(\Gamma\left(1+\frac{1}{k}\right)\right)^2\right]$

Distribución normal

La distribución normal también es llamada la distribución gaussiana, rindiendo homenaje al matemático alemán Carl Friedrich Gauss (1777-1855). Esta distribución es, sin duda, una de las distribuciones más importantes, y de uso más frecuente en la teoría estadística, puesto que una gran parte de la teoría estadística fue desarrollada inicialmente para variables con esta distribución. Por otra parte, gracias al teorema del límite central, muchas distribuciones ajenas a la normal, incluyendo las variables discretas, pueden ser aproximadas por ésta cuando el tamaño muestral es grande. Para detalles sobre la historia de la distribución normal, consulte a Stahl (2008).

Figura 1.24: *Carl Friedrich Gauss (1777-1855).*

Definición 1.1.10. *Una variable aleatoria X tiene distribución normal con parámetros μ y σ^2 si su función de densidad está dada por:*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} I_{\mathbb{R}}(x), \tag{1.1.12}$$

donde $\sigma > 0$ y se nota como $X \sim N(\mu, \sigma^2)$.

La distribución normal tiene dos parámetros, representado como $\boldsymbol{\theta} = (\mu, \sigma^2)$ y $\boldsymbol{\Theta} = \mathbb{R} \times (0, \infty)$. Cada uno de los dos parámetros de la distribución normal determina un aspecto específico de la distribución. En primer lugar, se puede ver fácilmente que para cualquier valor x, se tiene que $f(\mu + x) = f(\mu - x)$, de donde se deduce que la función de densidad es simétrica con respecto a μ. En segundo lugar, la función de

densidad toma el valor máximo en el punto $x = \mu$, y el máximo es igual a $(2\pi\sigma^2)^{-1/2}$, de donde se tiene que entre más pequeño sea el valor de σ^2, el área bajo la función de densidad está más concentrada alrededor del valor μ. En la Figura 1.25, se muestran algunas funciones de densidad para diferentes valores de μ y σ^2, donde se puede confirmar lo comentado anteriormente.

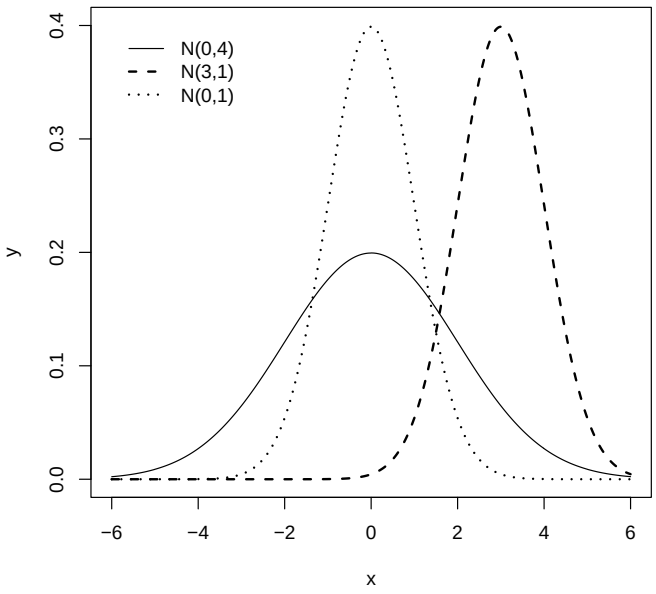

Figura 1.25: *Función de densidad de una distribución normal con diferentes parámetros.*

Ahora, haciendo una analogía entre el histograma de un conjunto de datos y una función de densidad, podemos sospechar que en primer lugar, si los datos provienen de una distribución normal, entonces el valor μ debe ubicarse alrededor del punto centro de los datos; y en segundo lugar, el valor σ^2 está asociado con la variación de los datos, ya que entre más pequeña sea ésta, más concentrados estarán los datos. El siguiente resultado confirma esta sospecha.

Resultado 1.1.19. *Si X es una variable aleatoria con distribución normal con parámetros μ y σ^2, entonces*

1. $E(X) = \mu$.
2. $Var(X) = \sigma^2$.
3. $m_X(t) = \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}$.

En la vida práctica, para que un conjunto de datos tenga distribución normal, una herramienta básica es examinar si el histograma de los datos tiene la forma llamada campana de Gauss, esto es, características similares a la función de densidad presentada en la Figura 1.25. La mayor frecuencia debe estar asociada con la clase media, y a medida que las clases se alejan de la clase media, la frecuencia debe disminuir simétricamente. En la Figura 1.26 se muestra el histograma de varios conjuntos de datos simulados usando la función `rnorm` de R. Obsérvese que la característica de la distribución normal es muy preeminente en muestras grandes, mientras que en muestras pequeñas, la detección de la normalidad mediante el histograma puede ser inadecuada.

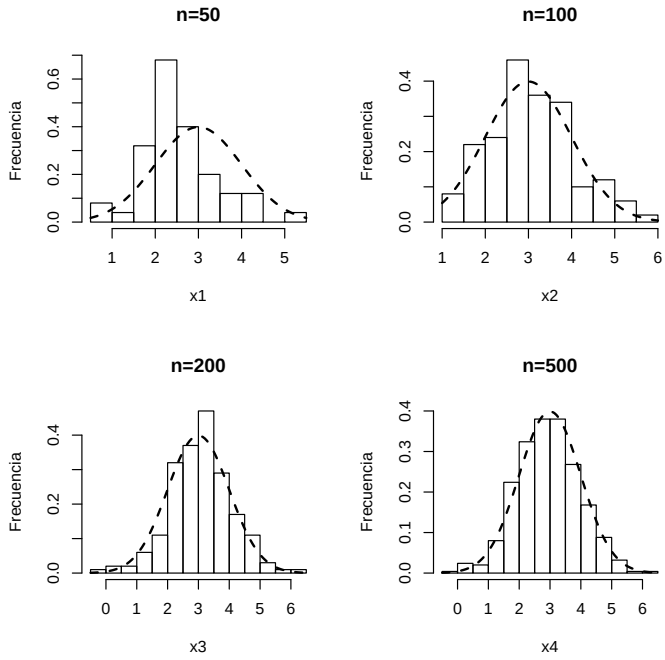

Figura 1.26: *Histograma de grupos de datos provenientes de la distribución normal con diferentes tamaños muestrales.*

Una propiedad muy particular de la distribución normal es que la distribución se conserva para transformaciones lineales, y se enuncia a continuación.

Resultado 1.1.20. *Si $X \sim N(\mu, \sigma^2)$, y α, β son constantes, entonces la variable $\alpha X + \beta$ tiene distribución $N(\alpha\mu + \beta, \alpha^2\sigma^2)$.*

Demostración. Se usará el hecho de que la función generadora de momentos caracteriza la distribución probabilística. Se tiene que:

$$\begin{aligned} m_{\alpha X+\beta}(t) &= E(e^{t(\alpha X+\beta)}) \\ &= E(e^{\alpha t X})e^{\beta t} \\ &= m_X(\alpha t)e^{\beta t} \\ &= e^{\mu\alpha t+\sigma^2\alpha^2 t/2}e^{\beta t} \\ &= e^{(\alpha\mu+\beta)t+\sigma^2\alpha^2 t/2} \end{aligned}$$

la cual es la función generadora de momentos de una distribución $N(\alpha\mu+\beta, \alpha^2\sigma^2)$, y el resultado queda demostrado. □

Como consecuencia inmediata del anterior resultado, se define la estandarización que es fundamental en la teoría relacionada con las distribuciones normales.

Definición 1.1.11. *Si $X \sim N(\mu, \sigma^2)$ con $\mu = 0$ y $\sigma = 1$, entonces se dice que X tiene distribución normal estándar y usualmente se denota por Z.*

Utilizando el Resultado 1.1.20, podemos comprobar que una variable $X \sim N(\mu, \sigma^2)$ puede ser transformada a una variable Z mediante una transformación lineal. Suponga que esta transformación se denota por $\alpha X + \beta$, entonces encontrar los valores de α y β para los cuales la variable transformada tenga distribución normal estándar equivale a solucionar las siguientes igualdades para α y β

$$\begin{cases} \alpha\mu + \beta = 0 \\ \alpha^2\sigma^2 = 1 \end{cases}$$

de donde se tienen dos soluciones $\alpha_1 = 1/\sigma$, $\beta_1 = -\mu/\sigma$ y $\alpha_2 = -1/\sigma$, $\beta_2 = \mu/\sigma$. Y de esta forma, hemos encontrado dos variables $Z_1 = \frac{X-\mu}{\sigma}$ y $Z_2 = \frac{\mu-X}{\sigma}$ con la distribución normal estándar. Sin embargo, se acostumbra a utilizar la transformación dada por la primera solución, y esta transformación se conoce como la estandarización, y la variable $Z = Z_1 = \frac{X-\mu}{\sigma}$ se conoce como la variable X estandarizada.

Análoga a las distribuciones Poisson, Gamma, la distribución normal también se conserva para suma de variables normales independientes, tal como lo muestra el siguiente resultado.

Resultado 1.1.21. *Sea $X_1, \cdots, X_n$ variables aleatorias independientes, donde $X_i \sim N(\mu_i, \sigma_i^2)$ con $i = 1, \cdots, n$, entonces la variable $\sum_{i=1}^n X_i \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.*

Demostración. Se deja como ejercicio (Ejercicio 1.13). □

Combinando los resultados 1.1.20 y 1.1.21, se puede establecer que el promedio de variables independientes con la misma distribución normal sigue teniendo distribución normal, y lo enunciamos a continuación.

Resultado 1.1.22. *Sea $X_1, \cdots, X_n$ variables aleatorias independientes e idénticamente distribuidas con distribución $N(\mu, \sigma^2)$, entonces la variable $\bar{X} = \sum_{i=1}^{n} X_i/n$ tiene distribución $N(\mu, \sigma^2/n)$.*

Demostración. Se deja como ejercicio (Ejercicio 1.14). □

Una de las razones por las que la distribución normal es de las más importantes en la teoría estadística radica en el hecho de que en un conjunto de variables independientes e idénticamente distribuidas no necesariamente con distribución normal, si el número de variables es grande, entonces la distribución del promedio de estas variables puede ser aproximada por la de una distribución normal. Lo anterior se conoce como el famoso ≪teorema del límite Central≫ y se enuncia a continuación.

Resultado 1.1.23. *Sea $X_1, X_2, \cdots$, una sucesión de variables aleatorias independientes e idénticamente distribuidas, suponga que las funciones generadoras de momentos $m_{X_i}(t)$ existen en una vecindad de 0, y la esperanza común se denota por μ y varianza común se denota por $\sigma^2 > 0$, y se define $\bar{X}_n$ como $\sum_{i=1}^{n} X_i/n$, y la función de distribución de la variable $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ se denota por $F_n(x)$, entonces se tiene que*

$$\lim_{n\to\infty} F_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy,$$

es decir, la distribución límite de $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ corresponde a la distribución normal estándar.

Demostración. Se probará que la función generadora de momentos de la variable $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converge a la función $e^{t^2/2}$ que corresponde a la función generadora de momentos de una distribución $N(0,1)$. Para eso primero se define Z_i como la variable X_i estandarizada, entonces $Z_i \sim N(0,1)$ para todo $i = 1, 2, \cdots$. Y podemos comprobar fácilmente que $\sqrt{n}(\bar{X}_n - \mu)/\sigma = \sum_{i=1}^{n} Z_i/\sqrt{n}$, entonces tenemos

$$\begin{aligned} m_{\sqrt{n}(\bar{X}_n-\mu)/\sigma}(t) &= m_{\sum_{i=1}^{n} Z_i/\sqrt{n}}(t) \\ &= \prod_{i=1}^{n} m_{Z_i}\left(\frac{t}{\sqrt{n}}\right) \qquad \text{por la independencia} \\ &= \left(m_Z\left(\frac{t}{\sqrt{n}}\right)\right)^n, \end{aligned}$$

donde Z denota una variable con distribución normal estándar. Ahora, usamos la expansión de Taylor para la función $m_Z\left(\frac{t}{\sqrt{n}}\right)$ alrededor del punto 0. Para eso recordamos que si $g(x)$ es una función derivable de cualquier orden, entonces se puede expandir $g(x)$ alrededor de un punto a como

$$g(x) = \sum_{i=0}^{\infty} \frac{g^{(i)}(a)(x-a)^i}{i!}, \tag{1.1.13}$$

donde $g^{(i)}(a)$ es la i-ésima derivada de $g(x)$ evaluada en $x = a$ y $g^{(0)}(a) = g(a)$.

De esta forma, tenemos que

$$\begin{aligned} m_Z\left(\frac{t}{\sqrt{n}}\right) &= \sum_{i=0}^{\infty} \frac{m_Z^{(i)}(0)\left(\frac{t}{\sqrt{n}}\right)^i}{i!} \\ &= m_Z(0) + m_Z'(0)\left(\frac{t}{\sqrt{n}}\right) + \frac{1}{2}m_Z''(0)\left(\frac{t}{\sqrt{n}}\right)^2 + \sum_{i=3}^{\infty} \frac{m_Z^{(i)}(0)\left(\frac{t}{\sqrt{n}}\right)^i}{i!} \\ &= 1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + R\left(\frac{t}{\sqrt{n}}\right), \end{aligned}$$

donde

$$R\left(\frac{t}{\sqrt{n}}\right) = \sum_{i=3}^{\infty} \frac{m_Z^{(i)}(0)\left(\frac{t}{\sqrt{n}}\right)^i}{i!}.$$

Y por consiguiente, tenemos que

$$\begin{aligned} \lim_{n\to\infty} m_{\sqrt{n}(\bar{X}_n-\mu)/\sigma}(t) &= \lim_{n\to\infty}\left(m_Z\left(\frac{t}{\sqrt{n}}\right)\right)^n \\ &= \lim_{n\to\infty}\left(1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + R\left(\frac{t}{\sqrt{n}}\right)\right)^n \\ &= \lim_{n\to\infty}\left[1 + \frac{1}{n}\left(\frac{t^2}{2} + nR\left(\frac{t}{\sqrt{n}}\right)\right)\right]^n. \end{aligned} \tag{1.1.14}$$

Por otro lado, el teorema de Taylor afirma que para la función g en (1.1.13), se tiene que

$$\lim_{x\to a} \frac{\sum_{i=r+1}^{\infty} \frac{g^{(i)}(a)(x-a)^i}{i!}}{(x-a)^r} = 0.$$

Aplicando lo anterior a $m_Z\left(\frac{t}{\sqrt{n}}\right)$ con $r = 2$, tenemos que

$$\lim_{\frac{t}{\sqrt{n}}\to 0} \frac{R\left(\frac{t}{\sqrt{n}}\right)}{\left(\frac{t}{\sqrt{n}}\right)^2} = 0,$$

la cual es equivalente a

$$\lim_{n\to\infty} \frac{R\left(\frac{t}{\sqrt{n}}\right)}{\left(\frac{1}{\sqrt{n}}\right)^2} = \lim_{n\to\infty} nR\left(\frac{t}{\sqrt{n}}\right) = 0,$$

para todo t. Y por consiguiente

$$\lim_{n\to\infty}\left(\frac{t^2}{2} + nR\left(\frac{t}{\sqrt{n}}\right)\right) = \frac{t^2}{2}. \tag{1.1.15}$$

Finalmente, combinando (1.1.14) y (1.1.15), y usando el hecho de que si una sucesión de números $a_n \to a$, entonces $\lim_{n\to\infty}\left(1+\frac{a_n}{n}\right)^n = e^a$, se tiene que

$$\lim_{n\to\infty}\left[1+\frac{1}{n}\left(\frac{t^2}{2}+nR\left(\frac{t}{\sqrt{n}}\right)\right)\right]^n = e^{t^2/2},$$

y en conclusión

$$\lim_{n\to\infty} m_{\sqrt{n}(\bar{X}_n-\mu)/\sigma}(t) = e^{t^2/2}.$$

En conclusión, la distribución límite de $\sqrt{n}(\bar{X}_n-\mu)/\sigma$ corresponde a la distribución normal estándar. □

En el anterior teorema se exige la existencia de la función generadora de momentos de las variables X_1, X_2, $\cdots$; en general, se puede demostrar la validez del resultado aún sin este supuesto, y la demostración se realiza por medio de funciones características de manera análoga.

El teorema del límite central es una herramienta muy poderosa en el sentido de que se aplica para la mayoría de las distribuciones de probabilidad; sin embargo, el teorema no nos brinda una medida de qué tan buena es la aproximación, y se debe examinar para cada distribución. En las figuras 1.27 y 1.28, se muestran las distribuciones muestrales del promedio en muestras simuladas distribuciones $Pois(3)$ y $Gamma(3,2)$ con tamaños de muestral 5, 10, 30, 50, 100 y 500 respectivamente. Podemos observar que efectivamente la distribución $\bar{X}$ se aproxima a una distribución normal, especialmente para muestras grandes.

Distribución Ji-cuadrado

Otra distribución como caso particular de la distribución Gamma es la distribución Ji-cuadrado con n grados de libertad es un caso particular de la distribución Gamma cuando el parámetro de forma k toma el valor $n/2$ para algún n entero, y el parámetro de escala θ toma el valor 2. La función de densidad de la distribución Ji-cuadrado se muestra en la siguiente definición.

Definición 1.1.12. *Una variable aleatoria X tiene distribución Ji-cuadrado con n grados de libertad, con n entero positivo, si su función de densidad está dada por:*

$$f_X(x) = \frac{x^{(n/2)-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}I_{(0,\infty)}(x), \tag{1.1.16}$$

y se nota como $X \sim \chi^2_n$.

Aunque la distribución Ji-cuadrado es un caso particular de la distribución Gamma, pero ésta está íntimamente relacionada con la distribución normal, tal como se muestra en la siguiente definición equivalente a la anterior, y resulta ser muy útil en el momento de demostrar que una variable tiene la distribución Ji-cuadrado.

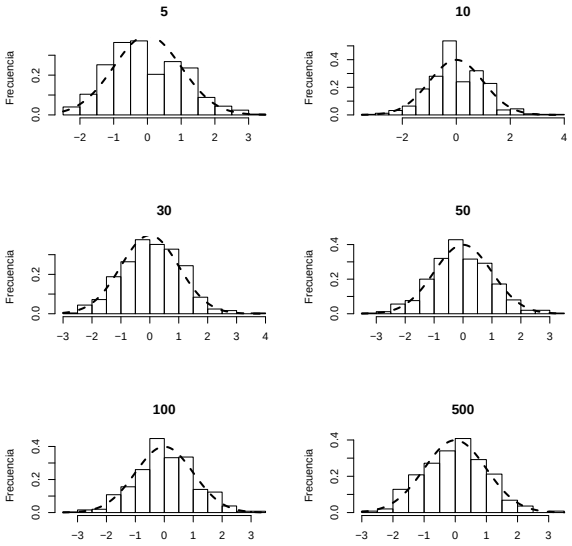

Figura 1.27: *Distribución muestral del promedio en muestras con distribución* $Pois(3)$ *de diferentes tamaños.*

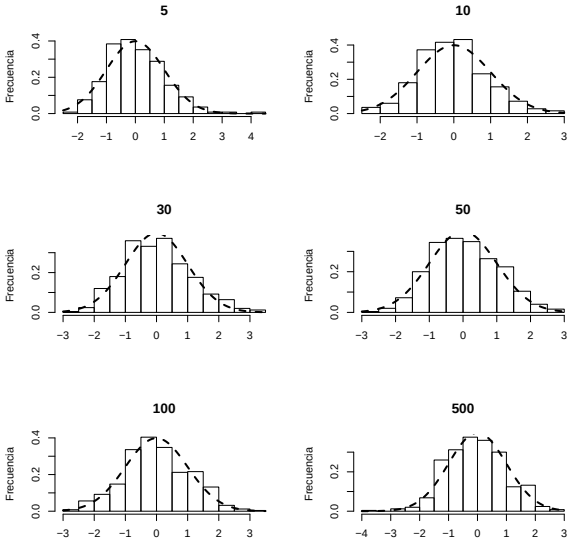

Figura 1.28: *Distribución muestral del promedio en muestras con distribución* $Gamma(3,2)$ *de diferentes tamaños.*

Definición 1.1.13. *Si $Z_1, \cdots, Z_n$ son variables aleatorias independientes e idénticamente distribuidas con distribución normal estándar, entonces la variable $\sum_{i=1}^{n} Z_i^2$ tiene distribución Ji-cuadrado con n grados de libertad.*

Dado que la distribución χ^2 es un caso particular de la distribución Gamma, la función de densidad también tiene facetas similares tales como la no simetría y la cola larga. En la Figura 1.29 se muestra la función de densidad de distribución χ^2_n para diferentes valores de n.

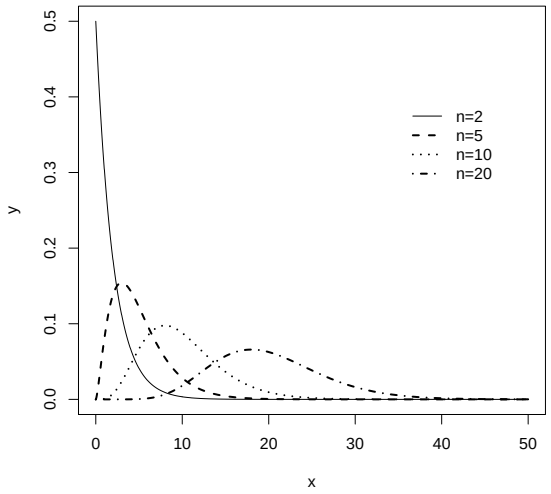

Figura 1.29: *Función de densidad de distribuciones Ji-cuadrado con diferentes grados de libertad.*

Usando el Resultado 1.1.12 para la distribución Gamma, se tiene fácilmente las siguientes propiedades de una variable con distribución Ji-cuadrado.

Resultado 1.1.24. *Si X es una variable aleatoria con distribución Ji-cuadro con n grados de libertad, entonces*

1. $E(X) = n$.
2. $Var(X) = 2n$.
3. $m_X(t) = \left(\frac{1}{1-2t}\right)^{n/2}$ *para* $t < 1/2$, *y no existe para otros valores de* t.

A continuación se presenta un resultado que nos será muy útil en los capítulos futuros.

Resultado 1.1.25. *Sea $X_1, \cdots, X_m$ variables aleatorias independientes con distribución $\chi^2_{n_i}$ para $i = 1, \cdots, m$, entonces la variable $X = \sum_{i=1}^m X_i$ tiene distribución Ji-cuadrado con $\sum_{i=1}^m n_i$ grados de libertad.*

Demostración. Se hará uso de la función generadora de momentos, tenemos que

$$\begin{aligned} m_X(t) &= E(e^{t\sum_{i=1}^m X_i}) \\ &= \prod_{i=1}^n E(e^{tX_i}) \\ &= \prod_{i=1}^n m_{X_i}(t) \\ &= \prod_{i=1}^n \left(\frac{1}{1-2t}\right)^{n_i/2} \\ &= \left(\frac{1}{1-2t}\right)^{\sum n_i/2}, \end{aligned}$$

la cual corresponde a la función generadora de momentos de una distribución χ^2 con grado de libertad $\sum_{i=1}^m n_i$, y el resultado queda demostrado. □

El anterior resultado establece que la suma de variables independientes con distribución χ^2 sigue teniendo la distribución χ^2. ¿Se puede afirmar que la resta de dos variables independientes χ^2 sigue manteniendo la distribución χ^2? Suponga que $X \sim \chi^2_{n_1}$ y $Y \sim \chi^2_{n_2}$ son independientes, y sea $Z = X - Y$, entonces

$$\begin{aligned} m_Z(t) &= E(e^{t(X-Y)}) \\ &= E(e^{tX})/E(e^{tY}) \qquad \text{Por ser } X \text{ y } Y \text{ independientes} \\ &= (1-2t)^{-n_1/2}/(1-2t)^{-n_2/2} \\ &= (1-2t)^{-(n_1-n_2)/2}, \end{aligned}$$

la cual corresponde a la función generadora de momentos de una distribución χ^2 con grado de libertad $n_1 - n_2$ siempre y cuando $n_1 > n_2$. Por lo anterior, podemos afirmar que la resta de dos variables independientes χ^2 sigue teniendo la distribución χ^2 siempre y cuando la resta de los dos grados de libertad sea positiva.

Distribución t-student

Otra distribución de vital importancia en la teoría estadística es la denominada distribución t-student. El descubrimiento de esta distribución fue publicado por el estadístico inglés William Sealy Gosset (1876-1937) en el año 1908 cuando trabajaba en la famosa empresa cervecera Guinness. La publicación la hizo de forma anónima bajo el nombre de Student, pues Guinness le prohibía la publicación por ser el descubrimiento parte de resultados de investigación realizada por la empresa. La definición de esta distribución se da a continuación.

Figura 1.30: *William Sealy Gosset (1876-1937)*

Definición 1.1.14. *Una variable aleatoria X tiene distribución t-student con n grados de libertad si su función de densidad está dada por:*

$$f_X(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\,\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2} I_{\mathbb{R}}(x), \tag{1.1.17}$$

donde $n > 0$ y se nota como $X \sim t_n$.

Otra definición que se encuentra frecuentemente en la literatura estadística es la siguiente, que es más útil que la definición anterior para demostrar que una variable tiene distribución t.

Definición 1.1.15. *Sea Z una variable aleatoria con distribución normal estándar y Y una variable aleatoria con distribución Ji-cuadrado con n grados de libertad, si Z y Y son independientes, entonces la variable $\frac{Z}{\sqrt{Y/n}}$ tiene distribución t-student con n grados de libertad.*

La función de densidad de la distribución t-student es muy parecida a la de distribución normal estándar, tiene la forma de campana de Gauss y simétrica con respecto al valor 0, además entre más grande sea el grado de libertad, más se parece a la distribución normal estándar. En la Figura 1.31 se muestra la función de densidad de la distribución normal estándar y la de distribución t con diferentes grados de libertad donde podemos observar la similitud entre estas distribuciones.

Algunas propiedades de la distribución t-student se muestran en el siguiente resultado.

Resultado 1.1.26. *Si X es una variable aleatoria con distribución t-student con n grados de libertad, entonces*

1. $E(X) = 0$ *para* $n > 1$.
2. $Var(X) = \frac{n}{n-2}$ *para* $n > 2$.

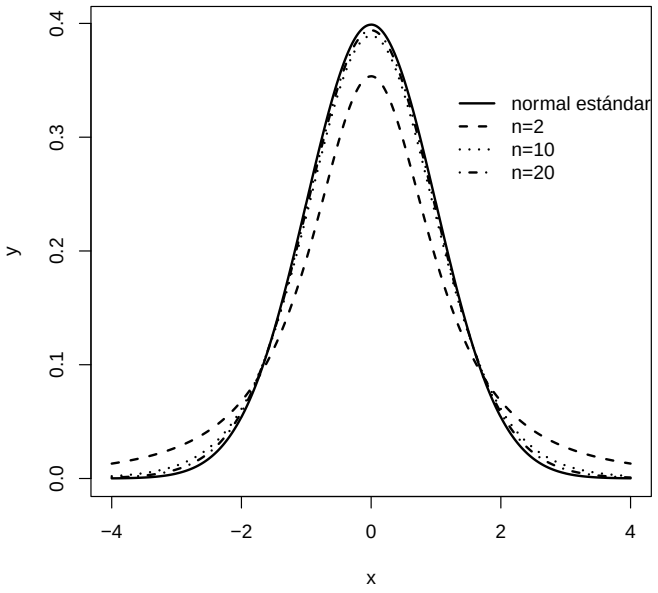

Figura 1.31: *Funciones de densidad de la distribución* $N(0,1)$ *y* t_n *con diferentes grados de libertad.*

En primer lugar, nótese que cuando n es grande, la varianza de X se aproxima al valor 1, la varianza de una distribución normal estándar. Por otro lado, cabe resaltar que la distribución t-student no tiene función generadora de momentos.

Distribución t-student no central

La distribución t student no central es una extensión natural de la t student descrita anteriormente[6], que permte que la esperanza de la distribución normal no sea cero, de allí el término ≪no central≫ para esta distribución. Análogo a la distribución t central, existen definiciones equivalentes para la distribución t-student no central, las cuales presentamos a continuación.

Definición 1.1.16. *Una variable aleatoria X tiene distribución* t*-student no central con grados de libertad* n*, y parámetro de no centralidad* μ *si su densidad es*

[6]La distribución con función de densidad (1.1.17) también se conoce como la t-student central. En este libro se utilizarán indistintamente estos dos nombres.

$$f_X(x) = \frac{n^{n/2}\exp\left\{-\frac{n\mu^2}{2(x^2+n)}\right\}}{\sqrt{\pi}\Gamma(n/2)2^{(n-1)/2}(x^2+n)^{(n+1)/2}}\int_0^\infty v^n \exp\left\{-\frac{1}{2}\left(v-\frac{\mu x}{\sqrt{x^2+n}}\right)^2\right\}dv$$

y se denota como $X \sim t^{nc}_{n,\mu}$.

La función de densidad en la anterior definición, es sin duda muy compleja para cualquier cálculo concerniente a esta distribución. Para efectos de este libro, solo haremos uso de su función de distribución cuyo cálculo se puede llevar a cabo usando la instrucción `pt` en R, más adelante explicaremos con detalles sobre el uso de esta función. La otra definición equivalente para la distribución t-student no central es por medio de construcción, y es útil a la hora de probar que cierta variable tiene esta distribución.

Definición 1.1.17. *Sea* X *una variable aleatoria con distribución* $N(\mu,1)$*, y* Y *una variable aleatoria con distribución Ji-cuadrado con* n *grados de libertad, si* X *y* Y *son independientes, entonces la variable* $\frac{X}{\sqrt{Y/n}}$ *tiene distribución* t*-student no central con grados de libertad* n*, y parámetro de no centralidad* μ

De la anterior definición, podemos ver que si X y Y son variables aleatorias independientes con $X \sim N(\mu,\sigma^2)$ y $Y \sim \chi^2_n$ entonces

$$\frac{X/\sigma}{\sqrt{Y/n}} \sim t^{nc}_{n,\frac{\mu}{\sigma}}.$$

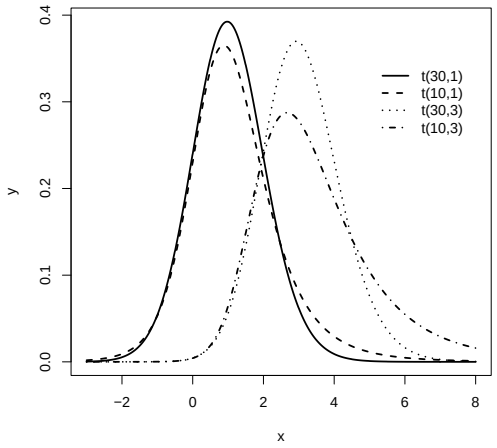

Figura 1.32: *Distintas funciones de densidad* t*-student no central.*

En la Figura 1.32 se grafica la función de densidad de la distribución t-student no central con diferentes grados de libertad y diferentes parámetros de no centralidad. Se puede observar que la esperanza de la distribución t no central es cercana, aunque no igual al parámetro de no centralidad μ; además la función de densidad de esta distribución no es simétrica, pero se torna más simétrica cuando el grado de libertad es más grande.

Distribución F

Otra distribución muy útil es la distribución F que también se conoce como la distribución F de Fisher o distribución de Fisher-Snedecor, haciendo referencia al gran estadístico Ronald Aylmer Fisher (1890-1962) y el fundador del primer departamento de estadística en los Estados Unidos, George Waddel Snedecor (1881-1974).

Figura 1.33: *Ronald Aylmer Fisher (1890-1962)*

La función de densidad de la distribución F se da en la siguiente definición.

Definición 1.1.18. *Una variable aleatoria X tiene distribución F con m grados de libertad en el numerador y n grados de libertado en el denominador si su función de densidad está dada por:*

$$f_X(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\left(\frac{m}{n}\right)^{m/2}\frac{z^{\frac{m}{2}-1}}{\left(1+\frac{m}{n}z\right)^{\frac{m+n}{2}}}I_{(0,\infty)}(x), \tag{1.1.18}$$

y se nota como $X \sim F_n^m$.

Otra definición equivalente pero más útil de la distribución F es como sigue:

Definición 1.1.19. *Sea X y Y variables aleatorias independientes con distribuciones Ji-cuadrado con m y n grados de libertad, respectivamente, entonces la variable $\dfrac{X/m}{Y/n}$ tiene distribución F con m grados de libertad en el numerador y n grados de libertado en el denominador.*

En la Figura 1.34 se observa la función de densidad de la distribución F para diferentes grados de libertad. Podemos observar que ésta es similar a la de una distribución Gamma, no simétrica con cola larga y en algunos casos similar a la distribución exponencial.

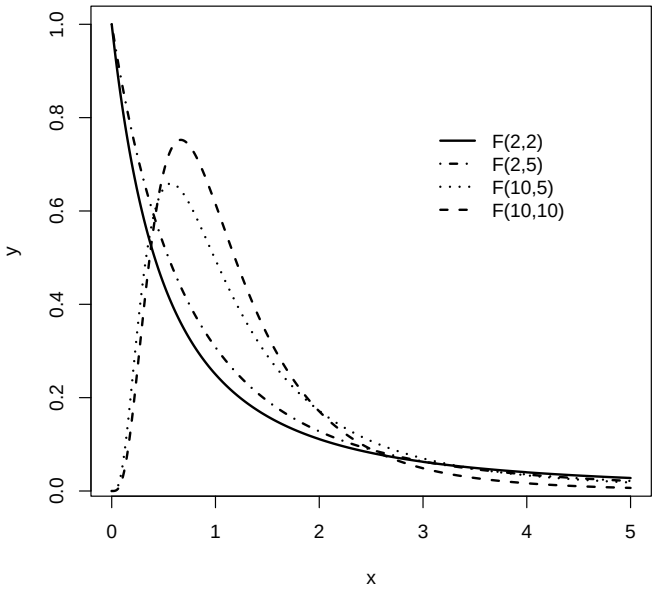

Figura 1.34: *Funciones de densidad de la distribución F con diferentes grados de libertad.*

Algunas propiedades de la distribución F se dan a continuación.

Resultado 1.1.27. *Si X es una variable aleatoria con distribución F con m grados de libertad en el numerador y n grados de libertado en el denominador, entonces*

1. $E(X) = \frac{n}{n-2}$ *para* $n > 2$.
2. $Var(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ *para* $n > 4$.
3. *La distribución F no tiene función generadora de momentos.*

Usando la Definición 1.1.19, se tiene fácilmente el siguiente resultado.

Resultado 1.1.28. *Si* $X \sim F_n^m$, *entonces* $1/X \sim F_m^n$.

Demostración. Trivial usando la Definición 1.1.19. □

Distribución Beta

La distribución Beta se difiere de las otras distribuciones continuas vistas anteriormente en el sentido de que la distribución Beta solo se usa para describir variables aleatorias que toman valores en el intervalo (0,1). Dado que los datos porcentuales tienen un rango entre 0 y 1[7], es común pensar que una distribución Beta puede ser apropiada para describir variables aleatorias que representan porcentajes. A continuación damos la definición de esta distribución en términos de su función de densidad.

Definición 1.1.20. *Una variable aleatoria X tiene distribución Beta con parámetros $a > 0$ y $b > 0$ si su función de densidad está dada por*

$$f_X(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} I_{(0,1)}(x) \tag{1.1.19}$$

donde $B(a,b)$ es la función Beta dada por

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 u^{a-1}(1-u)^{b-1} du.$$

Usaremos la notación $X \sim Beta(a,b)$ para una variable con esta distribución.

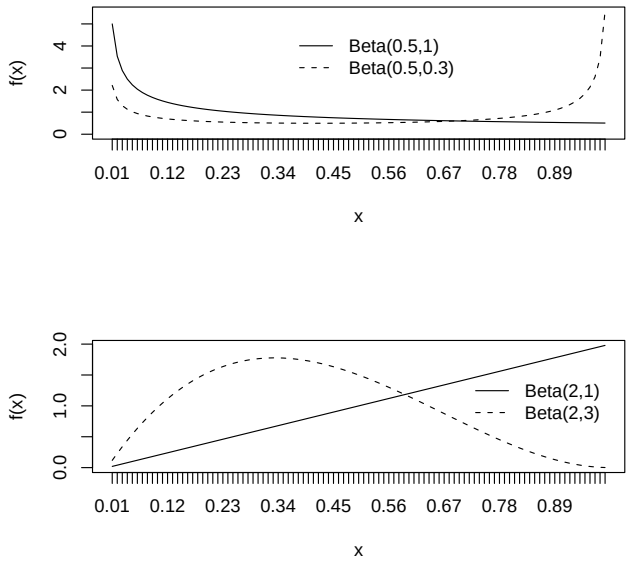

Figura 1.35: *Funciones de densidad de algunas distribuciones Beta.*

[7] Cuando un porcentaje es 100 %.

Podemos notar que cuando ambos parámetros a y b toman el valor 1, la distribución $Beta(a, b)$ se reduce a una distribución uniforme continua sobre el intervalo $(0, 1)$, esto es, $Beta(1, 1) = Unif(0, 1)$. Por otro lado, podemos ver que la función de densidad de una distribución Beta depende de x por medio de un polinomio de grado $a+b-2$, y la función de densidad toma diversas formas según cambian los parámetros a y b. En la Figura 1.35 se muestra la función de densidad de algunas distribuciones Beta. Como puede ver el lector, la función de densidad puede ser una línea recta, polinomial, y en algunos casos puede no tener un máximo o modas. Esta gran versatilidad permite que la distribución Beta sea apta para describir muchas variables, siempre y cuando éstas toman valores en el intervalo (0,1).

Algunas propiedades de la distribución Beta se enuncian a continuación.

Resultado 1.1.29. *Si X es una variable aleatoria con distribución $Beta(a, b)$, entonces*

1. $E(X) = \frac{a}{a+b}$.
2. $Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$.

Observando la forma de la función de densidad de una distribución Beta dada en (1.1.19) podemos observar cierta simetría para $f(x)$ y $f(1 - x)$, sugiriendo que la variable $1 - X$ puede también tener la distribución Beta si X la tiene. Tenemos el siguiente resultado.

Resultado 1.1.30. *Si $X \sim Beta(a, b)$ entonces $Y = 1 - X \sim Beta(b, a)$.*

Demostración. Una forma de probar lo anterior es encontrando la función de distribución de Y, tenemos que para $y \in (0, 1)$,

$$\begin{aligned} F_Y(y) = Pr(Y \leq y) &= Pr(1 - X \leq y) \\ &= Pr(X \geq 1 - y) \\ &= \int_{1-y}^{1} \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)} dx \end{aligned}$$

Haciendo un cambio de variables $x = 1 - u$, tenemos que

$$F_Y(y) = \int_0^y \frac{(1-u)^{a-1}u^{b-1}}{B(a, b)} dx$$

Y como $B(a, b) = B(b, a)$, podemos concluir que $Y \sim Beta(b, a)$. □

Relaciones de las distribuciones

John D. Cook ha elaborado la gráfica mostrada en la Figura 1.36 acerca de las relaciones de las distribuciones de probabilidad. Según Cook, el gráfico está inspirado en Leemis & McQueston (2008), en donde los autores muestran en un gráfico mucho más extenso y completo cómo cada una de las distribuciones de probabilidad univariadas se relacionan y forman familias según las propiedades que tengan.

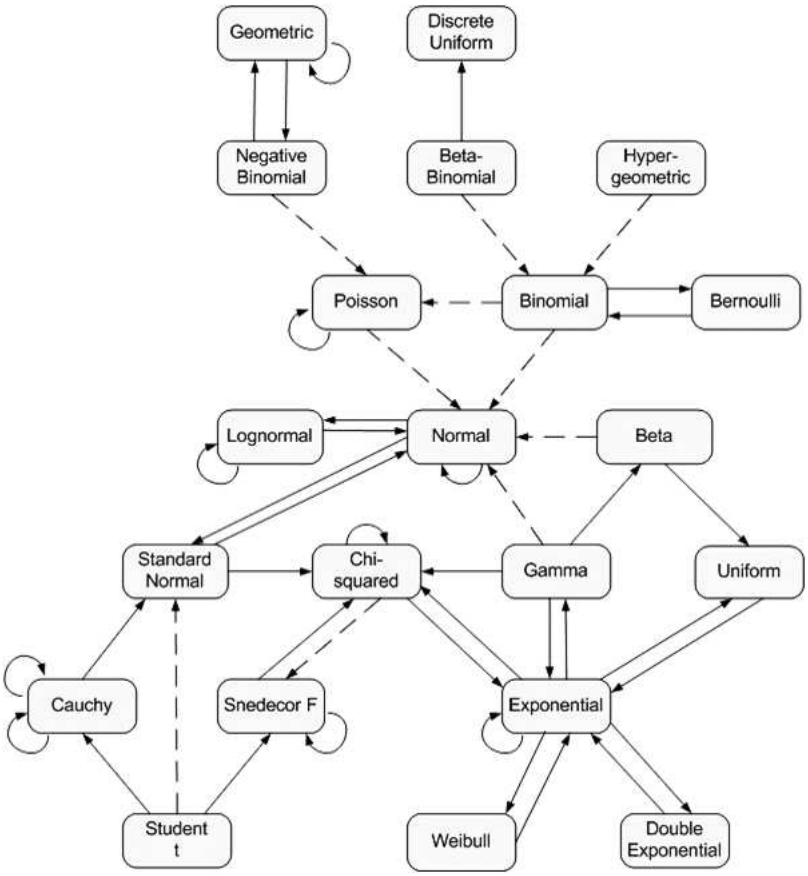

Figura 1.36: *Relaciones entre diferentes distribuciones.*

Entre las propiedades que parametrizan las distribuciones de probabilidad se encuentran las siguientes:

- Propiedad de combinación lineal: por ejemplo, la suma de variables aleatorias normales independientes resulta tener una distribución normal.

- Propiedad de convolución: por ejemplo, una variable aleatoria con distribución Ji-cuadrado y con n grados de libertad puede venir de la suma de n variables independientes con distribución Ji-cuadrado y con un grado de libertad.

- Propiedad del producto: por ejemplo, el producto de variables independientes con distribución lognormales resulta tener una distribución lognormal.

- Propiedad de la inversa: por ejemplo, si una variable aleatoria tiene distribución F, su inversa aritmética también tendrá distribución F.

1.1.3 Percentiles

Un concepto relacionado con las variables aleatorias que es muy importante para la inferencia estadística, específicamente la teoría de intervalo de confianza y las pruebas de hipótesis, es el concepto del percentil de una variable aleatoria. Definimos este concepto a continuación.

Definición 1.1.21. *Para una variable aleatoria X, el percentil p de X, con $0 < p < 1$, se define como $\inf\{x : \ F_X(x) \geq p\}$ y se denota como X_p. Esto es, X_p es el valor más pequeño que acumula una probabilidad no inferior de p.*

Para variables aleatorias continuas, la función de distribución correspondiente también es continua, y existe un único punto x con $F(x) = p$, de donde podemos ver que el percentil p de X es simplemente $F^{-1}(p)$.

Ejemplo 1.1.1. *Sea X una variable aleatoria con distribución $Exp(\theta)$, entonces la función de distribución de X está dada por*

$$F_X(x) = \begin{cases} 1 - e^{-x/\theta} & \text{si } x > 0 \\ 0 & \text{si no} \end{cases} \tag{1.1.20}$$

Podemos ver que F_X tiene inversa para valores de x en $(0, \infty)$ dada por $F^{-1}(x) = -\theta \ln(1 - x)$, con $0 < x < 1$. De donde podemos calcular el percentil p de la distribución $Exp(\theta)$ como $-\theta \ln(1 - p)$. Por ejemplo, el percentil 0.3 de $Exp(4)$ está dado por $-4 \ln(1 - 0.3) = 1.4267$.

Por otro lado, en R, podemos obtener los percentiles de una distribución exponencial con la instrucción `qexp` *teniendo en cuenta que R utiliza una parametrización diferente que la de este libro. El comando para calcular el percentil 0.3 de $Exp(4)$ y el resultado es como sigue*

```
> qexp(0.3,1/4)
[1] 1.426700
```

En algunas distribuciones continuas, como la distribución normal estándar, la función de distribución es de una forma muy complicada, y se requiere de algoritmos numéricos para calcular los percentiles de esta distribución. El comando en R para ese fin es `qnorm` y usa el algoritmo propuesto por Wichura (1988). En este texto, denotaremos el percentil p de una distribución normal estándar como z_p, y los percentiles comunes que se utilizarán a lo largo del texto que se encuentran en el apéndice G.

Para variables aleatorias discretas, el cálculo de los percentiles es un poco más complicado, puesto que la función de distribución es una función escalonada, y por consiguiente no tiene inversa. Ilustramos el cálculo en el siguiente ejemplo.

Ejemplo 1.1.2. *Sea X una variable aleatoria con distribución $Bin(8, 0.3)$, entonces tenemos que la función de densidad está dada por*

$$f(x) = Pr(X = x) = \begin{cases} 0.058 & \textit{si } x = 0 \\ 0.198 & \textit{si } x = 1 \\ 0.296 & \textit{si } x = 2 \\ 0.254 & \textit{si } x = 3 \\ 0.136 & \textit{si } x = 4 \\ 0.047 & \textit{si } x = 5 \\ 0.01 & \textit{si } x = 6 \\ 0.001 & \textit{si } x = 7 \\ 0.00006 & \textit{si } x = 8 \\ 0 & \textit{en otro caso} \end{cases} \tag{1.1.21}$$

Y su función de distribución está dada por

$$F(x) = Pr(X \leq x) = \begin{cases} 0 & \textit{si } x < 0 \\ 0.058 & \textit{si } 0 \leq x < 1 \\ 0.255 & \textit{si } 1 \leq x < 2 \\ 0.552 & \textit{si } 2 \leq x < 3 \\ 0.806 & \textit{si } 3 \leq x < 4 \\ 0.942 & \textit{si } 4 \leq x < 5 \\ 0.989 & \textit{si } 5 \leq x < 6 \\ 0.999 & \textit{si } 6 \leq x < 7 \\ 0.9999 & \textit{si } 7 \leq x < 8 \\ 1 & \textit{si } x \geq 8 \end{cases} \tag{1.1.22}$$

Suponga que se quiere hallar la mediana de la distribución $Bin(8, 0.3)$, esto es, el percentil 50 %, tenemos que

$$X_{50} = \text{ínf}\{x : \ F(x) \geq 0.5\} = \text{ínf}\{x : \ x \geq 2\} = 2$$

Nótese que en este caso, el valor 2 no solo es el percentil 50 %, también es el percentil 51 %, o 55 %.

El cómputo de los percentiles en R para la distribución binomial se lleva a cabo usando el comando `qbinom`

```
> qbinom(0.5,8,0.3)
[1] 2
```

En muchos textos estadísticos en la tabla concerniente a la distribución F, sólo se disponen los percentiles de 90 %, 95 %, 97.5 % y 99 % y no los percentiles 10 %,

5 %, 2.5 % y 1 %. Sin embargo, por el Resultado 1.1.28, se puede ver fácilmente que el percentil α de una distribución F_m^n es simplemente la inversa del percentil $1-\alpha$ de la distribución F_n^m. Para ver eso, suponga que $X \sim F_n^m$, y denotamos el percentil α de X como $f_{n,\alpha}^m$, entonces tenemos que

$$\alpha = Pr(X < f_{n,\alpha}^m) = Pr\left(\frac{1}{X} > \frac{1}{f_{n,\alpha}^m}\right) = 1 - Pr\left(\frac{1}{X} < \frac{1}{f_{n,\alpha}^m}\right)$$

De donde tenemos que $1-\alpha = Pr\left(\frac{1}{X} < \frac{1}{f_{n,\alpha}^m}\right)$, y vemos que $\frac{1}{f_{n,\alpha}^m}$ es el percentil de $1/X$ que se distribuye como F_m^n. De lo anterior, $\frac{1}{f_{n,\alpha}^m} = f_{m,1-\alpha}^n$, y finalmente tenemos que $f_{n,\alpha}^m = \frac{1}{f_{m,1-\alpha}^n}$.

En la Tabla 1.1 presentamos los comandos en R para calcular percentiles. También presentamos algunos percentiles usuales de las distribuciones t student, Ji-cuadrado y F al final de este capítulo.

Distribución	Comando	Ejemplo	
$Bin(n,p)$	`qbinom`	Percentil 0.1 de $Bin(12,0.4)$	`qbinom(0.1,12,0.4)`
$Pois(\lambda)$	`qpois`	Percentil 0.2 de $Pr(5)$	`qpois(0.2,5)`
$U(a,b)$	`qunif`	Percentil 0.3 de $U(3,6)$	`qunif(0.3,3,6)`
$Gammma(k,\theta)$	`qgamma`	Percentil 0.4 de $Gamma(2,8)$	`qgamma(0.4,2,8)`
$Exp(\theta)$	`qexp`	Percentil 0.5 de $Exp(2)$	`qexPr(0.5,1/2)`
$N(\mu,\sigma^2)$	`qnorm`	Percentil 0.6 de $N(2,9)$	`qnorm(0.6,2,sqrt(9))`
$N(0,1)$	`qnorm`	Percentil 0.7 de $N(0,1)$	`qnorm(0.7)`
t_n	`qt`	Percentil 0.8 de t_{20}	`qt(0.8,20)`
χ_n^2	`qchisq`	Percentil 0.9 de χ_{25}^2	`qchisq(0.9,25)`
F_n^m	`qf`	Percentil 0.95 de F_{12}^{11}	`qf(0.95,11,12)`

Tabla 1.1: *Comandos en R para cálculo de percentiles en las distribuciones de uso frecuente.*

1.2 Familia exponencial

En este apartado, se introduce el concepto de familia exponencial que es muy útil en algunos temas tratados en este libro, como son la teoría de estimación puntual y prueba de hipótesis; así mismo, resulta útil en la teoría bayesiana que, sin embargo, no será tratada en este libro.

1.2.1 Familia exponencial uniparamétrica

En esta parte, se introduce la familia exponencial para distribuciones que depende solamente de un parámetro que se denominará familia exponencial uniparamétrica, cuya definición se da a continuación.

Definición 1.2.1. *Una distribución de probabilidad con parámetro θ pertenece a la familia exponencial uniparamétrica si la función de densidad se puede escribir de la forma*

$$f_X(x,\theta) = h(x)c(\theta)\exp\{d(\theta)T(x)\} \tag{1.2.1}$$

donde $T(x)$ y $h(x)$ son funciones que dependen de x únicamente, y $d(\theta)$ y $c(\theta)$ son funciones que dependen de θ únicamente.

En algunos textos, la representación de la familia exponencial es $f_X(x,\theta) = \exp\{d(\theta)T(x) - c(\theta)\}h(x)$, la cual es equivalente a la definición anterior. A continuación se ilustra dos ejemplos de distribuciones pertenecientes a esta familia.

Ejemplo 1.2.1. *La distribución Poisson con parámetro θ pertenece a la familia exponencial uniparamétrica puesto que*

$$\begin{aligned} f(x,\theta) &= \frac{e^{-\theta}\theta^x}{x!} I_{\{0,1,\cdots\}}(x) \\ &= \exp\{x\ln\theta\}\exp\{-\theta\}\frac{I_{0,1,\ldots}(x)}{x!}, \end{aligned}$$

en conclusión $f(x,\theta)$ es de la forma (1.2.1) con $d(\theta) = \ln\theta$, $T(x) = x$, $c(\theta) = \exp\{-\theta\}$ y $h(x) = \dfrac{I_{\{0,1,\cdots\}}(x)}{x!}$.

Ejemplo 1.2.2. *La distribución Gamma con parámetro de forma k conocida pertenece a la familia exponencial uniparamétrica puesto que*

$$\begin{aligned} f(x,\theta) &= \frac{x^{k-1}e^{-x/\theta}}{\theta^k\Gamma(k)} I_{(0,\infty)}(x) \\ &= \exp\left\{-\frac{x}{\theta}\right\}\theta^{-k}\frac{x^{k-1}I_{(0,\infty)}(x)}{\Gamma(k)}, \end{aligned}$$

el cual es de la forma (1.2.1) con $d(\theta) = -1/\theta$, $T(x) = x$, $c(\theta) = \theta^{-k}$ y $h(x) = \dfrac{x^{k-1}I_{(0,\infty)}(x)}{\Gamma(k)}$.

Nótese que esta representación de la familia exponencial no es única, puesto que al definir $d(\theta) = 1/\theta$ y $T(x) = -x$, también se puede concluir que la distribución Gamma con k fijo pertenece a la familia exponencial uniparamétrica.

En casi toda la teoría estadística, se trata más de una variable aleatoria que en la práctica, se observan los datos que corresponden a realizaciones de estas variables. En este caso, se examina la pertenencia de la familia exponencial de la función de densidad conjunta, y tenemos el siguiente resultado.

Resultado 1.2.1. *Si $X_1, \cdots, X_n$ son variables aleatorias independientes e idénticamente distribuidas con función de densidad común perteneciente a la familia exponencial uniparamétrica, entonces la función de densidad conjunta $f(x_1,\cdots,x_n)$ también pertenece a la familia exponencial uniparamétrica.*

Demostración.

$$\begin{aligned} f(x_1, \cdots, x_n, \theta) &= \prod_{i=1}^{n} f(x_i, \theta) \\ &= \prod_{i=1}^{n} h(x_i) c(\theta) \exp\{d(\theta) T(x_i)\} \\ &= c(\theta)^n \left[\prod_{i=1}^{n} h(x_i)\right] \exp\left\{d(\theta) \sum_{i=1}^{n} T(x_i)\right\} \end{aligned} \tag{1.2.2}$$

el cual es de la forma (1.2.1). □

Usando el anterior resultado junto con el Ejemplo 1.2.1, donde se mostró que la distribución Poisson pertenece a la familia exponencial, podemos afirmar que cuando se tiene n variables independientes e idénticamente distribuidas con distribución Poisson, la función conjunta también pertenece a la familia exponencial.

Otra utilidad del resultado es que cuando se necesita ver que una función de densidad conjunta pertenece a la familia exponencial, basta ver para la función de densidad marginal.

La estadística $\sum_{i=1}^{n} T(x_i)$ en (1.2.2) será de gran interés en los futuros capítulos, y estamos interesados en conocer sus propiedades como la esperanza y la varianza, y para eso solo necesitamos $E(T(X))$ y $Var(T(X))$ para X con la misma distribución que X_1, $\cdots$, X_n. El siguiente resultado nos provee las respectivas fórmulas.

Resultado 1.2.2. *Si X es una variable aleatoria con función de densidad perteneciente a la familia exponencial de la forma (1.2.1), entonces*

1. $E\left(\dfrac{\partial d(\theta)}{\partial \theta} T(X)\right) = -\dfrac{\partial}{\partial \theta} \ln c(\theta)$
2. $Var\left(\dfrac{\partial d(\theta)}{\partial \theta} T(X)\right) = -\dfrac{\partial^2}{\partial \theta^2} \ln c(\theta) - E\left(\dfrac{\partial^2 d(\theta)}{\partial \theta^2} T(X)\right)$

Ahora, las distribuciones con dos parámetros, como la distribución $N(\mu, \sigma^2)$, pueden considerarse como distribución uniparamétrica cuando μ o σ^2 se considera fijo conocido. Y se puede ver que en ambos casos, la distribución pertenece a la familia exponencial uniparamétrica.

1.2.2 Familia exponencial multi-paramétrica

Las distribuciones Gamma, normal y beta dependen de dos parámetros, y para estas distribuciones, se puede generalizar la definición de la familia exponencial uniparamétrica para distribuciones dependientes de más de un parámetro. La definición correspondiente se da a continuación.

Definición 1.2.2. *Una distribución de probabilidad pertenece a la familia exponencial multi-paramétrica si la función de densidad se puede escribir de la forma*

$$f_X(x,\boldsymbol{\theta}) = c(\boldsymbol{\theta})h(x)\exp\{d(\boldsymbol{\theta})'T(x)\} \tag{1.2.3}$$

donde $T(x)$ y $d(\boldsymbol{\theta})$ son funciones vectoriales, $h(x)$ y $c(\boldsymbol{\theta})$ son funciones reales.

Ejemplo 1.2.3. *La distribución Gamma con parámetro de forma k y parámetro de escala θ pertenece a la familia exponencial multi-paramétrica pues*

$$\begin{aligned} f_X(x) &= \frac{x^{k-1}e^{-x/\theta}}{\theta^k\Gamma(k)}I_{(0,\infty)}(x) \\ &= \exp\left\{-\frac{x}{\theta} + (k-1)\ln x\right\}\frac{1}{\theta^k\Gamma(k)}I_{(0,\infty)}(x) \\ &= \exp\left\{\left(-\frac{1}{\theta}, k-1\right)\begin{pmatrix} x \\ \ln x\end{pmatrix}\right\}\frac{1}{\theta^k\Gamma(k)}I_{(0,\infty)}(x) \end{aligned}$$

el cual es de la forma (1.2.3) con $\boldsymbol{\theta} = (\theta, k)'$, $d(\boldsymbol{\theta}) = (-\frac{1}{\theta}, k-1)'$, $c(\boldsymbol{\theta}) = \frac{1}{\theta^k\Gamma(k)}$, $T(x) = (x, \ln x)'$ y $h(x) = I_{(0,\infty)}(x)$.

Nótese que al igual a la familia exponencial uniparamétrica, la representación de la familia exponencial multi-paramétrica tampoco es única, pues en el ejemplo anterior también se puede tomar $\eta(\theta) = (\frac{1}{\theta}, k-1)'$ y $T(x) = (-x, \ln x)'$.

Las distribuciones normal, gamma, exponencial, Ji-cuadrado, beta, Bernoulli, binomial, binomial negativa, multinomial, Poisson y geométrica, pertenecen todas a la familia exponencial. También la distribución Weibull pertenece a la familia exponencial cuando el parámetro de forma es conocido. Por otra parte, las distribuciones Cauchy, Laplace, uniforme y Weibull cuando el parámetro de forma es desconocido no pertenecen a la familia exponencial.

La razón por la que las distribuciones de la familia uniforme no pertenecen a la familia exponencial va más allá de los objetivos de este libro, pues se necesita conocimientos sobre teoría estadística basada en la teoría de la medida. La afirmación de que la distribución uniforme no pertenece a la familia exponencial porque no se puede escribir de la forma (1.2.3) no es una razón válida.

1.3 Ejercicios

1.1 Demuestre el Resultado 1.1.1.

1.2 La función de densidad de una variable X con distribución Bernoulli también se puede escribir como

$$f_X(x) = \begin{cases} p & \text{si } x = 1 \\ 1-p & \text{si } x = 0 \\ 0 & \text{si no} \end{cases}$$

Verifique que esta función coincide con la función de densidad dada en (1.1.2).

1.3 Un vendedor de seguros realiza en promedio 10 visitas por semana a los posibles clientes, él por experiencia sabe que la probabilidad de que un cliente compre el seguro en una visita es de 0.2, y suponga que el resultado de una visita no se ve afectado por resultados de visitas anteriores.

(a) ¿Cuál es la probabilidad de que el vendedor en una semana venda más de dos seguros?

(b) Suponga que el vendedor ha tenido semanas consecutivas con muy malas ventas, y que perderá el trabajo si en la próxima semana no vende por lo menos un seguro. Por lo tanto, el vendedor decide aumentar el número de visitas a la semana para tratar al menos un seguro, ¿por lo menos cuántas visitas debe realizar la próxima semana para que la probabilidad de vender al menos un seguro sea superior a 90 %?

1.4 Demuestre el Resultado 1.1.3.

1.5 Para la distribución exponencial, escriba cuál es el parámetro y cuál es el espacio paramétrico.

1.6 Demuestre el Resultado 1.1.7.

1.7 Demuestre el Resultado 1.1.9.

1.8 Demuestre el Resultado 1.1.11.

1.9 Demuestre el Resultado 1.1.13.

1.10 Demuestre que la función de densidad de la distribución $N(\mu, \sigma^2)$ cumple las siguientes propiedades

(a) simétrica con respecto a μ,

(b) es creciente para $x < \mu$ y decreciente para $x > \mu$ y por consiguiente, tiene un máximo en μ.

1.11 Demuestre que si $X \sim N(\mu, \sigma^2)$ entonces $E(X) = \mu$ y $Var(X) = \sigma^2$ usando la función generadora de momentos de X dada en el Resultado 1.1.19.

1.12 Sea $X \sim N(\mu, \sigma^2)$, y sea $Z_1 = \frac{X-\mu}{\sigma}$ y $Z_2 = \frac{\mu-X}{\sigma}$, usando el Resultado 1.1.20 compruebe que tanto Z_1 como Z_2 tienen distribución normal estándar, es decir, la forma de estandarizar una variable con distribución normal no es única.

1.13 Demuestre el Resultado 1.1.21.

1.14 Demuestre el Resultado 1.1.22.

1.15 Calcule las siguientes probabilidades:

(a) $Pr(X > 2)$, $Pr(X < -1)$ y $Pr(1 < X < 3)$ donde $X \sim N(1.5, 4)$.

(b) $Pr(\frac{X-2Y}{3} > 4)$ donde $X \sim N(1.5, 4)$ y $Y \sim N(-1, 2)$ son variables independientes.

1.16 Para $X_1, \cdots, X_n$ independientes e idénticamente distribuidas provenientes de las siguientes distribuciones, escriba cuál será la variable $\sqrt{n}(\bar{X}-\mu)/\sigma$ en el contexto del Resultado 1.1.23.

1.17 Para las distribución Ji-cuadrado, escriba cuál es el parámetro y cuál es el espacio paramétrico.

1.18 Si $X \sim \chi^2_n$, demuestre que $E(X) = n$ y $Var(X) = 2n$ usando la Definición 1.1.13.

1.19 (a) Calcule el percentil 5 %, 10 % y 98 % de una variable con distribución normal estándar.

(b) Calcule el percentil 5 % y 90 % de una variable con distribución $N(1, 3)$.

(c) Encuentre valores a y b tales que $Pr(a < Z < b) = 0.98$.

(d) Encuentre valores a y b tales que $Pr(a < X < b) = 0.95$ donde $X \sim N(-2, 5)$.

1.20 Comprobar que en una distribución normal estándar $z_p = -z_{1-p}$, corrobora lo anterior calculando estos percentiles con los comandos dados en la Tabla 1.1. con un valor p fijo.

1.21 Encuentre los percentiles 5 %, 98 % de una variable con distribución Ji-cuadrado con 8 y 15 grados de libertad, respectivamente.

1.22 Encuentre los percentiles 5 %, 98 % de una variable con distribución t-student con 10 y 20 grados de libertad, respectivamente.

1.23 Encuentre los percentiles 5 %, 98 % de una variable con distribución F^8_5 y F^{12}_8, respectivamente.

1.24 Demuestre que las siguientes distribuciones pertenecen a la familia exponencial identificando las funciones $d(\theta)$, $T(x)$, $c(\theta)$ y $h(x)$:

(a) Binomial con n conocido.

(b) Exponencial.

(c) Distribución normal con media μ conocido.

(d) Distribución normal con varianza σ^2 conocida.

(e) La función de densidad conjunta de $X_1, \cdots, X_n$, donde las variables X_i son independientes e idénticamente distribuidas con distribución común $N(\mu, \sigma^2)$ con μ conocido (primero directamente y luego usando el Resultado 1.2.1.).

(f) La función de densidad conjunta de $X_1, \cdots, X_n$ donde las variables X_i son independientes e idénticamente distribuidas con distribución común $Pois(\lambda)$ (primero directamente y luego usando el Resultado 1.2.1.).

1.25 Demuestre la distribución Weibull, cuando k es conocido, pertenece a la familia exponencial uniparamétrica. Identifique $d(\theta)$, $T(x)$, $c(\theta)$ y $h(x)$.

1.26 Demuestre que la distribución $N(\mu, \sigma^2)$ pertenece a la familia exponencial multiparamétrica. Identifique $d(\boldsymbol{\theta})$, $T(x)$, $c(\boldsymbol{\theta})$ y $h(x)$.

1.27 Demuestre que la distribución Beta pertenece a la familia exponencial multiparamétrica. Identifique $d(\boldsymbol{\theta})$, $T(x)$, $c(\boldsymbol{\theta})$ y $h(x)$.

1.28 Sean $X_1, \cdots, X_n$ variables aleatorias independientes e idénticamente distribuidas con distribución común, $N(\mu_1, \sigma_1^2)$ y $Y_1, \cdots, Y_m$ variables independientes e idénticamente distribuidas con distribución común $N(\mu_2, \sigma_2^2)$, además las variables X_i son independientes de las variables Y_j, demuestre que la función de densidad conjunta de $X_1, \cdots, X_n, Y_1, \cdots, Y_m$ pertenece a la familia exponencial.

1.29 Repita el punto anterior suponiendo que $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Capítulo 2

Estimación puntual

2.1 Introducción

Desde la revolución estadística de Pearson y Fisher, la inferencia estadística busca encontrar los valores que parametrizan a la distribución desconocida de los datos. El primer enfoque, propuesto por Pearson, afirmaba que si era posible observar a la variable de interés en todos y cada uno de los individuos de una población, entonces era posible calcular los parámetros de la distribución de la variable de interés; por otro lado, si sólo se tenía acceso a una muestra representativa, entonces era posible calcular una estimación de tales parámetros.

Fisher discrepó de tales argumentos, asumiendo que las observaciones están sujetas a un error de medición y por lo tanto, así se tuviese acceso a toda la población, es imposible calcular los parámetros de la distribución de la variable de interés.

Del planteamiento de Fisher resultaron una multitud de métodos estadísticos para la estimación de los parámetros teóricos. Es decir, si la distribución de X está parametrizada por $\theta \in \Theta$, con Θ el espacio paramétrico inducido por el comportamiento de la variable de interés, el objetivo de la teoría estadística inferencial es calcular una estimación $\hat{\theta}$ del parámetro θ por medio de los datos observados. En este enfoque, los parámetros se consideran cantidades fijas y constantes.

El anterior será el enfoque que seguiremos a lo largo del desarrollo de este libro, aunque el lector debe tener conocimiento de que no es el único enfoque que los estadísticos utilizan en términos de inferencia acerca de los parámetros de una distribución.

Nótese que en términos de inferencia estadística existen, por lo menos, el enfoque clásico, propuesto por Fisher y desarrollado en este libro, el enfoque bayesiano (Gelman, Carlin, Stern & Rubin 2004), el enfoque no paramétrico (Conover 1998) y el enfoque de inferencia en poblaciones finitas (Gutiérrez 2009).

2.2 Conceptos básicos

Muchos de los estudios estadísticos están enfocados en estudiar características de una población objetiva; por ejemplo, una empresa productora puede estar interesada en conocer el gasto promedio semanal en alimentos de las familias de estrato socioeconómico bajo, con el fin de diseñar una estrategia de mercadeo para promover la demanda en el mercado. Es claro que en la ciudad hay una gran cantidad de familias de este perfil, y por consiguiente resulta prácticamente imposible saber el gasto promedio semanal de cada una de estas familias.

En casos como el anterior, la solución es inferir acerca de la característica de la población usando información obtenida de un subconjunto o una muestra de la población, y técnicas de esta rama constituyen la inferencia estadística. A continuación se presentan algunos conceptos básicos para poder estudiar la teoría de la inferencia estadística.

Definición 2.2.1. *Una muestra aleatoria de tamaño n es un conjunto constituido por n variables aleatorias independientes e idénticamente distribuidas, $X_1, \cdots, X_n$.*

Una muestra aleatoria es utilizada para lograr el objetivo de estimar un parámetro teórico desconocido, y esto se logra usando una función o funciones de las variables aleatorias de la muestra, conocidas como estadísticas. La definición formal se enuncia a continuación.

Definición 2.2.2. *Una estadística es una función de variables aleatorias de una muestra aleatoria que no contiene parámetros desconocidos.*

Algunas estadísticas comunes basadas en una muestra aleatoria $X_1, \cdots$ m X_n son:

- El promedio muestral o la media muestral, definido como $\bar{X} = \sum_{i=1}^{n} X_i/n$.
- Las varianzas muestrales, definidas como $S_n^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$ y $S_{n-1}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$.
- Mínimo y máximo de la muestra, definidos como $X_{(1)} = \text{mín}\{X_1, \cdots, X_n\}$ y $X_{(n)} = \text{máx}\{X_1, \cdots, X_n\}$. Nótese que en general la estadística $X_{(1)}$, al igual que $X_{(n)}$, no es ninguna de las variables $X_1, \cdots, X_n$. Considere un experimento aleatorio con $\Omega = \{a, b, c\}$ y se definen dos variables aleatorias X_1 y X_2 sobre Ω con $X_1(a) = X_1(b) = 1$, $X_1(c) = 2$ y $X_2(a) = 0$, $X_2(b) = X_2(c) = 2$. Si denotamos al mínimo de X_1 y X_2 como Y, se puede ver que $Y(a) = 0$, $Y(b) = 1$ y $Y(c) = 2$, y claramente Y no es ninguna de las variables X_1 y X_2. De esta forma si la función de densidad de donde proviene la muestra es f_X, la función de densidad de $X_{(1)}$ y la de $X_{(n)}$ no corresponden a f_X. Veamos

$$\begin{aligned} Pr(X_{(1)} > x) &= Pr(X_1 > x, \cdots, X_n > x) \\ &= Pr(X_1 > x) \cdots Pr(X_n > x) \\ &= (1 - F_{X_1}(x)) \cdots (1 - F_{X_n}(x)) \\ &= (1 - F_X(x))^n \end{aligned}$$

donde F_X denota la función de distribución común de la muestra. Por otro lado, $Pr(X_{(1)} > x) = 1 - F_{X_{(1)}}(x)$, de esta forma, tenemos que

$$F_{X_{(1)}}(x) = 1 - (1 - F_X(x))^n,$$

y podemos hallar la función de densidad para encontrar la función de densidad de $X_{(1)}$ como

$$f_{X_{(1)}}(x) = nf_X(x)(1 - F_X(x))^{n-1}. \tag{2.2.1}$$

Usando lo anterior, podemos encontrar que la función de densidad de $X_{(1)}$ en una muestra con distribución exponencial de media θ, corresponde a

$$f_{X_{(1)}}(x) = \frac{n}{\theta} e^{-nx/\theta} I_{(0,\infty)}(x).$$

Usando un razonamiento análogo, se puede ver (Ejercicio 2.1) que para la estadística máximo $X_{(n)}$, se tiene que

$$F_{X_{(n)}}(x) = F_X(x)^n \tag{2.2.2}$$

y

$$f_{X_{(n)}}(x) = nf_X(x)F_X(x)^{n-1}. \tag{2.2.3}$$

Una vez definido el concepto de estadística, podemos definir lo que es un estimador. Aunque en muchos contextos, lo que se desea estimar es el parámetro de una distribución, θ, en algunos casos, lo que nos interesa es una función del parámetro $g(\theta)$. Por ejemplo, suponga que en la línea de atención al cliente de una empresa, para describir el tiempo de espera para ser atendido por un asesor se emplea una distribución $Exp(\theta)$. En este caso, θ describe el tiempo de espera promedio que es el parámetro de interés; otra cantidad que puede resultar interesante es, por ejemplo, la probabilidad de que un cliente tenga que esperar menos de 5 minutos antes de ser atendida, la cual es $1 - e^{-5/\theta}$, que es una función del parámetro θ. Por esta razón, a continuación se presenta la definición general para un estimador de $g(\theta)$.

Definición 2.2.3. *Una estadística T cuyos valores son utilizados para estimar una función del parámetro $g(\theta)$ es un estimador de $g(\theta)$ y las realizaciones del estimador se llaman estimaciones y se denotan por t.*

Nótese que un estimador es una función de variables aleatorias, de tal manera que cuando las variables aleatorias se cambian de valor, el estimador también. Por lo tanto, cuando la muestra aleatoria cambia, el valor que toma el estimador, es decir, la estimación también cambia. Por lo tanto, un mismo estimador puede producir diferentes estimaciones si se cambia la muestra aleatoria. Y de lo anterior, debe quedar claro que un estimador es aleatorio, mientras que una estimación es un número, puesto que es la realización numérica del estimador. En la literatura estadística, se acostumbra denotar a los estimadores con letras mayúsculas, y a las estimaciones, minúsculas. De esta forma, las realizaciones de $\bar{X}$ se denotan como $\bar{x}$, las de S_n^2 como s_n^2 y análogamente para cualquier otro estimador.

Pongamos un ejemplo: suponga que se desea conocer la cantidad promedio de dinero que se gasta semanalmente una familia de estrato 2 en la compra de arroz. Un estadístico A seleccionó una muestra de 20 hogares, y utilizó como estimador el promedio muestral y tuvo como resultado $\bar{x} = 5300$ pesos; ahora, otro estadístico B seleccionó una muestra de 30 personas, y utilizó el mismo estimador, es decir, $\bar{X}$ y tuvo como resultado $\bar{x} = 4700$. En este caso, los dos usaron el mismo estimador; sin embargo, las dos estimaciones que obtuvieron no son iguales.

2.3 Estimaciones puntuales

El tópico de estimación puntual consiste en encontrar estimadores para estimar un cierto parámetro θ o una función de este $g(\theta)$. Dada la definición de estimador, es claro que cualquier estadística puede ser un estimador, y por consiguiente cada persona puede usar cualquier estadística para estimar según su antojo.

Tome el ejemplo de estimar el gasto promedio semanal en alimentos de familias de estrato socioeconómico bajo, si la muestra aleatoria es de tamaño 10, y los valores numéricos (en miles de pesos) son: 50, 62, 53, 65, 70, 64, 60, 58, 62 y 65, una forma razonable de estimar es tomar el promedio muestral $\bar{X}$ como estimador del media teórica, y arroja como resultado $\bar{x} = 60900$ pesos, la que parece ser una estimación aceptable; pero esta no es la única forma de estimar, podemos definir otras estadísticas, por ejemplo: $\sum_{i=1}^{n} X_i$, $(X_{(N)} + X_{(1)})/2$, u otras estadísticas no tan lógicas como $\exp\{X_1 + \cdots + X_n\}$, $\sum_{i=1}^{n} X_i^2$ o cualquier otra estadística que se nos viene a la mente. Sin embargo, en la literatura estadística existen, por lo menos, dos métodos estándares que nos ayudan a construir estimadores: el método de máxima verosimilitud y el método de momentos que se estudiará a continuación.

2.3.1 Método de máxima verosimilitud

Suponga que se desea estimar $g(\theta)$ basada en una muestra observada $x_1, \cdots, x_n$. La idea del método de máxima verosimilitud se basa en encontrar el valor de $g(\theta)$ que maximiza la probabilidad de observar la muestra $x_1, \cdots, x_n$. Esto es, el valor de $g(\theta)$ que hace más creíble a la muestra observada, y de allí viene el nombre de máxima verosimilitud.

Introducimos este método con un ejemplo muy sencillo, suponga que la alcaldía local está interesada en conocer el número promedio de homicidios mensuales ocurridos en la localidad de Usaquén de Bogotá. Dadas las características de esta variable de estudio, se puede pensar que ésta sigue una distribución Poisson con el parámetro desconocido θ. Como θ es la esperanza de la distribución Poisson, entonces lo que interesa a la alcaldía es conocer el valor de θ. Además, suponga que durante los últimos tres meses, el número de homicidios fue: 11, 9 y 7 respectivamente.

En la anterior situación, θ es el parámetro del modelo probabilístico que rige en la población, donde sólo están disponibles las realizaciones de tres variables, que al suponer que el tiempo no es un factor importante, constituyen una muestra aleatoria

de tamaño 3, y los denotamos por X_1, X_2, X_3. Ahora podemos hacer la siguiente pregunta: ¿cuál es la probabilidad de que la muestra aleatoria tenga como realización los valores 11, 9 y 7? Es decir, ¿cuál es la probabilidad de observar lo que realmente sucedió? Es claro que esta probabilidad depende del parámetro desconocido θ. Tenemos

Si $\theta = 6$, entonces

$$Pr(X_1 = 11, X_2 = 9, X_3 = 7) = \frac{e^{-6}6^{11}}{11!}\frac{e^{-6}6^{9}}{9!}\frac{e^{-6}6^{7}}{7!} = 2.1 \times 10^{-4}.$$

Si $\theta = 8$, entonces

$$Pr(X_1 = 11, X_2 = 9, X_3 = 7) = \frac{e^{-8}8^{11}}{11!}\frac{e^{-8}8^{9}}{9!}\frac{e^{-8}8^{7}}{7!} = 1.3 \times 10^{-3}.$$

Si calculamos esta misma probabilidad para otros valores de θ, podemos obtener la Tabla 2.1, donde se observa que la probabilidad es más grande cuando $\theta = 9$.[1] Ahora, como ya se observaron los valores 11, 9 y 7, es natural pensar que la probabilidad de asociada a estos valores fuera grande, y esto nos conduce a que el valor más plausible para θ debe ser 9.

θ	5	6	7	8	9	10
Pr	3.1×10^{-5}	2.1×10^{-4}	6.8×10^{-4}	1.3×10^{-3}	1.5×10^{-3}	1.3×10^{-3}

Tabla 2.1: *Probabilidad de observar la muestra de tamaño 3 conformado por 11,9,7 provenientes de una distribución $Pois(\theta)$ para diferentes valores de θ.*

El anterior razonamiento induce el método de máxima verosimilitud. Para estudiar este método, primero se da la siguiente definición.

Definición 2.3.1. *Dadas n variables aleatorias $X_1, \cdots, X_n$, la función de verosimilitud se define como la función de densidad conjunta de las n variables, y se denota por $L(x_1, \cdots, x_n, \theta)$.*

Aunque en este texto se trabaja solamente muestras aleatorias, la definición de la función de verosimilitud presentada anteriormente es válida en cualquier conjunto de variables aleatorias. En particular, cuando las n variables conforman una muestra aleatoria, la función de verosimilitud queda expresada como

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

donde f es la función de densidad común para las n variables. También nótese que cuando solo se dispone de una variable aleatoria X, la función de verosimilitud es simplemente la función de densidad de X, esto es, $f_X(x)$.

[1] El parámetro θ puede tomar cualquier valor positivo, no necesariamente entero, aquí se consideró solo algunos valores para θ, con el fin de introducir el método de máxima verosimilitud, mas no es un procedimiento riguroso.

Dada la definición de la función de verosimilitud, el método de máxima verosimilitud para un parámetro θ consiste en encontrar el valor de θ que maximice esta función, éste será el estimador de máxima verosimilitud de θ y lo denotaremos por $\hat{\theta}_{MV}$. Cuando la función de verosimilitud es una función continua de θ y además derivable, entonces podemos usar la primera y la segunda derivada para encontrar el estimador de máxima verosimilitud. Lo ilustramos con el siguiente ejemplo:

Ejemplo 2.3.1. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con distribución $Pois(\theta)$, el estimador de máxima verosimilitud de θ es el promedio muestral $\bar{X}$. Para verificar esta afirmación, se calcula primero la función de verosimilitud:*

$$\begin{aligned} L(x_1, \cdots, x_n, \theta) &= f(x_1, \theta) \cdots f(x_n, \theta) \\ &= \frac{e^{-\theta}\theta^{x_1}}{x_1!} I_{\{0,1,\cdots\}}(x_1) \cdots \frac{e^{-\theta}\theta^{x_n}}{x_1!} I_{\{0,1,\cdots\}}(x_n) \\ &= \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \prod_{i=1}^n I_{\{0,1,\cdots\}}(x_i) \end{aligned}$$

Encontrar el valor de θ para maximizar la anterior expresión de $L(x_1, \cdots, x_n, \theta)$ es equivalente a maximizar $e^{-n\theta}\theta^{\sum_{i=1}^n x_i}$, pues ésta es la parte que depende de θ. Ahora, encontrar el valor que maximiza una función es equivalente a encontrar el valor que maximiza el logaritmo natural de esta función, pues la función logarítmico es estrictamente creciente. Por lo tanto, basta encontrar el valor de θ que maximiza

$$L'(\theta) = \ln(e^{-n\theta}\theta^{\sum_{i=1}^n x_i}) = -n\theta + \sum_{i=1}^n x_i \ln\theta$$

En la distribución $Pois(\theta)$, el espacio paramétrico es $(0, \infty)$, y la función $L'(\theta)$ es función derivable, entonces la forma de hallar el máximo de la función será resolver la ecuación $\frac{\partial L'(\theta)}{\partial\theta} = 0$. En este caso, la solución es $\theta = \sum x_i/n$ cuando $\sum x_i \neq 0$, es decir, por lo menos una de las observaciones debe ser estrictamente mayor a 0; cuando $\sum x_i = 0$, es decir, cuando $x_1 = \ldots = x_n = 0$, no existe el estimador de máxima verosimilitud de θ. Supongamos que $\sum x_i \neq 0$, calculamos la segunda derivada de $L(\theta)$ evaluada en la anterior solución, tenemos que

$$\left.\frac{\partial^2 L'(\theta)}{\partial\theta^2}\right|_{\theta=\sum x_i/n} = \left.\frac{-\sum x_i}{\theta^2}\right|_{\theta=\sum x_i/n} = \frac{-n^2}{\sum x_i}$$

las observaciones x_i provienen de distribuciones tipo Poisson, lo cual garantiza que toman valores no negativos, de donde concluimos que la anterior expresión es negativa, lo cual verifica que la solución hallada $\theta = \sum x_i/n$ efectivamente maximiza la función de verosimilitud.

En conclusión, el estimador de máxima verosimilitud del parámetro θ de una distribución Poisson es el promedio muestral $\bar{X}$, y lo anterior también concuerda con el razonamiento, pues θ es el valor esperado o el media teórica, y es lógico estimar el media teórica con el promedio muestral.

Veamos una aplicación del anterior estimador.

Ejemplo 2.3.2. *Suponga que una organización internacional de derechos humanos necesita conocer el número de muertes violentas que ocurren mensualmente en una determinada ciudad, y para eso se seleccionaron 15 de los 63 barrios donde los resultados son 1, 1, 5, 5, 2, 3, 3, 6, 4, 3, 2, 3, 2, 3 y 4. Si denotamos el número de muertes violentas mensuales en un barrio de la ciudad por X, entonces X toma valor en $\{0, 1, \cdots\}$; por consiguiente, una distribución apropiada para X puede ser la distribución $Pois(\theta)$. Y por el anterior ejemplo, tenemos que la estimación de máxima verosimilitud para θ es el promedio muestral, esto es, $\hat{\theta}_{MV} = \bar{x} = 3.13$.*

Aparte de estimar el parámetro θ que puede ser interpretado como el número promedio de muertes violentas mensuales que ocurren en un barrio de la ciudad, podemos estimar otras cantidades que permiten a la organización tener una mejor idea acerca de la ciudad. Por ejemplo, ¿cuál es la probabilidad de que en un barrio no ocurra ninguna muerte violenta durante un mes? Usando propiedades de la distribución Poisson, tenemos que esta probabilidad es igual a $Pr(X = 0) = e^{-\theta}$. Dado que ya se obtuvo una estimación para θ, podemos utilizarla para estimar $e^{-\theta}$ como $e^{-3.1} = 0.045$. Más aún, podemos afirmar que esta estimación es de máxima verosimilitud (la teoría se verá más adelante).

No solo podemos hacer inferencia al nivel de los barrios sino también al nivel de la ciudad. Dado que esta está compuesta por 63 barrios, entonces el número de muertes violentas mensuales en la ciudad es la suma de los 63 barrios. Si denotamos con Y_i el número de muertes violentas mensuales en el i-ésimo barrio, entonces $Y = \sum_{i=1}^{63} Y_i$ denota el número de muertes violentas en la ciudad. Y usando el Resultado 1.1.9, podemos ver que $Y \sim Pois(63\theta)$, por lo tanto, el número promedio mensual de muertes violentas es 63θ, y una estimación de ésta será $63 \times 3.13 \approx 197$. Y podemos usar esta estimación para estimar cantidades como probabilidad de que en un mes en la ciudad ocurra menos de 100 muertes violentas u otras probabilidades de interés.

En muestras provenientes de distribuciones como $Exp(\theta)$ o $Ber(\theta)$, el procedimiento para encontrar el estimador $\hat{\theta}_{MV}$ es similar al ejemplo anterior y se puede ver fácilmente que $\hat{\theta}_{MV} = \bar{X}$, de nuevo se encuentra que el estimador de máxima verosimilitud del media teórica es el promedio muestral en estas dos distribuciones (Ejercicios 2.3 y 2.6).

Ejemplo 2.3.3. *Suponga que una empresa de EPS para mascotas que cuenta con sedes en diferentes ciudades en Colombia tiene vendedores que hacen visita a clientes potenciales, estos son, los hogares que tienen mascota, para ofrecer los productos. Es claro que para fijar metas de venta, el gerente de la empresa debe conocer el rendimiento de los vendedores y así fijar un número de visitas que éstos deben realizar con el fin de lograr la meta de venta. En este caso, el gerente necesita conocer cuál es la probabilidad de que un vendedor logre obtener una venta exitosa en una visita.*

Para economizar los recursos, el gerente hace seguimiento a 18 visitas, y en cada visita denota el éxito con 1 y fracaso con 0, de esta forma la muestra observada está constituida por 18 números de la forma 0 y 1, y por el contexto del problema, podemos identificar la distribución Bernoulli, y así estimar la probabilidad de éxito en

cada ensayo usando $\bar{X}$, que en este caso corresponde al número de éxitos dividido por el número total de visitas. Así que si en las 18 visitas registradas el número de ventas exitosas es 3, la probabilidad estimada de que un vendedor de esta empresa logre una venta exitosa en una visita será $\hat{p}_{MV} = 3/18 \approx 0.167$.

Ahora, suponga que un vendedor en un día promedio realiza 5 visitas, y el gerente está interesado en conocer qué tan probable es que en las 5 visitas el vendedor logre por lo menos una venta exitosa. Si denotamos el número de ventas exitosas en las 5 visitas por X, podemos calcular esta probabilidad como

$$Pr(X > 1) = 1 - Pr(X = 0) = 1 - (1 - p)^5.$$

Para encontrar una estimación de esta probabilidad, podemos pensar en usar la estimación de máxima verosimilitud de p encontrada anteriormente, de esta forma, tenemos que una estimación de $Pr(X_1)$ será $1 - (1 - 0.167)^5 = 0.598$. Análogamente, también podemos estimar la probabilidad de vender un seguro en las cinco visitas. En este caso, esta probabilidad está dada por $5Pr(1-p)^4$, que puede ser estimada como $5 * 0.167 * (1 - 0.167)^4 = 0.402$.

La pregunta interesante es ¿se puede afirmar que estas dos últimas estimaciones siguen siendo de máxima verosimilitud? Esta pregunta la responderemos más adelante.

Ejemplo 2.3.4. *En muchas aerolíneas, se pueden comprar tiquetes por medio de llamadas telefónicas atendidas por operadores de la aerolínea. Si un cliente debe esperar mucho tiempo en la línea para ser atendido, es más probable que el cliente desista, con lo cual la aerolínea perdería un cliente potencial. Por lo tanto la aerolínea desea conocer el rendimiento de los operadores que atienden estas llamadas. Para eso, se observan aleatoriamente 20 llamadas y se registra el tiempo transcurrido antes de que fueran atendidas por un operador. Estos tiempos en minutos son 0.13, 0.06, 0.50, 0.41, 1.44, 0.60, 0.22, 1.08, 0.78, 0.92, 2.73, 0.83, 0.19, 0.21, 1.75, 0.79, 0.02, 0.05, 2.30 y 1.03. Dado el contexto, se desea estimar el tiempo promedio que debe esperar un cliente antes de ser atendido, es decir, el media teórica. Para eso, necesitamos, en primer lugar, suponer una distribución adecuada para los datos. Los datos a la mano son del tipo continuo, además solo toma valores positivos, de donde podemos proponer una distribución exponencial, gamma o una distribución normal [2].*

Una forma de verificar la distribución de los datos es observar el histograma, el cual está dado en la Figura 2.1, donde podemos ver que la forma de las barras se asemeja a la función de densidad de una distribución exponencial. Otra forma de ver la distribución de los datos es usando las gráficas de QQ plot, y la presentamos en la Figura 2.2 para la distribución exponencial y la distribución normal [3]. Podemos ver que una vez más, la distribución exponencial parece ser apropiada para los datos. Entonces el problema se convierte en estimar el parámetro θ de una distribución $Exp(\theta)$, puesto

[2] Aunque una distribución normal toma valores en todos los números reales, pero se concentra alrededor de la media, por lo tanto, una muestra de valores positivos también pueden provenir de la distribución normal.

[3] La inversa de la función de distribución de la distribución Gamma es difícil de hallar y por consiguiente no es posible encontrar el QQ plot para verificar que un conjunto de datos provienen de una distribución Gamma.

que la esperanza de la distribución es θ. Y como se observó anteriormente, $\hat{\theta}_{MV} = \bar{X}$, podemos tener que la estimación de máxima verosimilitud de θ es $\bar{x} = 0.8$ minutos.

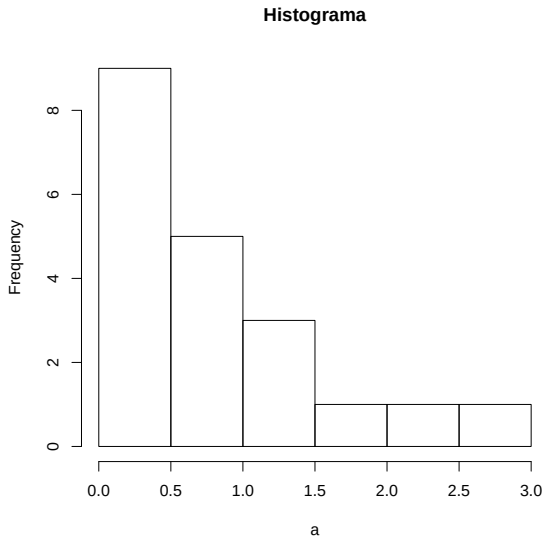

Figura 2.1: *Histograma de los datos del Ejemplo 2.3.4.*

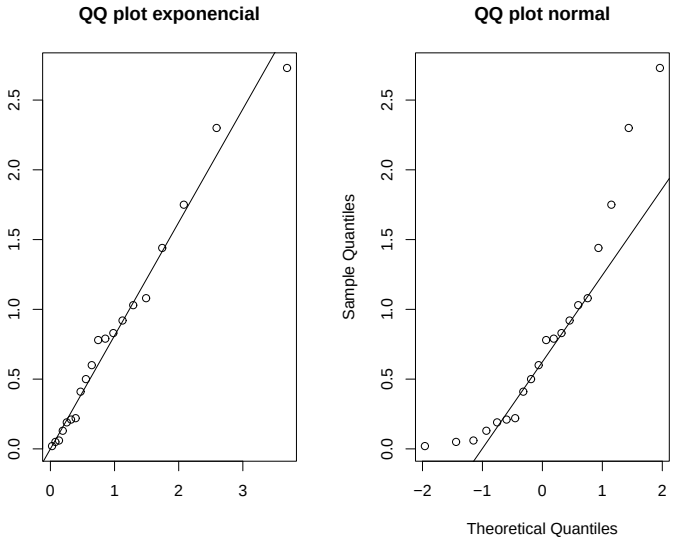

Figura 2.2: *QQ plot para verificar la distribución de los datos del Ejemplo 2.3.4.*

Ahora suponga que los directivos de la aerolínea han observado que si un cliente tiene que esperar más de 2 minutos, con toda seguridad cuelga la llamada; el 30 % de los clientes que tienen que esperar entre 1 minuto y medio y 2 minutos cuelgan la llamada y ningún cliente cuelga antes del minuto y medio. Entonces podemos estimar el porcentaje de clientes que cuelgan antes de ser atendidos, esto es, clientes potenciales que la aerolínea pierde. Para eso, se debe estimar el porcentaje de llamadas que necesitan más de 2 minutos para ser atendidas, éste se puede expresar como $Pr(X > 2)$*, donde* X *denota el tiempo de espera de una llamada y* $X \sim Exp(\theta)$*. Entonces debe estimar* $Pr(X > 2) = e^{-2/\theta}$*, la cual es una función de* θ*, y como ya se ha encontrado una estimación de* θ *dada por* $\hat{\theta}_{MV} = 0.8$*, podemos simplemente estimar* $Pr(X > 2)$ *como* $e^{-2/0.8} = 0.08$*, esto es, se estima que el 8 % de llamadas necesitan más de 2 minutos para ser atendidas, y por consiguiente este 8 % de clientes cuelga antes de ser atendido. Ahora, para estimar el porcentaje de llamadas que necesitan entre 1 minuto y medio y 2 minutos para ser atendidas como* $e^{-1.5/0.8} - e^{-2/0.8} = 0.07$*, es decir, 7 % de llamadas requieren entre 1.5 y 2 minutos para ser atendidas, y por consiguiente* $7\,\% \times 0.3 = 0.021 = 2.1\,\%$ *de clientes cuelgan la llamada antes de ser atendida. Sumando el 8 % hallado anteriormente, podemos afirmar que se estima que la aerolínea pierde el 10.1 % de los clientes potenciales por no ser atendidos oportunamente. Más adelante, se verá que esta estimación sigue siendo de máxima verosimilitud.*

Las distribuciones consideradas anteriormente tienen sólo un parámetro desconocido. Para distribuciones que tienen dos parámetros desconocidos, el procedimiento es levemente distinto, como lo ilustra el siguiente ejemplo con la distribución normal.

Ejemplo 2.3.5. *Dada una muestra aleatoria* $X_1, \cdots, X_n$ *con distribución* $N(\mu, \sigma^2)$*, el estimador de máxima verosimilitud del vector de parámetros* $\boldsymbol{\theta} = (\mu, \sigma^2)'$ *es* $(\bar{X}, S_n^2)'$*. Tenemos las siguientes expresiones para la función de verosimilitud:*

$$\begin{aligned} L(\boldsymbol{\theta}, x_1, \cdots, x_n) &= \frac{1}{\sqrt{\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_1 - \mu)^2\right\} \cdots \frac{1}{\sqrt{\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_n - \mu)^2\right\} \\ &= (\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \end{aligned}$$

Para facilitar la maximización de L*, se calcula* $\ln(L)$*:*

$$\ln(L) = -\frac{n}{2}\ln(\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Ahora para obtener valores de μ *y* σ^2 *que maximicen a* $\ln(L)$*, se procede a resolver las dos siguientes ecuaciones:*

$$\frac{\partial \ln(L)}{\partial \mu} = 0 \tag{2.3.1}$$

y

$$\frac{\partial \ln(L)}{\partial \sigma^2} = 0 \tag{2.3.2}$$

Resolviendo la ecuación (2.3.1), se obtiene la solución de $\mu = \bar{x}$ y resolviendo (2.3.2), se obtiene la solución de $\sigma^2 = \sum(x_i - \mu)^2/n$, donde al reemplazar $\mu = \bar{x}$, se tiene que $\sigma^2 = \sum(x_i - \bar{x})^2/n = s_n^2$.

Ahora debemos verificar que las anteriores soluciones halladas efectivamente maximicen la función $\ln(L)$. Dado que esta función tiene dos argumentos, es necesario hacer uso de la matriz Hessiana. La matriz se calcula de la siguiente manera:

$$H(\ln(L)) = \begin{bmatrix} \dfrac{\partial^2 \ln(L)}{\partial \mu^2} & \dfrac{\partial^2 \ln(L)}{\partial \mu \partial \sigma^2} \\ \dfrac{\partial^2 \ln(L)}{\partial \sigma^2 \partial \mu} & \dfrac{\partial^2 \ln(L)}{\partial (\sigma^2)^2} \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{-n}{\sigma^2} & \dfrac{n\mu - \sum x_i}{\sigma^4} \\ \dfrac{n\mu - \sum x_i}{\sigma^4} & \dfrac{n}{2\sigma^4} - \dfrac{\sum(x_i - \mu)^2}{\sigma^6} \end{bmatrix}. \tag{2.3.3}$$

Ahora reemplazamos las soluciones halladas $\mu = \bar{x}$ y $\sigma^2 = s_n^2$, se tiene que la matriz Hessiana es:

$$H = \begin{bmatrix} \dfrac{-n^2}{\sum(x_i - \bar{x})^2} & 0 \\ 0 & \dfrac{-n^3}{2(\sum(x_i - \bar{x})^2)^2} \end{bmatrix}.$$

Obsérvese que la matriz Hessiana es una matriz diagonal con valores negativas en la diagonal, lo cual demuestra que es definida negativa, con eso se concluye que las soluciones halladas efectivamente maximizan la función $\ln(L)$. En conclusión, los estimadores de máxima verosimilitud del vector de parámetros $\boldsymbol{\theta} = (\mu, \sigma^2)'$ son $(\bar{X}, S_n^2)'$.

A continuación, se presenta una aplicación del anterior ejemplo.

Ejemplo 2.3.6. *Suponga que una fábrica de vidrios tiene una línea de producción de láminas de vidrio templado de grosor de 3 cm. Para controlar la calidad de los vidrios producidos por esta línea, se seleccionan 12 láminas para inspección. Estas 12 láminas midieron (en cm) 3.56, 3.36, 2.99, 2.71, 3.31, 3.68, 2.78, 2.95, 2.82, 3.45, 3.42 y 3.15. Estos datos son, aparentemente, continuos, y podemos pensar que una distribución normal puede ser apropiada para los datos. Podemos, en primer lugar, observar la forma del histograma de estos datos presentando la Figura 2.3, donde aparentemente no se observa una forma similar a la función de densidad de una distribución normal.*

Sin embargo, como el número de datos es relativamente pequeño, el histograma puede no reflejar la distribución verdadera de los datos, y por esta razón, usamos la gráfica de QQ plot para ver qué tan adecuada es la distribución normal. El comando en R está dado a continuación

```
> vidrio<-c(3.56, 3.36, 2.99, 2.71, 3.31,3.68, 2.78, 2.95,
2.82, 3.45, 3.42 ,3.15)
> qqnorm(vidrio,main="QQ plot para distribución normal",xlab=
"Cuantiles teoricos",ylab="Cuantiles muestrales")
> qqline(vidrio)
```

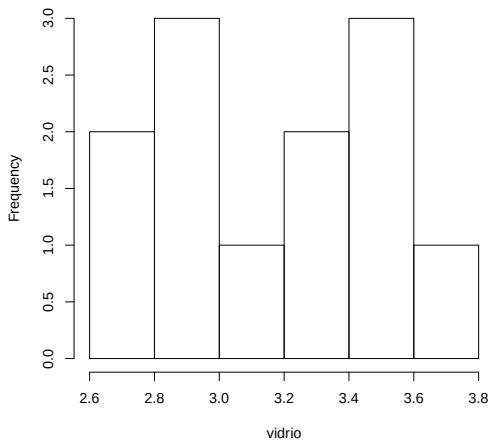

Figura 2.3: *Histograma de los datos del Ejemplo 2.3.6.*

Esta gráfica se muestra en la Figura 2.4, donde podemos ver que una distribución normal parece ser apropiada. Por lo tanto, usando el anterior ejemplo, podemos estimar el grosor promedio de las láminas de esta línea como $\hat{\mu}_{MV} = \bar{x} = 3.18\ cm$ *y la varianza estimada en este caso es* $\hat{\sigma}^2_{MV} = s^2_n = 0.097\ cm^2$*. Sin embargo, es difícil dar interpretación práctica a la varianza puesto que la unidad de ésta es la unidad de los datos al cuadrado, por esta razón en la práctica se usa con más frencuencia* σ *como la medida de dispersión. En este caso, tenemos que* $\hat{\sigma} = \sqrt{0.097\ cm^2} = 0.31\ cm$.

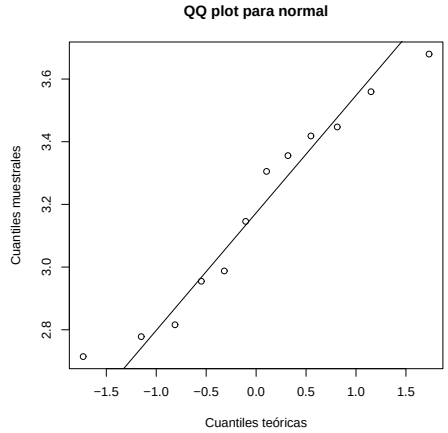

Figura 2.4: *QQ plot para verificar la distribución normal de los datos del Ejemplo 2.3.6.*

Otra cantidad interesante que se quiere conocer es σ/μ, que puede ser vista como una medida de dispersión teórica que está libre de las unidades de medición y por consiguiente, es útil en la práctica para comparar dos poblaciones. Esta cantidad se puede ver como una función del vector de parámetros (μ, σ^2), razón por la cual puede ser estimada por $\frac{\sqrt{S_n^2}}{\bar{X}}$ que en este ejemplo da como resultado 9.7 %.

Ahora, suponga que las láminas de grosor entre 2.8 cm y 3.2 cm son vendidas al mercado, las de grosor menor de 2.8 cm son desechadas y las de grosor mayor de 3.2 son usadas como materia prima para futuras producciones. Usando las estimaciones de μ y σ podemos estimar las proporciones de láminas que serán vendidas, desechadas y usadas como materia prima. Si denotamos el grosor de una lámina como X, tenemos que la proporción de láminas que serán vendidas es igual a

$$\begin{aligned} Pr(2.8 < X < 3.2) &= Pr\left(\frac{2.8-\mu}{\sigma} < \frac{X-3.18}{0.31} < \frac{3.2-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{3.2-\mu}{\sigma}\right) - \Phi\left(\frac{2.8-\mu}{\sigma}\right). \end{aligned}$$

Usando $\hat{\mu}_{MV} = 3.18$ y $\hat{\sigma} = 0.31$, podemos estimar la proporción de láminas que serán vendidas como $\Phi\left(\frac{3.2-3.18}{0.31}\right) - \Phi\left(\frac{2.8-3.18}{0.31}\right)$, el cual es igual a 0.42. Es decir, se estima que sólo el 42 % de las láminas producidas serán vendidas.

Análogamente, se puede encontrar que el 11 % serán desechadas y el 47 % serán usadas como materia prima. Es claro que según los datos muestrales y las estimaciones obtenidas de éstos, el uso de esta línea de producción no parece ser muy rentable, puesto que menos de la mitad de las láminas producidas pueden ser vendidas. Para aumentar la proporción de láminas que son aptas para la venta, la fábrica debe mejorar la línea de producción con la ayuda de los expertos para

- *Disminuir el grosor promedio de las láminas, puesto que en la muestra se observó un promedio de 3.18 cm, y se podrá pensar que el grosor real de las láminas es superior al valor especificado de 3 cm. Suponga que después de una mejora de la línea de producción, el promedio muestral fuera $\bar{x} = 3.05$ y $\hat{\sigma}$ se mantiene igual. Se puede ver que en este caso, la proporción estimada de láminas para venta se aumentará a 48 %.*
- *Estabilizar las láminas en término del grosor; de esta forma, la estimación de σ será más pequeña y la proporción de láminas para venta se incrementará. Suponga que después de una mejora de la línea de producción, $\hat{\sigma} = 0.2\ cm$ y $\hat{\mu}$ se mantiene igual. Se puede ver que la proporción estimada de láminas para venta se aumentará a 51 %.*

Finalmente, si se puede lograr que μ sea más cercano a 3 cm y al mismo tiempo disminuir el valor de σ, la proporción de láminas para venta será aún mayor, y la línea de producción será más rentable.

En el anterior ejemplo, el estimador de máxima verosimilitud de μ es $\bar{X}$, y esto es válido aún cuando la varianza teórica σ^2 es conocida; por otro lado, cuando μ es

conocido, el estimador de máxima verosimilitud de σ^2 ya no es S_n^2 sino $\sum_{i=1}^n (X_i - \mu)^2/n$, esto es, se mide la dispersión tomando las diferencia entre cada variable con la media teórica μ (Ejercicio 2.5).

El problema de maximizar una función puede, en algunos casos, resultar complicado, y peor aún, puede no encontrar una solución explícita. Un ejemplo de ello es la distribución gamma cuando ambos parámetros son desconocidos. Para estos casos, es necesario usar métodos numéricos para encontrar el máximo de la función de verosimilitud.

Ahora, para las distribuciones con dos parámetros como normal o gamma, cuando uno de los dos parámetros es fijo conocido, entonces sólo habrá necesidad de estimar el otro parámetro y el procedimiento es similar al presentado anteriormente, y lo ilustramos en el siguiente ejemplo.

Ejemplo 2.3.7. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con distribución Gamma con parámetro de forma k conocido y parámetro de escala θ desconocido, se tiene que el estimador de máxima verosimilitud de θ es $\frac{\sum_{i=1}^n X_i}{nk}$. Para la verificación, primero calculamos $\ln(L)$ que es la función que se necesita maximizar:*

$$\begin{aligned}\ln(L) &= \ln\left(\frac{\prod x_i^{k-1} e^{-\sum x_i/\theta}}{\theta^{nk}\Gamma(k)^n}\right) \\ &= (k-1)\sum \ln(x_i) - \sum x_i/\theta - nk\ln\theta - n\ln(\Gamma(k))\end{aligned}$$

cuya primera derivada parcial con respecto a θ está dada por:

$$\frac{\partial \ln(L)}{\partial \theta} = \frac{\sum x_i}{\theta^2} - \frac{nk}{\theta},$$

el cual al igualar a 0, se obtiene la solución de $\theta = \frac{\sum x_i}{nk}$. Ahora, al calcular la segunda derivada de $\ln(L)$ y evaluar en la anterior solución, se tiene que

$$\frac{\partial^2 \ln(L)}{\partial \theta^2} = \frac{-nk}{\theta^2} < 0,$$

con lo cual se concluye que el estimador de máxima verosimilitud de θ es $\frac{\sum X_i}{nk}$.

Teniendo en cuenta que la distribución exponencial es un caso particular de la distribución Gamma cuando $k = 1$, entonces del anterior ejemplo se puede concluir que el estimador de máxima verosimilitud del parámetro θ de una distribución exponencial es $\bar{X} = \frac{\sum X_i}{n}$, estimador que también se puede obtener maximizando directamente la función de verosimilitud.

Cuando la función de verosimilitud no es función continua o derivable del parámetro θ, el problema de maximizar no se puede llevar a cabo haciendo el uso de la derivada de la manera habitual, los dos siguientes ejemplos ilustran tales situaciones.

Ejemplo 2.3.8. *En situaciones donde la característica de interés es el tamaño de una población N, puede ser, por ejemplo, la cantidad de cierto tipo de animales en un determinado lugar (puede ser un bosque). Una forma de determinar N es, en primer lugar, identificar R de los N individuos ($R < N$); en el caso de los animales, puede ser conveniente capturar R de ellos y marcarlos de alguna forma. Después de eso, se espera que los R individuos se mezclen bien con los otros $N - R$, y se seleccionan aleatoriamente n individuos ($n < N$), y se cuenta cuántos de los R individuos fueron seleccionados. Para ver cómo este procedimiento nos puede ayudar a encontrar el valor de N, primero identifiquemos el contexto en términos de las distribuciones probabilísticas.*

Sea X la variable aleatoria que denota el número de los R individuos que fueron seleccionados en la muestra de tamaño n, entonces podemos ver que $X \sim Hg(n, R, N)$, donde n y R son conocidos, y se quiere encontrar el estimador de máxima verosimilitud de N. En este caso la función de verosimilitud es la misma función de densidad de la variable X, esto es:

$$L(N, x) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}} I_{\{0,1,\cdots,n\}}(x). \tag{2.3.4}$$

En esta función, el argumento N toma valores discretos y no se puede derivar L con respecto a N para hallar el máximo. La forma de encontrar el valor de N que maximiza a L es encontrar para qué valores de N, la función L es creciente y para qué valores de N es decreciente. Para lograr este fin, se despeja el valor de N en $L(N)/L(N-1) > 1$, como sigue:

$$\begin{gathered}
\frac{L(N)}{L(N-1)} = \frac{\binom{N-R}{n-x}\binom{N-1}{n}}{\binom{N}{n}\binom{N-R-1}{n-x}} > 1 \\
\frac{(N-R)!(N-1)!(N-n)!(N-R-1-n+x)!}{(N-R-1)!N!(N-n-1)!(N-R-n+x)!} > 1 \\
(N-R)(N-n) > N(N-R-n+x) \\
Rn/x > N.
\end{gathered}$$

Análogamente, se obtiene que $L(N)/L(N-1) < 1$ cuando $N > Rn/x$. Lo anterior indica que la función L es creciente para valores de N menores que Rn/x y decreciente para valores mayores que Rn/x. Pero no podemos afirmar que Rn/X es el estimador de máxima verosimilitud para L puesto que este cociente puede no ser entero, y en este caso la estimación no se ubicaría dentro del espacio paramétrico de N. Lo que sí se puede afirmar es que el estimador de máxima verosimilitud de N es $[Rn/X]$ o $[Rn/X] + 1$ donde $[\cdot]$ es la función parte entera. Shao (2003) afirma que $\hat{N}_{MV} = [Rn/X]$. Esto es verdadero, puesto que hemos encontrado anteriormente que $L(N) < L(N-1)$ si $N > Rn/x$, entonces podemos concluir que como $[Rn/x] + 1 > Rn/x$, entonces $L([Rn/x]+1) < L([Rn/x])$. Y podemos afirmar que el estimador de máxima verosimilitud de N es $[Rn/X]$.

Otra aplicación interesante de la distribución hipergeométrica es el caso donde se conoce el tamaño poblacional N, y se desea estimar el número de individuos que tienen cierta característica basada en una muestra de tamaño n; por ejemplo, se conoce que

en un estanque hay N peces, y se sabe que una parte de ellos están infectados por un tipo de parásito, y se quiere saber cuántos peces tienen dicho parásito con base en la observación de una muestra de tamaño n. En estos casos, estamos interesados en estimar R con N y n conocidos. Un razonamiento análogo al caso de estimar N conduce al siguiente estimador de R [4]*.*

$$\hat{R}_{MV} = \begin{cases} \frac{x(N+1)}{n} - 1 \text{ ó } \frac{x(N+1)}{n} & \text{si } \frac{x(N+1)}{n} \text{ es entero} \\ [\frac{x(N+1)}{n}] & \text{si } \frac{x(N+1)}{n} \text{ no es entero} \end{cases}$$

Retomando el problema de estimar el tamaño de un subgrupo considerado en el Capítulo 1, donde se supone que en una ciudad existen 2396 empresas que pueden clasificar en empresas grandes, medianas o pequeñas según el número de empleados, si en una muestra aleatoria simple sin reemplazos de tamaño 200 se encuentran 28 empresas grandes, para tener una estimación del número total de empresas grandes en la ciudad, calculamos en primer lugar $28 * (2396 + 1)/200 = 335.58$. Este número no es entero, de donde concluimos que la estimación de máxima verosimilitud del número de empresas grandes en la ciudad es de 335.

Finalmente, consideramos las distribuciones donde los valores que toma la variable aleatoria X depende del parámetro θ; por ejemplo, las distribuciones del tipo uniforme. Para este tipo de distribuciones, el procedimiento para encontrar $\hat{\theta}_{MV}$ en general se puede resumir en los siguientes pasos:

(1) Calcular la función de verosimilitud, sin omitir las funciones indicadoras, pues éstas dependen de θ.

(2) Encontrar el rango de valores de θ donde la función de verosimilitud no toma el valor 0. Generalmente este rango depende del máximo y/o el mínimo de la muestra: $x_{(n)}$ y $x_{(1)}$.

(3) Dentro del rango encontrado en el paso anterior, mediante empleo de derivadas o simplemente observación directa, buscar el valor de θ que maximice la función de verosimilitud.

Ilustramos el anterior procedimiento en el siguiente ejemplo.

Ejemplo 2.3.9. *Dada una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una población con distribución uniforme continua sobre el intervalo $[0, \theta]$, se quiere encontrar el estimador de máxima verosimilitud del parámetro θ. En primer lugar, obsérvese que en la función de verosimilitud L existe un término de la función indicadora que depende de θ, puesto que $L(\theta) = \theta^{-n} \prod_{i=1}^{n} I_{[0,\theta]}(x_i)$.*

Tenemos que

$$L \neq 0 \Leftrightarrow 0 \leq x_i \leq \theta \text{ para todo } i = 1, \cdots, n \Leftrightarrow \theta \geq x_{(n)}$$

[4] Para más detalles, consulte Zhang (2009).

donde $x_{(n)} = \max\{x_1, \cdots, x_n\}$. *Entonces se concluye que el rango de valores de* θ *para que la función de verosimilitud sea diferente de 0 es* $[x_{(n)}, \infty)$. *Ahora, observe que dentro de este rango,* $L = \theta^{-n}$, *que es una función decreciente de* θ, *entonces para valores pequeños de* θ, L *toma valores grandes. Pero el valor más pequeño que puede tomar* θ *dentro del rango* $[x_{(n)}, \infty)$ *es* $x_{(n)}$, *entonces se concluye que* $\hat{\theta}_{MV} = X_{(n)}$.

Ahora, retomando situaciones donde la cantidad que se quiere estimar es una función del parámetro, $g(\theta)$, textos como Mood, Graybill & Boes (1974) establecen que el estimador de máxima verosimilitud de $g(\theta)$ es simplemente $g(\hat{\theta}_{MV})$ siempre y cuando g es una función uno a uno.

Otra forma de ver esto es mediante la reparametrización de la función $L(\theta)$. Por ejemplo, suponga que se desea estimar $\lambda = e^{-\theta}$ en una muestra proveniente de una distribución $Pois(\theta)$. Podemos escribir la función de verosimilitud en término de λ como

$$L(\lambda) = \frac{\lambda^n(-n\ln\lambda)^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i} \prod_{i=1}^n I_{\{0,1,\cdots\}}(x_i),$$

de donde

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = \frac{n}{\lambda} + \frac{\sum_{i=1}^n x_i}{\ln\lambda}\frac{1}{\lambda},$$

igualando la anterior expresión a 0 y resolviendo para λ, se tiene que $\hat{\lambda}_{MV} = e^{-\bar{X}}$. Ahora, recordando que $\hat{\theta}_{MV} = \bar{X}$, lo cual coincide con la conclusión dada anteriormente.

Aunque lo planteado es válido para el caso cuando la función g es una función uno a uno, existe el siguiente resultado que establece la invarianza del estimador de máxima verosimilitud para cualquier función g.

Resultado 2.3.1. *Dada una muestra aleatoria* $X_1, \cdots, X_n$ *proveniente de una población con distribución* $f(x, \boldsymbol{\theta})$ *donde* $\boldsymbol{\theta}$ *es el vector de parámetros, y suponga que* $T = T(X_1, \cdots, X_n)$ *es el estimador de máxima verosimilitud de* $\boldsymbol{\theta}$, *y* g *es una función del vector de parámetros, entonces el estimador de máxima verosimilitud de* $g(\boldsymbol{\theta})$ *es la estadística* $g(T)$.

Demostración. Casella & Berger (2002, p. 172 y p. 321). □

Dado el anterior resultado, podemos ver que todas las estimaciones de los Ejemplos 2.3.2, 2.3.3, 2.3.4 y 2.3.6 son estimaciones de máxima verosimilitud.

Otra situación interesante es cuando se dispone de dos muestras aleatorias independientes, esto es, cualquier conjunto de variables de la primera muestra es independiente de cualquier conjunto de variables de la segunda, y el objetivo es estimar los parámetros concernientes a las dos muestras.

Consideramos, en primer lugar, dos muestras provenientes de distribuciones normales.

Suponga que se tienen dos muestras aleatorias independientes $X_1, \ldots, X_{n_X}$ y $Y_1, \ldots, Y_{N_Y}$ provenientes de $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$, respectivamente. Y se desean estimar algunos de los parámetros μ_X, μ_Y, σ_X^2 y σ_Y^2. En primer lugar, calculamos la función de verosimilitud, la cual está dada por

$$L = (2\pi\sigma_X^2)^{-n_X/2}(2\pi\sigma_Y^2)^{-n_Y/2}\exp\left\{-\frac{1}{2\sigma_X^2}\sum_{i=1}^{n_X}(x_i-\mu_X)^2 - \frac{1}{2\sigma_Y^2}\sum_{j=1}^{n_Y}(y_j-\mu_Y)^2\right\}$$

En el caso de que las dos muestras provenientes de la misma distribución, esto es, si $\mu_X = \mu_Y = \mu$ y $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, entonces el proceso de la estimación de máxima verosimilitud de μ y σ^2 se llevan a cabo, simplemente, usando conjuntamente las variables de las dos muestras. Y tenemos que

$$\hat{\mu}_{MV} = \frac{\sum_{i=1}^{n_X} X_i + \sum_{j=1}^{n_Y} Y_j}{n_X + n_Y}, \tag{2.3.5}$$

y

$$\hat{\sigma}_{MV}^2 = \frac{\sum_{i=1}^{n_X}(X_i - \hat{\mu}_{MV})^2 + \sum_{j=1}^{n_Y}(Y_j - \hat{\mu}_{MV})^2}{n_X + n_Y}.$$

Ejemplo 2.3.10. *Retomamos el Ejemplo 2.3.6, donde se disponía una muestra de 12 láminas. Ahora suponga que se selecciona una nueva muestra de 10 láminas de la misma línea de producción con grosor 3.56, 3.17, 2.98, 2.95, 3.03, 2.87, 3.58, 3.73, 2.83 y 3.43. Dado que las dos muestras son productos de una misma línea de producción, entonces podemos afirmar que las dos muestras provienen de una misma distribución normal $N(\mu, \sigma^2)$, y podemos estimar el grosor promedio de esta línea como 3.2 cm y la desviación estándar como 0.31 cm.*

Otra situación que puede surgir en la práctica es cuando las dos muestras provienen de distribuciones con la misma esperanza, pero diferentes varianzas, esto es, $\mu_X = \mu_Y = \mu$ y $\sigma_X^2 \neq \sigma_Y^2$. Supongamos, en primer lugar, que σ_X^2 y σ_Y^2 son conocidas, y tenemos que la función de verosimilitud está dada por

$$L(\mu) = (2\pi\sigma_X^2)^{-n_X/2}(2\pi\sigma_Y^2)^{-n_Y/2}\exp\left\{-\frac{1}{2\sigma_X^2}\sum_{i=1}^{n_X}(x_i-\mu)^2 - \frac{1}{2\sigma_Y^2}\sum_{j=1}^{n_Y}(y_j-\mu)^2\right\}$$

de donde

$$\frac{\partial \ln L(\mu)}{\partial \mu} = \frac{\sum_{i=1}^{n_X}(x_i-\mu)}{\sigma_X^2} + \frac{\sum_{j=1}^{n_Y}(y_j-\mu)}{\sigma_Y^2}.$$

Igualando la anterior derivada a cero y despejando μ, se encuentra que la solución está dada por $\mu = \dfrac{\sigma_Y^2 n_X \bar{x} + \sigma_X^2 n_Y \bar{y}}{n_X\sigma_Y^2 + n_Y\sigma_X^2}$. Ahora, es claro que

$$\frac{\partial^2 \ln L(\mu)}{\partial \mu^2} = -\frac{n_X}{\sigma_X^2} - \frac{n_Y}{\sigma_Y^2} < 0,$$

y en conclusión, se tiene que

$$\hat{\mu}_{MV} = \frac{\sigma_Y^2 n_X \bar{X} + \sigma_X^2 n_Y \bar{Y}}{n_X \sigma_Y^2 + n_Y \sigma_X^2} = \frac{n_X \bar{X} + n_Y \bar{Y} \frac{\sigma_X^2}{\sigma_Y^2}}{n_X + n_Y \frac{\sigma_X^2}{\sigma_Y^2}}. \tag{2.3.6}$$

Nótese que la anterior expresión se asemeja a un promedio ponderado, donde entre más grande sea la varianza teórica de la segunda población σ_Y^2, menos peso tienen las variables de la muestra correspondiente. Esto es muy natural, puesto que en una distribución normal, una varianza grande indica que los valores de la distribución tienden a tomar valores muy alejados a la media. Entonces, si $\sigma_X^2/\sigma_Y^2 < 1$, los valores de las variables $Y_1, \cdots, Y_{n_Y}$ tienden a estar más lejos de μ que las variables $X_1, \cdots, X_{n_X}$, lo cual las hacen menos confiables, y por esta razón se les asigna un peso menor. Cuando $\sigma_X^2 = \sigma_Y^2$, la anterior estimación de μ se reduce a (2.3.5).

Ejemplo 2.3.11. *Considera la fábrica vidrios del Ejemplo 2.3.6, y suponga que hay, en total, dos líneas de producción de láminas de vidrio templado de 3 cm, y además por ajuste inapropiado de temperatura, la línea A tiene una desviación estándar del 0.6 cm, mucho mayor que la línea B cuya desviación estándar es del 0.3 cm. Si se desea estimar el grosor promedio de las láminas de vidrio del grosor nominal del 3 cm, se debe seleccionar una muestra de las láminas de la línea A, y una muestra de la línea B. Suponga que el grosor de 10 láminas de cada línea corresponde a 3.80, 2.81, 2.98, 2.97, 3.69, 2.77, 3.08, 2.98, 2.37, 3.00 y 2.87, 3.48, 2.65, 3.38, 2.75, 2.99, 2.81, 2.54, 2.84, 2.79, respectivamente, entonces asignando un peso mayor a las observaciones de la línea B según la teoría expuesta anteriormente, se tiene que* $\hat{\mu}_{MV} = 2.955$ *cm.*

Finalmente, consideramos el caso donde se supone que las dos varianzas teóricas coinciden, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ y $\mu_X \neq \mu_Y$. En este caso, tenemos que los estimadores de máxima verosimilitud de μ_X, μ_Y y σ^2 son $\bar{X}$, $\bar{Y}$ y $[\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + \sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2]/(n_X + n_Y)$, respectivamente (Ejercicio 2.25), esto es, las observaciones de las dos muestras se utilizan separadamente para estimar las medias teóricas, mientras que la varianza se estima usando las muestras conjuntamente con la misma ponderación. En general, cuando no se puede asumir que $\sigma_X^2 = \sigma_Y^2$, los estimadores de máxima verosimilitud de μ_X y μ_Y siguen siendo $\bar{X}$ y $\bar{Y}$, respectivamente.

Ejemplo 2.3.12. *Suponga que se desea comparar dos institutos de capacitación en término de calificación obtenida por sus respectivos alumnos. Las calificaciones (sobre 100 puntos) para 15 alumnos del centro A es: 75, 87, 83, 73, 74, 88, 88, 74, 64, 92, 73, 87, 91, 83 y 84; y para 13 alumnos del centro B es: 64, 85, 72, 64, 74, 93, 70, 79, 79, 75, 66, 83 y 74. Antes de entrar a a analizar los datos, consideremos una distribución que puede ser apropiada para estos datos. Dada la naturaleza del problema, los datos son enteros entre 0 y 100, y por consiguiente debe ser realización de una variable discreta; sin embargo, a veces, una distribución continua también puede ser apropiada para describir datos discretos. En la Figura 2.5, se muestran las gráficas QQ plot de la distribución normal para estos dos conjuntos de datos, donde podemos ver que la distribución normal puede ser apropiada para estos datos.*

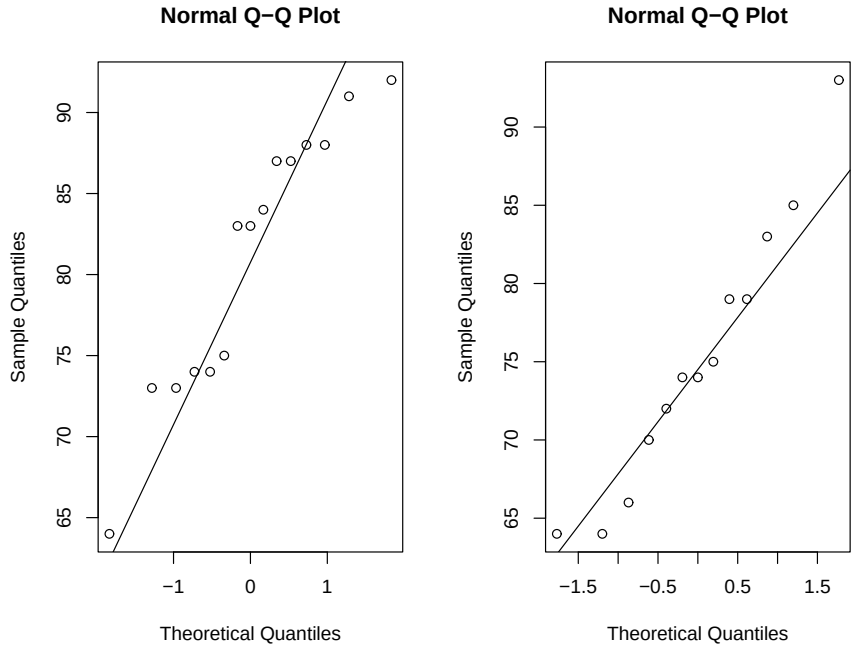

Figura 2.5: *Gráficas de QQ plot para los datos del Ejemplo 2.3.12.*

Ahora, dado que el objetivo es comparar los dos institutos, no se puede asumir la igualdad entre las dos medias teóricas; de esta forma, las dos medias teóricas se estiman mediante las medias muestrales dadas por $\hat{\mu}_A = 81$ *y* $\hat{\mu}_B = 75$*. Observamos que la estimación para el media teórica del centro A es superior a la del centro B. En los capítulos 3 y 4 se estudiarán herramientas que nos permiten concluir si esta diferencia es significativa o puede considerarse como insignificante, dado que las estimaciones no se pueden tratar como si fueran los valores verdaderos de los parámetros.*

En algunas situaciones, se desea comparar dos poblaciones en término de la dispersión tal como lo muestra el siguiente ejemplo.

Ejemplo 2.3.13. *Suponga que se desea estudiar el precio por metro cuadrado de viviendas en el centro de la capital de los países Colombia y Ecuador. La información disponible para el caso de Colombia basada en 80 viviendas es:* $\bar{x} = 700$ *(miles de pesos colombianos) y* $s_{x,n} = 95$ *(miles de pesos colombianos); y para el caso ecuatoriano basado en 50 viviendas es:* $\bar{y} = 1023$ *(bolívares venezolanos) y* $s_{y,m} = 300$ *(bolívares venezolanos).*

Dados los anteriores datos, no se puede comparar la dispersión de los dos países usando directamente las desviaciones estándares, puesto que éstas tienen unidades diferentes. Una alternativa es calcular los respectivos coeficientes de variación dados por $\rho_x = 95/700 = 13.57\,\%$ *y* $\rho_y = 300/1023 = 29.32\,\%$*. Estos coeficientes de*

variación están libres de unidad de los datos originales y pueden ser usados directamente para comparar la dispersión. Y podemos concluir que el precio de la vivienda del centro del capital de país vecino es mucho más inestable que en el caso colombiano.

2.3.2 Método de los momentos

Otro método común para encontrar estimadores de un parámetro es el método de los momentos. Para estudiar este método, primero introducimos algunas definiciones útiles.

Definición 2.3.2. *Dada una variable aleatoria X, se define el k-ésimo momento de X como $\mu_k = E(X^k)$.*

La anterior definición es al nivel poblacional. Cuando se dispone de una muestra aleatoria, se definen los momentos muestrales como sigue.

Definición 2.3.3. *Dada una muestra aleatoria $X_1, \cdots, X_n$, se define el k-ésimo momento muestral como $M_k = \sum_{i=1}^n X_i^k/n$.*

Nótese que dada una muestra aleatoria, los momentos muestrales son variables aleatorias; más aun, son estadísticas. Y podemos utilizarlos para estimar los respectivos momentos teóricos. Ahora, si logramos escribir a los parámetros desconocidos en términos de los momentos teóricos, podemos obtener fácilmente estimadores de los parámetros simplemente reemplazando los momentos teóricos por los muestrales. Los estimadores obtenidos de esta manera se llaman estimadores de momentos, se denotará por $\hat{\theta}_{mom}$. En particular, tenemos el siguiente resultado que es válido en muestras provenientes de cualquier distribución.

Resultado 2.3.2. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con esperanza común μ y varianza común σ^2, se tiene que $\hat{\mu}_{mom} = \bar{X}$ y $\hat{\sigma^2}_{mom} = S_n^2$.*

Demostración. En primer lugar, nótese que la esperanza μ es simplemente el primer momento teórico, es decir, $\mu = \mu_1$, el cual se estima con el primer momento muestral M_1, entonces se tiene que $\hat{\mu}_{mom} = M_1 = \bar{X}$.

Ahora, la varianza teórica σ^2 se puede escribir en términos de los dos primeros momentos teóricos, $\sigma^2 = \mu_2 - (\mu_1)^2$, que se estimará con $M_2 - (M_1)^2$, esto es: $\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2$, y con un poco de operación algebraica, se puede ver que ésta es $S_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2$. En conclusión, $\hat{\sigma^2}_{mom} = S_n^2$. □

Una aplicación inmediata del anterior resultado es en una muestra proveniente de una distribución normal donde los parámetros μ y σ^2 corresponden a la esperanza y la varianza de la distribución.

Ejemplo 2.3.14. *Dada una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una distribución $N(\mu, \sigma^2)$, el estimador de momentos de μ y σ^2 es simplemente la media y la varianza muestral: $\bar{X}$ y $S_n^2 = \sum(X_i - \bar{X})^2/n$. Nótese que en este caso, los estimadores de momentos coinciden con los de máxima verosimilitud.*

Otra utilidad del Resultado 2.3.2 es la siguiente forma para encontrar un estimador de momentos.

(a) Cuando hay que estimar un sólo parámetro θ, escribir a θ en término de la media teórica μ: $\theta = g(\mu)$, y al estimar μ con la media muestral: $\bar{X}$, se obtiene un estimador de momentos para θ. Esto es, $\hat{\theta}_{mom} = g(M_1)$.[5] (también se puede escribir θ en término de la varianza teórica σ^2, y al estimar σ^2 con S_n^2, se tiene un estimador de momentos de θ).

(b) Cuando hay que estimar dos parámetros θ_1 y θ_2, escribir a cada uno de ellos en término de la media y la varianza teórica $\theta_1 = g_1(\mu, \sigma^2)$ y $\theta_2 = g_2(\mu, \sigma^2)$, y al estimar μ y σ^2 con la media muestral $\bar{X}$ y la varianza muestral S_n^2, se obtiene un estimador de momentos para θ_1 y θ_2. Esto es, $\hat{\theta}_{1,mom} = g_1(\bar{X}, S_n^2)$ y $\hat{\theta}_{2,mom} = g_2(\bar{X}, S_n^2)$.

Nota: la parte (a) nos ilustra que el estimador de momentos al igual que el estimador de máxima verosimilitud, puede no ser único. Un ejemplo es cuando la muestra aleatoria proviene de la distribución $Pois(\lambda)$, se sabe que $\lambda = E(X)$, entonces un estimador de momentos de λ es, naturalmente, $\bar{X}$. Pero también se sabe que $\lambda = Var(X)$; de esta manera, se tiene otro estimador de λ que es S_n^2.

De esta forma, para los datos del Ejemplo 2.3.2, podemos tener dos estimaciones de momentos para el número promedio de muertes violentas por ciudad: $\bar{x} = 3.13$, la cual coincide con la estimación de máxima verosimilitud; y la otra estimación de momentos corresponde a $s_n^2 = 1.98$, que es muy diferente a la estimación obtenida usando $\bar{x}$. La pregunta natural ahora es ¿cuál de las dos estimaciones es mejor?, es decir, ¿cuál valor se acerca más al valor verdadero de λ? No podemos responder esta pregunta directamente, puesto que no conocemos el valor de λ. En la siguiente sección, se introducen conceptos que nos permiten evaluar la calidad de los estimadores. A pesar de que hasta ahora no tenemos herramientas para escoger entre las dos estimaciones, un simple ejercicio de simulación nos permite escoger de forma empírica. Se simula, en primer lugar, muestras provenientes de una distribución $Pois(5)$ y $Pois(15)$ con tamaño de muestra $n = 5, \cdots, 300$, y en cada muestra simulada, se calculan las dos estimaciones de momentos $\bar{x}$ y s_n^2, y se observa cuál es más cercano al valor verdadero del parámetro. Los resultados se visualizan en la gráfica superior de la Figura 2.6, donde la línea negra horizontal denota el valor verdadero del parámetro λ. El comando en R de estas simulaciones es como sigue

```
> set.seed(123)
> n<-5:300
> est.mean<-rep(NA,length(n))
> est.var<-rep(NA,length(n))
> for(i in 1:length(n)){
+ x<-rpois(n[i],5)
```

[5] Este método es válido siempre y cuando se pueda escribir al parámetro en término de la media teórica. En casos donde esto no es posible, por ejemplo en distribuciones donde no existe la media teórica o ésta no depende del parámetro, se debe recurrir a momentos teóricos superiores.

```
+ est.mean[i]<-mean(x)
+ est.var[i]<-var(x)*(n[i]-1)/n[i]
+ }
>
> est1.mean<-rep(NA,length(n))
> est1.var<-rep(NA,length(n))
> for(i in 1:length(n)){
+ y<-rpois(n[i],15)
+ est1.mean[i]<-mean(y)
+ est1.var[i]<-var(y)*(n[i]-1)/n[i]
+ }
>
>
> par(mfrow=c(2,1))
> plot(n,est.var,xlab="n",ylab="Estimación",main="Población P(5)",
type="l",col="blue",ylim=c(2,9))
> abline(5,0)
> lines(n,est.mean,col="red")
> legend(200,9.2,c("Media","Varianza"),lty=c(1,1),col=c("red","blue"),
+  box.col=0)
>
> plot(n,est1.var,xlab="n",ylab="Estimación",main="Población P(15)",
type="l",col="blue")
> abline(15,0)
> lines(n,est1.mean,col="red")
> legend(200,45,c("Media","Varianza"),lty=c(1,1),col=c("red","blue"),
+  box.col=0)
```

Se puede ver claramente, de la Figura 2.6, que la media $\bar{X}$ comparada con S_n^2 estima mejor el parámetro de la distribución, puesto que las estimaciones de $\bar{X}$ parecen estar más cercanas del valor de λ en ambas gráficas. Y por consiguiente, podemos intuir que para los datos del Ejemplo 2.3.2, la estimación $\bar{x} = 3.13$ debe ser más cercana al valor verdadero de λ. En la siguiente sección, se discutirán métodos formales acerca de escogencia entre estimadores, y se verá que el estimador $\bar{X}$ tiene mejores propiedades que S_n^2.

En los siguientes ejemplos, se ilustra la forma general de encontrar estimadores de momentos para distribuciones con dos parámetros desconocidos.

Ejemplo 2.3.15. *Dada una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una distribución gamma con parámetro de forma k y parámetro de escala θ, entonces la media y la varianza teórica están dadas por $\mu = k\theta$ y $\sigma^2 = k\theta^2$, de donde podemos expresar k y θ en término de μ y σ^2 como $k = \frac{\mu^2}{\sigma^2}$ y $\theta = \frac{\sigma^2}{\mu}$. Por lo tanto, los estimadores de momentos serán*

$$\hat{k}_{mom} = \frac{\bar{X}^2}{S_n^2} \tag{2.3.7}$$

y

$$\hat{\theta}_{mom} = \frac{S_n^2}{\bar{X}} \tag{2.3.8}$$

.

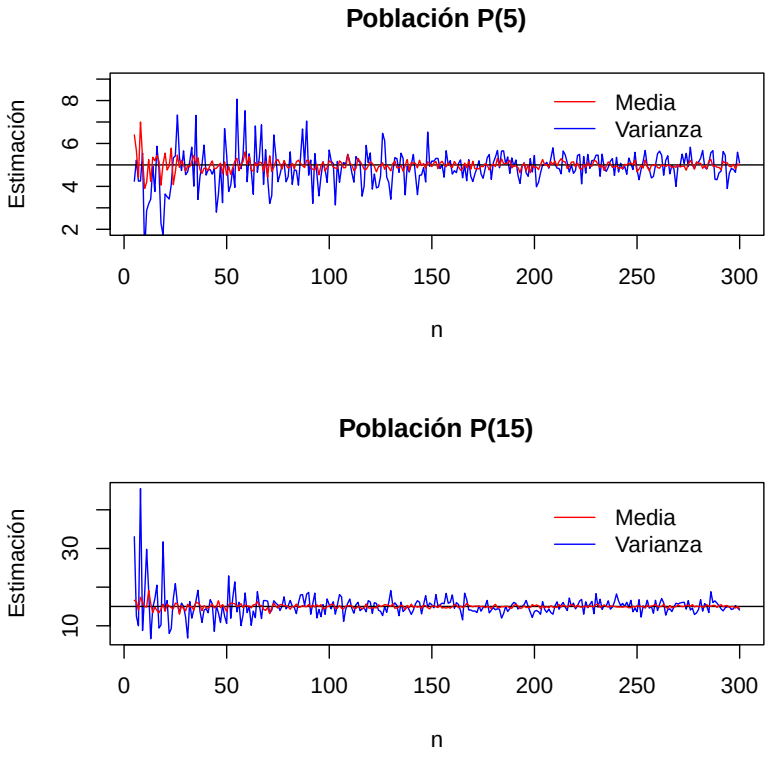

Figura 2.6: *Comparación entre la media y la varianza muestral como estimador de λ en muestras provenientes de dos distribuciones Poisson.*

Como una aplicación de lo anterior, suponga que un instituto técnico enfrenta problemas financieros y plantea un aumento en el costo de las matrículas. Es claro que ante un aumento del valor de matrícula, los estudiantes que no tengan la capacidad de pago pueden retirarse del instituto, lo que representa una pérdida económica para éste. Por lo anterior es necesario que el instituto conozca el nivel de ingreso de los estudiantes con el fin de decidir sobre el aumento de las matrículas que no cause la deserción estudiantil y a la vez pueda representar una ganancia económica para el instituto.

Con el fin de conocer el nivel de ingreso de los estudiantes, el instituto planeó una encuesta donde 127 estudiantes del instituto suministraron el valor de sus ingresos mensuales. Es usual pensar que una distribución Gamma puede ser apropiada para

variables como ingreso puesto que en primer lugar, esta variable toma valores positivos, y en segundo lugar, es altamente no simétrica, ya que para la mayoría de la población la variable ingreso toma valores intermedios o bajos, pero existe una minoría de la población que tiene ingresos bastantes altos. Para observar el comportamiento del ingreso en los 127 datos muestrales, observamos el histograma presentado en la Figura 2.7, donde se puede ver estas características reflejas, y finalmente asumimos la distribución Gamma como la distribución teórica.

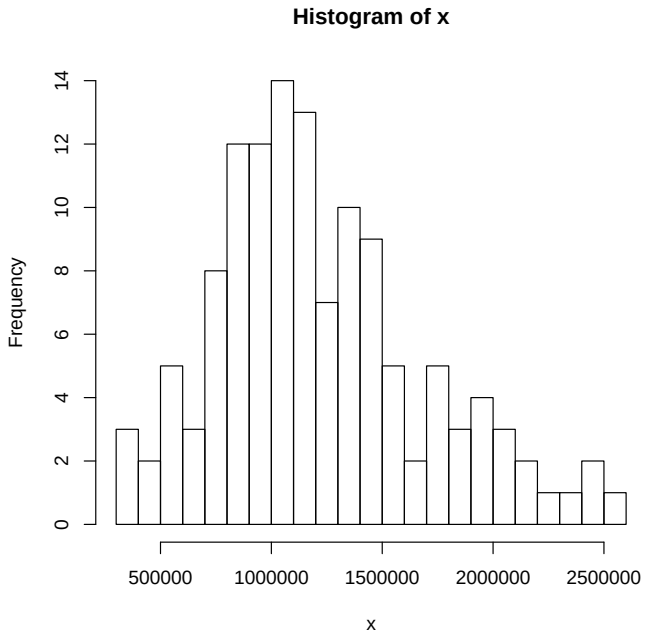

Figura 2.7: *Histograma de los datos del Ejemplo 2.3.15.*

Para estimar los parámetros de la distribución Gamma, calculamos los valores de las estadísticas $\bar{X}$ y S_n, éstas son $1.220.195 y $479.670.8, y usando éstas calculamos los parámetros estimados usando (2.3.7) y (2.3.8), y tenemos que $\hat{k} = 6.96$ y $\hat{\theta} = 175094.6$. Para visualizar el ajuste de la distribución $Gamma(6.96, 175094.6)$ a los datos, graficamos la función de densidad de esta distribución usando los siguientes códigos en R y la gráfica resultante se muestra en la Figura 2.8 donde podemos ver que la función de densidad se asemeja bastante al histograma de los datos presentado anteriormente.

```
> gama.densidad<-function(x){
+ fx<-dgamma(x,shape=k,scale=the)
+ }
>
> hist(x,breaks=20,freq=F)
```

```
> curve(gama.densidad(x),add=T)
## donde x contiene los 127 datos muestrales
```

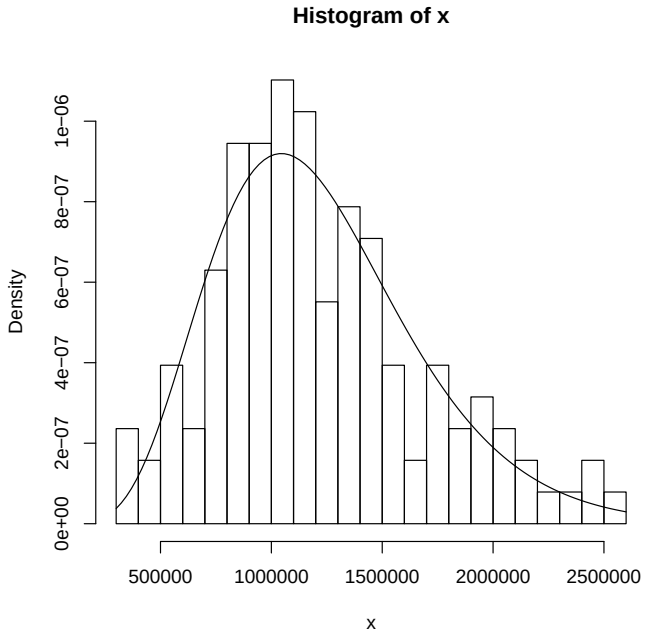

Figura 2.8: *Función de densidad de la distribución Gamma estimada de los datos del Ejemplo 2.3.15.*

Ahora, suponga que el instituto conoce que con el aumento los estudiantes con ingreso superior a $2.000.000 no tendrán dificultad para pagar el costo de la nueva matrícula, entonces una estimación del porcentaje de estos estudiantes puede ayudar a los directivos del instituto a prever los efectos del aumento. Como la distribución teórica identificada es la distribución $Gamma(6.96, 175094.6)$*, podemos calcular este porcentaje como* $Pr(X > 1500000)$ *donde* $X \sim Gamma(6.96, 175094.6)$*. Esta probabilidad se puede calcular usando el comando*

```
> pgamma(2000000,shape=6.96, scale=175094.6,lower.tail = F)
[1] 0.06111038
```

Y tenemos que aproximadamente el 6.1 % de estudiantes de este instituto tiene un ingreso mensual superior a $2.000.000. Dado que este porcentaje es bastante pequeño, los directivos deberían plantear un aumento menos drástico para evitar el efecto de deserción estudiantil.

Ejemplo 2.3.16. *Dada una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una distribución $Beta(a,b)$, entonces la esperanza y la varianza de la distribución teórica están dadas por*

$$\mu = \frac{a}{a+b} \tag{2.3.9}$$

y

$$\sigma^2 = \frac{ab}{(a+b+1)(a+b)^2}$$

Para encontrar los estimadores de momentos de a y b se debe escribir a estos en términos de μ y σ^2. Para eso tenemos que

$$\begin{aligned}\sigma^2 &= \mu\frac{b}{(a+b-1)(a+b)} \\ &= \mu(1-\mu)\frac{1}{a+b-1}\end{aligned}$$

De donde

$$a+b = \frac{\mu(1-\mu)}{\sigma^2} - 1$$

Reemplazando lo anterior en (2.3.9) tenemos que

$$a = \mu(a+b) = \frac{\mu^2(1-\mu)}{\sigma^2} - \mu$$

y

$$b = \frac{\mu(1-\mu)}{\sigma^2} - 1 - a = (1-\mu)\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right)$$

Y utilizando los principios del método de los momentos de estimar μ y σ^2 con $\bar{X}$ y S_n^2, tenemos los siguientes estimadores de momentos de a y b

$$\hat{a}_{mom} = \frac{\bar{X}^2(1-\bar{X})}{S_n^2} - \bar{X}$$

y

$$b = (1-\bar{X})\left(\frac{\bar{X}(1-\bar{X})}{S_n^2} - 1\right)$$

Como una aplicación de la distribución Beta, suponga que un almacén de cadena de ropa femenina investiga acerca de las prendas que son devueltas por los clientes después de la venta. En el caso de que se observe un gran porcentaje de devoluciones, los directivos del almacén estudiarán las causas que pueden ser mala calidad de las prendas por parte del proveedor, precios muy altos comparados con productos de la misma categoría o inclusive vendedores muy hábiles pueden disuadir a los clientes aún cuando no quieren realizar la compra y éstos pueden arrepentir posteriormente y efectuar la devolución de la compra.

Con el fin de llevar a cabo la investigación, los directivos disponen de porcentaje de prendas devueltas en un mes para diferentes sucursales. Estos datos son: 0.7 %,

0.14 %, 19.7 %, 0.1 %, 12.4 %, 1.1 %, 0.5 %, 18.9 %, 5.0 %, 0.3 %, 0.6 %, 5.4 %, 6.7 % y 0.9 %. Dado que los datos son porcentajes y los porcentajes siempre están dentro del intervalo $[0, 1]$*, podemos pensar que una distribución Beta puede ser apropiada para describir a estos datos. Observemos el histograma en la Figura 2.9 de los datos para verificar si esta distribución es adecuada o no.*

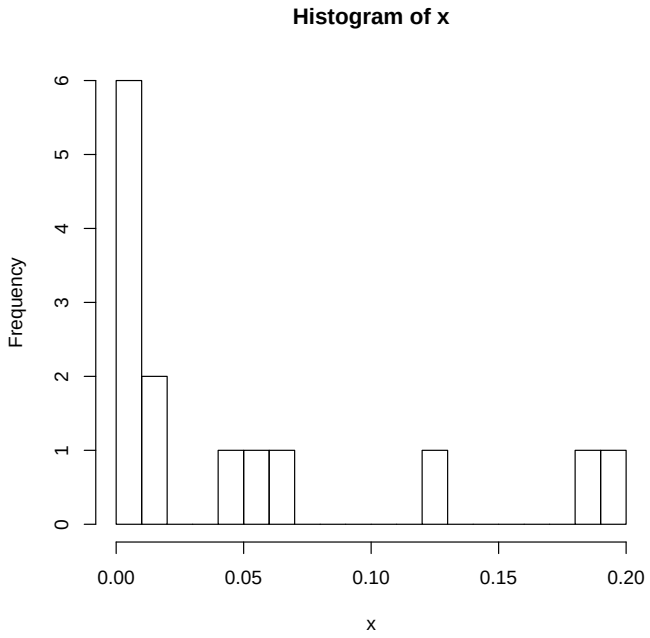

Figura 2.9: *Histograma de los datos del Ejemplo 2.3.16.*

Podemos ver que la gráfica es altamente no simétrica, de donde se descarta la distribución normal para describir a los datos. Sin embargo, podríamos pensar que una distribución exponencial puede ser apropiada dado que el histograma de los datos se asemeja a la función de densidad de esta distribución. Tenemos una herramienta útil para verificar si una distribución exponencial puede ser apropiada para los datos que es la gráfica QQ plot. Esta gráfica para los datos de este ejemplo se muestra en la Figura 2.10 donde se observa que la mayoría de datos están situados por debajo de la recta, lo cual indica que la distribución exponencial no describe bien el comportamiento de los datos, y por consiguiente también está descartada.

Dado lo anterior, podemos intentar utilizar la distribución Beta para describir este conjunto de datos. Primero calcularemos las estimaciones de los parámetros de la distribución Beta mediante el siguiente código.

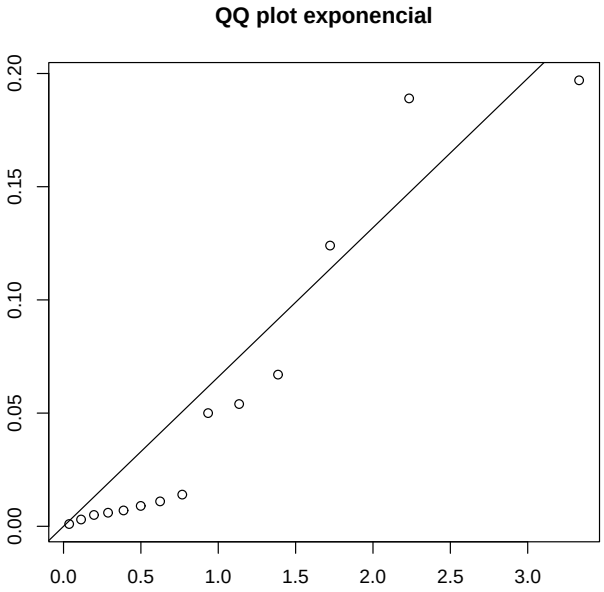

Figura 2.10: *QQ plot para verificar la distribución exponencial para los datos del Ejemplo 2.3.16.*

```
> x<-c(0.7, 1.4, 19.7, 0.1, 12.4, 1.1, 0.5, 18.9, 5.0, 0.3,
+   0.6, 5.4, 6.7, 0.9)/100
> n<-length(x)
> va<-var(x)*(n-1)/n
> bar<-mean(x)
> a<-bar^2*(1-bar)/va-bar
> a
[1] 0.5447021
> b<-(1-bar)*(bar*(1-bar)/va-1)
> b
[1] 9.80242
```

De lo anterior, tenemos las estimaciones de 0.03 y 0.46 para los parámetros de la distribución Beta. Podemos visualizar la forma de la función de densidad Beta con estos parámetros y ver que tenga aspectos similares con el histograma de los datos. Lo anterior se puede llevar a cabo usando los siguientes códigos.

```
> hist(x,breaks=20,freq=F)
> curve(dbeta(x,a,b),add=T)
```

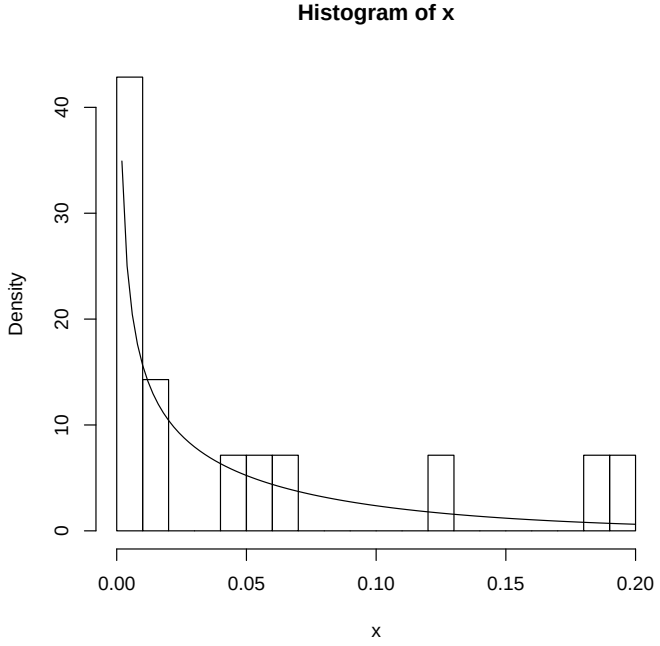

Figura 2.11: *Histograma de los datos (a) y la función de densidad estimada (b) de los datos del Ejemplo 2.3.16.*

Y como se puede observar en la Figura 2.11, la función de densidad de la distribución $Beta(0.03, 0.46)$ *parece ser apropiada para los datos observados. Ahora, si se quiere estimar el porcentaje promedio de prendas devueltas, esto es, la esperanza de la distribución teórica, podemos utilizar simplemente* $\bar{x} = 0.052 = 5.2\,\%$*, o equivalentemente la expresión (2.3.9) dada por* $\hat{a}/(\hat{a} + \hat{b}) = 0.052 = 5.2\,\%$.

En los ejemplos anteriores, se vio que en la distribución normal, los estimadores de momentos coinciden con los de máxima verosimilitud, situación que también ocurre en la distribución Poisson, exponencial y Bernoulli[6] (Ejercicios 2.3 y 2.6). Sin embargo, en muestras provenientes de algunas distribuciones del tipo uniforme, el estimador de momentos puede no coincidir con el estimador de máxima verosimilitud, como lo ilustra el siguiente ejemplo.

Ejemplo 2.3.17. *Dada una muestra aleatoria* $X_1, \cdots, X_n$ *proveniente de una distribución uniforme continua sobre* $(0, \theta)$*. Para encontrar el estimador de momentos de* θ*, se tiene en cuenta que* $\mu = \theta/2$*, de donde* $\theta = 2\mu$*, entonces se concluye que* $\hat{\theta}_{mom} = 2\bar{X}$*, el cual es diferente que el estimador de máxima verosimilitud dada por*

[6] En general, el estimador de momentos no es único, por ejemplo en la distribución Poisson, el estimador de momentos $\bar{X}$ coincide con el obtenido con el método de máxima verosimilitud, pero puede haber otros estimadores de momentos diferentes que $\bar{X}$

$\hat{\theta}_{MV} = X_{(n)}$. En la siguiente sección, se estudiará cuál de estos dos estimadores es mejor. Sin embargo, podemos realizar un pequeño estudio simulación: se simulan muestras de tamaño 5, $\cdots$, 300 de las distribuciones $Unif(0,3)$ y $Unif(0,5)$, en cada muestra simulada se calculan el estimador de momentos y de máxima verosimilitud. Estas simulaciones se pueden llevar a cabo modificando levemente los códigos en R presentados en la página 90 . Las estimaciones resultantes se observan en la Figura 2.12, donde la línea negra horizontal denota el valor verdadero del parámetro. Podemos ver que con el estimador de máxima verosimilitud siempre se obtuvieron valores más cercanos al parámetro sin importar el tamaño muestral, aunque las estimaciones de máxima verosimilitud parecen estar por debajo del θ verdadero, situación que no sucede con las estimaciones de momentos. Este hecho se confirmará en la siguiente sección mediante desarrollos teóricos.

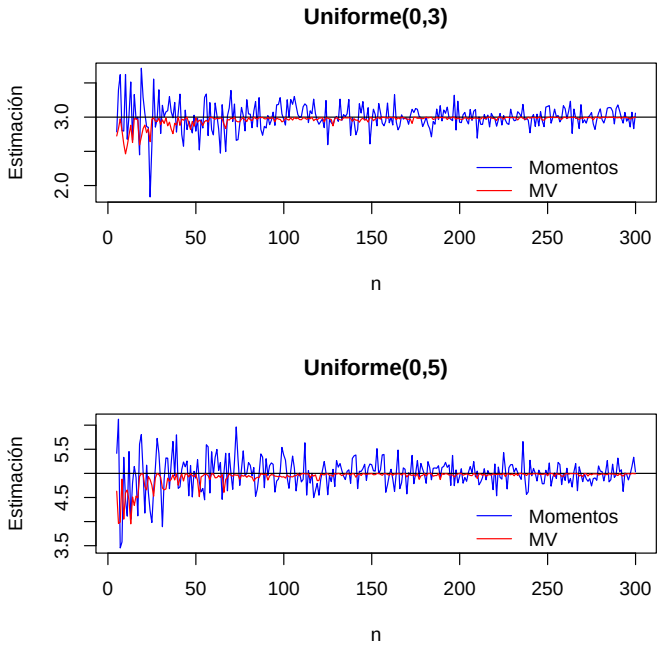

Figura 2.12: *Comparación entre el estimador de máxima verosimilitud y el de momentos en muestras provenientes de distribuciones $Unif(0,3)$ y $Unif(0,5)$.*

Análogo al método de máxima verosimilitud, en el método de los momentos también podemos garantizar la invarianza del estimador obtenido para cualquier función del parámetro $g(\theta)$. Suponga que el tiempo de llegada de un bus a cierta estación contada desde las ocho de la mañana sigue una distribución $Unif(0,\theta)$. Con esta distribución, estamos suponiendo que el tiempo de llegada toma valor en cualquier intervalo de longitud fijo de $(0,\theta)$ con la misma probabilidad. Se ha visto en el ejem-

plo anterior que $\hat{\theta}_{mom} = 2\bar{X}$, y si la cantidad que se desea estimar es la probabilidad de que el bus llegue en menos de un minuto, entonces estamos interesados en estimar $p = 1/\theta$, y p puede ser escrito en función del primer momento, puesto que $p = 1/(2\mu)$. De esta forma, aplicando los principios del método de los momentos, se tiene que $\hat{p}_{mom} = 1/(2\bar{X})$.

Para concluir el método de los momentos, damos dos ejemplos interesantes acerca de la distribución uniforme.

Ejemplo 2.3.18. *Suponga que una muestra aleatoria $X_1, \cdots, X_n$ proviene de la distribución $U[-\theta, \theta]$, en este caso la esperanza es 0, y no depende del parámetro θ, por lo tanto, no hay forma de escribir a θ en función de la esperanza. Pero podemos recurrir a la varianza de la distribución uniforme, notando que la varianza es $\theta^2/3$, de donde se tiene que $\theta = \sqrt{3\sigma^2}$ (la solución $\theta = -\sqrt{3\sigma^2}$ claramente no puede ser el parámetro de la distribución, puesto que θ debe ser positivo), en conclusión un estimador de momentos de θ es $\sqrt{3S_n^2}$. Por otro lado, se puede ver que el estimador de máxima verosimilitud de θ está dado por $\max\{-X_{(1)}, X_{(n)}\}$ (Ejercicio 2.11).*

En la Figura 2.13, se observan resultados de muestras de tamaño 5, $\cdots$, 300 simulados de una distribución $Unif[-3, 3]$ y en cada muestra simulada se calcula el estimador de momentos y de máxima verosimilitud. Podemos observar que en muestras pequeñas los resultados obtenidos con el estimador de máxima verosimilitud casi siempre están por debajo del parámetro causando el problema de subestimación. A medida que las muestras se hacen grandes el método de máxima verosimilitud parece funcionar mejor; por otro lado, aunque los valores obtenidos con el método de los momentos no parecen tener el problema de subestimación que sí tiene el de máxima verosimilitud, los valores obtenidos con el método de los momentos son muy dispersos, y hay muestras donde la estimación puede estar realmente lejos del parámetro.

En el anterior ejemplo se vio que en algunas ocasiones en el método de los momentos puede no ser útil evaluar el primer momento sino usando el segundo momento (o equivalentemente la varianza de la distribución teórica). En el siguiente ejemplo, ilustramos un caso donde el procedimiento estándar del método de los momentos arroja dos soluciones y se debe tener en cuenta el espacio paramétrico de los parámetros de interés para escoger la solución apropiada.

Ejemplo 2.3.19. *Suponga que una muestra aleatoria $X_1, \cdots, X_n$ proviene de la distribución $U[\theta_1, \theta_2]$, donde θ_1 y θ_2 son desconocidos. Para estimar estos parámetros vía el método de los momentos, necesitamos escribirlos en término de la media μ y la varianza σ^2. Usando el Resultado 1.1.11, tenemos que*

$$\begin{cases} \mu = \dfrac{\theta_1 + \theta_2}{2} \\ \sigma^2 = \dfrac{(\theta_2 - \theta_1)^2}{12}, \end{cases}$$

y tenemos dos soluciones para θ_1 y θ_2, estas son $\begin{cases} \theta_1 = \mu - \sqrt{3}\sigma \\ \theta_2 = \mu + \sqrt{3}\sigma \end{cases}$ *o* $\begin{cases} \theta_1 = \mu + \sqrt{3}\sigma \\ \theta_2 = \mu - \sqrt{3}\sigma \end{cases}$. *Nótese que en la segunda solución $\theta_1 > \theta_2$, no cumple con el supuesto de una distribución $U[\theta_1, \theta_2]$, y por consiguiente, usaremos la primera solución, y los estimadores de*

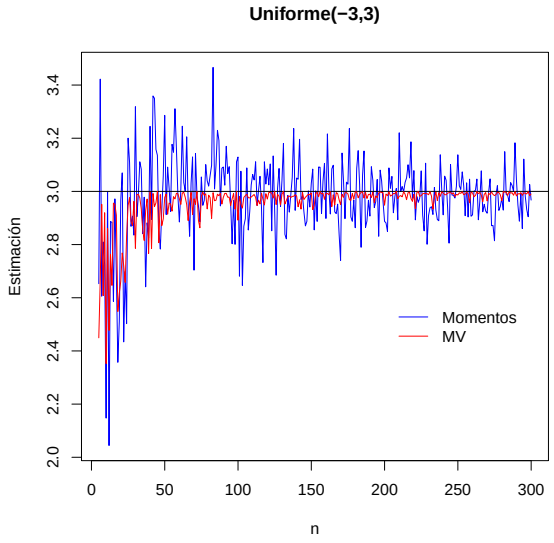

Figura 2.13: *Comparación entre el estimador de máxima verosimilitud y el de momentos en muestras provenientes de la distribución* $Unif(-3,3)$.

momentos de θ_1 *y* θ_2 *serán* $\bar{X} - \sqrt{3}S_n$ *y* $\bar{X} + \sqrt{3}S_n$*, respectivamente. Se puede ver fácilmente que los estimadores de máxima verosimilitud de* θ_1 *y* θ_2 *son* $X_{(1)}$ *y* $X_{(n)}$ *(Ejercicio 2.13).*

2.3.3 Método de mínimos cuadrados

El método de mínimos cuadrados es un método muy común en la teoría del modelamiento estadístico, aquí se hace una breve introducción utilizando este método para estimar la esperanza de una distribución teórica.

Definición 2.3.4. *Dado un conjunto de datos* $x_1, \cdots, x_n$*, una medida de dispersión es una función que satisface las siguientes propiedades*

1. $D(\mathbf{x} + \mathbf{1}_n c) = D(\mathbf{x})$
2. $D(-\mathbf{x}) = D(\mathbf{x})$

donde $\mathbf{x}$ *es el vector conformado por los* n *datos,* $\mathbf{1}_n$ *es el vector de unos de tamaño* n *y* c *es cualquier número real.*

Ejemplo 2.3.20. *Las varianzas* S_n^2*,* S_{n-1}^2 *y las respectivas desviaciones estándares son medidas de dispersión. Puesto que en primer lugar, si se traslada a un conjunto de datos agregando una constante* c S_n^2 *y* S_{n-1}^2 *no cambian de valor; en segundo lugar, estas varianzas tampoco cambian de valor si se multiplica por* -1 *a todos los datos.*

Ahora introducimos el método de mínimos cuadrados para estimar la esperanza de una distribución teórica μ basándonos en una muestra $X_1, \cdots, X_n$. Dado que μ es la media teórica, se espera que los valores de la muestra observada estén cercanos a μ, por lo tanto, podemos proponer estimar μ como el valor que más se acerque a los datos. Una forma de medir la distancia entre dos puntos es la distancia euclidiana, así podemos estimar μ como aquel que minimiza la cantidad

$$Q = \sum_{i=1}^{n} (x_i - \mu)^2.$$

De lo anterior, es claro el origen del nombre de estimador de mínimos cuadrados.

Resultado 2.3.3. *Sea $X_1, \cdots, X_n$ una muestra aleatoria proveniente de una distribución con media teórica μ, entonces el estimador de mínimos cuadrados de μ es $\bar{X}$.*

Demostración. Derivando Q con respecto a μ e igualando a cero, tenemos que

$$\frac{\partial Q}{\partial \mu} = -2 \sum_{i=1}^{n} (x_i - \mu) = 0 \tag{2.3.10}$$

el cual conduce a la solución de $\mu = \bar{X}$, de donde tenemos que el estimador de la media teórico bajo cualquier distribución está dado por

$$\hat{\mu}_{MC} = \bar{X}.$$

□

2.4 Propiedades de estimadores puntuales

En la anterior sección, se observó que para estimar un parámetro θ, el método de máxima verosimilitud y el de momentos pueden conducir a estimadores diferentes; más aún, se pueden crear muchos otros tipos de estimadores para θ, pues un estimador es simplemente una estadística que es usada para estimar. Por ejemplo, dada una muestra aleatoria $X_1, \cdots, X_{20}$ con media teórica μ desconocido, un estimador razonable para μ es la media muestral $\bar{X}$; sin embargo, alguien puede querer usar la estadística $\sum X_i$ para estimar μ, otro puede preferir algo como $\exp\{\sum X_i\}$, otra persona puede inventar su propio estimador, entonces ¿cómo se puede escoger el mejor estimador entre un conjunto de estimadores?, ¿qué aspectos y propiedades se deben tener en cuenta para esa escogencia? El objetivo de este capítulo es introducir conceptos que contestan estas preguntas.

2.4.1 Error cuadrático medio

Consideramos la siguiente situación hipotética. Suponga que para estimar un parámetro θ, se disponen de tres estimadores T_1, T_2 y T_3, y además suponga que las respectivas

Muestra	T_1	T_2	T_3
1	4.1	5.5	5.1
2	4.3	5.6	5.0
3	5.6	5.4	4.8
4	5.3	5.5	4.9
5	4.5	5.4	5.2
6	4.7	5.6	5.0
7	5.7	5.5	4.9
promedio	4.88	5.5	4.99
desviación	0.64	0.08	0.13

Tabla 2.2: *Valores de tres estimadores en 7 muestras diferentes.*

estimaciones en 7 muestras observadas de la población son los valores dados en la Tabla 2.2.

Y además suponga que el valor verdadero de θ es 5, ¿cuál estimador es mejor dadas las anteriores estimaciones? Para responder esta pregunta, observamos lo siguiente con respecto a los tres estimadores:

- Los valores que toma T_1 en promedio están cerca del 5, pero estos están muy alejados entre sí, es decir, tienen una dispersión grande. Esta dispersión grande es una propiedad indeseada del estimador, pues generalmente en la práctica, tenemos solo una muestra, una dispersión grande entre los valores de T_1 implica que hay mayor probabilidad de que T_1 tome un valor alejado del parámetro en una muestra.

- Los valores que toma T_2 están alrededor del 5.5, muy por encima del valor verdadero de θ, 5, esta situación se llama la sobreestimación. Por otro lado, en términos de la dispersión, se observa que los valores están altamente concentrados.

- Los valores que toma T_3, en primer lugar, están alrededor del 5, además de tener una dispersión pequeña. Lo anterior indica que en todas las muestras, el valor de T_3 está cercano del valor de θ. Y podemos concluir que el mejor estimador de los tres es el T_3.

La anterior situación nos ilustra que un buen estimador T debe tener dos propiedades

1. Los valores que toma T en promedio deben ser cercanos al parámetro θ. Teniendo en cuenta que la esperanza de una variable puede ser interpretada como un promedio ponderado de todos los valores que toma la variable, podemos concluir que T debe cumplir con $E(T) = \theta$,

2. La varianza de T debe ser pequeña.

Ahora damos la siguiente definición que describe la propiedad 1.

Definición 2.4.1. *Dada una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una distribución con parámetro desconocido θ, y sea T un estimador de θ, se define el sesgo de T como $B_T = E(T) - \theta$.*

Cuando $B_T = 0$ o equivalentemente $E(T) = \theta$, se dice que el estimador T es insesgado para θ. Cuando $B_T > 0$ o equivalentemente $E(T) > \theta$, se dice que T sobreestima a θ, es decir, en promedio la estimación obtenida usando T es mayor que θ. Análogamente se dice que T subestima a θ cuando $B_T < 0$.

Dada la anterior definición, en primera instancia, se necesitan estimadores con sesgo pequeño, y si es posible, insesgados. Adicionalmente, se espera que un buen estimador tenga varianza pequeña. De esta forma, si entre dos estimadores T_1 y T_2, T_1 tiene sesgo y varianza ambos menores que T_2, podemos concluir fácilmente que T_1 es mejor que T_2. Pero cuando T_1 tiene sesgo menor, pero varianza mayor que T_2, no es fácil determinar cuál es mejor. En este caso, podemos usar el siguiente criterio que combina tanto al sesgo como a la varianza de un estimador.

Definición 2.4.2. *Dada una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una distribución con parámetro desconocido θ, y sea T un estimador de θ, se define el error cuadrático medio de T como $ECM_T = E(T-\theta)^2$.*

Nótese que en la anterior definición, cuando un estimador T es insesgado para θ, se tiene que $ECM_T = E(T-E(T))^2$, esto es, el error cuadrático medio es la varianza del estimador T.

Más aun, el criterio del error cuadrático medio combina al sesgo y la varianza, tenemos que

$$\begin{aligned} ECM_T &= E[T - E(T) + E(T) - \theta]^2 \\ &= E[(T-E(T))^2 + 2(T-E(T))(E(T)-\theta) + (E(T)-\theta)^2] \\ &= E[(T-E(T))^2] + 2(E(T)-\theta)\underbrace{E[T-E(T)]}_{\text{igual a 0}} + (E(T)-\theta)^2 \\ &= Var(T) + B_T^2 \end{aligned}$$

Entonces un buen estimador debe tener el error cuadrático medio pequeño, y para los estimadores insesgados, se necesita que la varianza sea pequeña.

Ahora, al principio del capítulo, afirmaba que es natural estimar la media teórica μ con la media muestral $\bar{X}$, ¿qué tan buena es esta idea? El siguiente resultado nos permite examinar el comportamiento de $\bar{X}$ como estimador de μ.

Resultado 2.4.1. *Sea una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una distribución con media μ y varianza σ^2, entonces*

1. *si se considera a $\bar{X}$ como el estimador de μ, se tiene que $\bar{X}$ es insesgado para μ, es decir, $E(\bar{X}) = \mu$ y además $Var(\bar{X}) = \sigma^2/n$.*
2. *si se considera a S_n^2 y S_{n-1}^2 como estimadores de σ^2, se tiene que S_n^2 es sesgado para σ^2 donde el sesgo es $-\dfrac{\sigma^2}{n}$; y S_{n-1}^2 es insesgado para σ^2.*

Demostración. La demostración de la parte 1 es trivial, y se deja como ejercicio. Para ver la parte 2, tenemos que

$$\begin{aligned} E(S_n^2) &= \frac{1}{n} E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) \\ &= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) \\ &= \frac{1}{n}\left\{\sum_{i=1}^{n}\left[Var(X_i) + (E(X_i))^2\right] - nE(\bar{X})^2\right\} \\ &= \frac{1}{n}\left\{n\sigma^2 + n\mu^2 - n\left[\frac{\sigma^2}{n} + \mu^2\right]\right\} \\ &= \frac{n-1}{n}\sigma^2, \end{aligned}$$

de donde se concluye que S_n^2 es sesgado para σ^2 con sesgo $-\dfrac{\sigma^2}{n}$. Con respecto a S_{n-1}^2, al observar que

$$S_{n-1}^2 = \frac{n}{n-1} S_n^2,$$

se tiene que $E(S_{n-1}^2) = \sigma^2$ y por consiguiente es insesgado para σ^2. □

Nótese en primer lugar que en el anterior resultado, no se ha especificado la distribución de probabilidad en la población, entonces podemos aplicarlo para muestras que provienen de cualquier distribución de probabilidad. En particular, en muestras provenientes de la distribución $Exp(\theta)$, $Pois(\theta)$, $Bernoulli(\theta)$, $N(\theta, \sigma^2)$, el estimador de máxima verosimilitud coincide con el de momentos $\bar{X}$. Usando el anterior resultado, podemos concluir que $\bar{X}$ es insesgado para el parámetro θ en cualquiera de estas cuatro distribuciones.

Por otro lado, observe que $Var(\bar{X})$ es inversamente proporcional al tamaño muestral n, es decir, a medida que la muestra crece, las estimaciones son más concentradas alrededor de μ. En la Figura 2.14, se muestra un estudio de simulación donde se simularon muestras de tamaños 1, $\cdots$, 300, provenientes de distribución normal y exponencial, y en cada muestra se calcula el promedio muestral. Se observa que entre más grande sea el valor de n, más concentradas están las estimaciones alrededor de la media teórica $\mu = 5$.

Otra observación interesante es que en el contexto del resultado anterior, la variable X_1, vista como un estimador de μ también es insesgada, puesto que por definición de μ, se tiene que $E(X_1) = \mu$; la misma conclusión se tiene para X_i para $i = 2, \cdots, n$. Es decir, el estimador insesgado para un parámetro puede no ser único [7]. Pero la varianza de X_i con $i = 1, \cdots, n$ es σ^2, que es mayor que $Var(\bar{X})$, de donde se concluye que estos no son tan buenos estimadores como $\bar{X}$.

[7] De hecho, si tomamos cualquier subconjunto de la muestra aleatoria, el promedio muestral de este subconjunto será un estimador insesgados para μ

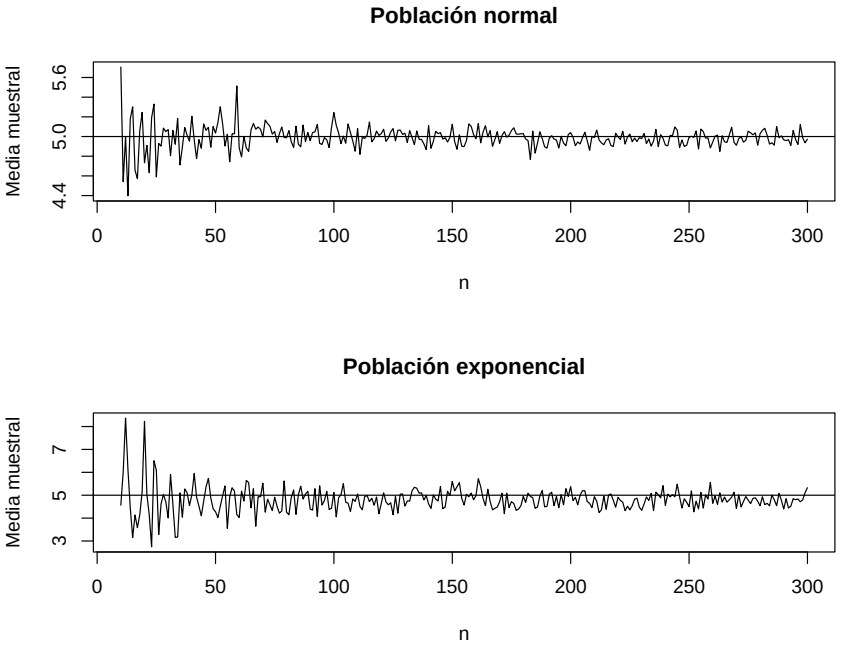

Figura 2.14: *Relación entre la estimación de μ y el tamaño muestral n.*

Ahora, revisamos los estimadores S_n^2 y S_{n-1}^2 como estimadores de σ^2. Aunque S_n^2 resulta ser sesgado para σ^2 , el sesgo se hace pequeño cuando el tamaño muestral crece; más aun,

$$\lim_{n\to\infty} B_{S_n^2} = \lim_{n\to\infty} -\frac{\sigma^2}{n} = 0.$$

Estimadores sesgados que cumplen la propiedad $\lim_{n\to\infty} B_{S_n^2} = 0$ son llamados *asintóticamente insesgados*. Para ilustrar los estimadores S_n^2 y S_{n-1}^2 en término de estimación, podemos simular muestras provenientes de una distribución normal de media 5 y desviación estándar 9 con tamaños de muestral $n = 2, 20, 30, 50, 300$, y en cada muestra calculamos los dos estimadores. El comando utilizado en R es

```
>  set.seed(1)
>  n<-c(2,10,30, 50,100, 300, 1000,5000)
>  var1<-rePr(NA,length(n))
>  var2<-rePr(NA,length(n))
>  for(k in 1:length(n))
+   {
+   data<-rnorm(n[k],5,9)
+   var1[k]<-var(data)
+   var2[k]<-(n[k]-1)*var(data)/n[k] }
```

```
>   plot(var1,type="b", col=4,ylim=c(min(var2),130),xlab="Tamaño de
+   muestra", ylab="Estimación de la varianza", xaxt="n")
>   lines(var2,type="b", col=2, pch=2)
>   abline(h=81)
>   axis(1, 1:length(n), n)
>   legend(3,120,c("Insesgado","Sesgado"), col=c(4,2), lty=c(1,1),
+   pch=c(1,2))
```

Y como resultado, obtenemos la Figura 2.15, donde podemos observar que las estimaciones del estimador sesgado S_n^2 siempre son inferiores que los del estimador insesgado S_{n-1}^2; en segundo lugar, la diferencia entre los dos estimadores se hace cada vez más pequeña y los valores de ambos estimadores se acercan al parámetro teórico a medida que el tamaño muestral crece. Por otro lado, aunque S_n^2 subestima la varianza teórica, en la gráfica podemos observar que en la muestra simulada del tamaño 300, 1000 y 5000, las estimaciones de S_n^2 estuvieron por encima de la varianza teórica, esto no es ninguna contradicción con el hecho de que S_n^2 subestima a σ^2, ya que el concepto de subestimación de un estimador indica que promediando todos los valores del estimador, da un valor inferior al parámetro, mas no indica que todas las estimaciones obtenidas son inferiores al parámetro.

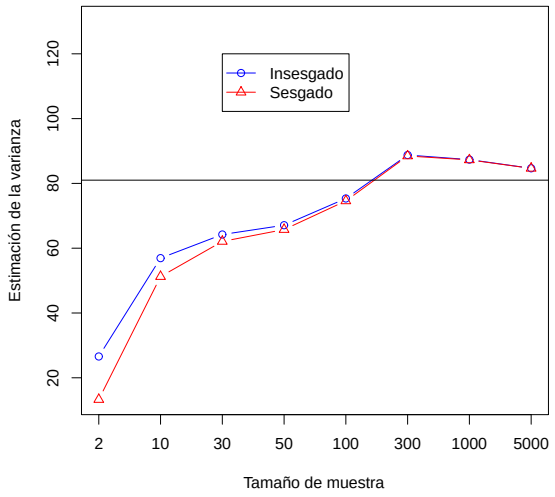

Figura 2.15: *Ilustración de las estimaciones de* S_n^2 *y* S_{n-1}^2 *como estimadores de* σ^2 *en muestras provenientes de* $N(5, 9^2)$.

Finalmente, de la parte dos del Resultado 2.4.1, también se puede concluir que los estimadores obtenidos mediante el método de máxima verosimilitud o el de momentos pueden ser sesgados, puesto que una muestra proveniente de la distribución normal, se ha visto que cuando μ es desconocido, $\hat{\sigma^2}_{MV} = \hat{\sigma^2}_{mom} = S_n^2$, y ésta es sesgada para

σ^2. Sin embargo, en la demostración del Resultado 2.4.1, se vio que en algunos casos, una pequeña modificación al estimador de máxima verosimilitud o el de momentos puede corregir el sesgo y obtener un estimador insesgado.

El Resultado 2.4.1 es válido para muestras provenientes de cualquier distribución; sin embargo, cuando la muestra proviene de una distribución normal, existe el siguiente resultado que nos permite ver que S_n^2 es sesgado para σ^2. Lo presentamos, pues es de gran utilidad para la teoría desarrollada en los capítulos siguientes.

Resultado 2.4.2. *Sea $X_1, \cdots, X_n$ una muestra aleatoria proveniente de $N(\mu, \sigma^2)$, y sea $Y = \dfrac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$, entonces se tiene que $Y \sim \chi^2_{n-1}$.*

Demostración. En primer lugar, consideramos la variable $\sum_{i=1}^n (X_i - \mu)^2$, tenemos

$$\begin{aligned}\sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 \\ &= \sum_{i=1}^n (X_i - \bar{X})^2 + 2(\bar{X} - \mu) \underbrace{\sum_{i=1}^n (X_i - \bar{X})}_{=0} + n(\bar{X} - \mu)^2 \\ &= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2\end{aligned}$$

Dividiendo σ^2 en ambos lados, se tiene que

$$\underbrace{\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2}}_{A} = \underbrace{\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}}_{Y} + \underbrace{\frac{n(\bar{X} - \mu)^2}{\sigma^2}}_{B}.$$

Si podemos suponer que las variables S_n^2 y $\bar{X}$ son independientes, podemos concluir que $\sum_{i=1}^n \dfrac{(X_i - \bar{X})^2}{\sigma^2}$ y $\dfrac{n(\bar{X} - \mu)^2}{\sigma^2}$ también son independientes. Entonces existe la siguiente relación entre las funciones generadora de momentos de las variables A, Y y B: $m_A(t) = m_Y(t) m_B(t)$, de donde se obtiene que

$$m_Y(t) = m_A(t)/m_B(t). \tag{2.4.1}$$

Ahora, $\dfrac{X_i - \mu}{\sigma}$ son variables con distribución normal estándar para $i = 1, \cdots, n$, entonces $\sum_{i=1}^n \dfrac{(X_i - \mu)^2}{\sigma^2}$ tiene distribución χ^2_n con función generadora de momentos $(1-2t)^{-n/2}$. Por el otro lado $\bar{X} \sim N(\mu, \sigma^2/n)$, entonces $\dfrac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)$, de donde se tiene que $\dfrac{n(\bar{X} - \mu)^2}{\sigma^2} \sim \chi^2_1$ cuya función generadora de momentos es $(1 - 2t)^{-1/2}$. Reemplazando lo anterior en (2.4.1), se tiene que $m_Y(t) = (1-2t)^{-(n-1)/2}$, lo cual indica que $Y \sim \chi^2_{n-1}$. □

Para completar la demostración del anterior resultado, es necesitario probar la independencia entre $\bar{X}$ y S_n^2. Tenemos el siguiente resultado.

Resultado 2.4.3. *Dada $X_1, \cdots, X_n$ una muestra aleatoria proveniente de $N(\mu, \sigma^2)$, se tiene que $\bar{X}$ y S_n^2 son independientes.*

Demostración. La demostración de este resultado es tomada de Casella & Berger (2002). Se probará que $\bar{X}$ y $\sum_{i=1}^n (X_i - \bar{X})^2$ son independientes. Tenemos

$$\begin{aligned}\sum_{i=1}^n (X_i - \bar{X})^2 &= (X_1 - \bar{X})^2 + \sum_{i=2}^n (X_i - \bar{X})^2 \\ &= \left[\sum_{i=1}^n (X_i - \bar{X}) - \sum_{i=2}^n (X_i - \bar{X})\right]^2 + \sum_{i=2}^n (X_i - \bar{X})^2 \\ &= \left[\sum_{i=2}^n (X_i - \bar{X})\right]^2 + \sum_{i=2}^n (X_i - \bar{X})^2.\end{aligned}$$

De lo anterior, se observa que $\sum_{i=1}^n (X_i - \bar{X})^2$ puede verse como una función de las variables $X_2 - \bar{X}, \cdots, X_n - \bar{X}$, por lo tanto, basta ver que estas variables son independientes de $\bar{X}$. Sin embargo, las variables $X_1, \cdots, X_n$ tienen distribución $N(\mu, \sigma^2)$, y la presencia de estos dos parámetros complica un poco los cálculos, por lo que se trabajará con las variables estandarizadas, $Z_1, \cdots, Z_n$, donde el promedio está dado por

$$\begin{aligned}\bar{Z} &= \frac{1}{n}\sum_{i=1}^n Z_i \\ &= \frac{1}{n}\sum_{i=1}^n \frac{X_i - \mu}{\sigma} \\ &= \frac{1}{n\sigma}\sum_{i=1}^n X_i - \frac{\mu}{\sigma} \\ &= \frac{\bar{X}}{\sigma} - \frac{\mu}{\sigma},\end{aligned}$$

además $Z_i - \bar{Z} = \dfrac{X_i - \bar{X}}{\sigma}$ para todo $i = 2, \cdots, n$. Por lo tanto, para ver que las variables $X_2 - \bar{X}, \cdots, X_n - \bar{X}$ son independientes de $\bar{X}$, basta ver que $Z_2 - \bar{Z}, \cdots, Z_n - \bar{Z}$ son independientes de $\bar{Z}$.

Para esto, utilizamos la función de densidad conjunta de las variables $Z_1, \cdots, Z_n$ dada por

$$f(z_1, \cdots, z_n) = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^n z_i^2\right\},$$

Ahora, se define la transformación $Y_1 = \bar{Z}$, y $Y_i = Z_i - \bar{Z}$ para $i = 2, \cdots, n$, con jacobiano igual a n^{-1}. Podemos ver que $Z_1 = Y_1 - \sum_{i=2}^n Y_i$ y $Z_i = Y_i + \bar{Y}$

para $i = 2, \cdots, n$. Usando el teorema de transformación, se tiene que la función de densidad conjunta de Y_1, $\cdots$, Y_n está dada por

$$f(y_1, \cdots, y_n) = n(2\pi)^{-n/2} \exp\left\{-\frac{1}{2}(y_1 - \sum_{i=2}^{n} y_i)^2\right\} \exp\left\{-\frac{1}{2}\sum_{i=2}^{n}(y_i + \bar{y})^2\right\}$$
$$= n(2\pi)^{-n/2} \exp\left\{-\frac{n}{2}y_1^2\right\} \exp\left\{\sum_{i=2}^{n} y_i^2 + \left[\sum_{i=2}^{n} y_i\right]^2\right\},$$

la cual es producto entre dos funciones, una que depende solo de y_1 y la otra de y_i con $i = 2, \cdots, n$[8]. Entonces podemos concluir que Y_2, $\cdots$, Y_n y Y_1 son independientes y el resultado queda demostrado.

Existe otra forma de probar esta independencia utilizando el denominado teorema de Basu; sin embargo, no hemos introducido algunos conceptos necesarios para este teorema, por esta razón, será presentado más adelante. □

Usando el Resultado 2.4.2 y propiedades de la distribución χ^2, se tiene que

$$E\left(\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}\right) = n - 1,$$

de donde

$$E(S_n^2) = E\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}\right) = \frac{n-1}{n}\sigma^2. \tag{2.4.2}$$

Es decir, el estimador S_n^2 es sesgado para σ^2, mientras que S_{n-1}^2 es insesgado para σ^2.

Ahora, recordamos que en muestras aleatorias provenientes de una distribución exponencial, Poisson o normal, el estimador de máxima verosimilitud es igual al estimador de momentos, pero no siempre es así, como ocurre en muestras provenientes de distribuciones uniformes continuas. Considere una muestra proveniente de $Unif[0, \theta]$, el estimador de máxima verosimilitud de θ es $\hat{\theta}_{MV} = X_{(n)}$ y el de momentos está dado por $\hat{\theta}_{mom} = 2\bar{X}$. Para saber cuál de estos dos estimadores es mejor, comparamos los dos estimadores en términos del sesgo y la varianza en el siguiente ejemplo.

Ejemplo 2.4.1. *Sea $X_1, \cdots, X_n$ una muestra aleatoria proveniente de una distribución uniforme continua sobre $[0, \theta]$, el estimador de máxima verosimilitud de θ es $\hat{\theta}_{MV} = X_{(n)}$ y el estimador de momentos es $\hat{\theta}_{mom} = 2\bar{X}$. Primero revisamos el desempeño de los estimadores en término del sesgo, es decir, calcularemos la esperanza de ambos estimadores. Para calcular $E(X_{(n)})$ es necesario conocer la función de densidad de probabilidad o la función de distribución de X_n. Para eso, usamos la propiedad (2.2.2), de donde para $x \in [0, \theta]$ tenemos:*

$$F_{X_{(n)}}(x) = \frac{x^n}{\theta^n}.$$

[8] ver el teorema 4.6.11 de Casella & Berger (2002)

Dada la función de distribución de $X_{(n)}$, podemos obtener la función de densidad dada por

$$f_{X_{(n)}}(x) = \frac{nx^{n-1}}{\theta^n} I_{[0,\theta]}(x).$$

Ahora calculamos $E(X_{(n)})$ como

$$E(X_{(n)}) = \int_0^\theta \frac{nx^{n-1}}{\theta^n} dx = \frac{n}{n+1}\theta. \tag{2.4.3}$$

Lo anterior concluye que $X_{(n)}$ como estimador de θ, es sesgado. Más aun, subestima a θ, hecho que se había observado en la Figura 2.12. También nótese que en la expresión (2.4.3), cuando el tamaño de la muestra $n \to \infty$, el sesgo tiende a cero, esto es, $X_{(n)}$ es un estimador asintóticamente insesgado. Ahora, miramos cómo es el sesgo del estimador de momentos, tenemos

$$E(2\bar{X}) = 2E(\bar{X}) = 2\frac{\theta}{2} = \theta,$$

pues la esperanza de $\bar{X}$ es igual a la esperanza de la distribución (ver Resultado 2.4.1). En conclusión, el estimador $2\bar{X}$ es insesgado para θ. En el término del sesgo, el estimador $2\bar{X}$ es mejor que $X_{(n)}$, aunque cuando n es grande, los dos son muy similares. Ahora miramos cuál es mejor en término de la varianza. Tenemos

$$\begin{aligned} Var(X_{(n)}) &= E(X_{(n)})^2 - (EX_{(n)})^2 \\ &= \int_0^\theta \frac{nx^{n+1}}{\theta^n} dx - \left(\frac{n\theta}{n+1}\right)^2 \\ &= \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} \\ &= \frac{n\theta^2}{(n+2)(n+1)^2}. \end{aligned}$$

Por el otro lado,

$$Var(2\bar{X}) = 4Var(\bar{X}) = 4\frac{\theta^2}{12n} = \frac{\theta^2}{3n}.$$

Algunas operaciones algebraicas indican que $Var(X_{(n)})$ es mucho más pequeña que la de $2\bar{X}$, y este aspecto ventajoso de $X_{(n)}$ puede recompensar con su sesgo que de todas formas es despreciable para valores grandes de n. Por lo tanto, se recomienda usar $X_{(n)}$ para estimar θ. Nótese que la anterior observación con respecto a la varianza también es reflejada en la Figura 2.12.

Hasta este punto, hemos concluido que en muchas situaciones, se prefiere en primera instancia a los estimadores insesgados (o por lo menos asintóticamente insesgados) y entre ellos, aquel que tiene menor varianza. Una pregunta interesante que surge ahora es si se dispone de una estimador insesgado para una función del parámetro $g(\theta)$, cómo podemos modificarlo para que siga siendo insesgado, pero con varianza menor. Para eso necesitamos el concepto de suficiencia de un estimador.

2.4.2 Suficiencia

El concepto de la suficiencia de un estimador está ligado con la idea de reducción de datos. Una muestra aleatoria provee información acerca del parámetro desconocido que se desea estimar, pero esta información está contenida en un conjunto de variables aleatorias. Si hay una manera de encontrar una función de estas variables, que contiene la misma cantidad de información para el propósito de estimación, se lograría la reducción de datos. Una variable que logra esta reducción y que además es usada para estimar el parámetro es un estimador suficiente para el parámetro. Siguiendo a esta idea, es natural pensar que toda la información contenida en la muestra $X_1, \cdots, X_n$ está contenida en el estimador suficiente (T), entonces una vez conocido el valor que toma T, la muestra ya no provee ninguna información acerca del parámetro.

La definición rigurosa de un estimador suficiente se presenta a continuación.

Definición 2.4.3. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con función de densidad $f(x_i, \theta)$, y sea $T = T(X_1, \cdots, X_n)$ un estimador de θ, se dice que T es suficiente para θ si la distribución condicional de X dados valores de T no depende de θ.*

En algunos textos, establecen que un estimador T es suficiente para θ si $Pr(X_1 = x_1, \cdots, X_n = x_n | T = t)$ no depende de θ, lo cual no es del todo riguroso, puesto que cuando las variables X_i con $i = 1, \cdots, n$ son continuas, la anterior probabilidad condicional (cuando está bien definida) siempre es igual a cero, que no depende de θ. Por otra part, también el estimador T como función de las variables de la muestra puede ser continuo, entonces $Pr(T = t) = 0$ y no se puede definir la esperanza condicional. Claro que cuando las variables X_i y T son discretas, podemos usar esta definición sin problema e ilustramos la forma de verificar que un estimador sea suficiente en el siguiente ejemplo.

Ejemplo 2.4.2. *Sea $X_1, \cdots, X_n$ una muestra aleatoria con distribución $Pois(\lambda)$, se ha visto que el estimador de máxima verosimilitud y de momentos de λ está dado por $\bar{X}$. Además éste es insesgado para λ por el Resultado 2.4.1. Ahora veamos que también es un estimador suficiente para λ. Como la distribución Poisson es discreta, entonces para verificar la suficiencia de $\bar{X}$ podemos verificar que $Pr(X_1 = x_1, \cdots, X_n = x_n | \bar{X} = x)$ no depende de λ, tenemos:*

$$
\begin{aligned}
& Pr(X_1 = x_1, \cdots, X_n = x_n | \bar{X} = x) \\
= & \frac{Pr(X_1 = x_1, \cdots, X_n = x_n, \bar{X} = x)}{Pr(\bar{X} = x)} \\
= & \frac{Pr(X_1 = x_1, \cdots, X_n = x_n, \sum_{i=1}^n X_i = nx)}{Pr(\sum_{i=1}^n X_i = nx)} \\
= & \frac{Pr(X_1 = x_1, \cdots, X_{n-1} = x_{n-1}, X_n = nx - \sum_{i=1}^{n-1} x_i)}{Pr(\sum_{i=1}^n X_i = nx)} \\
= & \frac{Pr(X_1 = x_1) \cdots Pr(X_{n-1} = x_{n-1}) Pr(X_n = nx - \sum_{i=1}^{n-1} x_i)}{Pr(\sum_{i=1}^n X_i = nx)}
\end{aligned}
$$

$$= \frac{\frac{e^{-n\lambda}\lambda^{x_1}\cdots\lambda^{x_{n-1}}\lambda^{nx-\sum_{i=1}^{n-1}x_i}}{x_1!\cdots x_{n-1}!(nx-\sum_{i=1}^{n-1}x_i)!}}{\frac{e^{-n\lambda}(n\lambda)^{nx}}{(nx)!}}$$

$$= \frac{(nx)!}{n^{nx}x_1!\cdots x_{n-1}!(nx-\sum_{i=1}^{n-1}x_i)!},$$

claramente la anterior probabilidad condicional no depende de λ, de donde se concluye que $\bar{X}$ es suficiente para λ. Utilizando un razonamiento completamente análogo, se puede ver que $\sum_{i=1}^{n} X_i$ también es suficiente para λ.

Ahora, como se vio en el anterior ejemplo, utilizar la definición para demostrar que un estimador es suficiente puede resultar un poco tedioso, pues el cómputo de una probabilidad condicional, en general, no es sencillo. El siguiente teorema, conocido como el criterio o el teorema de factorización de Fisher-Neyman, es útil para verificar que un estimador es suficiente para el parámetro desconocido.

Resultado 2.4.4. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con función de densidad $f(x_i, \theta)$, y sea $T = T(X_1, \cdots, X_n)$ un estimador de θ, entonces T es suficiente para θ, si y solo si, se puede factorizar la función de verosimilitud $L(\theta, x_1, \cdots, x_n)$ como*

$$L(\theta, x_1, \cdots, x_n) = g(t(x_1, \cdots, x_n), \theta)h(x_1, \cdots, x_n)$$

Demostración. Se hará la prueba para cuando las variables $X_1, \cdots, X_n$ y T son discretas, la demostración es como sigue:

($\Leftarrow$) Primero supongamos que se tiene la factorización

$$L(\theta, x_1, \cdots, x_n) = g(t(x_1, \cdots, x_n), \theta)h(x_1, \cdots, x_n), y$$

veamos que T es suficiente para θ, es decir, veamos que $Pr(X_1 = x_1, \cdots, X_n = x_n | T = t)$ no depende de θ. En primer lugar, si $t \neq T(x_1, \cdots, x_n)$ entonces la probabilidad vale 0 y por consiguiente no depende de θ. Ahora si $t = T(x_1, \cdots, x_n)$, tenemos:

$$\begin{aligned} Pr(X_1 = x_1, \cdots, X_n = x_n | T = t) &= \frac{Pr(X_1 = x_1, \cdots, X_n = x_n, T = t)}{Pr(T = t)} \\ &= \frac{Pr(X_1 = x_1, \cdots, X_n = x_n)}{Pr(T = t)} \\ &= \frac{g(t(x_1, \cdots, x_n), \theta)h(x_1, \cdots, x_n)}{Pr(T = t)} \end{aligned}$$

Al definir A como el conjunto de valores de $x_1, \cdots, x_n$ que son enviados al valor t mediante la variable T, tenemos que

$$\begin{aligned} Pr(X_1 = x_1, \cdots, X_n = x_n | T = t) &= \frac{g(t(x_1, \cdots, x_n), \theta) h(x_1, \cdots, x_n)}{\sum_A Pr(X_1 = x_1, \cdots, X_n = x_n)} \\ &= \frac{g(t(x_1, \cdots, x_n), \theta) h(x_1, \cdots, x_n)}{\sum_A g(t(x_1, \cdots, x_n), \theta) h(x_1, \cdots, x_n)} \\ &= \frac{g(t, \theta) h(x_1, \cdots, x_n)}{\sum_A g(t, \theta) h(x_1, \cdots, x_n)} \\ &= \frac{h(x_1, \cdots, x_n)}{\sum_A h(x_1, \cdots, x_n)}, \end{aligned}$$

el cual no depende del valor θ.

($\Rightarrow$) Ahora supongamos que T es suficiente para θ, veamos que se tiene la factorización $L(\theta, x_1, \cdots, x_n) = g(t(x_1, \cdots, x_n), \theta) h(x_1, \cdots, x_n)$ para algunas funciones g y h. Tenemos que

$$\begin{aligned} & L(\theta, x_1, \cdots, x_n) \\ &= Pr(X_1 = x_1, \cdots, X_n = x_n) \\ &= Pr(X_1 = x_1, \cdots, X_n = x_n, T = t(x_1, \cdots, x_n)) \\ &= \underbrace{Pr(T = t(x_1, \cdots, x_n))}_{g} \underbrace{Pr(X_1 = x_1, \cdots, X_n = x_n | T = t(x_1, \cdots, x_n))}_{h}, \end{aligned}$$

la primera probabilidad no depende de θ por la suficiencia de T, y la segunda probabilidad depende de $t(x_1, \cdots, x_n)$ y de θ, y hemos logrado obtener la factorización de $L(\theta, x_1, \cdots, x_n)$.

La prueba para cuando X_1, $\cdots$, X_n y T son continuas es más complicada, el lector puede consultarla en Lehmann & Romano (2005, p. 20). □

Ahora, retomamos el Ejemplo 2.4.2. utilizando el criterio de factorización para ilustrar la utilidad del resultado. Tenemos:

$$\begin{aligned} L(\lambda, x_1, \cdots, x_n) &= \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \prod_{i=1}^n I_{\{0,1,\cdots\}}(x_i) \\ &= \underbrace{e^{-n\lambda} \lambda^{n\bar{x}}}_{g(\bar{x}, \lambda)} \underbrace{\frac{\prod_{i=1}^n I_{\{0,1,\cdots\}}(x_i)}{\prod_{i=1}^n x_i!}}_{h(x_1, \cdots, x_n)}, \end{aligned}$$

con la anterior expresión se logra escribir la función de verosimilitud en forma del Resultado 2.4.4., de donde se concluye que $\bar{X}$ es suficiente para λ. Nótese que la anterior factorización no es única, pues también se tiene que:

$$L(\lambda, x_1, \cdots, x_n) = \underbrace{e^{-n\lambda} \lambda^{\sum x_i}}_{g(\sum x_i, \lambda)} \underbrace{\frac{\prod_{i=1}^n I_{\{0,1,\cdots\}}(x_i)}{\prod_{i=1}^n x_i!}}_{h(x_1, \cdots, x_n)},$$

de donde se concluye que también $\sum_{i=1}^{n} X_i$ es suficiente para λ. Utilizando este criterio, se puede verificar fácilmente que en muestras provenientes de las distribuciones $Exp(\theta)$, $Bernoulli(\theta)$, $N(\theta, \sigma^2)$ con σ^2 conocida, las estadísticas $\bar{X}$ y $\sum_{i=1}^{n} X_i$ son suficientes para θ.

El criterio de factorización de Fisher-Neyman presentado en el Resultado 2.4.4. cubre solamente a las distribuciones con un parámetro desconocido, también existe la versión general para distribuciones con más de un parámetro. Dado que en la mayoría de los casos no se trabaja con distribuciones con más de dos parámetros, se presenta únicamente la versión para distribuciones con dos parámetros.

Resultado 2.4.5. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con función de densidad $f(x_i, \theta_1, \theta_2)$, y sea $T_1 = T_1(X_1, \cdots, X_n)$ y $T_2 = T_2(X_1, \cdots, X_n)$ son estimadores de θ_1 y θ_2, entonces T_1 y T_2 son suficientes para θ_1 y θ_2, si y solo si, se puede factorizar la función de verosimilitud $L(\theta, x_1, \cdots, x_n)$ como*

$$L(\theta, x_1, \cdots, x_n) = g(t_1(x_1, \cdots, x_n), \theta_1, t_2(x_1, \cdots, x_n), \theta_2) h(x_1, \cdots, x_n)$$

La utilidad del resultado se ilustra en el siguiente ejemplo.

Ejemplo 2.4.3. *En una distribución beta, la función de densidad de probabilidad está dada por:*

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0,1)}(x).$$

Dada una muestra aleatoria de tamaño n, la función de verosimilitud está dada por

$$\begin{aligned} L(\alpha, \beta) &= \frac{\Gamma(\alpha+\beta)^n}{\Gamma(\alpha)^n \Gamma(\beta)^n} \prod_{i=1}^{n} x_i^{\alpha-1} \prod_{i=1}^{n} (1-x_i)^{\beta-1} \prod_{i=1}^{n} I_{(0,1)}(x_i) \\ &= \underbrace{\frac{\Gamma(\alpha+\beta)^n}{\Gamma(\alpha)^n \Gamma(\beta)^n} \left(\prod_{i=1}^{n} x_i \right)^{\alpha-1} \left(\prod_{i=1}^{n} (1-x_i) \right)^{\beta-1}}_{g(t_1, t_2, \alpha, \beta)} \underbrace{\prod_{i=1}^{n} I_{(0,1)}(x_i)}_{h(x_1, \cdots, x_n)}. \end{aligned}$$

Y podemos concluir que las estadísticas $\prod_{i=1}^{n} X_i$ y $\prod_{i=1}^{n} (1-X_i)$ son suficientes para α y β.

Para las distribuciones pertenecientes a la familia exponencial, siempre podemos encontrar estadísticas suficientes para el parámetro, el resultado se da a continuación.

Resultado 2.4.6. *Dada una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una distribución $f(x, \theta)$ perteneciente a la familia exponencial, es decir,*

$$f(x, \theta) = h(x) c(\theta) \exp\{d(\theta) T(x)\},$$

entonces la estadística $\sum_{i=1}^{n} T(X_i)$ es una estadística suficiente para θ.

Demostración. El resultado es trivial usando el criterio de factorización de Fisher-Neyman. Por (1.2.2), tenemos que la función de verosimilitud de una muestra aleatoria con función de densidad perteneciente a la familia exponencial está dada por

$$L(\theta, x_1, \cdots, x_n) = c(\theta)^n \left[\prod_{i=1}^{n} h(x_i)\right] \exp\left\{d(\theta) \sum_{i=1}^{n} T(x_i)\right\}.$$

Al tomar $\sum_{i=1}^{n} T(X_i)$ como la estadística T y $c(\theta)^n \exp\{d(\theta) \sum_{i=1}^{n} T(x_i)\}$ como la función $g(t(x_1, \cdots, x_n), \theta)$, y el restante como $h(x_1, \cdots, x_n)$, se tiene que $\sum_{i=1}^{n} T(X_i)$ es suficiente para θ. □

Para ilustrar la utilidad del resultado, consideramos una muestra proveniente de la distribución $Ber(p)$, esta distribución pertenece a la familia exponencial, puesto que

$$\begin{aligned} f(x,p) &= p^x (1-p)^{1-x} I_{\{0,1\}}(x) \\ &= (\frac{p}{1-p})^x I_{\{0,1\}}(x) \\ &= (1-p) I_{\{0,1\}}(x) \exp\left\{x \ln \frac{p}{1-p}\right\}, \end{aligned}$$

entonces $T(x) = x$, y por el anterior resultado, se tiene que $\sum_{i=1}^{n} T(X_i) = \sum_{i=1}^{n} X_i$ es una estadística suficiente para p. Teniendo en cuenta que una estadística suficiente resume toda la información contenida en una muestra acerca de un parámetro θ, lo anterior nos indica que en un conjunto de observaciones del tipo 0 y 1 provenientes de $Ber(p)$, para el efecto de estimación de p, basta con observar la suma de las observaciones, de esta forma podemos reducir un gran volumen de datos en solo un dato, y la estimación obtenida de p no se ve afectada [9].

Por otro lado, como la presentación de una función de densidad de la familia exponencial no es única, entonces podemos encontrar diferentes estadísticas suficientes para un mismo parámetro. En efecto, la densidad de la distribución $Ber(p)$ también puede escribirse como:

$$f(x,p) = (1-p) I_{\{0,1\}}(x) \exp\left\{\frac{x}{n}\left[n \ln \frac{p}{1-p}\right]\right\},$$

de esta manera, $T(x) = x/n$, así también se probó que $\bar{X} = \sum_{i=1}^{n} X_i/n$ es una estadística suficiente para p.

Ahora, en distribuciones biparamétricas también podemos encontrar fácilmente estadísticas suficientes si éstas pertenecen a la familia exponencial. El siguiente resultado es el análogo al Resultado 2.4.6. para distribuciones biparamétricas.

Resultado 2.4.7. *Dada una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una distribución $f(x_i, \theta_1, \theta_2)$ perteneciente a la familia exponencial biparamétrica de la forma*

$$f(x_i, \theta_1, \theta_2) = c(\boldsymbol{\theta}) h(x) \exp\{d(\boldsymbol{\theta})' T(x)\},$$

[9] Vea el Ejemplo 2.3.3. donde la estimación de p se llevó a cabo usando solamente la suma de las observaciones.

donde $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $d(\boldsymbol{\theta}) = (d_1(\boldsymbol{\theta}), d_2(\boldsymbol{\theta}))'$ *y* $T(x) = (T_1(x), T_2(x))'$, *entonces las estadísticas* $\sum_{i=1}^n T_1(X_i)$ *y* $\sum_{i=1}^n T_2(X_i)$ *son estadísticas suficientes para* θ_1 *y* θ_2.

Demostración. La prueba es análoga al caso para distribuciones uniparamétricas, usando el criterio de factorización de Fisher-Neyman. Tenemos que la función de verosimilitud está dada por

$$\begin{aligned}
&L(\theta_1, \theta_2, x_1, \cdots, x_n) \\
&= c(\boldsymbol{\theta})^n \left\{ \prod_{i=1}^n h(x_i) \right\} \exp\left\{ d(\boldsymbol{\theta})' \sum_{i=1}^n T(x_i) \right\} \\
&= c(\theta_1, \theta_2)^n \left\{ \prod_{i=1}^n h(x_i) \right\} \exp\left\{ (d_1(\boldsymbol{\theta}), d_2(\boldsymbol{\theta}))' \sum_{i=1}^n \begin{pmatrix} T_1(x_i) \\ T_2(x_i) \end{pmatrix} \right\} \\
&= \underbrace{\left\{ \prod_{i=1}^n h(x_i) \right\}}_{h(x_1, \cdots, x_n)} \underbrace{c(\theta_1, \theta_2)^n \exp\left\{ d_1(\boldsymbol{\theta}) \sum_{i=1}^n T_1(x_i) + d_2(\boldsymbol{\theta}) \sum_{i=1}^n T_2(x_i) \right\}}_{g(t_1, \theta_1, t_2, \theta_2)},
\end{aligned}$$

de esta forma, tenemos que $\sum_{i=1}^n T_1(X_i)$ y $\sum_{i=1}^n T_2(X_i)$ son suficientes para θ_1 y θ_2. □

Ilustramos la aplicación del resultado en el siguiente ejemplo.

Ejemplo 2.4.4. *Sea* $X_1, \cdots, X_n$ *una muestra aleatoria con distribución* $N(\mu, \sigma^2)$, *el anterior resultado servirá para encontrar estadísticas suficientes para* μ *y* σ^2. *La función de densidad pertenece a la familia exponencial biparamétrica pues se puede escribir de la forma*

$$f(x, \mu, \sigma^2) = \exp\left\{ \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right) \begin{pmatrix} x \\ x^2 \end{pmatrix} \right\} \exp\left\{ -\frac{\mu^2}{2\sigma^2} \right\} (2\pi\sigma^2)^{-1/2},$$

entonces $T_1(x) = x$ *y* $T_2(x) = x^2$, *entonces el resultado anterior indica que las estadísticas* $\sum_{i=1}^n X_i$ *y* $\sum_{i=1}^n X_i^2$ *son suficientes y completas para* μ *y* σ^2.

Volviendo al tópico de la evaluación de la calidad de un estimador, una inquietud que había surgido al tener en cuenta que un buen estimador debe ser insesgado con varianza pequeña es: "dado un estimador insesgado, cómo construir otro insesgado con varianza menor". El siguiente teorema de Rao-Blackwell [10] afirma que al combinar un estimador insesgado con una estadística suficiente, se puede lograr un estimador insesgado con una varianza menor.

Resultado 2.4.8. *Sea* $X_1, \cdots, X_n$ *una muestra aleatoria con función de densidad* $f(x_i, \theta)$, *si* T_1 *es un estimador insesgado para una función de* θ, $g(\theta)$, *y* T_2 *es suficiente para* θ, *entonces el estimador* $T = E(T_1|T_2)$ *es insesgado para* $g(\theta)$ *y tiene varianza menor que* T_1.

[10] El teorema fue establecido por el estadístico hindú Calyampudi Radhakrishna Rao y por el americano David Blackwell

Demostración. En primer lugar $E(T) = E(E(T_1|T_2)) = E(T_1) = g(\theta)$. Ahora, en término de varianza, tenemos

$$\begin{aligned} Var(T_1) &= Var(E(T_1|T_2)) + E(Var(T_1|T_2)) \\ &= Var(T) + E(Var(T_1|T_2)) \\ &\geq Var(T) \end{aligned}$$

□

La importancia de este teorema radica en que para estimar una función de un parámetro desconocido $g(\theta)$ si tenemos un estimador insesgado T_1 podemos, con base en este, construir un mejor estimador que T_1, siempre y cuando se disponga de un estimador suficiente para θ.

Para un mejor entendimiento del teorema revisamos, en primer lugar, el concepto de la esperanza condicional. Lo más importante que hay que aclarar es que la expresión $E(T_1|T_2)$ en el resultado anterior no es una constante, sino una variable aleatoria. Para ilustrar esto, considere el siguiente ejemplo:

Ejemplo 2.4.5. *Dadas variables aleatorias X e Y con función de densidad de probabilidad conjunta dada por:*

$$f(x,y) = \begin{cases} e^{-y} & \text{si } 0 < x < y \\ 0 & \text{en otro caso} \end{cases},$$

Para calcular $E(X|Y)$ primero recordamos que ésta es una función con dominio igual al rango de Y y a cada valor y lo envía a la esperanza $E(X|Y = y)$. Entonces dado y, para calcular $E(X|Y = y)$, primero se calcula la función de densidad condicional $f_{X|Y}(x|y) = f(x,y)/f_Y(y)$, en nuestro caso,

$$f_{X|Y}(x|y) = \begin{cases} y^{-1} & \text{si } 0 < x < y \\ 0 & \text{en otro caso} \end{cases},$$

de donde para un valor particular que toma la variable Y, se puede calcular $E(X|Y = y)$ como

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx,$$

nótese que la anterior esperanza condicional es un número, función de y. En nuestro caso, tenemos que $E(X|Y = y) = y/2$. Entonces $E(X|Y)$ envía cada valor y a $y/2$, es decir $E(X|Y) = Y/2$, la cual claramente es una variable aleatoria.

En general, calcular una esperanza condicional $E(X|Y)$ puede implicar cálculos tediosos, pero en algunos casos puede ser trivial como lo indica el siguiente resultado.

Resultado 2.4.9. *Si X e Y son variables aleatorias, y X puede escribirse como una función de Y, entonces $E(X|Y) = X$.*

Aclarado el concepto de la esperanza condicional, volvemos al teorema de Rao Blackwell. En la demostración no se utilizó el hecho de que T_2 sea una estadística suficiente, por lo tanto, podemos intuir que al condicionar T_1 en cualquier otra estadística, digamos Q, también se puede mejorar la calidad del estimador, en el término de que $Var(T_1|S)$ será menor que $Var(T_1)$. Lo anterior es cierto, pero puede suceder que $T_1|S$ dependa del parámetro θ y deja de ser un estimador. El lector puede consultar a Casella & Berger (2002, Ejemplo 7.3.18, p. 343) para un ejemplo donde $T_1|S$ depende de θ. Por esta razón, se necesita que el condicionamiento sea sobre una estadística suficiente para garantizar que la resultante esperanza condicional no dependa del parámetro y pueda ser usada como un estimador.

2.4.3 Estimadores UMVUE

El teorema de Rao Blackwell plantea la posibilidad de un proceso continuo de construcción de estimadores insesgados con varianzas cada vez menores, la inquietud que surge ahora es si podemos construir estimadores de varianza cada vez menor o podemos encontrar un estimador insesgado T de tal forma que ya no existe ningún otro estimador insesgado con varianza menor que $Var(T)$. Si existe alguna cota inferior para la varianza de los estimadores, y se encuentra un estimador insesgado T con varianza igual a esta cota, se podrá concluir que no habrá otro estimador insesgado con varianza más pequeña que ésta, y se podrá afirmar que T es el mejor de todos los estimadores insesgados. Esta cota existe efectivamente y se denomina la cota de Cramer Rao, y no solo es la cota inferior para la varianza de los estimadores insesgados, sino también puede ser cota inferior para la varianza de todos los estimadores. Para estudiar la cota de Cramer Rao, introducimos algunos conceptos preliminares.

Definición 2.4.4. *Dada X una variable aleatoria con función de densidad $f(x,\theta)$, donde θ es el parámetro de la distribución, y además existe $\dfrac{\partial}{\partial\theta}\ln f(x,\theta)$, entonces se define la información de Fisher contenida en X acerca de θ como*

$$I_X(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\ln f(X,\theta)\right]^2\right\}.$$

Nótese que en la anterior definición, $f(X,\theta)$ no es la función de densidad $f(x,\theta)$, sino la variable X transformada a través de la función f, es decir, $f(X,\theta)$ es una variable aleatoria, y por consiguiente, tiene sentido calcular la esperanza. Ahora, en algunas situaciones existe una definición equivalente que mide esta cantidad de información, y puede resultar más fácil el cálculo.

Resultado 2.4.10. *En la anterior definición, si además existe $\dfrac{\partial^2}{\partial\theta^2}\ln f(x,\theta)$, entonces se tiene que*

$$I_X(\theta) = -E\left\{\frac{\partial^2}{\partial\theta^2}\ln f(X,\theta)\right\}.$$

Demostración. Tenemos:

$$\begin{aligned}
-E\left\{\frac{\partial^2}{\partial\theta^2}\ln f(X,\theta)\right\} &= -E\left\{\frac{\partial}{\partial\theta}\left[\frac{1}{f(X,\theta)}\frac{\partial f(X,\theta)}{\partial\theta}\right]\right\} \\
&= -E\left\{-\frac{1}{f^2(X,\theta)}\left[\frac{\partial f(X,\theta)}{\partial\theta}\right]^2+\frac{1}{f(X,\theta)}\frac{\partial^2 f(X,\theta)}{\partial\theta^2}\right\} \\
&= E\left\{\frac{1}{f^2(X,\theta)}\left[\frac{\partial f(X,\theta)}{\partial\theta}\right]^2\right\}-E\left\{\frac{1}{f(X,\theta)}\frac{\partial^2 f(X,\theta)}{\partial\theta^2}\right\},
\end{aligned}$$

La última esperanza vale cero, puesto que si X es una variable continua,

$$\begin{aligned}
E\left\{\frac{1}{f(X,\theta)}\frac{\partial^2 f(X,\theta)}{\partial\theta^2}\right\} &= \int_{\mathbb{R}}\frac{1}{f(x,\theta)}\frac{\partial^2 f(x,\theta)}{\partial\theta^2}f(x,\theta)dx \\
&= \int_{\mathbb{R}}\frac{\partial^2 f(x,\theta)}{\partial\theta^2}dx \\
&= \frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}}f(x,\theta)dx \\
&= \frac{\partial^2}{\partial\theta^2}(1)=0.
\end{aligned}$$

Y si X es discreta, suponga que los valores que toma son $x_1, x_2\cdots$, entonces,

$$\begin{aligned}
E\left\{\frac{1}{f(X,\theta)}\frac{\partial^2 f(X,\theta)}{\partial\theta^2}\right\} &= \sum_i\frac{1}{f(x_i,\theta)}\frac{\partial^2 f(x_i,\theta)}{\partial\theta^2}Pr(X=x_i) \\
&= \sum_i\frac{\partial^2 f(x_i,\theta)}{\partial\theta^2} \\
&= \frac{\partial^2}{\partial\theta^2}\sum_i f(x_i,\theta) \\
&= \frac{\partial^2}{\partial\theta^2}(1)=0.
\end{aligned}$$

En conclusión,

$$-E\left\{\frac{\partial^2}{\partial\theta^2}\ln f(X,\theta)\right\} = E\left\{\frac{1}{f^2(X,\theta)}\left[\frac{\partial f(X,\theta)}{\partial\theta}\right]^2\right\}.$$

Ahora,

$$\begin{aligned} I_X(\theta) &= E\left\{\left[\frac{\partial}{\partial\theta}\ln f(X,\theta)\right]^2\right\} \\ &= E\left\{\left[\frac{1}{f(X,\theta)}\frac{\partial f(X,\theta)}{\partial\theta}\right]^2\right\} \\ &= E\left\{\frac{1}{f^2(X,\theta)}\left[\frac{\partial f(X,\theta)}{\partial\theta}\right]^2\right\} \\ &= -E\left\{\frac{\partial^2}{\partial\theta^2}\ln f(X,\theta)\right\}, \end{aligned}$$

y así el resultado queda demostrado. □

Las anteriores definiciones introducen la información contenida en una variable; sin embargo, cuando tenemos disponible una muestra aleatoria, es necesario definir la información contenida en una muestra aleatoria acerca de algún parámetro.

Definición 2.4.5. *Dada $X_1, \cdots, X_n$ variables aleatorias con función de densidad $f(x_i,\theta)$, donde θ es el parámetro de la distribución, y además existe $\frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f(x_i,\theta)$, entonces se define la información de Fisher contenida en la muestra aleatoria acerca de θ como*

$$I_{X_1,\cdots,X_n}(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f(X_i,\theta)\right]^2\right\}.$$

Recordemos que en una muestra aleatoria, las variables tienen la misma distribución de probabilidad, además son independientes, entonces es natural pensar que cada variable debe aportar la misma cantidad de información, es decir, la información contenida en una muestra de tamaño n debe ser igual a n veces la información contenida en cualquier variable de la muestra. El siguiente resultado confirma esta intuición.

Resultado 2.4.11. *Dada $X_1, \cdots, X_n$ una muestra aleatoria, entonces*

$$I_{X_1,\cdots,X_n}(\theta) = nI_X(\theta),$$

donde $I_X(\theta) = I_{X_i}(\theta)$, con $i = 1,\cdots,n$. Es decir, en una muestra aleatoria, cada variable aporta la misma cantidad de información, y la cantidad total de información en la muestra es la suma de la información en cada variable.

Demostración.

$$\begin{aligned}
I_{X_1,\cdots,X_n}(\theta) &= E\left\{\left[\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(X_i,\theta)\right]^2\right\} \\
&= E\left\{\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right]^2\right\} \\
&= E\left\{\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right]^2\right\} + \\
&\qquad \underbrace{E\left\{\sum_{\substack{i,j=1\\ i\neq j}}^{n}\left[\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\frac{\partial}{\partial\theta}\ln f(X_j,\theta)\right]\right\}}_{=0,\ \text{por la independencia entre } X_i \text{ y } X_j} \\
&= \sum_{i=1}^{n} E\left\{\left[\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right]^2\right\} \\
&= \sum_{i=1}^{n} I_X(\theta) = nI_X(\theta).
\end{aligned}$$

□

Ilustramos el cálculo de la información contenida en una muestra en el siguiente ejemplo, y posteriormente presentamos cómo este concepto resulta útil en la definición de la cota de Cramer Rao.

Ejemplo 2.4.6. *Sea $X_1, \cdots, X_n$ una muestra aleatoria proveniente de la distribución $N(\mu,\sigma^2)$, la información contenida en la muestra acerca de μ es n/σ^2. Para verificar esta afirmación, calculamos la información acerca de μ en una variable X con distribución $N(\mu,\sigma^2)$. Tenemos:*

$$\begin{aligned}
I_X(\mu) &= -E\left\{\frac{\partial^2}{\partial\mu^2}\ln f(X,\theta)\right\} \\
&= -E\left\{\frac{\partial^2}{\partial\mu^2}\left[-\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(X-\mu)^2\right]\right\} \\
&= -E\left\{\frac{\partial}{\partial\mu}\left[\frac{X-\mu}{\sigma^2}\right]\right\} \\
&= -E\left\{-\frac{1}{\sigma^2}\right\} \\
&= \frac{1}{\sigma^2}.
\end{aligned}$$

Ahora, usando el Resultado 2.4.11, se tiene que $I_{X_1,\cdots,X_n}(\mu) = n/\sigma^2$.

Nótese que esta información, en primer lugar, depende del tamaño n de manera que entre más grande sea la muestra, hay mayor información acerca de μ; en segundo lugar, entre más pequeña sea la varianza σ^2, la cantidad de información acerca de μ también incrementa. Esto es natural, puesto que si σ^2 es pequeña, los datos de la muestra están muy concentrados alrededor de μ, entonces estos datos aportan más información que otros datos con más dispersión.

Para muestras provenientes de otras distribuciones como la binomial, exponencial y Poisson, también se puede hallar la información de Fisher de manera análoga. Ilustramos estos casos a continuación.

Ejemplo 2.4.7. *Si X es una variable aleatoria con distribución $Bin(n, \theta)$, entonces para calcular la información de X acerca de θ, tenemos que*

$$\ln f(X) = \ln \binom{n}{X} + X \ln \theta + (n - X) \ln(1 - \theta)$$

y

$$\frac{\partial^2 \ln f(X)}{\partial \theta^2} = -\frac{X}{\theta^2} - \frac{n - X}{(1 - \theta)^2}$$

Por lo tanto al calcular la esperanza, y por consiguiente la información de Fisher, se tiene que

$$I_X(\theta) = -E\left[\frac{\partial^2 \ln f(X)}{\partial \theta^2}\right] = \frac{n\theta}{\theta^2} + \frac{n - n\theta}{(1 - \theta)^2} = \frac{n}{\theta(1 - \theta)}$$

De donde vemos que al tener un número mayor de ensayos, también se obtiene mayor información acerca de θ.

Ahora consideramos una muestra con distribución Poisson.

Ejemplo 2.4.8. *Si $X_1,\ldots,X_n$ es una muestra aleatoria de variables con distribución $Pois(\theta)$, la información de Fisher contenida en la muestra acerca de θ está dada por $I(\theta) = n/\theta$ puesto que*

$$\ln f(X_1, \ldots, X_n) = -n\theta + \sum_{i=1}^{n} X_i \ln \theta - \sum_{i=1}^{n} \ln(X_i!)$$

y

$$\frac{\partial^2 \ln f(X_1, \ldots, X_n)}{\partial \theta^2} = -\frac{\sum_{i=1}^{n} X_i}{\theta^2}$$

Por lo tanto al calcular la esperanza, se tiene que

$$I_{X_1,\cdots,X_n}(\theta) = -E\left[\frac{\partial^2 \ln f(X_1, \ldots, X_n)}{\partial \theta^2}\right] = \frac{\sum_{i=1}^{n} E(X_i)}{\theta^2} = \frac{n}{\theta}$$

Ahora, cuando la distribución teórica tiene más de un parámetro, entonces la información contenida en una variable acerca del vector de parámetros va a ser una matriz. Presentamos la definición correspondiente a continuación.

Definición 2.4.6. *Dada una variable aleatoria X con función de densidad $f(x,\boldsymbol{\theta})$, la matriz de información contenida en X acerca de $\boldsymbol{\theta}$ se define como*

$$I_X(\boldsymbol{\theta}) = E\left\{\frac{\partial \ln f(X,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\left(\frac{\partial \ln f(X,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right\}$$

En la anterior definición, $\boldsymbol{\theta}$ es un vector columna, y por consiguiente $\frac{\partial \ln f(X,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ también lo es, y podemos ver que $I_X(\boldsymbol{\theta})$ es una matriz cuadrada de dimensión $r \times r$, donde r denota el número de parámetros en el vector $\boldsymbol{\theta}$.

En el caso de que se dispone de una muestra aleatoria, el concepto de información se extiende de forma análoga al caso de uniparamétrica.

Definición 2.4.7. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con función de densidad $f(x_i,\boldsymbol{\theta})$, la matriz de información contenida en la muestra acerca de $\boldsymbol{\theta}$ se define como*

$$I_{X_1,\cdots,X_n}(\boldsymbol{\theta}) = E\left\{\frac{\partial \ln \prod_{i=1}^n f(X_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\left(\frac{\partial \ln \prod_{i=1}^n f(X_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right\}$$

Y se deja como ejercicio verificar que en una muestra aleatoria $I_{X_1,\cdots,X_n}(\boldsymbol{\theta}) = nI_X(\boldsymbol{\theta})$ donde X tiene la misma distribución que las variables $X_1, \cdots, X_n$ (Ejercicio 2.17).

Ejemplo 2.4.9. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con distribución común $N(\mu,\sigma^2)$, vamos a hallar la matriz de información contenida en la muestra acerca del vector de parámetros (μ,σ^2). Tenemos que*

$$\begin{aligned}
&I_{X_1,\cdots,X_n}(\mu,\sigma^2)\\
&= E\left\{\begin{pmatrix}\frac{\partial \ln \prod_{i=1}^n f(X_i,\mu,\sigma^2)}{\partial \mu}\\ \frac{\partial \ln \prod_{i=1}^n f(X_i,\mu,\sigma^2)}{\partial \sigma^2}\end{pmatrix}\begin{pmatrix}\frac{\partial \ln \prod_{i=1}^n f(X_i,\mu,\sigma^2)}{\partial \mu} & \frac{\partial \ln \prod_{i=1}^n f(X_i,\mu,\sigma^2)}{\partial \sigma^2}\end{pmatrix}\right\}\\
&= E\left\{\begin{pmatrix}\frac{\sum_{i=1}^n X_i - n\mu}{\sigma^2}\\ \frac{\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2}{2\sigma^4}\end{pmatrix}\begin{pmatrix}\frac{\sum_{i=1}^n X_i - n\mu}{\sigma^2} & \frac{\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2}{2\sigma^4}\end{pmatrix}\right\}\\
&= E\begin{pmatrix}\frac{(\sum_{i=1}^n X_i - n\mu)^2}{\sigma^4} & \frac{(\sum_{i=1}^n X_i - n\mu)(\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2)}{2\sigma^6}\\ \frac{(\sum_{i=1}^n X_i - n\mu)(\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2)}{2\sigma^6} & \frac{(\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2)^2}{4\sigma^8}\end{pmatrix}
\end{aligned}$$

Donde el primer elemento diagonal de la anterior matriz está dado por

$$\begin{aligned}
E\left\{\frac{(\sum_{i=1}^n X_i - n\mu)^2}{\sigma^4}\right\} &= \left[Var\left(\sum_{i=1}^n X_i - n\mu\right) + (E\left(\sum_{i=1}^n X_i - n\mu\right))^2\right]/\sigma^4\\
&= n\sigma^2/\sigma^4 = n/\sigma^2.
\end{aligned}$$

El segundo elemento diagonal está dado por

$$E\left\{\frac{(\sum_{i=1}^{n}(X_i-\mu)^2-n\sigma^2)^2}{4\sigma^8}\right\} \tag{2.4.4}$$

$$=\frac{1}{4\sigma^8}E\left\{\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]^2+n^2\sigma^4-2n\sigma^2\sum_{i=1}^{n}(X_i-\mu)^2\right\} \tag{2.4.5}$$

$$=\frac{1}{4\sigma^8}\left\{Var(\sum_{i=1}^{n}(X_i-\mu)^2)+\left[E(\sum_{i=1}^{n}(X_i-\mu)^2)\right]^2+n^2\sigma^4-2n\sigma^2E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]\right\} \tag{2.4.6}$$

Usando el hecho de que

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

y la esperanza y varianza de la distribución χ_n^2, tenemos que la expresión (2.4.4) está dada por

$$\frac{1}{4\sigma^8}\left\{2n\sigma^4+\left[n\sigma^2\right]^2+n^2\sigma^4-2n\sigma^2n\sigma^2\right\}=\frac{n}{2\sigma^4}.$$

Finalmente, el elemento fuera de la diagonal de la matriz $I_{X_1,\cdots,X_n}(\mu,\sigma^2)$ está dado por

$$E\left\{\left(\sum_{i=1}^{n}X_i-n\mu\right)\left(\sum_{i=1}^{n}(X_i-\mu)^2-n\sigma^2\right)\right\}$$

$$=E\left\{\sum_{i=1}^{n}X_i\left(\sum_{i=1}^{n}(X_i-\mu)^2-n\sigma^2\right)-n\mu\left(\sum_{i=1}^{n}(X_i-\mu)^2-n\sigma^2\right)\right\}$$

$$=E\left\{\sum_{i=1}^{n}X_i\sum_{i=1}^{n}(X_i-\mu)^2\right\}-n\sigma^2E\left(\sum_{i=1}^{n}X_i\right)-n\mu E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right)+n^2\mu\sigma^2$$

$$=E\left(\sum_{i=1}^{n}X_i\sum_{i=1}^{n}X_i^2\right)-2\mu E\left[(\sum_{i=1}^{n}X_i)^2\right]+n^2\mu^3-n^2\mu\sigma^2-n^2\mu\sigma^2+n^2\mu\sigma^2$$

$$=E\left(\sum_{i=1}^{n}X_i^3+\sum_{i\neq j}X_iX_j^2\right)-2\mu(n\sigma^2+n^2\mu^2)+n^2\mu^3-n^2\mu\sigma^2$$

$$=\sum_{i=1}^{n}\left[3\mu E(X_i^2)-2\mu^3\right]+\sum_{i\neq j}E(X_i)E(X_j^2)-2n\mu\sigma^2-2n^2\mu^3+n^2\mu^3-n^2\mu\sigma^2$$

$$=3n\mu(\sigma^2+\mu^2)-2n\mu^3+\mu(\sigma^2+\mu^2)(n^2-n)-2n\mu\sigma^2-2n^2\mu^3+n^2\mu^3-n^2\mu\sigma^2$$

$$=0$$

De donde obtenemos finalmente la matriz de información $I_{X_1,\cdots,X_n}(\mu,\sigma^2)$ dada por

$$I_{X_1,\cdots,X_n}(\mu,\sigma^2)=\begin{pmatrix}\frac{n}{\sigma^2} & 0\\ 0 & \frac{n}{2\sigma^4}\end{pmatrix}$$

El hecho de que esta matriz de información sea diagonal nos ilustra que la información total en la muestra con respecto a μ no tiene relación con la información contenida acerca de σ^2. Y podemos confirmar este hecho con el Resultado 2.4.3 donde muestra que los estimadores de μ y σ^2 son efectivamente independientes.

Cuando se introdujo el concepto de una estadística suficiente, su interpretación es que contiene toda la información de la muestra acerca de algún parámetro, esta información se puede entender como la información de Fisher, y el siguiente resultado provee la respectiva sustentación.

Resultado 2.4.12. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con función de densidad $f(x_i,\theta)$, y sea T un estimador suficiente para θ, entonces la información contenida en T acerca de θ es la misma información contenida en la muestra aleatoria acerca de θ*

Demostración. Lo que probaremos es $I_T(\theta)=I_{X_1,\cdots,X_n}(\theta)$. Tenemos, en primer lugar,

$$\begin{aligned}f(X_1,\cdots,X_n|T)&=\frac{f(X_1,\cdots,X_n,T)}{f(T)}\\&=\frac{g(T,\theta)h(X_1,\cdots,X_n)}{f(T)}\quad\text{usando el criterio de factorización,}\end{aligned}$$

de donde tenemos que

$$f(T)=\frac{g(T,\theta)h(X_1,\cdots,X_n)}{f(X_1,\cdots,X_n|T)}.$$

Ahora usando la función de densidad de T calculamos la información contenida en T como

$$\begin{aligned}I_T(\theta)&=E\left\{\left[\frac{\partial}{\partial\theta}\ln f(T)\right]^2\right\}\\&=E\left\{\left[\frac{\partial}{\partial\theta}\ln\frac{g(T,\theta)h(X_1,\cdots,X_n)}{f(X_1,\cdots,X_n|T)}\right]^2\right\}\\&=E\left\{\left[\frac{\partial}{\partial\theta}\left(\ln g(T,\theta)+\ln h(X_1,\cdots,X_n)-\ln f(X_1,\cdots,X_n|T)\right)\right]^2\right\}\\&=E\left\{\left[\frac{\partial}{\partial\theta}\ln g(T,\theta)\right]^2\right\},\end{aligned}$$

pues $h(X_1, \cdots, X_n)$ no depende de θ, y tampoco $f(X_1, \cdots, X_n|T)$ por la definición de suficiencia de T.

Ahora, calculamos la información contenida en la muestra acerca de θ, tenemos

$$\begin{aligned} I_{X_1,\cdots,X_n}(\theta) &= E\left\{\left[\frac{\partial}{\partial\theta}\ln f(X_1,\cdots,X_n)\right]^2\right\} \\ &= E\left\{\left[\frac{\partial}{\partial\theta}\ln g(T,\theta)h(X_1,\cdots,X_n)\right]^2\right\} \\ &= E\left\{\left[\frac{\partial}{\partial\theta}(\ln g(T,\theta)-\ln h(X_1,\cdots,X_n))\right]^2\right\} \\ &= E\left\{\left[\frac{\partial}{\partial\theta}\ln g(T,\theta)\right]^2\right\}. \end{aligned}$$

Y podemos concluir que $I_T(\theta) = I_{X_1,\cdots,X_n}(\theta)$ de donde se concluye que la información contenida en T con respecto a θ es la misma información contenida en la muestra $X_1, \cdots, X_n$. □

Ahora, como se mencionaba anteriormente, el concepto de información de Fisher permite encontrar una cota inferior para la varianza de los estimadores. Este se enuncia en la famosa desigualdad de información, y se presenta a continuación:

Resultado 2.4.13. *Dada X_1, $\cdots$, X_n variables aleatorias con distribución de probabilidad $f(x_i, \theta)$, y T es un estimador para $g(\theta)$, si*

$$\frac{\partial}{\partial\theta}E(T) = \int \frac{\partial}{\partial\theta}t(x_1,\cdots,x_n)f(x_1,\cdots,x_n)dx_1\cdots dx_n \tag{2.4.7}$$

y $Var(T) < \infty$, entonces

$$Var(T) \geq \frac{(\frac{\partial}{\partial\theta}E(T))^2}{I_{X_1,\cdots,X_n}(\theta)}.$$

Y $\dfrac{(\frac{\partial}{\partial\theta}E(T))^2}{I_{X_1,\cdots,X_n}(\theta)}$ es llamado la cota de Cramer Rao.

Demostración. La demostración del resultado se basa en el hecho de que el coeficiente de correlación entre dos variables es siempre menor o igual a 1. Entonces para las variables T y $\frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f(X_i,\theta)$, se tiene que

$$1 \geq Corr(T, \frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f(X_i,\theta)) = \frac{Cov(T, \frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f(X_i,\theta))}{\sqrt{Var(T)Var(\frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f(X_i,\theta))}},$$

el cual es equivalente a

$$1 \geq \frac{Cov(T, \frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f(X_i,\theta))^2}{Var(T)Var(\frac{\partial}{\partial\theta}\ln\prod_{i=1}^n f(X_i,\theta))}$$

de donde se tiene

$$Var(T) \geq \frac{Cov(T, \frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(X_i, \theta))^2}{Var(\frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(X_i, \theta))}. \tag{2.4.8}$$

Ahora

$$\begin{aligned} Var(\frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(X_i, \theta)) &= E\left\{ \left[\frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(X_i, \theta) \right]^2 \right\} - \underbrace{\left(E\left[\frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(X_i, \theta) \right] \right)^2}_{=0} \\ &= E\left\{ \left[\frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(X_i, \theta) \right]^2 \right\} \\ &= I_{X_1, \cdots, X_n}(\theta), \end{aligned}$$

puesto que

$$\begin{aligned} E\left[\frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(X_i, \theta) \right] &= \sum_{i=1}^{n} \int_{\mathbb{R}} \frac{\partial}{\partial\theta} \ln f(x_i, \theta) f(x_i, \theta) dx_i \\ &= \sum_{i=1}^{n} \int_{\mathbb{R}} \frac{1}{f(x_i, \theta)} \frac{\partial f(x_i, \theta)}{\partial\theta} f(x_i, \theta) dx_i \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \int_{\mathbb{R}} f(x_i, \theta) dx_i \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta}(1) = 0, \end{aligned}$$

si las variables $X_1, \cdots, X_n$ son continuas. Cuando son discretas, se tiene análogamente.

Por otro lado,

$$\begin{aligned} &Cov\left(T, \frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(X_i, \theta) \right) \\ &= E\left[T \frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(X_i, \theta) \right] \qquad \text{pues } E\left(\frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f(X_i, \theta) \right) = 0 \\ &= \int_{\mathbb{R}^n} t(x_1, \cdots, x_n) \frac{\partial \ln f_{X_1, \cdots, X_n}(x_1, \cdots, x_n)}{\partial\theta} f_{X_1, \cdots, X_n}(x_1, \cdots, x_n) dx_1 \cdots dx_n \\ &= \int_{\mathbb{R}^n} t(x_1, \cdots, x_n) \frac{\partial}{\partial\theta} f_{X_1, \cdots, X_n}(x_1, \cdots, x_n) dx_1 \cdots dx_n \\ &= \frac{\partial}{\partial\theta} \int_{\mathbb{R}^n} t(x_1, \cdots, x_n) f_{X_1, \cdots, X_n}(x_1, \cdots, x_n) dx_1 \cdots dx_n = \frac{\partial}{\partial\theta} E(T). \end{aligned}$$

Reemplazando las anteriores expresiones en (2.4.8), se tiene el resultado. □

La condición en (2.4.7) es parte de las condiciones denominadas condiciones de regularidad, y se debe garantizar que esta condición se cumple para tener la validez de la desigualdad de Cramer Rao. Para las distribuciones pertenecientes a la familia exponencial, esta condición siempre se tiene; sin embargo, para distribuciones donde el rango de la variable depende del parámetro como la distribución uniforme, la desigualdad de Cramer Rao puede no ser cierta. El lector puede consultar Casella & Berger (2002, Ejemplo 7.3.13, p. 339) para ver un ejemplo donde ocurre esta situación.

Ahora, la cota de Cramer Rao dada en el anterior resultado no asume muchas condiciones como la independencia acerca de las variables, y tampoco características especiales del estimador T. Sin embargo, la mayor utilidad de la cota de Cramer Rao se da cuando las variables constituyen una muestra aleatoria y el estimador T sea insesgado para $g(\theta)$. En este caso $\frac{\partial}{\partial\theta}E(T) = \frac{\partial}{\partial g(\theta)}\theta = g'(\theta)$ y la desigualdad se convierte en

$$Var(T) \geq \frac{g'(\theta)}{nI_X(\theta)}.$$

Y en el caso cuando lo que se desea estimar es simplemente el parámetro de la distribución θ, $g(\theta) = \theta$, y la desigualdad se convierte en

$$Var(T) \geq \frac{1}{nI_X(\theta)}. \tag{2.4.9}$$

Ahora, si un estimador insesgado T tiene varianza igual a la cota de Cramer Rao, entonces cualquier otro estimador insesgado necesariamente tendrá varianza más grande que T, es decir, T será el mejor entre todos los estimadores insesgados, y existe un nombre especial para estos estimadores que se presenta en la siguiente definición.

Definición 2.4.8. *Dada $X_1, \cdots, X_n$ una muestra aleatoria con función de densidad $f(x_i, \theta)$, y T un estimador insesgado para θ, si $Var(T) \leq Var(T^*)$ para todo θ para cualquier otro estimador T^* insesgado de θ, entonces se dice que T es un estimador insesgado de varianza uniformemente mínima, UMVUE* [11].

Con respecto a estimadores UMVUE, tenemos varios comentarios.

- Como lo mencionado anteriormente, cuando un estimador insesgado tiene varianza igual a la cota de Cramer Rao, entonces este es un UMVUE, pero no necesariamente sucede lo contrario. Es decir, un UMVUE no necesariamente tiene varianza igual a la cota de Cramer Rao. Más adelante en el Ejemplo 2.4.12, se mostrará un caso de estos.
- Una forma para verificar que un estimador sea UMVUE es ver que la varianza es igual a la cota de Cramer Rao, además hay que ver que es un estimador insesgado. En otras palabras, un estimador sesgado que tenga varianza igual a la cota de Cramer Rao no es UMVUE.
- Un estimador UMVUE tiene la varianza más pequeña entre todos los estimadores insesgados, pero puede existir un estimador sesgado T^* con varianza aún más

[11] Por su sigla en inglés *Uniformly Minimum Variance Unbiased Estimator*.

pequeña. En este caso, se debe escoger entre el UMVUE y T^*, pues puede suceder que la ganancia en la varianza sea considerable y que T^* sea asintóticamente insesgado y de esta forma, corregir el sesgo aumentando el tamaño de muestra n.

Consideremos un ejemplo para ilustrar los estimadores UMVUE. En la sección anterior se mencionaba que en una muestra proveniente de la distribución $Pois(\theta)$, existen dos estimadores de momentos $\bar{X}$ y S_n^2. Mediante estudios de simulación, se vio que $\bar{X}$ es mejor que S_n^2. La razón teórica es que $\bar{X}$ es UMVUE para θ, lo cual mostramos en el siguiente ejemplo.

Ejemplo 2.4.10. *Dada $X_1, \cdots, X_n$ una muestra aleatoria con distribución $Pois(\theta)$, el estimador $\bar{X}$ como estimador de θ es UMVUE. En primer lugar, se vio en el Resultado 2.4.1. $\bar{X}$ siempre es insesgado para la media teórica, el cual en la distribución Poisson es igual al parámetro θ. Entonces tenemos que $\bar{X}$ es insesgado para θ. Resta verificar que la varianza de $\bar{X}$ es mínima entre todos los estimadores insesgados. Para eso, podemos calcular la cota de Cramer Rao, y ver que ésta es igual a $Var(\bar{X})$.*

Por el Resultado 2.4.1, $Var(\bar{X}) = \sigma^2/n$ donde σ^2 es la varianza de la distribución de probabilidad, que en este ejemplo es la distribución $Pois(\theta)$ cuya varianza es θ. En conclusión

$$Var(\bar{X}) = \frac{\theta}{n}.$$

Ahora calculamos la cota de Cramer Rao, tenemos

$$\begin{aligned} I_{X_1,\cdots,X_n}(\theta) &= nI_X(\theta) \\ &= -nE\left\{\frac{\partial^2}{\partial\theta^2}\ln f(X,\theta)\right\} \\ &= -nE\left\{\frac{\partial^2}{\partial\theta^2}\left[-\theta + X\ln\theta - \ln X!\right]\right\} \\ &= -nE\left\{\frac{\partial}{\partial\theta}\left[-1 + \frac{X}{\theta}\right]\right\} \\ &= -nE\left\{-\frac{X}{\theta^2}\right\} \\ &= \frac{n}{\theta}, \end{aligned}$$

de donde se concluye que la cota de Cramer Rao es θ/n, que es igual a la varianza de $\bar{X}$. En conclusión, $\bar{X}$ es UMVUE para θ.

Por otro lado, de nuevo usando el Resultado 2.4.1, el otro estimador de momentos S_n^2 es insesgado para la varianza teórica que en una distribución Poisson también corresponde al parámetro θ. Este insesgamiento se puede corroborar con la Figura 2.6 donde muestran las estimaciones obtenidas usando $\bar{X}$ y S_n^2 en muestras de distribución Poisson, donde es claro que las estimaciones de S_n^2 estuvieron siempre alrededor del valor de θ; sin embargo, como $\bar{X}$ es UMVUE, siempre será mejor que S_n^2 como estimador de θ.

Como se vio en el anterior ejemplo, para ver que la varianza de un estimador sea igual a la cota de Cramer Rao, se necesita calcular la información de Fisher contenida en la muestra, y esto puede ser tedioso. Sin embargo, existe un resultado que requiere tal vez menos operaciones algebraicas, que nos permite saber cuándo un estimador tiene la varianza igual a la cota de Cramer Rao.

Resultado 2.4.14. *Dada $X_1, \cdots, X_n$ una muestra aleatoria con distribución de probabilidad $f(x_i, \theta)$, y $T = T(X_1, \cdots, X_n)$ un estimador de $g(\theta)$. Si se tiene la siguiente factorización*

$$\frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} f(x_i, \theta) = K(\theta)(t(x_1, \cdots, x_n) - g(\theta)),$$

entonces la varianza de T es igual a la cota de Cramer Rao.

Demostración. La condición

$$\frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} f(x_i, \theta) = K(\theta)(t(x_1, \cdots, x_n) - g(\theta))$$

es equivalente a

$$\frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} f(x_i, \theta) = \alpha t(x_1, \cdots, x_n) + \beta$$

para algunos constantes α y β que no depende de la muestra aleatoria, sino únicamente del parámetro θ. Y en consecuencia, tenemos que

$$\frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} f(X_i, \theta) = \alpha T + \beta.$$

Ahora recordando propiedades del coeficiente de correlación, tenemos que

$$Corr\left(T, \frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} f(X_i, \theta)\right) = 1,$$

y retomando la demostración del Teorema 2.3.12, tenemos que la varianza de T es igual a la cota de Cramer Rao.

□

Ahora, volvemos al Ejemplo 2.4.8 para ilustrar la utilidad del anterior resultado.

Tenemos:

$$\begin{aligned}\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n}f(x_i,\theta) &= \frac{\partial}{\partial\theta}\ln\left\{\frac{e^{-n\theta}\theta^{\sum x_i}}{\prod x_i!}\prod I_{\{0,1,\cdots\}}(x_i)\right\}\\ &= \frac{\partial}{\partial\theta}\left\{-n\theta+\sum_{i=1}^{n}x_i\ln\theta\right\}\\ &= -n+\frac{\sum_{i=1}^{n}x_i}{\theta}\\ &= \frac{-n\theta+n\bar{x}}{\theta}\\ &= \frac{n}{\theta}(\bar{x}-\theta),\end{aligned}$$

la anterior expresión es de la forma $K(\theta)(t(x_1,\cdots,x_n)-\theta)$ donde $K(\theta)=n/\theta$ y $t(x_1,\cdots,x_n)=\bar{x}$. Así se concluye que el estimador $\bar{X}$ tiene varianza igual a la cota de Cramer Rao, teniendo en cuenta que es también insesgado para θ, se puede concluir que $\bar{X}$ es UMVUE para θ.

Finalmente, resaltamos el procedimiento para comprobar que un estimador sea UMVUE. Primero se debe demostrar que el estimador es insesgado y luego comprobar que la varianza del estimador es igual a la cota de Cramer Rao. Para este último hay dos formas de hacerlo: (1) usar directamente la definición, en este caso se debe calcular la varianza del estimador y también calcular la cota de Cramer Rao; (2) usar el Resultado 2.4.14, en este caso no es necesario el cálculo de la varianza del estimador, ni la cota de Cramer Rao.

2.4.4 Completez

Otro concepto útil en la construcción de un buen estimador es el concepto de la completez. Este concepto, a diferencia de conceptos como sesgo, varianza o suficiencia, carece de una interpretación clara y de fácil entendimiento, pero resulta ser muy útil. Como se verá más adelante, este concepto nos permite encontrar los estimadores UMVUE.

Definición 2.4.9. *Dada una muestra aleatoria $X_1,\cdots,X_n$ con función de densidad $f(x_i,\theta)$, y T una estadística, se dice que T es completo para θ, si para cualquier función $g(\cdot)$, el hecho de $E(g(T))=0$ para todo θ, implica que $g(T)=0$.*

En algunos casos, no es muy complicado demostrar que un estimador es completo, como lo ilustra el siguiente ejemplo.

Ejemplo 2.4.11. *Dada una muestra aleatoria $X_1,\cdots,X_n$ una muestra aleatoria con distribución $Ber(p)$, entonces el estimador $\sum_{i=1}^{n}X_i$ es un estimador completo para p. Para ver eso, tomamos una función $g(\cdot)$ cualquiera, y supongamos que $E(g(\sum_{i=1}^{n}X_i))=0$, y veamos que $g(\sum_{i=1}^{n}X_i)=0$. Recordando que la distribución*

de $\sum_{i=1}^{n} X_i$ *tiene distribución* $Bin(n,p)$*, tenemos:*

$$\begin{aligned} 0 &= E(g(\sum_{i=1}^{n} X_i)) \\ &= \sum_{i=0}^{n} g(i)\binom{n}{i}p^i(1-p)^{n-i} \\ &= (1-p)^n \sum_{i=0}^{n} \binom{n}{i} g(i)(\frac{p}{1-p})^i, \end{aligned}$$

de donde se tiene que $\sum_{i=0}^{n} \binom{n}{i} g(i)(\frac{p}{1-p})^i = 0$*, nótese que el lado izquierdo de la igualdad es un polinomio en* $\frac{p}{1-p}$ *de grado* n*. Recordando que un polinomio es igual a 0 si cada coeficiente del polinomio es 0, entonces se puede concluir que* $\binom{n}{i} g(i) = 0$ *para todo* $i = 0, \cdots, n$*, de donde se tiene que* $g(i) = 0$ *para* $i = 0, \cdots, n$*. Ahora la estadística* $\sum_{i=1}^{n} X_i$ *toma valores 0,* $\cdots$*,* n*, de donde se concluye finalmente que* $g(\sum_{i=1}^{n} X_i) = 0$*.*

En el anterior ejemplo, la muestra aleatoria proviene de una distribución discreta, cuando se trata de una distribución continua, la forma de proceder es diferente, como se ilustra en el siguiente ejemplo.

Ejemplo 2.4.12. *Sea* $X_1, \cdots, X_n$ *una muestra aleatoria con distribución uniforme sobre* $[0, \theta]$*, el estimador por el método de máxima verosimilitud es* $X_{(n)}$ *el máximo de la muestra, veamos que este estimador es completo. Para cualquier función* g*, tenemos*

$$\begin{aligned} 0 &= E(g(X_{(n)})) \\ &= \int_0^{\theta} g(x)\frac{nx^{n-1}}{\theta^n}dx, \end{aligned}$$

utilizando el teorema fundamental de cálculo, se tiene que

$$0 = \frac{ng(\theta)}{\theta}$$

para todo $\theta > 0$*, es decir,* $g(\theta) = 0$ *para todo* $\theta > 0$*, y como la estadística* $X_{(n)}$ *toma valores positivos, entonces* $g(X_{(n)}) = 0$*.*

Para distribuciones pertenecientes a la familia exponencial, es muy fácil encontrar una estadística completa. Para distribuciones en la familia exponencial uniparamétrica tenemos el siguiente resultado.

Resultado 2.4.15. *Dada una muestra aleatoria* $X_1, \cdots, X_n$ *proveniente de una distribución* $f(x, \theta)$ *perteneciente a la familia exponencial, es decir,*

$$f(x, \theta) = h(x)c(\theta)\exp\{d(\theta)T(x)\},$$

entonces la estadística $\sum_{i=1}^{n} T(X_i)$ *es una estadística completa para* θ*.*

La versión equivalente para distribuciones en la familia exponencial biparamétrica se da a continuación.

Resultado 2.4.16. *Dada una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una distribución $f(x_i, \theta_1, \theta_2)$ perteneciente a la familia exponencial biparamétrica de la forma*

$$f(x_i, \theta_1, \theta_2) = c(\boldsymbol{\theta})h(x)\exp\{d(\boldsymbol{\theta})'T(x)\},$$

donde $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $d(\boldsymbol{\theta}) = (d_1(\boldsymbol{\theta}), d_2(\boldsymbol{\theta}))'$ y $T(x) = (T_1(x), T_2(x))'$, entonces las estadística $\sum_{i=1}^n T_1(X_i)$ y $\sum_{i=1}^n T_2(X_i)$ son estadísticas completas para θ_1 y θ_2.

Los dos anteriores resultados, en conjunto con los resultados 2.4.6 y 2.4.7, nos permiten encontrar fácilmente estadísticas que sean a la vez suficientes y completas para distribuciones pertenecientes a la familia exponencial.

Una utilidad de las estadísticas completas es que, anteriormente se había visto que para un parámetro puede haber más de un estimador insesgado, pero cuando el estimador insesgado es función se combina con completez, sí se tiene la unicidad, como lo ilustra el siguiente resultado.

Resultado 2.4.17. *Dada una muestra aleatoria $X_1, \cdots, X_n$ una muestra aleatoria con parámetro θ, si T es una estadística completa, $g_1(T)$ y $g_2(T)$ son estimadores insesgados de θ, entonces $g_1(T) = g_2(T)$.*

Demostración. Por hipótesis, se tiene que $E(g_1(T)) = E(g_2(T)) = 0$, de donde, $E((g_1 - g_2)(T)) = 0$, por definición de estimador completo, se sigue que $(g_1 - g_2)(T) = 0$, es decir, $g_1(T) = g_2(T)$. □

En la demostración del Resultado 2.4.3, se mencionó el teorema de Basu. Este teorema necesita los conceptos de suficiencia y completez de una estadística y también el concepto de estadística auxiliar, esto es, estadística cuya distribución no depende del parámetro de la distribución. Dado este concepto, presentamos a continuación el teorema de Basu.

Resultado 2.4.18. *Si T es una estadística suficiente y completa, entonces $T(X)$ es independiente de toda estadística auxiliar.*

Demostración. La demostración para el caso discreto se encuentra en Casella & Berger (2002, p. 287). □

El Resultado 2.4.3 puede ser fácilmente demostrado usando el teorema de Basú, puesto que en primer lugar, en una muestra proveniente de la distribución $N(\mu, \sigma^2)$, $\bar{X}$ es una estadística suficiente y completa para μ. Por otro lado, $\dfrac{\sum(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$, de donde podemos tener que la distribución de S_n^2 no depende de μ, al igual que la distribución de S_{n-1}^2; de esta forma, podemos concluir que $\bar{X}$ y S_n^2 (o S_{n-1}^2) son independientes.

La mayor importancia de las propiedades de suficiencia y completez es que aparte de ser propiedades deseables para los estimadores, nos permiten construir estimadores UMVUE, como lo ilustra el siguiente resultado.

Resultado 2.4.19. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con parámetro θ, si T_1 es un estimador insesgado para una función del parámetro $g(\theta)$ y T_2 es suficiente y completo para θ, entonces se tiene que $E(T_1|T_2)$ es el único estimador UMVUE para $g(\theta)$.*

El anterior resultado nos brinda una herramienta poderosa para encontrar estimadores UMVUE, que consiste en los siguientes pasos:

(1) Encontrar un estimador insesgado para $g(\theta)$, la función del parámetro θ que se desea estimar. Nótese que cuando $g(\theta)$ es el media teórica, entonces un estimador insesgado es el promedio muestral.

(2) Encontrar un estimador suficiente y completo para θ, que para distribuciones pertenecientes a la familia exponencial resulta bastante útil.

Ahora aplicamos las anteriores herramientas a un problema de estimación.

Ejemplo 2.4.13. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con distribución $Pois(\theta)$, se quiere encontrar el estimador UMVUE para θ. Para eso, primero se debe encontrar un estimador insesgado para θ, que es $\bar{X}$ por ser θ el media teórica. En segundo lugar, la distribución Poisson pertenece a la familia exponencial con $T(x) = x$, entonces los Resultados 2.4.5 y 2.4.15 establecen que $\sum_{i=1}^n X_i$ es una estadística suficiente y completa para θ. De esta manera, el estimador UMVUE es $E(\bar{X}|\sum_{i=1}^n X_i) = \bar{X}$.*

Para distribuciones con dos parámetros, existe la siguiente generalización:

Resultado 2.4.20. *Dada una muestra aleatoria $X_1, \cdots, X_n$ una muestra aleatoria con parámetro $\boldsymbol{\theta} = (\theta_1, \theta_2)$, si T_1, T_2 son estimadores insesgados para θ_1, θ_2, y S_1, S_2 son suficientes y completos para θ_1, θ_2, entonces se tiene que $E(T_1, T_2|S_1, S_2)$ son UMVUE para θ_1 y θ_2, además se tiene la unicidad.*

Ejemplo 2.4.14. *Sea $X_1, \cdots, X_n$ una muestra aleatoria con distribución $N(\mu, \sigma^2)$, $\sum_{i=1}^n X_i$ y $\sum_{i=1}^n X_i^2$ son estimadores suficientes y completos para μ y σ^2, y también se vio anteriormente que $\bar{X}$ es insesgado para μ y S_{n-1}^2 insesgado para σ^2, entonces el resultado anterior indica que las estadísticas $E(\bar{X}, S_{n-1}^2|\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ son UMVUE único para μ y σ^2. Para calcular la esperanza condicional, se tiene en cuenta que en primer lugar $\bar{X}$ es función de $\sum_{i=1}^n X_i$ y $S_{n-1}^2 = (\sum_{i=1}^n X_i^2 - n\bar{X}^2)/(n-1)$ es función de $\sum_{i=1}^n X_i$ y $\sum_{i=1}^n X_i^2$, entonces se tiene que los UMVUE para μ y σ^2 son $\bar{X}$ y S_{n-1}^2.*

Nótese en primer lugar que por la forma como están definidos S_n^2 y S_{n-1}^2, se tiene que $Var(S_n^2) < Var(S_{n-1}^2)$, es decir, el estimador S_n^2 tiene una varianza más pequeña que la del estimador UMVUE; sin embargo, nótese que como estimador de σ^2, S_n^2 es sesgado. Lo anterior muestra el hecho de que puede un estimador UMVUE

tener varianza más pequeña sólo entre los estimadores insesgados; sin embargo, puede existir un estimador sesgado con varianza aún más pequeña que la de UMVUE.

También observemos que en el Ejemplo 2.4.9, se encontró que la matriz de información en la muestra acerca de los parámetros μ y σ^2 es

$$I_{X_1,\cdots,X_n}(\mu,\sigma^2)=\begin{pmatrix}\frac{n}{\sigma^2} & 0\\ 0 & \frac{n}{2\sigma^4}\end{pmatrix}$$

cuya inversa está dada por

$$I^{-1}_{X_1,\cdots,X_n}(\mu,\sigma^2)=\begin{pmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{2\sigma^4}{n}\end{pmatrix}$$

la cual corresponde a la cota de Cramer-Rao para el vector de estimadores $(\bar{X},S^2_{n-1})$.

Ahora, a pesar de que ya se comprobó que $(\bar{X},S^2_{n-1})$ es UMVUE para (μ,σ^2), la matriz de varianzas de $(\bar{X},S^2_{n-1})$ no es la cota de Cramer-Rao, puesto que

$$Var(S^2_{n-1})=\frac{\sigma^4}{(n-1)^2}Var\left(\frac{(n-1)S^2_{n-1}}{\sigma^2}\right)=\frac{\sigma^4}{(n-1)^2}2(n-1)=\frac{2\sigma^4}{n-1}.$$

la cual es más grande que el elemento $2\sigma^4/n$ en la cota de Cramer-Rao.

Hasta este punto del libro, se han expuesto varios aspectos que se deben tener en cuenta al momento de escoger un estimador; más aún, al momento de escoger uno de varios posibles estimadores. Y al momento de realizar comparaciones, es necesario calcular la esperanza y la varianza de los estimadores, y estos cálculos pueden ser muy difíciles cuando los estimadores toman formas complicadas; por ejemplo, cuando se trata de una muestra proveniente de $Unif[\theta_1,\theta_2]$ se vio anteriormente que los estimadores de momentos de los parámetros son $\bar{X}-\sqrt{3}S_n$ y $\bar{X}+\sqrt{3}S_n$, respectivamente, mientras que los estimadores de momentos son las estadísticas de orden $X_{(1)}$ y $X_{(n)}$. Para tener una idea en este caso de cuál método arrojó mejores estimadores, deberíamos calcular el sesgo y la varianzas de estos estimadores, pero es claro que no es nada trivial calcular estas cantidades para los estimadores de momentos. En casos como el anterior, no tenemos otra alternativa que utilizar las simulaciones. Las simulaciones se pueden llevar a cabo simulando muestras de $Unif(\theta_1,\theta_2)$, y en cada muestra simulada se calculan los estimadores de momentos y de máxima verosimilitud.

En la Figura 2.16 se muestran los resultados de estos estimadores para muestras provenientes de $Unif(3,5)$. Podemos observar que los estimadores de máxima verosimilitud parecen sobreestimar a θ_1 y subestimar θ_2 en muestras pequeñas; en muestras grandes, el efecto de sobreestimación o subestimación tiende a desaparecer. Por otra parte, con respecto a los estimadores de momentos, en general no se detectan problemas de sobreestimación o subestimación, aunque la varianza es considerablemente mayor que los estimadores de máxima verosimilitud, aún para muestras

grandes. Por consiguiente, se recomienda usar los estimadores de máxima verosimilitud en muestras provenientes de la distribución $Unif(\theta_1, \theta_2)$.

El lector puede notar que la teoría expuesta anteriormente es válida para cuando se quiere estimar una función del parámetro $g(\theta)$. Sin embargo, poco se ha discutido acerca de la evaluación de estos estimadores. La razón es que, en primer lugar, en general no es fácil determinar el sesgo de estos estimadores. Sobre este aspecto, se considera el Ejemplo 2.3.3, donde en un conjunto de ensayos del tipo Bernoulli, el parámetro de interés es la probabilidad de éxito p que se estima mediante $\bar{X}$, pero si la cantidad que se desea estimar se cambia a $3p(1-p)^4$, aunque es fácil encontrar el estimador de máxima verosimilitud $3\bar{X}(1-\bar{X})^4$, no es fácil determinar si este es insesgado para $3p(1-p)^4$, y por consiguiente tampoco se puede determinar si este es el estimador UMVUE para $3p(1-p)^4$. Lo anterior indica que en general no se puede establecer que para cualquier función g, si T es un estimador insesgado para θ, entonces $g(T)$ es insesgado para $g(\theta)$, esto es, no se tiene en general que $E(g(T)) = g(\theta)$ suponiendo que $E(T) = \theta$.

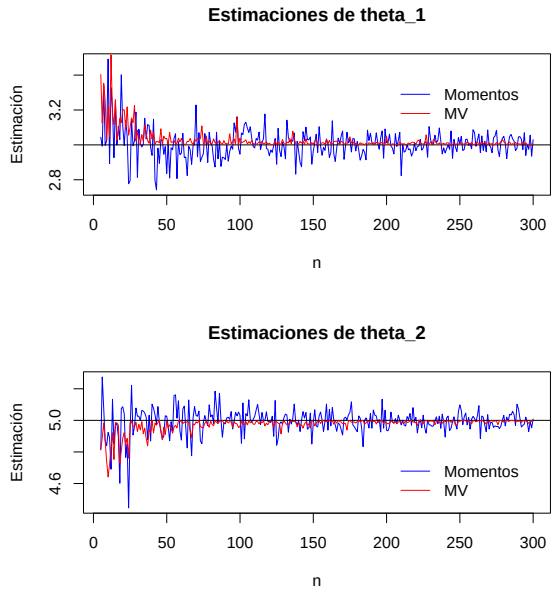

Figura 2.16: *Comparación empírico entre los estimadores de máxima verosimilitud y de momentos en una muestra proveniente de* $Unif(\theta_1, \theta_2)$

Es claro que cuando la función g es una función lineal, esto es $g(\theta) = a\theta + b$, para constantes a y b, entonces $E(g(T)) = E(aT + b) = aE(T) + b$ y si T es insesgado para θ, $g(T)$ también lo es para $g(\theta)$. Cuando g toma otras formas, podemos calcular $E(g(T))$ e inclusive $Var(g(T))$ de manera aproximada usando el método de Delta presentado a continuación.

Resultado 2.4.21. *Dado T un estimador de θ, y g una función, entonces se tiene que*

$$E(g(T)) \approx g(\theta) + g'(\theta)E(T-\theta)$$

y

$$Var(g(T)) \approx (g'(\theta))^2 Var(T)$$

De esta forma, se tiene que cuando T es insesgado para θ, $E(g(T)) \approx g(\theta)$.

Demostración. Para cada valor t que toma el estimador T, se hará uso de la expansión de Taylor de primer orden para $g(t)$ alrededor del punto $t = \theta$, tenemos que $g(t) \approx g(\theta) + g'\theta(t-\theta)$. De esta forma, se tiene que $g(T) \approx g(\theta) + g'(\theta)(T-\theta)$. Tomando la esperanza, se tiene que $E(g(T)) \approx g(\theta) + g'(\theta)E(T-\theta)$. Es claro que cuando T es insesgado para θ, $E(g(T)) \approx g(\theta)$.

Por otro lado, tomando varianza en $g(T) \approx g(\theta) + g'(\theta)(T-\theta)$, se tiene que

$$\begin{aligned} Var(g(T)) &\approx (g'(\theta))^2 Var(T-\theta) \\ &= (g'(\theta))^2 Var(T) \end{aligned}$$

□

Del anterior resultado, se puede concluir que $g(T)$ es un estimador *aproximadamente* insesgado para $g(\theta)$. Pero no podemos saber qué tan buena (o qué tan mala) resulta esta aproximación, y tampoco saber, en general, si con el posible aumento del tamaño muestral, se puede hacer que $E(g(T))$ se acerque más a $g(\theta)$. Sin embargo, mediante el uso de simulaciones, podemos hacernos una idea de la bondad de esta aproximación con diferentes funciones g para diferentes tamaños muestrales n.

Suponga que en una muestra proveniente de la distribución exponencial con parámetro $\theta = 1$, se desea estimar $p_1 = Pr(X < 1)$ y $p_2 = Pr(1 < X < 2)$. Estas dos probabilidades se pueden expresar como funciones del parámetro como $p_1 = e^{-1/\theta}$ y $p_2 = e^{-1/\theta} - e^{-2/\theta}$, respectivamente. Y los estimadores de máxima verosimilitud de p_1 y p_2 son $T_1 = e^{-1/\bar{X}}$ y $T_2 = e^{-1/\bar{X}} - e^{-2/\bar{X}}$ respectivamente, para estudiar el sesgo de T_1 y T_2 como estimadores de p_1 y p_2. Se simulan 1000 muestras de tamaño 5, 10, 30, 50, 100, 500 de una distribución $Exp(1)$ y $Exp(5)$, y en cada muestra simulada se calculan los valores que toman T_1 y T_2. Y para cada n fijo se calcula el promedio de los 1000 valores de T_1 y T_2, éstos se pueden tomar como estimaciones de $E(T_1)$ y $E(T_2)$. Y por consiguiente, restando p_1 y p_2 a estas estimaciones, se pueden obtener estimaciones del sesgo de T_1 y T_2. En la Figura 2.17, se muestra el comportamiento de estos dos estimadores en término del sesgo. Podemos observar que en primer lugar, a medida que el tamaño muestral crece, el sesgo de ambos estimadores se acerca al valor 0, es decir, T_1 y T_2 parecen ser asintóticamente insesgados. Otro aspecto interesante es que la sobreestimación o la subestimación de los estimadores depende del valor del parámetro θ, pues cuando $\theta = 1$, T_2 siempre tuvo un sesgo estimado negativo, pero al cambiar el valor de θ a 5, se encuentra que ahora su sesgo estimado es siempre positivo.

Repitamos la simulación para muestras provenientes de la distribución Bernoulli con parámetro p. Suponga que se desea estimar la varianza teórica dada por $p(1-$

p), la probabilidad de obtener un éxito en cuatro ensayos dada por $4p(1-p)^3$ y la probabilidad de obtener más de un éxito en cuatro ensayos dada por $1-(1-p)^4$. Como el estimador de máxima verosimilitud de p es $\bar{X}$, tenemos que los estimadores de máxima verosimilitud de estas tres cantidades que se desean estimar son $T_1 = \bar{X}(1-\bar{X})$, $T_2 = 4\bar{X}(1-\bar{X})^3$ y $T_3 = 1-(1-\bar{X})^4$, respectivamente. Para estudiar el sesgo de estos tres estimadores, se simulan 1000 muestras de tamaño 5, 10, 30, 50, 100, 500 de una distribución $Ber(0.3)$ y $Ber(0.7)$, y en cada muestra simulada se calculan los valores que toman T_1, T_2 y T_3. Y se obtienen las estimaciones de los sesgos de los tres estimadores análogamente al caso de la distribución exponencial. En la Figura 2.18, se muestra el comportamiento de estos dos estimadores en término del sesgo, y podemos observar comportamientos análogos al caso de la distribución exponencial, esto es, al incrementar n los estimadores son asintóticamente insesgados, y la sobreestimación o la subestimación puede depender del valor del parámetro p.

2.4.5 Consistencia

El concepto de consistencia, al igual que el concepto del insesgamiento asintótico, es una propiedad asintótica, es decir, se considera el caso cuando el tamaño muestral $n \to \infty$. El concepto de insesgamiento asintótico establece que cuando el tamaño muestral es suficientemente grande, los valores que toma el estimador están alrededor de la cantidad que se desea estimar, $g(\theta)$; mientras que el concepto de consistencia va un paso más allá, y estudia características del estimador visto como una variable aleatoria. Más específicamente, estudia la probabilidad de que el estimador esté cercano de $g(\theta)$. La definición de consistencia se da a continuación.

Definición 2.4.10. *Dada una muestra aleatoria con parámetro desconocido θ, y T un estimador de $g(\theta)$ para alguna función g, se dice que T es un estimador consistente si T converge en probabilidad a $g(\theta)$ ($T_n \xrightarrow{P} g(\theta)$), esto es, para todo $\varepsilon > 0$, se tiene que*

$$\lim_{n\to\infty} Pr(|T - g(\theta)| > \varepsilon) = 0.$$

Como se había mencionado anteriormente, es natural estimar la media teórica μ utilizando el promedio muestral $\bar{X}$, y en distribuciones como la Bernoulli, Poisson, exponencial y normal $\bar{X}$ es UMVUE para μ, es decir, entre todos los estimadores insesgados, $\bar{X}$ tiene menor varianza. Adicionalmente, la siguiente ley débil de los grandes números nos ilustra que además es un estimador consistente.

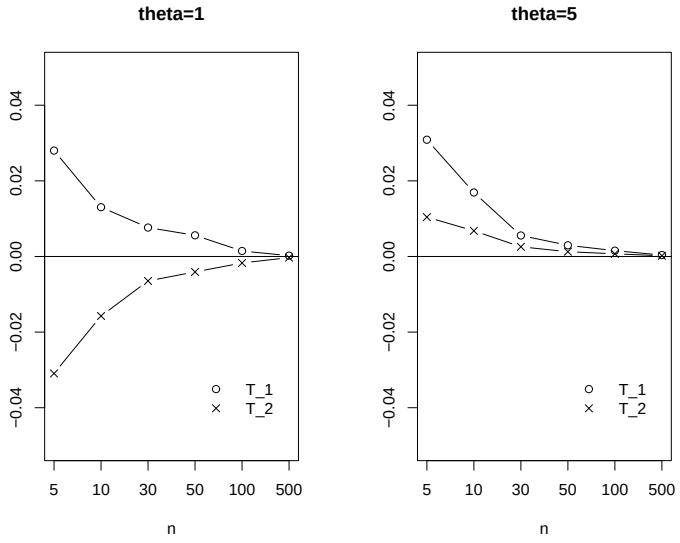

Figura 2.17: *Sesgo estimado de los estimadores T_1 y T_2 para diferentes tamaños de muestra.*

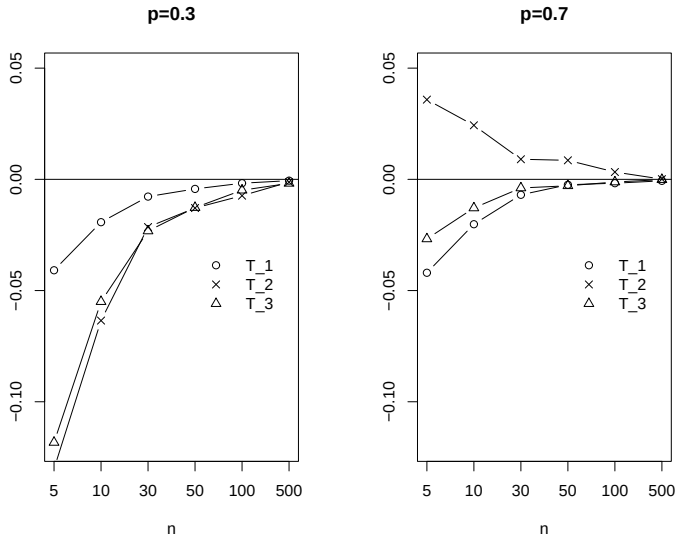

Figura 2.18: *Sesgo estimado de los estimadores T_1, T_2 y T_3 para diferentes tamaños de muestra.*

Resultado 2.4.22. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con media común μ y $E(|X_i|) < \infty$ para todo i, y se define $\bar{X}_n = \sum_{i=1}^{n} X_i/n$, entonces $\bar{X}_n$ convergen en probabilidad a μ.*

Demostración. Esta versión de la ley débil de los grandes números se conoce con el nombre de Khintchin, y no se requiere de la existencia de la varianza teórica. Cuando ésta existe, el resultado se puede mostrar fácilmente aplicando la desigualdad de Chebychev. En el caso general sin ningún supuesto acerca de la varianza teórica, la prueba es más complicada, el lector interesado puede consultar Resnick (1999, p. 205). □

Dado que el concepto de la convergencia está directamente relacionado con la convergencia en probabilidades, las propiedades deseables de esta convergencia nos permiten obtener conclusiones interesantes acerca de la consistencia. Específicamente, se conoce en la teoría estadística que si $X_n \xrightarrow{P} a$ y g es una función continua en a, entonces $g(X_n) \xrightarrow{P} g(a)$. De esta forma, no sólo podemos afirmar que en una muestra del tipo Bernoulli, $\bar{X}$ es consistente para estimar la probabilidad de éxito p, sino también $\bar{X}(1-\bar{X})$ es consistente para la varianza teórica $p(1-p)$, y los estimadores mencionados en el Ejemplo 2.3.3 también lo son (aunque no podemos afirmar lo mismo acerca de las propiedades de insesgamiento y varianza mínima).

El concepto de consistencia, por su definición, describe la propiedad de que a medida que el tamaño muestral crece, el estimador se hace cada vez más cercano a la cantidad que se desea estimar $g(\theta)$. Este concepto está ligado, naturalmente, con la varianza estimador, pues entre más pequeña sea ésta, más concentrados están los valores que toma el estimador, y si adicionalmente el estimador es insesgado, los valores que toman el estimador deben estar muy cercanos a $g(\theta)$. El siguiente resultado confirma esta relación entre los conceptos del insesgamiento, varianza pequeña y consistencia.

Resultado 2.4.23. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con parámetro desconocido θ, si T es un estimador asintóticamente insesgado y su $Var(T) \to 0$ cuando $n \to \infty$, entonces T es consistente.*

Demostración. Consideramos el error cuadrático medio de T, este se puede escribir en término del sesgo y la varianza de T de la forma $ECM(T) = B_T^2 + Var(T)$. Dada la hipótesis del resultado, tenemos que

$$\lim_{n\to\infty} ECM(T) = 0.$$

Ahora,

$$\begin{aligned}
\lim_{n\to\infty} Pr(|T - g(\theta)| > \varepsilon) &\leq \lim_{n\to\infty} \frac{1}{\varepsilon^2} E[(T - g(\theta))^2] \qquad \text{Desigualdad de Chebychev} \\
&= \lim_{n\to\infty} \frac{1}{\varepsilon^2} ECM(T) \\
&= 0
\end{aligned}$$

y concluimos que T es consistente. □

Nótese que usando el anterior resultado junto con el Resultado 2.4.1, es otra forma de probar que en muestras provenientes de cualquier distribución, se tiene que $\bar{X}$ es un estimador consistente para μ, aunque en este caso, se exige que la varianza teórica sea finita para la prueba.

Dada la anterior relación entre la consistencia y la varianza de un estimador, podemos establecer alguna relación entre los estimadores consistentes y los UMVUE. Cuando un estimador T es insesgado para $g(\theta)$, tenemos que la cota de Cramer-Rao dada en (2.4.9), donde $I_X(\theta)$ es la información contenida en una variable, y por consiguiente no depende de n. De esta forma, si un estimador insesgado tiene varianza igual a la cota de Cramer-Rao, su varianza converge a 0 y por el Resultado 2.4.23, podemos afirmar que también es consistente. Pero si podemos afirmar la igualdad entre la cota de Cramer-Rao y la varianza del estimador, tampoco podemos concluir la consistencia.

Para finalizar el concepto de consistencia, el lector puede visualizar este concepto en la Figura 2.14, donde se observa claramente que cuando n crece, los valores de $\bar{x}$ se acercan cada vez más a μ.

2.5 Comparación empírica de algunas propiedades

Hasta este punto, hemos introducido varias propiedades de un estimador para tener en cuenta a la hora de escoger un buen estimador. Sin embargo, muchos estadísticos y/o profesores han manifestado la dificultad de ≪traducir≫ estos conceptos abstractos en la práctica. Una manera de solucionar este problema es ilustrar gráficamente estos conceptos como lo hemos venido haciendo en este texto. Aquí ilustramos otros aspectos y conceptos adicionales que pueden ayudar a los lectores a entender mejor estos conceptos abstractos.

En primer lugar, suponga que tenemos dos estimadores para un mismo parámetro tales que uno es insesgado y el otro es levemente sesgado. ¿Cuál estimador se debe escoger? Más específicamente, suponga que se desean comparar dos estimadores de la varianza de una muestra aleatoria de variables con distribución Normal de media 5 y varianza 81; por ejemplo, el estimador de máxima verosimilitud $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$ y el clásico estimador insesgado $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$.

La propiedad del insesgamiento está relacionada directamente con la esperanza matemática del estimador, en términos de su distribución de muestreo, y por consiguiente, para un estimador sesgado es frecuente que ese sesgo dependa del parámetro teórico, por lo que no se puede cuantificar en la práctica. Sin embargo, podemos usar simulaciones para tener una estimación de su esperanza. La Figura 2.19 nos muestra una estimación de la esperanza de los estimadores S_n^2 y S_{n-1}^2. Esta figura fue realizada de la siguiente manera: para un tamaño de muestra fijo $n = 10$ generada de la distribución $N(5, 81)$, se estima el parámetro de interés con los dos estimadores. Ahora, este proceso se realiza 1, 10, 100, 1000, 10000, 100000 y 1000000 veces. En cada repetición se calcula el promedio de las estimaciones y se grafica. Nótese que en un momento dado cada una de las dos líneas parece converger a un valor. Por supuesto, el estimador insesgado S_n^2 converge a 81, el verdadero valor, mientras que

el sesgado converge a un valor inferior, lo cual coincide con el hecho de que S^2_{n-1} subestima a la varianza teórica.

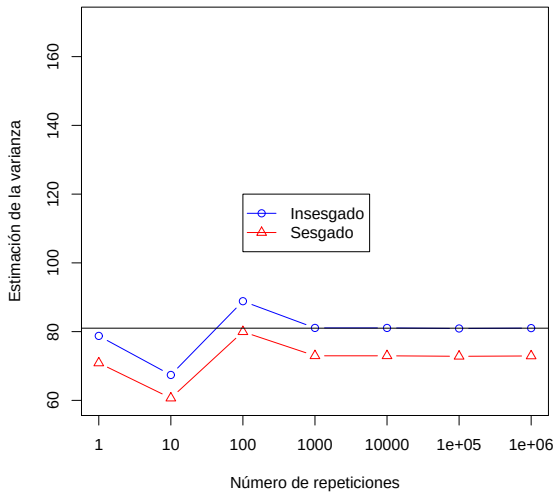

Figura 2.19: *Esperanza simulada de los estimadores* S^2_n *y* S^2_{n-1} *con* $n = 10$.

```
> # PROPIEDAD DE INSESGAMIENTO

>  N<-c(1,10,100,1000,10000,100000,1000000)
>  n<-10
>  meanvar1<-rep(NA,length(N))
>  meanvar2<-rep(NA,length(N))
>  for(k in 1:length(N))
+  {
+  var1<-rep(NA,N[k])
+  var2<-rep(NA,N[k])
+  for(l in 1:(N[k])){
+  data<-rnorm(n,5,9)
+  var1[l]<-var(data)
+  var2[l]<-(n-1)*var(data)/n
+  }
+  meanvar1[k]<-mean(var1)
+  meanvar2[k]<-mean(var2)
+  }
>  meanvar1
[1] 78.75174 67.39362 88.84951 81.07334 81.07010 80.91384 81.01865
```

```
>  meanvar2
[1] 70.87657 60.65425 79.96456 72.96601 72.96309 72.82246 72.91679

>  plot(meanvar1,type="b", col=4,ylim=c(60,170),xlab="Número de
+  repeticiones", ylab="Estimación de la varianza", xaxt="n")
>  lines(meanvar2,type="b", col=2, pch=2)
>  abline(h=81)
>  axis(1, 1:length(N), N)
>  legend(3,120,c("Insesgado","Sesgado"), col=c(4,2), lty=c(1,1),
+  pch=c(1,2))
```

Ahora, podemos preguntarnos si los resultados de la Figura 2.19 son una prueba fehaciente de que el estimador insesgado resulta mejor que su competidor. Sin embargo, el hecho de que un estimador sea insesgado indica que ≪en promedio≫ sus valores son iguales al parámetro, pero no garantiza que las estimaciones individuales sean buenas. Es posible que un estimador insesgado arroje estimaciones individuales ridículas pero en promedio sea igual al parámetro. Así que la propiedad de insesgamiento no basta para escoger un estimador.

Existe otro criterio que podemos tener en cuenta para escoger entre dos estimadores de un mismo parámetroθ. Al denotar estos dos estimadores por T_1 y T_2, se dice que T_1 domina a T_2 si el ECM de T_1 siempre es menor o igual al ECM de T_2 para todo θ, esto es, $E[(T_1-\theta)^2] \leq E[(T_1-\theta)^2]$. Y como consecuencia se define la eficiencia relativa como la cociente entre los ECM, dada por

$$e(T_1,T_2) = \frac{E[(T_1-\theta)^2]}{E[(T_2-\theta)^2]}.$$

De esta forma, un valor de $e(T_1,T_2)$ inferior a 1 para todo θ muestra la superioridad de T_1 comparado con T_2. Ahora, análogo al sesgo o la esperanza de un estimador, el ECM y la eficiencia tampoco se pueden conocer en la práctica pues también dependen del parámetro teórico. Podemos adoptar un procedimiento similar al presentado anteriormente para tener una estimación de ECM de los dos estimadores. El siguiente código nos permite calcular ECM de los dos estimadores S_n^2 y S_{n-1}^2.

```
> # PROPIEDAD DE EFICIENCIA

>  N<-c(2,10,100,1000,10000,100000,1000000)
>  n<-10
>  msevar1<-rep(NA,length(N))
>  msevar2<-rep(NA,length(N))
>  for(k in 1:length(N))
+  {
+  var1<-rep(NA,N[k])
+  var2<-rep(NA,N[k])
+  for(l in 1:(N[k])){
+  data<-rnorm(n,5,9)
```

```
+  var1[l]<-var(data)
+  var2[l]<-(n-1)*var(data)/n
+  }
+  msevar1[k]<-var(var1)
+  msevar2[k]<-var(var2)+(mean(var1)-81)^2
+  }
> msevar1
[1]  965.0842 1529.4889 1460.5530 1651.2588 1420.0702 1461.4549 1458.1439
>  msevar2
[1] 1194.755 1244.103 1183.322 1337.701 1150.396 1183.819 1181.098

>  plot(msevar1,type="b", col=4,ylim=c(1000,3100),xlab="Número de
+  repeticiones", ylab="Eficiencia de las estimaciones", xaxt="n")
>  lines(msevar2,type="b", col=2, pch=2)
>  axis(1, 1:length(N), N)
>  legend(3,2500,c("Insesgado","Sesgado"), col=c(4,2), lty=c(1,1),
+  pch=c(1,2))
```

El anterior código arroja la Figura 2.20 donde se aprecia que el ECM del estimador insesgado está alrededor de 1500, siendo más alto que el ECM del estimador sesgado, que se encuentra alrededor de 1200. Las anteriores cantidades se pueden calcular teóricamente: para el estimador insesgado, resulta ser igual a 1458 y para el sesgado resulta ser 1246. Y podemos concluir que no siempre es mejor un estimador insesgado frente a un sesgado, puesto que el ECM del estimador sesgado puede ser menor.

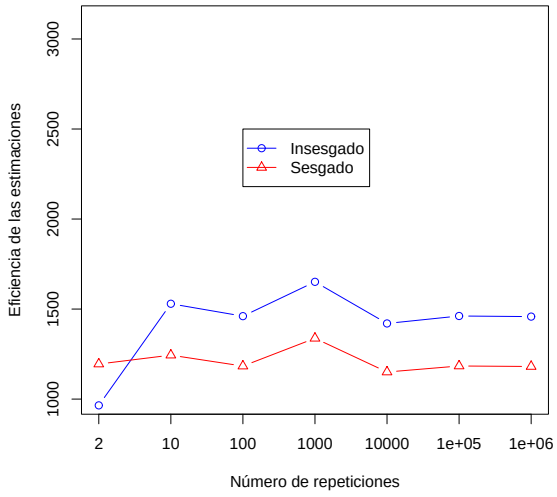

Figura 2.20: *ECM simulado de los estimadores* S_n^2 *y* S_{n-1}^2 *con* $n = 10$.

2.6 Ejercicios

2.1 En una muestra aleatoria $X_1, \cdots, X_n$ con función de densidad y de distribución f_X y F_X, respectivamente, demuestre que la función de densidad y de distribución del máximo $X_{(n)}$ están dadas por (2.2.2) y (2.2.3).

2.2 En la literatura estadística también es común definir la distribución exponencial con la siguiente función de densidad

$$f_X(x) = \lambda e^{-\lambda x} I_{(0,\infty)}(x) \tag{2.6.1}$$

Se puede ver que la anterior función de densidad utiliza la reparametrización $\lambda = 1/\theta$ en la función (1.1.10). Demuestre que en una muestra aleatoria con la función de densidad (2.6.1), $\hat{\lambda}_{MV} = 1/\bar{X}$.

2.3 En efecto de la estimación, observar una muestra aleatoria $X_1, \cdots, X_n$ de tamaño n proveniente de una distribución $Ber(p)$ equivale a observar el valor de la variable aleatoria $S = \sum X_i$ la cual tiene distribución $Bin(n, p)$ (esto es, S es una estadística suficiente).

(a) Demuestre que en la muestra $X_1, \cdots, X_n$, $\hat{p}_{MV} = \hat{p}_{mom} = \bar{X}$,

(b) Demuestre que cuando se usa S para estimar p, $\hat{p}_{MV} = S/n = \bar{X}$,

(c) Demuestre que S es suficiente.

2.4 Un almacén de ropas femeninas, después de la navidad, lanza la promoción del descuento de hasta 60 % en todo el almacén. El gerente desea conocer qué tan efectiva es la promoción; para ello, él tuvo en cuenta que en un determinado día entraron al almacén 40 clientes, y 25 de ellos hicieron alguna compra. Cómo puede estimar la probabilidad de que

(a) ¿Un cliente realice alguna compra?

(b) ¿Ninguna venta sea exitosa en cinco clientes consecutivos?

2.5 Considere una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una distribución $N(\mu, \sigma^2)$. Demuestre que

(a) El estimador de máxima verosimilitud de μ siempre es $\bar{X}$ sin importar si σ^2 es conocida o no, esto es, demuestre que el estimador de máxima verosimilitud de μ cuando $\sigma^2 = \sigma_0^2$ es $\bar{X}$

(b) El estimador de máxima verosimilitud de σ^2 puede variar dependiendo si μ es conocida o no, esto es, demuestre que el estimador de máxima verosimilitud de σ^2 cuando $\mu = \mu_0$ es $\sum_{i=1}^{n}(X_i - \mu_0)^2/n$.

2.6 Sea $X_1, \cdots, X_n$ una muestra aleatoria proveniente de distribución exponencial con media θ,

(a) Encuentre el estimador de máxima verosimilitud de θ.

(b) Encuentre el estimador de máxima verosimilitud de $Pr(X > 1)$, donde X tiene la misma distribución exponencial.

(c) Encuentre dos estimadores de momentos de θ.

(d) Encuentre un estimador de momentos de $Pr(X < 2)$, donde X tiene la misma distribución exponencial.

2.7 Se desea conocer el funcionamiento de un cierto dispositivo electrónico fabricado por una empresa, se seleccionan 20 dispositivos terminados al final de la línea de producción, y se ponen en funcionamiento. La vida útil de los dispositivos (en horas) fue: 6, 23, 1, 38, 43, 149, 2, 32, 41, 23, 10, 47, 46, 111, 43, 30, 21, 3, 96 y 53.

(a) Elabore una gráfica QQ plot para verificar si los datos provienen de una distribución exponencial.

(b) ¿Cómo se puede obtener un estimativo de la vida útil del dispositivo electrónico basándose en la muestra seleccionada?

(c) ¿Cuál es la probabilidad estimada de que un dispositivo funcione por más de 55 horas?

(d) Suponga que un cliente compra dos de esos dispositivos, y ambos se dañaron antes de las 30 horas, ¿cuál es la probabilidad de que esto ocurra?

2.8 Dentro del contexto del Ejemplo 2.3.8,

(a) Grafique la función $L(N)$ para $R = 8$, $n = 5$ y $x = 4$, y verifique en la gráfica que $L(N)$ tiene un máximo en $[Rn/x]$.

(b) Considerando a (2.3.4) como una función de R, grafique la función $L(R)$ para $N = 12$, $n = 8$, $x = 4$. Verifique en la gráfica que $L(R)$ tiene un máximo en $[x(N + 1)/n]$

(c) Grafique la función $L(R)$ para $N = 11$, $n = 8$, $x = 4$. Verifique en la gráfica que $L(R)$ tiene un máximo en $x(N + 1)/n - 1$ y $x(N + 1)/n$.

2.9 En una época de brote de una enfermedad respiratoria, se quiere conocer el número de colegios en una determinada ciudad que cuenta con condiciones sanitarias adecuadas. Suponga que la ciudad tiene 1325 colegios, y 86 de los 1400 colegios seleccionados al azar tienen buenas condiciones sanitarias. Basado en lo anterior, ¿cuántos colegios de esta ciudad se estima que tienen buenas condiciones sanitarias, y cuántos no lo hacen? Y ¿cuál es el porcentaje de colegios en la ciudad que tienen buenas condiciones sanitarias?

2.10 El estimador de máxima verosimilitud puede no ser único, como lo ilustra esta situación: suponga que $X_1, \cdots, X_n$ constituyen una muestra aleatoria proveniente de la distribución uniforme continua sobre el intervalo $[\theta - 0.5, \theta + 0.5]$, demuestre que cualquier estadística T con $X_{(n)} - 0.5 \leq T \leq X_{(1)} + 0.5$ es un estimador de máxima verosimilitud de θ.

	Distancia recorrida (en Km) por galón
Marca A	39.4, 41.1, 39.5, 40.0, 43.7, 46.0, 43.5, 42.1
Marca B	52.6, 49.4, 49.4, 46.4, 51.2, 49.2, 55.0, 53.6, 55.7, 57.4

Tabla 2.3: *Datos del Ejercicio 2.14.*

2.11 Dada una muestra aleatoria proveniente de una distribución $Unif(-\theta, \theta)$, demuestre que el estimador de máxima verosimilitud de θ está dado por $\hat{\theta}_{MV} = \text{máx}\{-X_{(1)}, X_{(n)}\}$.

2.12 Suponga que un sistema masivo de transporte urbano comienza a funcionar desde las 6 am, y la hora de llegada de la ruta A1 a la primera parada de buses puede ser cualquier hora a partir de las 6 am. Para tener un control acerca de la calidad de servicio de este sistema, se observa el tiempo de llegada de esta ruta en 8 días. Estos tiempos son 6:01 am, 6:07 am, 6:03 am, 6:07 am, 6:10 am, 6:11 am, 6:05 am y 6:03 am. Basado en estas observaciones

(a) Encuentre una estimación para la hora máxima de llegada de esta ruta.

(b) Encuentre una estimación para la probabilidad de que en un día determinado, la ruta llegue antes de las 6:05 am.

2.13 Dada una muestra aleatoria $X_1, \cdots, X_n$ con distribución $U[\theta_1, \theta_2]$,

(a) Demuestre que los estimadores de máxima verosimilitud de θ_1 y θ_2 son $X_{(1)}$ y $X_{(n)}$, respectivamente.

(b) Calcule el ECM de los estimadores de máxima verosimilitud y de momentos, y concluya cuáles son mejores en términos del menor ECM.

2.14 Suponga que se quiere comparar dos marcas de carros con respecto al rendimiento en términos de la distancia recorrida por galón de dos marcas de automóviles en referencias con especificaciones similares bajo circunstancias similares con respecto a la carretera, clima y demás condiciones controlables por los técnicos y expertos automovilísticos. Los datos se muestran en la Tabla 2.3

(a) Elabore la gráfica QQ para cada uno de los dos conjuntos datos y verifique que la distribución normal puede ser apropiada para describirlos.

(b) Estime el número de kilómetros recorridos por galón de gasolina y la respectiva desviación estándar en cada una de las dos marcas.

(c) Estime el coeficiente de variación para cada número de kilómetros recorridos por galón de gasolina en cada una de las dos marcas

(d) Estime el porcentaje de automóviles que recorren más de 50 kilómetros por galón para cada una de las dos marcas.

(e) Dado lo anterior, compare las dos marcas en términos de rendimiento y de estabilidad.

(f) Si suponemos que los automóviles de las dos marcas son igualmente estables en términos del rendimiento de gasolina, estime una medida que describa la diferencia promedio entre los automóviles de las dos marcas en términos del rendimiento de gasolina.

2.15 Sea $X_1, \cdots, X_n$ una muestra aleatoria con distribución $Exp(\theta)$

(a) Demuestre que las estadísticas $\sum_{i=1}^{n} X_i$ y $\bar{X}$ son suficientes para θ.

(b) Encuentre la información de Fisher contenida en la muestra acerca de θ.

2.16 Sea $X_1, \cdots, X_n$ una muestra aleatoria con distribución $Pois(\theta)$, un estimador suficiente para θ es $T = \sum_{i=1}^{n} X_i$. Compruebe que $I_T(\theta) = I_{X_1,\cdots,X_n}(\theta)$ usando la definición.

2.17 En una muestra aleatoria $X_1, \cdots, X_n$ donde $\boldsymbol{\theta}$ denota el vector de parámetros desconocidos, demostrar que $I_{X_1,\cdots,X_n}(\boldsymbol{\theta}) = nI_X(\boldsymbol{\theta})$ donde X tiene la misma distribución que las variables $X_1, \cdots, X_n$.

2.18 Se ha visto que en una muestra aleatoria proveniente de una distribución $Pois(\theta)$, el estimador de máxima verosimilitud y el estimador de momentos es el mismo: $\theta_{MV} = \theta_{mom} = \bar{X}$, ¿Este estimador es $UMVUE$ para θ?

2.19 Dada $X \sim Bin(n, p)$, con n conocido y p desconocido

(a) Encuentre el estimador de máxima verosimilitud de p.

(b) Encuentre el estimador de momentos de p.

(c) ¿Son iguales los dos estimadores hallados anteriormente? En caso afirmativo, comente sobre el desempeño del estimador. En caso negativo, ¿cuál estimador es mejor?

2.20 Considere una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una distribución Gamma con parámetro de forma k fijo y parámetro de escala θ desconocido.

(a) Calcular la información de Fisher contenida en la muestra aleatoria acerca del parámetro θ.

(b) Encuentra el estimador de máxima verosimilitud para θ.

(c) Encuentra el estimador de momentos para θ.

(d) ¿Son iguales los dos estimadores hallados anteriormente? En caso afirmativo, comente sobre el desempeño del estimador. En caso negativo, ¿cuál estimador es mejor?

2.21 Sea $X_1, \cdots, X_n$ una muestra aleatoria proveniente de $N(\mu_0, \sigma^2)$ con μ_0 conocido, se ha visto en el Ejercicio 2.5 que ahora el estimador de máxima verosimilitud de σ^2 es $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2$, mientras que cuando μ es desconocido el estimador de máxima verosimilitud de σ^2 es $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Compare los dos estimadores en término de sesgo y varianza y decida cuál es mejor (*Ayuda: usar* $\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_n$).

2.22 La función de densidad de la distribución Pareto con parámetro θ está dada por:

$$f(x,\theta) = \theta c^{\theta} x^{-(\theta+1)},$$

para $x \geq c$, $c > 0$ y $\theta > 0$. Encuentre los estimadores de máxima verosimilitud y de momentos para el parámetro θ.

2.23 La función de densidad de la distribución Rayleigh está dada por:

$$f(x,\theta) = \frac{x}{\theta^2} \exp\left\{-\frac{x^2}{2\theta^2}\right\},$$

para $x > 0$ y $\theta > 0$. Encuentre el estimador de máxima verosimilitud de θ.

2.24 La función de densidad de la distribución Weibull está dada por:

$$f(x,\theta) = \theta c x^{c-1} \exp\{-\theta x^c\},$$

para $x \geq 0$, $c > 0$ y $\theta > 0$. Encuentre el estimador de máxima verosimilitud de θ.

2.25 Sea $X_1, \cdots, X_{n_X}$ y $Y_1, \cdots, Y_{n_Y}$ dos muestras aleatorias independientes provenientes de una distribución $N(\mu_X, \sigma^2)$ y $N(\mu_Y, \sigma^2)$ respectivamente. Demuestre que el estimador de máxima verosimilitud de μ_X, μ_Y y σ^2 es $\bar{X}$, $\bar{Y}$ y $[\sum_{i=1}^{n_X}(X_i - \bar{X}_2)^2 + \sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2]/(n_X + n_Y)$.

2.26 Dada una muestra aleatoria proveniente de una distribución con parámetro desconocido θ, si T es un estimador insesgado de θ, y $g(\cdot)$ es una función, entonces en general no se tiene que $g(T)$ es un estimador insesgado para $g(\theta)$. Lo anterior se ilustra con la siguiente situación: X es una variable aleatoria con distribución $Bin(n,p)$, demuestra que

(a) X/n es un estimador insesgado para p.

(b) $(X/n)(1 - X/n)$ no es un estimador insesgado para $p(1-p)$.

2.27 En general, el estimador de máxima verosimilitud y el estimador de momentos para los parámetros en una distribución uniforme no son iguales. Considere $X_1, \cdots, X_n$ una muestra aleatoria con distribución $Unif[\theta_1, \theta_2]$. Encuentre los estimadores de MV y de momentos de θ_1 y θ_2.

2.28 Dada una muestra aleatoria $X_1, \cdots, X_n$ con distribución Gamma con parámetro de forma k y parámetro de escala θ, encuentre una estadística suficiente y completa para el vector de parámetros $\theta = (k, \theta)'$.

2.29 Dada una muestra aleatoria $X_1, \cdots, X_n$ con distribución $Bin(n,p)$, encuentre el estimador UMVUE para p.

2.30 Dada una muestra aleatoria $X_1, \cdots, X_n$ con distribución $N(\mu, \sigma_0^2)$, donde σ_0^2 es conocido,

(a) Demuestre que las estadísticas $\sum_{i=1}^{n} X_i$ y $\bar{X}$ son suficientes y completas para μ.

(b) Encuentre el estimador UMVUE para μ.

2.31 Dada una muestra aleatoria $X_1, \cdots, X_n$ con distribución $N(\mu_0, \sigma^2)$, donde μ_0 es conocido,

(a) Encuentre una estadística suficiente y completa para σ^2.

(b) Encuentre el estimador UMVUE para σ^2.

2.32 Dada una muestra aleatoria $X_1, \cdots, X_n$ con distribución Gamma con parámetro de forma k conocido y parámetro de escala θ desconocido, encuentre el estimador UMVUE para θ.

2.33 En una muestra aleatoria con distribución $N(\mu, \sigma^2)$,

(a) Calcule la información de Fisher contenida en la muestra acerca del parámetro σ^2.

(b) Calcule la cota de Cramer Rao para S^2_{n-1} como estimador de σ^2

(c) Verifique que la desigualdad de información se cumple para S^2_{n-1}

(d) Repita (b) y (c) para S^2_n como estimador de σ^2.

(e) Compare S^2_n y S^2_{n-1} en términos del ECM.

2.34 Realice un ejercicio de simulación donde muestra que $Var(\bar{X})$ disminuye al hacer crecer a n en muestras provenientes de

(a) Bernoulli.

(b) Poisson.

2.35 Utilizar el teorema 2.4.23 para ver que en una muestra aleatoria con distribución $N(\mu, \sigma^2)$, tanto S^2_n como S^2_{n-1} son estimadores consistentes de σ^2.

Capítulo 3

Estimación por intervalo de confianza

3.1 Introducción

En el capítulo anterior, estudiamos el problema de estimar parámetros de una distribución teórica usando datos muestrales, pero es claro que el valor de la estimación es sólo un acercamiento de lo que es el parámetro verdadero, entonces cuando una estimación $\hat{\theta} = 5$, no estamos asegurando que θ es igual a 5, sino que puede estar cercano a este valor, sea mayor o menor que 5, de donde podemos ver que el parámetro está en un rango alrededor del valor de la estimación. En algunas situaciones, es de interés conocer el rango de posibles valores para el parámetro de interés; por ejemplo, si necesitamos conocer acerca del peso corporal de niños entre 11 y 14 años, podemos obtener una estimación de 43 kilos, pero este valor no nos da información acerca de si todos los niños tienen un peso muy cercano a 43 kilos, o si hay niños con peso muy superior de 43 kilos, y niños con peso infereior a 43 kilos. Dicho de otra forma, el valor de la estimación puntual no nos da información acerca de si existe fenómeno de obesidad o mala nutrición en los niños de esta edad, o si todos los niños tienen el peso ideal en promedio. En cambio, si podemos encontrar un límite inferior y un límite inferior, es decir, un intervalo para el peso promedio de los niños de este rango de edad, podemos tener más información acerca de nuestro parámetro de interés. En este capítulo estudiamos métodos para encontrar intervalos donde puede estar el parámetro. Estos intervalos contienen al parámetro de interés con un grado de confiabilidad, y se denominan intervalo de confianza. Primero damos la definición de este concepto.

Definición 3.1.1. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con función de densidad de probabilidad $f(x_i, \theta)$, un intervalo de nivel de confianza(ó probabilidad de cobertura) de $(1-\alpha) \times 100\,\%$ para una función del parámetro $g(\theta)$ es un intervalo aleatorio (T_1, T_2) con $Pr(T_1 < g(\theta) < T_2) = 1 - \alpha$. $1-\alpha$ se denomina el nivel de confianza o la probabilidad de cobertura.*

En algunas situaciones, no estamos interesados en hallar ambos límites inferior y superior, sino solamente el límite superior. Por ejemplo, en un estudio de emisión de gas dióxido de carbono en un cierto modelo de auto, estamos interesados en saber cuánto dióxido de carbono produce el auto puesto en funcionamiento en un determinado periodo del tiempo, y no es de interés saber cuál es el límite inferior, pues dadas las consideraciones ecológicas, entre menos dióxido de carbono produzca, mejor. Otro ejemplo se encuentra en la industria, donde en las líneas de producción de una fábrica, se necesita que la variación de alguna característica de los productos fabricados no se sobrepase de cierto límite superior, pues si la variación fuera grande, es un indicio de que la línea de producción no es estable. En estos casos, el intervalo de interés no es bilateral como en la Definición 3.1.1, sino unilateral. Tenemos la siguiente definición.

Definición 3.1.2. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con función de densidad de probabilidad $f(x_i, \theta)$, un intervalo de nivel de confianza unilateral superior de $(1-\alpha)\times 100\,\%$ para una función del parámetro $g(\theta)$ está conformado por una estadística T que satisface $Pr(g(\theta) < T) = 1-\alpha$.*

Ahora, considere el estudio de la vida útil de algún tipo de motor; entre más larga sea la vida útil, mejor es el motor. Por lo tanto, en un estudio de inferencia, estaríamos interesados en conocer cuál es la vida útil mínima del motor, mas no la vida útil máxima, y estaríamos interesados en hallar el límite inferior. En situaciones como esta, nos es útil el intervalo de confianza unilateral inferior definido a continuación.

Definición 3.1.3. *Dada una muestra aleatoria $X_1, \cdots, X_n$ con función de densidad de probabilidad $f(x_i, \theta)$, un intervalo de nivel de confianza unilateral inferior de $(1-\alpha)\times 100\,\%$ para una función del parámetro $g(\theta)$ está conformado por una estadística T que satisface $Pr(T < g(\theta)) = 1-\alpha$.*

En la siguiente sección se introducirán métodos para encontrar intervalos de confianza. Como se observa en las tres anteriores definiciones, encontrar un intervalo es equivalente a encontrar estadísticas que nos sirven como límites inferiores o superiores[1], y debido a la aleatoriedad de las estadísticas, los intervalos de confianza son realmente intervalos aleatorios en el sentido de que cuando la muestra observada cambia, los intervalos también toman diferentes valores. En la literatura estadística no se hace una distinción entre un intervalo de confianza que está conformado por estadísticas y un intervalo de confianza que está conformado por valores numéricos, pero el lector debe estar consciente en cada ocasión de cuál es el intervalo al que se hace referencia.

Como se verá en el siguiente capítulo, para un parámetro puede haber varios intervalos de confianza (similar al caso de estimación puntual, donde para un parámetro puede haber más de un estimador). En estos casos, es necesario conocer cuáles son los criterios que se deben tener en cuenta para escoger el mejor intervalo. En general, los aspectos más importantes que determinan la calidad de un intervalo de confianza son el nivel de confianza y la longitud del intervalo. Se espera, en primer lugar, que la probabilidad de cobertura sea alta, los valores comunes en la práctica estadística oscilan

[1] Se verá que estas estadísticas se pueden obtener modificando estimadores del parámetro de interés.

entre los 90 y 99 %; en segundo lugar, esperamos que la longitud del intervalo no sea muy grande, pues de lo contrario, el intervalo puede no aportar ningún conocimiento nuevo acerca del parámetro. Por ejemplo, si un laboratorio médico está interesado en conocer la tasa de curación de un nuevo medicamento para cierta enfermedad, encontrar un intervalo como $(0.01, 0.99)$ para esta tasa de curación puede no ser muy útil, pues el rango es demasiado grande, y realmente el intervalo aporta información casi nula acerca de qué valores puede estar tomando la verdadera tasa de curación. En cambio, un intervalo como $(0.2, 0.35)$ da una información mucho más precisa acerca de dónde se ubica la tasa de curación del medicamento. De lo anterior observamos que se buscan intervalos con una alta precisión, esto es, intervalos de longitud pequeña.

Ahora, es claro que en intervalos unilaterales no se puede definir la longitud del intervalo puesto que en un intervalo unilateral superior (inferior) el límite inferior (superior) es infinito. En un intervalo bilateral, la longitud se define, naturalmente, como el límite superior menos el límite inferior, esto es

$$l = T_2 - T_1.$$

Como se verá a lo largo de este capítulo, en general, cuando la longitud del intervalo es pequeña, el nivel de confianza es bajo; y cuando el nivel de confianza es alto, la longitud es grande. Por lo tanto, no se puede maximizar el nivel de confianza y minimizar la longitud al mismo tiempo, y entonces el procedimiento usado es fijar el nivel de confianza, y una vez fijado el nivel de confianza, se busca el intervalo con menor longitud. Para eso, observe que tanto T_2 como T_1 son estadísticas y son aleatorias, por lo tanto, la longitud del intervalo l también es aleatoria, entonces cuando se comparan dos intervalos, la comparación se debe llevar a cabo usando la longitud esperada, esto es, $E(l)$, y escogeremos el intervalo que, en promedio, tiene la longitud más corta. Otra característica que esperamos es que la varianza de l sea pequeña, puesto que un intervalo de confianza con varianza grande indica que puede tener longitudes muy grandes en algunas muestras y muy pequeñas en otras. Y como en la práctica muchas veces solo disponemos de una muestra, es más probable que el intervalo tenga longitud grande si $Var(l)$ es grande.

En general, el problema de encontrar un intervalo de confianza es más fácil cuando la muestra aleatoria proviene de una distribución normal y la teoría también está más unificada; mientras que para las otras distribuciones hay diferentes enfoques y en algunos casos aún en la literatura estadística hace falta más investigación. Por esta razón, introducimos primero los intervalos de confianza bajo la distribución normal, y posteriormente introducimos los intervalos de confianza cuando la muestra aleatoria proviene de otra distribución.

3.2 Bajo normalidad

En esta parte, estudiamos la estimación por intervalo de confianza para los parámetros de una distribución normal. Cuando se trata de una población de estudio, generalmente disponemos de una muestra aleatoria y estamos interesados en encontrar un intervalo

de confianza para la media y la varianza teórica; mientras que cuando se trata de comparar dos poblaciones independientes, estamos interesados en utilizar los intervalos de confianza para comparar las medias teóricas y las varianzas teóricas. El método presentado es el método de la variable pivote que también es apta para distribuciones diferentes a la distribución normal.

3.2.1 Problemas de una muestra

En esta parte, estudiamos intervalos de confianza para μ y σ^2 de una distribución normal $N(\mu, \sigma^2)$ basados en una muestra aleatoria.

Intervalos de confianza para la media μ

Método de la variable pivote

Existen muchas formas de encontrar un intervalo de confianza para un parámetro, el más sencillo y popular se llama el método de la variable pivote. Para estudiar este método, primero se introduce el concepto de las variables pivotes.

Definición 3.2.1. *Dada $X_1, \cdots, X_n$ una muestra aleatoria cuya distribución es $f(x_i, \theta)$ y sea Q una función de variables aleatorias de la muestra, entonces Q es una variable pivote para θ si*

1. *Q es una función no constante de θ y*
2. *la distribución de Q no depende de θ.*

Para encontrar una variable pivote para un parámetro, se puede comenzar, generalmente, con un estimador de este parámetro, pues en muchos casos, se puede obtener una variable pivote modificando el estimador. Considere el siguiente ejemplo.

Intervalo bilateral para μ cuando σ^2 es conocida

Ejemplo 3.2.1. *Sea $X_1, \cdots, X_n$ una muestra aleatoria proveniente de $N(\mu, \sigma^2)$ con $\sigma^2 = \sigma_0^2$ conocido, y se quiere encontrar una variable pivote para μ. En primer lugar, el estimador más conocido para μ es el promedio muestral $\bar{X}$. Se ha visto en el capítulo anterior que $\bar{X} \sim N(\mu, \sigma_0^2/n)$, de donde estandarizando $\bar{X}$, se tiene que*

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0} \sim N(0, 1).$$

Es claro que la variable $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}$ depende del parámetro μ, y su distribución no depende de μ, y por consiguiente ésta es una variable pivote para μ. Nótese que para μ, puede existir más de una variable pivote. Un ejemplo simple es que cualquiera de las $(X_i - \mu)/\sigma_0$ con $i = 1, \cdots, n$ es una variable pivote pues también se distribuye como normal estándar, la cual no depende de μ.

Hay que hacer la aclaración de que una variable pivote no es una estadística, puesto que una variable pivote depende del parámetro. Ahora, en el ejemplo anterior, la varianza es conocida, entonces el único parámetro desconocido es μ. Sin embargo, cuando la varianza también es desconocida, tenemos dos parámetros desconocidos y la definición de la variable pivote para cualquiera de los dos parámetros es diferente, como se explica más adelante.

Suponga que la varianza $\sigma^2 = \sigma_0^2$ es conocida. Una vez encontrada una variable pivote para μ, se pueden aplicar los siguientes pasos del método de la variable pivote para encontrar un intervalo bilateral para cualquier parámetro desconocido θ (siempre y cuando θ sea el único parámetro desconocido):

1. Encontrar una variable pivote Q para el parámetro θ,
2. Encontrar percentiles de la distribución de Q, a y b tales que $Pr(a < S < b) = 1 - \alpha$,
3. Despejar θ en la igualdad del anterior paso.

Como se ha visto anteriormente, cuando $\sigma^2 = \sigma_0^2$ es conocida, una variable pivote para μ es $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}$, la cual tiene distribución $N(0,1)$. Procedemos con el segundo paso del método de la variable pivote descrito anteriormente, para eso se necesita encontrar percentiles de la distribución $N(0,1)$: a y b tales que

$$Pr\left(a < \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0} < b\right) = 1 - \alpha. \tag{3.2.1}$$

Con el fin de ilustrar el proceso de encontrar a y b, adicionalmente se define $Pr(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0} < a) = A_1$ y $Pr(b < \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}) = A_2$, y si encontramos los valores de A_1 y A_2, podemos encontrar los valores de a y b.

Ahora, recordando que para una distribución de probabilidad continua como lo es la distribución normal, la probabilidad en (3.2.1) puede ser representada gráficamente como el área bajo la curva de la función de densidad de la distribución normal estándar entre los números a y b como lo ilustra la Figura 3.1.

Ahora, dado que toda el área bajo la curva de cualquier función de densidad es 1, podemos establecer la siguiente relación

$$A_1 + A_2 = \alpha, \tag{3.2.2}$$

la anterior ecuación contiene dos incógnitas $A_1 + A_2$, y por consiguiente, hay infinitas soluciones para A_1 y A_2, y así mismo, infinitas soluciones para a y b, y esto nos conduce a infinitos intervalos de confianza para μ. Una forma de resolver este problema es considerando que el intervalo resultante debe tener la longitud lo más pequeña posible. Entonces primero se busca cómo es el intervalo para μ en función de las incógnitas a y b, y luego encontrar los valores de a y b que minimizan la longitud del intervalo si ésta es constante, o la longitud esperada si ésta es aleatoria. Para eso

despejamos μ de (3.2.1), y tenemos que:

$$Pr\left(\bar{X} - b\frac{\sigma_0}{\sqrt{n}} < \mu < \bar{X} - a\frac{\sigma_0}{\sqrt{n}}\right) = 1 - \alpha, \tag{3.2.3}$$

cuya longitud está dada por $(b-a)\sigma_0/\sqrt{n}$ la cual es una constante. Entonces se buscan los valores de a y b (determinan las áreas A_1 y A_2 de forma única) de tal manera que minimicen esta longitud o equivalentemente los que minimicen $b-a$. Esta minimización se debe llevar a cabo teniendo en cuenta que el intervalo resultante debe tener probabilidad de cobertura $1-\alpha$, o equivalentemente $A_1 + A_2 = \alpha$. Nótese que $A_1 = \Phi(a)$ y $A_2 = 1 - \Phi(b)$, donde $\Phi(\cdot)$ denota la función de distribución de la distribución normal estándar. Entonces la minimización se lleva a cabo teniendo en cuenta que $\Phi(a) + 1 - \Phi(b) = \alpha$.

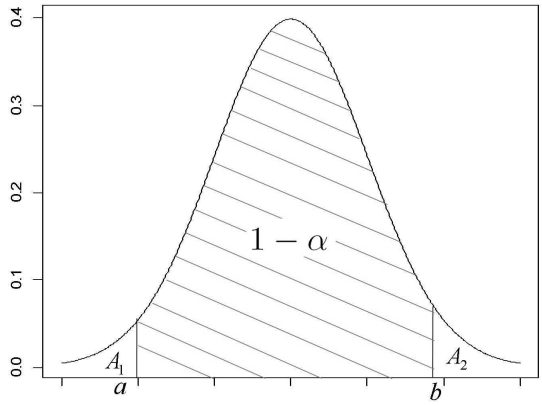

Figura 3.1: *Ilustración de los percentiles de una distribución normal estándar.*

Recurriendo a la técnica del multiplicador de Lagrange, se construye la siguiente función

$$g = b - a - \lambda(\Phi(a) + 1 - \Phi(b) - \alpha),$$

Al derivar g con respecto a a y b, e igualar a cero, se tiene que

$$-1 - \lambda f(a) = 0$$

y

$$1 + \lambda(f(b)) = 0,$$

donde f denota la función de densidad e la distribución normal estándar. De las dos anteriores ecuaciones se tiene que $f(b) = f(a)$, la única pareja de valores de a y b diferentes[2] que cumple esta igualdad es cuando $a = -b$, en este caso, $A_1 = A_2$

[2]Pues si $a = b$, en el intervalo (3.2.3), el límite inferior coincide con el superior, y el intervalo se reduce a $\bar{X}$ que es el estimador puntual de μ.

por la simetría de la función de densidad de la distribución normal estándar. Y como $A_1 + A_2 = \alpha$, se tiene que $A_1 = A_2 = \alpha/2$. Recordando la definición de A_1 y A_2, tenemos que

$$Pr\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0} < a\right) = \alpha/2$$

y

$$Pr\left(b < \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}\right) = \alpha/2,$$

de donde se concluye que $b = z_{1-\alpha/2}$ y $a = z_{\alpha/2} = -z_{1-\alpha/2}$. Reemplazando estos valores en (3.2.3), se tiene que

$$Pr\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) = 1-\alpha. \tag{3.2.4}$$

En conclusión, el intervalo de confianza bilateral de $(1-\alpha)\times 100\,\%$ usando como variable pivote $\dfrac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}$ de más alta precisión para la media μ, cuando la varianza teórica es conocida, está dado por

$$IC(\mu) = \left(\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right). \tag{3.2.5}$$

En referencia a la notación, haremos distinción entre las letras mayúsculas y las minúsculas. El intervalo conformado por estadística será denotado con letras mayúsculas como en (3.2.5). Cuando la muestra aleatoria $X_1, \ldots, X_n$ toma valores numéricos $x_1, \ldots, x_n$, el intervalo aleatorio también toma valores numéricos y los límites del intervalo se tornan números, y denotaremos el intervalo como $\left(\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right)$.

Otra forma para encontrar el intervalo con menor longitud se encuentra en Casella & Berger (2002). Consideremos el intervalo (3.2.3), la longitud de este intervalo está dada por $\sigma_0(b-a)/\sqrt{n}$ que depende de $b-a$. Por lo tanto, se deben buscar valores de a y b que minimizan $b-a$ y que cumplen la condición (3.2.1). El teorema 9.3.2 de Casella & Berger (2002) establece condiciones para encontrar un intervalo de menor longitud para una variable pivote con distribución unimodal. En primer lugar, recordamos que una función de densidad $f(x)$ es unimodal si existe un valor x^* tal que $f(x)$ es no decreciente para $x \leq x^*$ y no creciente para $x \geq x^*$. Las distribuciones normal, t, χ^2 para algunos grados de libertad están dentro de las distribuciones unimodales. Para estas distribuciones unimodales, el teorema 9.3.2 de Casella & Berger (2002) afirma lo siguiente.

Resultado 3.2.1. *Sea X una variable aleatoria con distribución unimodal $f(x)$, si el intervalo $[a,b]$ satisface*

1. $Pr(a < X < b) = 1-\alpha$
2. $f(a) = f(b) > 0$
3. $a \leq x^* \leq b$ *donde* x^* *es una moda de* $f(x)$

entonces, $[a, b]$ *es el intervalo de longitud más corto que satisface 1.*

Demostración. Se toma cualquier intervalo $[a', b']$ con longitud menor a $[a, b]$, esto es, $b' - a' < b - a$, veamos que $Pr(a' < X < b') < 1 - \alpha$. El intervalo $[a', b']$ puede ser de diferentes formas con respecto al $[a, b]$,

i. Si $a' \leq a$ y $b' \leq a$, entonces se tiene que

$$\begin{aligned} Pr(a' < X < b') &= \int_{a'}^{b'} f(x)dx \\ &\leq f(b')(b' - a') \quad (f(b') \geq x,\ \forall x \leq b') \\ &\leq f(a)(b' - a') \quad (f(b') \leq f(a), \text{ por ser } f \text{ no decreciente}) \\ &\leq f(a)(b - a) \quad (b' - a' < b - a) \\ &\leq \int_{a}^{b} f(x)dx \\ &= 1 - \alpha \end{aligned}$$

ii. Si $a' \leq a$ y $b' > a$, entonces $a' < a < b' < b$. Tenemos que

$$\begin{aligned} Pr(a' < X < b') &= \int_{a'}^{b'} f(x)dx \\ &= \int_{a}^{b} f(x)dx + \int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx \\ &= (1 - \alpha) + \int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx. \end{aligned}$$

Veamos $\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx < 0$, tenemos que

$$\int_{a'}^{a} f(x)dx \leq f(a)(a - a'),$$

por ser f no decreciente en $[a', a]$; y por otro lado, tenemos que

$$\int_{b'}^{b} f(x)dx \geq f(b)(b - b'),$$

por ser f no creciente en $[b', b]$. Usando estas dos desigualdades se tiene que

$$\begin{aligned} \int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx &\leq f(a)(a - a') - f(b)(b - b') \\ &= f(a)[(a - a') - (b - b')] \qquad (f(a) = f(b)) \\ &= f(a)[(b' - a') - (b - a)] \\ &\leq 0 \quad (b' - a' < b - a) \end{aligned}$$

en conclusión $\int_{a'}^{b'} f(x)dx < 1 - \alpha$.

iii. Si $a \leq a' \leq b' \leq b$, en este caso, el intervalo $[a', b']$ está contenido dentro del intervalo $[a, b]$, entonces por ser $f(x)$ no negativa, se tiene que

$$\int_{a'}^{b'} f(x)dx \leq \int_{a}^{b} f(x)dx = 1 - \alpha,$$

en conclusión $\int_{a'}^{b'} f(x)dx < 1 - \alpha$.

iv. Si $a \leq a' \leq b \leq b'$, este caso es análogo al caso II, y se tiene que

$$Pr(a' < X < b') = (1 - \alpha) + \int_{b}^{b'} f(x)dx - \int_{a}^{a'} f(x)dx.$$

Tenemos que

$$\int_{b}^{b'} f(x)dx \leq f(b)(b' - b)$$

por ser f no creciente en $[b, b']$, y

$$\int_{a}^{a'} f(x)dx \geq f(a)(a' - a)$$

por ser f no decreciente en $[a, a']$. Usando lo anterior se tiene que

$$\begin{aligned}\int_{b}^{b'} f(x)dx - \int_{a}^{a'} f(x)dx &\leq f(b)(b' - b) - f(a)(a' - a) \\ &= f(a)(b' - b) - f(a)(a' - a) \\ &= f(a)[(b' - a') - (b - a)] \\ &\leq 0\end{aligned}$$

en conclusión $\int_{a'}^{b'} f(x)dx < 1 - \alpha$.

Lo anterior muestra que cualquier intervalo que cumpla la condición 1 tiene longitud mayor que $[a, b]$, entonces $[a, b]$ es el intervalo más corto que cumple la condición 1. □

Usando el anterior resultado, se encuentra que el intervalo de la forma (a, b) más corto para la variable pivote $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}$ es $(-z_{1-\alpha/2}, z_{1-\alpha/2})$, despejando μ, se tiene el intervalo (3.2.5), y éste es el de menor longitud.

Como se vio anteriormente en el Resultado 3.2.1, en algunas situaciones puede resultar muy fácil para encontrar el intervalo más preciso; sin embargo, hay que tener en cuenta que este , permite encontrar el intervalo de menor longitud para la variable pivote, mas no directamente para el parámetro de interés. Entonces al utilizar este resultado se debe garantizar que al despejar el parámetro del intervalo de menor

longitud para la variable pivote, el intervalo resultante para el parámetro sigue siendo de menor longitud. Más adelante se verá una situación donde esto no ocurre, y por consiguiente, no se puede hacer uso de este resultado.

Ahora, analicemos la forma del intervalo (3.2.5).

1. En primer lugar, nótese que el límite superior es simplemente el estimador puntual $\bar{X}$ desplazado por la cantidad $z_{1-\alpha/2}\sigma_0/\sqrt{n}$ a la derecha, y el límite inferior es el estimador $\bar{X}$ desplazado por la misma cantidad a la izquierda. Esta cantidad $z_{1-\alpha/2}\sigma_0/\sqrt{n}$ puede ser interpretada como una medición de la incertidumbre. De lo anterior, podemos ver que el estimador puntual se ubica en el centro del intervalo, los intervalos que cumplen esta propiedad serán llamados intervalos simétricos. En la práctica, estos intervalos tienen la ventaja de que una vez conozcamos el intervalo calculado, podemos conocer el valor de la estimación del parámetro. Adicionalmente, es muy lógico que el intervalo (3.2.5) resulte ser simétrico, puesto que el estimador $\bar{X}$ es insesgado.

2. En segundo lugar, la cantidad en que $\bar{X}$ es desplazada a la derecha e izquierda para construir el intervalo depende de la desviación estándar conocida σ_0, de manera que entre más grande sea ésta, más ancho es el intervalo. Esto también es muy lógico, puesto que entre más grande sea la desviación estándar, menos información acerca de μ contiene la muestra aleatoria (ver Ejemplo 2.4.6), de manera que hay más incertidumbre, y por consiguiente, el intervalo resultante será más ancho y menos preciso. De la misma manera, entre más pequeña sea la desviación estándar teórica, más preciso será el intervalo.

Los intervalos para μ no sólo nos dan un rango de posibles valores para μ, también pueden ser una herramienta útil para verificar si una cierta afirmación acerca de μ es apoyada por la muestra observada. Por ejemplo, suponga que se cree que $\mu = \mu_0$ (el tópico de prueba de hipótesis se tratará en el siguiente capítulo con más detalles), y un intervalo bilateral para μ calculado sobre una muestra observada es (t_1, t_2), entonces se puede afirmar que los datos no muestran evidencia para rechazar que $\mu = \mu_0$ si efectivamente $\mu_0 \in (t_1, t_2)$; de lo contrario, los datos sugieren que μ posiblemente no toma el valor μ_0.

Ejemplo 3.2.2. *Retomando los datos del Ejemplo 2.3.6 donde se tiene la medición del grosor de 12 láminas de vidrio producidos por cierta línea de producción y se vio que la distribución normal es apropiada. Suponga que los valores nominales de la línea de producción, es decir, los valores estándares que describen los productos de la línea, corresponden a un grosor promedio de 3 cm y una desviación estándar de 0.8 cm.*

Suponga que se desea verificar que el valor nominal del grosor promedio es, realmente, 3 cm, se utiliza el cálculo de un intervalo de confianza. En este caso $\bar{x} = 3.18$*, si se calcula, en primer lugar, un intervalo del 95 %, tenemos el percentil* $z_{1-\alpha/2} = z_{0.975} = 1.96$*, y el intervalo calculado para* μ *basado en estas 12 láminas es* $(2.73, 3.63)$*. Podemos ver que el intervalo contiene el valor nominal de 3 cm, indicando que los datos están a favor de que la línea de producción sí produce láminas de un grosor de 3 cm.*

Finalmente, consideramos la interpretación del intervalo $(2.73, 3.63)$ *como un intervalo de 95 % para* μ*. Una interpretación popular es afirmar que «la probabilidad de que* μ *se encuentre en el intervalo* $(2.73, 3.63)$ *es de 0.95». A pesar de la simplicidad y la popularidad de este tipo de interpretaciones, no son teóricamente correctas. Puesto que* μ *es una constante fija, no es una variable aleatoria, por consiguiente la probabilidad de que* μ *se encuentra en* $(2.73, 3.63)$ *es 0 o 1, y no puede ser 0.95.*

Una forma de interpretar el intervalo $(2.73, 3.63)$ *es tener en cuenta que éste es una realización del intervalo (3.2.5) que está conformado por dos estadísticas y por consiguiente el intervalo toma valores distintos en muestras diferentes. Y el hecho de que la probabilidad de cobertura es 95 %, implica que de posibles 100 muestras distintas del mismo tamaño de la población donde en cada muestra se calcula el intervalo, entonces aproximadamente 95 de estos 100 intervalos contienen el parámetro verdadero* μ*.*

Como se menciona al principio del capítulo, un buen intervalo debe tener una longitud pequeña o equivalentemente tener una precisión alta. Se ha visto que el intervalo (3.2.5) es el de menor longitud usando la variable pivote encontrada usando el mejor estimador $\bar{X}$, y la longitud del intervalo l está dada por

$$l = \frac{2z_{1-\alpha/2}\sigma_0}{\sqrt{n}}, \tag{3.2.6}$$

la cual es una constante, y podemos observar directamente a l para ver cuándo ésta se hace pequeña. Dada la forma de l y teniendo en cuenta que σ_0 es la desviación estándar de la población, la cual está fija y conocida, entonces para que l sea pequeña hay las siguientes dos alternativas:

- Disminuir el nivel de confianza $1-\alpha$, esto es, aumentar el valor de α, de esta manera $1-\alpha/2$ se hace más pequeño y el percentil $z_{1-\alpha/2}$ también disminuye. De lo anterior podemos afirmar que entre más pequeño sea el nivel de confianza, menor longitud tendrá el intervalo y más preciso será el intervalo. En la práctica, hay que tener cuidado con lo anterior, puesto que si un intervalo tiene nivel de confianza o probabilidad de cobertura muy baja, no será muy útil por más preciso que sea. Por otro lado, también podemos ver que al aumentar el nivel de confianza, la longitud del intervalo se hace cada vez más grande, y es cada vez menos precisa.

- Aumentar el tamaño muestral n. Este aumento puede implicar más esfuerzo en la recolección de datos y en algunos casos costos económicos más altos para el investigador. Para hacer una idea sobre el efecto que tiene sobre l cuando se incrementa n, en la Figura 3.2, se muestra la gráfica de la función $1/\sqrt{n}$. Se observa que para valores de n aproximadamente desde 20, la disminución en la longitud por cada incremento de unidad en n empieza a ser pequeña. Esta puede ser la razón por la que muchos textos estadísticos afirman que un tamaño muestral superior a 30 es suficientemente grande, pero esta recomendación es simplemente heurística, y no debe ser usada sin considerar el contexto de los problemas prácticos en la mano. Adicionalmente, podemos calcular el tamaño

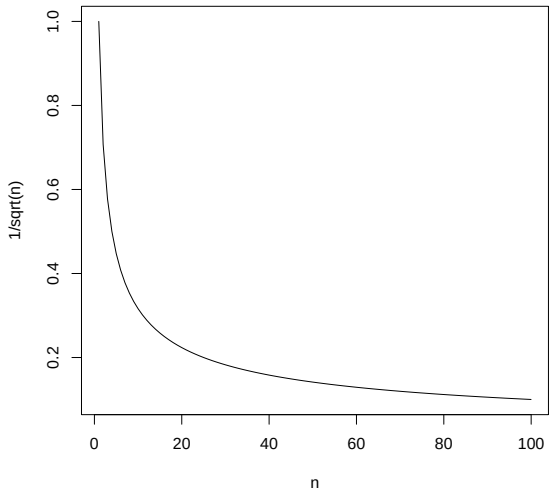

Figura 3.2: *Función* $1/\sqrt{n}$.

muestral mínimo necesario para tener una longitud previamente especificada. De (3.2.6), tenemos que

$$n = \left(\frac{z_{1-\alpha/2}\sigma_0}{l/2} \right)^2 \tag{3.2.7}$$

donde $l/2$ satisface

$$Pr\left(|\bar{X} - \mu_0| < \frac{l}{2} \right) = 1 - \alpha.$$

Es decir, $l/2$ se puede interpretar como el error máximo permitido para la diferencia entre el parámetro μ y su estimador $\bar{X}$ y en muchos casos se denomina margen de error. Y para un margen de error establecido de antemano, podemos calcular el tamaño muestral necesario n usando (3.2.7). Nótese que el tamaño de muestral también depende del nivel de confianza, de tal forma que al aumentar el nivel de confianza, naturalmente el intervalo se hace más ancho, y por consiguiente se necesita una muestra aún más grande para tener un margen de error específico.

En el Ejemplo 3.2.2, donde en una muestra de 12 láminas de vidrio, se encontró un intervalo de confianza del 95 % para el grosor promedio de las láminas, al asumir que $\sigma_0 = 0.8\ cm$, si se especifica que el margen de error máximo debe ser 0.3 cm, entonces el tamaño de muestra necesario puede ser calculado como

$$n = \left(\frac{1.96 * 0.8}{0.3} \right)^2 = 27.32$$

Por consiguiente, se debe tener una muestra de por lo menos 28 láminas para tener este margen de error de 0.3 cm.

Intervalo de confianza para una función del parámetro

En el Ejemplo 2.3.6, se vio que no sólo se puede encontrar la estimación de máxima verosimilitud de μ y σ, sino que además es posible estimar probabilidades en función de estos dos parámetros; por ejemplo, se estimó el porcentaje de vidrios que serán desechados, puestos a la venta y usados como materia prima. La pregunta ahora es si podemos construir intervalos de confianza para estas probabilidades. Esta pregunta, en el contexto general, es equivalente a cómo construir un intervalo de confianza para una función del parámetro $g(\theta)$ usando un intervalo para θ. La respuesta está dada en el siguiente resultado, y solo podemos encontrar un intervalo de confianza para $g(\theta)$ para algunas funciones g.

Resultado 3.2.2. *Dada una muestra aleatoria con parámetro desconocido θ, entonces*

- *si (T_1, T_2) es un intervalo de confianza de $100 \times (1-\alpha)\,\%$ para θ, entonces un intervalo de confianza de $100 \times (1-\alpha)\,\%$ para $g(\theta)$ es $(g(T_1), g(T_2))$ si la función g es estrictamente creciente para θ y $(g(T_2), g(T_1))$ si la función g es estrictamente decreciente*

- *si $(-\infty, T)$ es un intervalo de confianza unilateral superior de $100 \times (1-\alpha)\,\%$ para θ, entonces un intervalo de confianza unilateral superior de $100 \times (1-\alpha)\,\%$ para $g(\theta)$ es $(-\infty, g(T))$ si la función g es estrictamente creciente para θ. Si la función g es estrictamente decreciente, entonces un intervalo unilateral inferior para $g(\theta)$ será $(g(T), \infty)$.*

- *si (T, ∞) es un intervalo de confianza unilateral inferior de $100 \times (1-\alpha)\,\%$ para θ, entonces un intervalo de confianza unilateral inferior de $100 \times (1-\alpha)\,\%$ para $g(\theta)$ es $(g(T), \infty)$ si la función g es estrictamente creciente para θ. Si la función g es estrictamente decreciente, entonces un intervalo unilateral superior para $g(\theta)$ será $(\infty, g(T))$.*

Demostración. Se probará para el intervalo bilateral (T_1, T_2) y se dejan como ejercicio los otros dos casos. Tenemos que si (T_1, T_2) es un intervalo de confianza de $100 \times (1-\alpha)\,\%$ para θ entonces

$$1 - \alpha = Pr(T_1 < \theta < T_2) = Pr(g(T_1) < g(\theta) < g(T_2))$$

si g es estrictamente creciente. Para ver la última igualdad con más rigurosidad matemática, recordemos que detrás de las variables aleatorias, existe un espacio de probabilidad $(\Omega, \mathfrak{F}, P)$, donde para cualquier variable aleatoria X, la probabilidad de que X tome valor en cierto conjunto A se define $Pr(X \in A) = Pr(\{\omega :\ X(\omega) \in A\})$. De esta forma, tenemos que

$$Pr(T_1 < \theta < T_2) = Pr(\{\omega :\ T_1(\omega) < \theta < T_2(\omega)\})$$

y análogamente

$$Pr(g(T_1) < g(\theta) < g(T_2)) = Pr(\{\omega : \ g(T_1(\omega)) < \theta < g(T_2(\omega))\}).$$

Entonces para demostrar que $Pr(T_1 < \theta < T_2) = Pr(g(T_1) < g(\theta) < g(T_2))$, basta ver que $\{\omega : \ T_1(\omega) < \theta < T_2(\omega)\} = \{\omega : \ g(T_1(\omega)) < \theta < g(T_2(\omega))\}$. Para eso, tomamos un $\omega \in \{\omega : \ T_1(\omega) < \theta < T_2(\omega)\}$, entonces $T_1(\omega) < \theta$, y como g es estrictamente creciente, $g(T_1(\omega)) < g(\theta)$, análogamente, se tiene que $g(\theta) < g(T_2(\omega))$, esto es $g(T_1(\omega)) < g(\theta) < g(T_2(\omega))$ y concluimos que $\omega \in \{\omega : g(T_1(\omega)) < \theta < g(T_2(\omega))\}$, de donde

$$\{\omega : \ T_1(\omega) < \theta < T_2(\omega)\} \subseteq \{\omega : \ g(T_1(\omega)) < \theta < g(T_2(\omega))\}.$$

Para ver la otra contenencia, se utiliza un razonamiento similar, teniendo en cuenta que g tiene inversa por ser estrictamente creciente. □

Utilizando el anterior resultado, volvemos al contexto del Ejemplo 2.3.6, donde se desea encontrar un intervalo de confianza para el porcentaje de vidrios que serán desechados, puestos a la venta y usados como materia prima. Estos porcentajes dependen de μ y σ; por ahora, supongamos que $\sigma = 0.8 \ cm$ es conocida, entonces tenemos que el porcentaje de vidrios desechados se puede escribir como $\Phi(\frac{2.8-\mu}{0.8})$, la cual es una función estrictamente decreciente de μ. Ahora, en el Ejemplo 3.2.2 se encontró el intervalo $(2.73, 3.63)$ para μ, entonces aplicando el anterior resultado, un intervalo de confianza para el porcentaje de vidrios desechados será $(\Phi(\frac{2.8-3.63}{0.8}), \Phi(\frac{2.8-2.73}{0.8})) = (0.15, 0.53)$. De manera análoga, para el porcentaje de vidrios que serán usados como materia prima se tiene el intervalo de confianza $(1 - \Phi(\frac{3.2-2.73}{0.8}), 1 - \Phi(\frac{3.2-3.63}{0.8})) = (0.28, 0.70)$. Con respecto al porcentaje de vidrios que serán vendidos, ésta se puede escribir como $\Phi(\frac{3.2-\mu}{0.8}) - \Phi(\frac{2.8-\mu}{0.8})$. Una simple gráfica de esta función revela que ésta es creciente para valores de μ menores de 3 y decreciente para μ mayor de 3, y por consiguiente no es posible hallar el intervalo para este porcentaje usando el anterior resultado.

Intervalo unilateral para μ cuando σ^2 es conocida

En algunas situaciones prácticas, no es necesario encontrar tanto el límite inferior como el límite superior para el parámetro de interés, sino solo uno de ellos. Por esta razón, ahora construimos intervalos unilaterales para la media teórica μ cuando la varianza es conocida.

Para encontrar un intervalo unilateral superior para μ, el método de pivote enunciado anteriormente para encontrar un intervalo bilateral se modifica en el segundo paso, y se convierte en

1. Encontrar una variable pivote Q para el parámetro de interés θ,
2. Si la relación que guarda entre la variable pivote Q y el parámetro θ es proporcional, entonces se encuentra el percentil de la distribución de Q denotado por b, tal que $Pr(S < Q) = 1 - \alpha$; si la relación es inversamente proporcional,

entonces se encuentra el percentil de la distribución de Q denotada por a, tal que $Pr(a < Q) = 1 - \alpha$.

3. Despejar μ en la igualdad del anterior paso para encontrar un intervalo unilateral superior para μ.

El primer paso del anterior procedimiento ya está completado, pues se vio que una variable pivote para μ es $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}$ cuando $\sigma^2 = \sigma_0^2$ es conocida. Esta variable guarda una relación inversamente proporcional con μ, puesto que para valores muy grandes de μ, el valor de la estadística es muy pequeño. Entonces se debe encontrar un número a tal que

$$Pr\left(a < \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}\right) = 1 - \alpha,$$

de donde

$$Pr\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0} < a\right) = \alpha,$$

recordando la definición de percentil y que la distribución de $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0} \sim N(0,1)$, se tiene que a es el percentil α de la distribución $N(0,1)$, esto es, $a = z_\alpha$. Entonces se tiene que

$$Pr\left(z_\alpha < \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}\right) = 1 - \alpha.$$

Ahora siguiendo al tercer paso del método de pivote, se despeja el parámetro μ, de donde se tiene que:

$$Pr\left(\bar{X} - z_\alpha \frac{\sigma_0}{\sqrt{n}} > \mu\right) = 1 - \alpha.$$

En conclusión, un intervalo unilateral superior para μ es $(-\infty, \bar{X} - z_\alpha \frac{\sigma_0}{\sqrt{n}})$. Ahora, $z_\alpha = -z_{1-\alpha}$ por la simetría de la distribución normal, entonces, se tiene el siguiente intervalo unilateral superior para μ:

$$IC(\mu) = \left(-\infty, \bar{X} + z_{1-\alpha} \frac{\sigma_0}{\sqrt{n}}\right). \tag{3.2.8}$$

Análogamente, para encontrar un intervalo de confianza unilateral inferior para un parámetro de interés θ, el método de la variable pivote es

1. Encontrar una variable pivote Q para el parámetro de interés θ,

2. Si la relación que guarda entre la variable pivote Q y el parámetro θ son proporcionales, entonces se encuentra el percentil de la distribución de Q: a, tal que $Pr(a < S) = 1 - \alpha$; si la relación es inversamente proporcional, entonces se encuentra el percentil de la distribución de Q: b, tal que $Pr(S < b) = 1 - \alpha$.

3. Despejar μ en la igualdad del anterior paso.

Aplicando el anterior procedimiento, se puede ver que un intervalo unilateral inferior para μ es

$$IC(\mu) = \left(\bar{X} - z_{1-\alpha} \frac{\sigma_0}{\sqrt{n}}, \infty \right). \tag{3.2.9}$$

Es claro que cuando se trata de los intervalos unilaterales, no se puede considerar a la longitud del intervalo como un criterio, pues éste es infinito.

En algunas situaciones, se tienen afirmaciones como $\mu \geq \mu_0$; por ejemplo, se cree que la vida útil de un tipo de bombillo no debe ser inferior a 7000 horas. El uso del intervalo unilateral (3.2.8) puede resultar útil en este caso, si el intervalo calculado en una muestra observada es $(-\infty, t)$, y ocurre que $t < \mu_0$, esto implica que ni siquiera el límite superior de μ supera a μ_0, entonces se concluye que los datos sugieren que la afirmación $\mu \geq \mu_0$ debe ser falsa. Por otro lado, si la afirmación de interés es del tipo $\mu \leq \mu_0$, entonces el intervalo usado debe ser (3.2.9). De tal forma que si el valor observado del límite inferior es mayor que μ_0, se puede concluir que los datos muestran evidencia de rechazo hacia la afirmación $\mu \leq \mu_0$.

Ejemplo 3.2.3. *Para el ejemplo de láminas de vidrios del Ejemplo 2.3.6, suponga que se afirma que máximo el 15 % de las láminas producidas son desechadas, y se desea usar la información suministrada por las 12 láminas para verificar esta afirmación. Para eso se debe calcular el intervalo de confianza unilateral inferior para este porcentaje, el cual al suponer $\sigma = 0.8\ cm$, está dado por $\Phi(\frac{2.8-\mu}{0.8})$ y es una función decreciente de μ. Entonces por el Resultado 3.2.2, se debe encontrar un intervalo unilateral superior para μ, el cual está dado en (3.2.8) y para los datos observados da como resultado $(-\infty, 3.56)$ con un nivel de confianza del 95 %. De esta forma, un intervalo inferior para $\Phi(\frac{2.8-\mu}{0.8})$ está dado por $(\Phi(\frac{2.8-3.56}{0.8}), \infty) = (0.17, \infty)$. Y por consiguiente podemos afirmar que los datos observados no apoyan la afirmación de que el porcentaje de láminas desechadas es inferior a 15 %.*

Finalmente, hacemos la aclaración de que los anteriores intervalos para μ no son únicos. Por ejemplo, los intervalos (3.2.5), (3.2.8) y (3.2.9) fueron hallados usando como variable pivote $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}$, pero ésta no es la única variable pivote para el caso cuando $\sigma^2 = \sigma_0^2$ es conocida. Considere la variable $Q' = \frac{X_2+\cdots+X_n}{n-1}$, es decir, el promedio muestral sin incluir la primera variable. Se tiene que $Q' \sim N(\mu, \frac{\sigma_0^2}{n-1})$, de donde $\frac{\sqrt{n-1}(Q'-\mu)}{\sigma_0} \sim N(0,1)$. Entonces ésta también es una variable pivote, y usando esta variable pivote, se puede construir el intervalo bilateral de $(1-\alpha)\times 100\,\%$ para μ como $(Q' - z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n-1}}, Q' + z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n-1}})$. Sin embargo, la longitud de este intervalo es $\frac{2z_{1-\alpha/2}\sigma_0}{\sqrt{n-1}}$ que siempre es mayor a la longitud del intervalo obtenido usando la variable pivote $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}$. Lo anterior es lógico, puesto que en el capítulo anterior se ha visto que $\bar{X}$ es el mejor estimador insesgado para μ, y es natural pensar que el intervalo de confianza construido usando este estimador también debe ser el mejor en término de precisión.

Intervalo bilateral para μ cuando σ^2 es desconocida

Ahora consideramos el caso general cuando la varianza teórica σ^2 no es conocida; en este caso, la distribución $N(\mu, \sigma)^2$ de donde proviene la muestra aleatoria tiene dos parámetros y ambos son desconocidos, y para este tipo de distribuciones, la definición de una variable pivote es diferente, como enuncia la siguiente definición.

Definición 3.2.2. *Dada $X_1, \cdots, X_n$ una muestra aleatoria proveniente de la distribución $f(x_i, \theta_1, \theta_2)$ donde θ_1 y θ_2 son parámetros desconocidos y sea Q una función de variables aleatorias de la muestra, entonces Q es una variable pivote para θ_1 si*

1. *Q depende de θ_1*
2. *Q no depende de θ_2*
3. *la distribución de Q no depende de θ_1 ni de θ_2.*

Ahora, aplicamos el anterior procedimiento para encontrar intervalos de confianza para μ en una muestra aleatoria proveniente de $N(\mu, \sigma^2)$ con σ^2 desconocida. Como siempre, la variable pivote puede ser encontrada modificando un estimador del parámetro de interés. Por esta razón, primero consideramos el estimador de μ: $\bar{X}$ cuya distribución es $N(\mu, \sigma^2/n)$. De donde $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$, aunque esta variable depende del parámetro μ y su distribución no depende de μ y σ^2, no es una variable pivote para μ, puesto que depende de σ^2 que es un parámetro desconocido, entonces no cumple con la segunda condición de la definición anterior. Una solución natural es reemplazar σ por su estimador S_{n-1}, es decir, la variable que podría ser pivote es $\frac{\sqrt{n}(\bar{X}-\mu)}{S_{n-1}}$. Nótese que esta variable depende de μ y no de σ^2, entonces falta ver que su distribución no depende de μ y σ^2. Para encontrar la distribución de esta variable, se tienen en cuenta las siguientes propiedades:

- $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$,
- $\frac{1}{\sigma^2}(n-1)S_{n-1}^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1}$
- $\bar{X}$ y $\sum_{i=1}^{n}(X_i - \bar{X})^2$ son variables independientes (ver el Resultado 2.4.3).

Recordando la definición de una distribucion t-student[3] dada en la Definición 1.1.14,

[3]Tal como se comentó en el primer capítulo, William Gosset descubrió la distribución t mientras trabajaba para la compañía cervecera Guinness. Este relato se puede volver más interesante si nos preguntamos lo siguiente: ¿por qué razón tal descubrimiento surgió de las entrañas de una cervecera y no de una compañía vinícola (fabricante de vinos)? John Cook afirma que los cerveceros siempre se han enorgullecido de la consistencia de sus cervezas, mientras que los productores de vino se enorgullecen de la variedad de sus cosechas. Por esta razón nunca escucharemos a ningún amante de la cerveza exclamar que 1998 fue un ≪buen año≫, de la manera como lo haría un sommeliér (experto en vinos) refiriéndose a alguna cosecha de alguna cepa de algún país. De hecho, la variedad de las cepas es en gran parte la culpable de que una botella de vino de la misma marca, pero de diferente cosecha, tenga un sabor distinto en el paladar. Por otro lado, el sabor de una cerveza destapada hoy será el mismo sabor que el de una cerveza destapada hace un año. Por tanto, los cerveceros valoran tanto la consistencia que invierten dinero y recursos en departamentos de investigación en control de calidad, y de allí surgió la famosa distribución t.

se tiene que

$$\frac{\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}}{\sqrt{\frac{S_{n-1}^2}{\sigma 2}}} \sim t_{n-1},$$

esto es,

$$\frac{\sqrt{n}(\bar{X}-\mu)}{S_{n-1}} \sim t_{n-1}, \tag{3.2.10}$$

de donde se observa que la distribución de $\frac{\sqrt{n}(\bar{X}-\mu)}{S_{n-1}}$ no depende de μ, y tampoco de σ^2, y por consiguiente hemos encontrado una variable pivote para μ. Entonces siguiendo los lineamientos del método de la variable pivote se procede a buscar valores a y b con

$$Pr\left(a < \frac{\sqrt{n}(\bar{X}-\mu)}{S_{n-1}} < b\right) = 1-\alpha. \tag{3.2.11}$$

Análogo al caso cuando σ^2 es conocido, puede haber un número infinito de soluciones para a y b que satisface (3.2.11), y debe buscar la solución que arroja el intervalo más preciso para μ, esto es, el intervalo con menor longitud. Para eso, notemos en primer lugar que el intervalo resultante de (3.2.11) para μ será $(\bar{X} - bS_{n-1}/\sqrt{n}, \bar{X} - aS_{n-1}/\sqrt{n})$, y la longitud de este intervalo está dada por

$$l = S_{n-1}(b-a)/\sqrt{n}.$$

Es claro que esta longitud depende de S_{n-1}, por consiguiente es una variable aleatoria, y su magnitud no puede ser medida directamente, sino mediante su esperanza, $E(l) = E(S_{n-1})(b-a)/\sqrt{n}$ que depende de $b-a$. Por tanto, se buscan valores de a y b que minimizan $b-a$ y que satisfacen $Pr(a < \frac{\sqrt{n}(\bar{X}-\mu)}{S_{n-1}} < b) = 1-\alpha$. Dadas las características de la función de densidad de la distribución t, se puede aplicar el Resultado 3.2.1, y se tiene que $a = -t_{n-1,1-\alpha/2}$ y $b = t_{n-1,1-\alpha/2}$. En conclusión, el intervalo de menor longitud para μ cuando σ^2 es desconocida está dado por

$$IC(\mu) = \left(\bar{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S_{n-1}}{\sqrt{n}}, \bar{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S_{n-1}}{\sqrt{n}}\right). \tag{3.2.12}$$

Nótese que, análogo al caso cuando $\sigma^2 = \sigma_0^2$ es conocida, el anterior intervalo se construye desplazando el estimador de máxima verosimilitud $\bar{X}$ una cantidad hacia la izquierda y la misma cantidad hacia la derecha, y por consiguiente, también es un intervalo simétrico.

Y la longitud del intervalo está dada por

$$l = \frac{2t_{n-1,1-\alpha/2}S_{n-1}}{\sqrt{n}}, \tag{3.2.13}$$

la cual es una variable aleatoria, y para cada muestra observada, toma un valor diferente, por lo tanto, para medir la precisión del intervalo, tomamos en cuenta $E(l)$,

más aún, calculamos estimaciones de $E(l)$ para diferentes tamaños de muestra. Se simulan 1000 muestras de tamaño $n = 3, \cdots, 100$ de una distribución normal estándar, y en cada muestra simulada se calcula el intervalo (3.2.12) y el valor que toma l. Y para cada valor de n, se calcula el promedio de los 1000 valores de l, ésta puede ser vista como una estimación de $E(l)$. En la Figura 3.3 se observan estas estimaciones. Se observa un comportamiento similar a la Figura 3.2 relacionada con la longitud del intervalo para μ cuando σ^2 es conocida, y podemos ver que la longitud se disminuye a medida que el tamaño de muestra n crece, y para n mayores a valores entre 30 y 40, la disminución en la longitud de intervalo es relativamente pequeña.

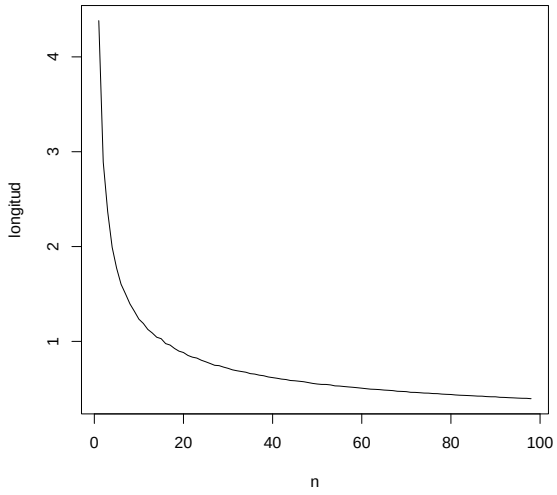

Figura 3.3: *Estimación de la longitud esperada $E(l)$ del intervalo (3.2.12) para diferentes tamaños de muestra n.*

En términos de tener un valor específico para el margen de error $l/2$, se necesita que el tamaño muestral sea por lo menos $\left(\frac{t_{n-1,1-\alpha/2}s_{n-1}}{l/2}\right)^2$. Sin embargo, esta cantidad no es directamente calculable, puesto que el percentil $t_{n-1,1-\alpha/2}$ y s_{n-1} depende del tamaño muestral n y de los valores observados en la muestra. Una solución rápida es reemplazarlo por el percentil $z_{1-\alpha/2}$ usando la similitud entre la distribución t-student y la distribución normal estándar, y mediante una muestra pivotal tener un acercamiento al valor de σ al cual denotaremos por $\tilde{\sigma}$. De esta forma, el tamaño muestral para alcanzar un margen de error $l/2$ se puede calcular como

$$n = \left(\frac{z_{1-\alpha/2}\tilde{\sigma}}{l/2}\right)^2$$

En R, el cálculo del intervalo (3.2.12) se lleva a cabo usando la función `t.test` [4].

[4] Esta función realiza los procedimientos de la prueba de hipótesis e intervalo de confianza, aquí solo

Para los datos del Ejemplo 2.3.6, el comando y salida se muestran a continuación.

```
>vidrio<-c(3.56, 3.36, 2.99, 2.71, 3.31,3.68, 2.78, 2.95, 2.82,
 3.45, 3.42, 3.15)
> t.test(vidrio)
95 percent confidence interval:
 2.974047 3.389287
sample estimates:
mean of x
 3.181667
```

Con la función `t.test`, se calcula automáticamente el intervalo de 95 % del nivel de confianza, en el caso de que se desee cambiar el nivel, se debe usar la opción `conf.level` dentro del comando. Para los mismos datos `vidrio`, si se desea calcular un intervalo del 90 %, el comando y la salida son

```
> t.test(vidrio,conf.level=0.9)
90 percent confidence interval:
 3.012260 3.351073
sample estimates:
mean of x
 3.181667
```

Y podemos ver que el intervalo del 90 % tiene longitud más corta comparado con el del 95 %, es decir, es más preciso. Esto coincide con la observación que en un intervalo de confianza al aumentar el nivel de confianza se disminuye el grado de precisión medido a través de la longitud.

Para estos mismos datos, si se exige un margen de error de 0.1 cm, podemos ver que para un intervalo de confianza del 95 %, el tamaño muestral necesario es 42, mientras que para un intervalo de confianza del 90 %, el tamaño muestral es de solo 30.

Intervalo unilateral para μ cuando σ^2 es desconocida

Usando la variable pivote dada en (3.2.10) y el procedimientos del método de la variable pivote para el intervalo unilateral, se pueden encontrar los dos intervalos unilaterales dados a continuación.

$$IC(\mu) = \left(-\infty, \bar{X} + t_{n-1,1-\alpha}\frac{S_{n-1}}{\sqrt{n}}\right),$$

y

$$IC(\mu) = \left(\bar{X} - t_{n-1,1-\alpha}\frac{S_{n-1}}{\sqrt{n}}, \infty\right).$$

Ilustramos el cómputo de estos intervalos en R en el siguiente ejemplo.

presentamos la parte de salida correspondiente al intervalo de confianza.

Ejemplo 3.2.4. *El cómputo de los dos anteriores intervalos unilaterales en R hace uso nuevamente de la instrucción* `t.test` *especificando* `greater` *para la opción* `alternative` *si se necesita calcular el intervalo inferior, y especificando* `less` *para la opción* `alternative` *si se necesita calcular el intervalo superior*[5]*. Para los datos del Ejemplo 2.3.6, calculamos los dos intervalos unilaterales. Los comandos y salidas en R se muestran a continuación.*

```
> t.test(vidrio,alternative="greater")
95 percent confidence interval:
 3.01226     Inf
sample estimates:
mean of x
 3.181667
```

Observamos que el intervalo inferior para el grosor promedio de las láminas producidas está dado por $(3.01, \infty)$*. Aunque no hay un límite superior para* μ*, hay que evitar conclusiones como «*μ *puede tomar CUALQUIER valor mayor que 3.01», puesto que dado el contexto del problema y los datos observados, es claro que* μ *difícilmente puede ser mayor de los 4 cm. La utilidad del intervalo* $(3.01, \infty)$ *es el límite inferior y permite afirmar que los datos indican que el grosor promedio está por encima de los 3.01 cm. Para calcular el intervalo superior, tenemos el siguiente comando y salida en R.*

```
> t.test(vidrio,alternative="less")
95 percent confidence interval:
     -Inf 3.351073
sample estimates:
mean of x
 3.181667
```

Observamos que el intervalo superior para el grosor promedio de las láminas producidas está dado por $(-\infty, 3.35)$*. De la misma manera, aclaramos que se debe evitar conclusiones como «*μ *puede tomar CUALQUIER valor menor que 3.35» pues es claro que* μ *debe ser, por lo menos, positivo, y por consiguiente no puede tomar valores negativos. Y en lo que se debe enfatizar al interpretar este intervalo es que el grosor promedio está por debajo de los 3.35 cm.*

De lo visto anteriormente, podemos observar que si se desea calcular un intervalo de confianza para la media teórica μ en una distribución normal, se debe tener en cuenta si la varianza teórica σ^2 es conocida o no. En el caso afirmativo el intervalo se construye con base en una distribución normal estándar, mientras que en el caso negativo, la distribución que se usa es la t student. Entonces podemos pensar qué pasa cuando la varianza es conocida, pero por alguna razón ignora este hecho, y en vez de usar el intervalo (3.2.5) usa el intervalo (3.2.12). En la Figura 3.4 se muestra la

[5]Estas opciones se refieren al uso de esta función para los procedimientos de pruebas de hipótesis que se discutirá en el siguiente capítulo.

longitud del intervalo con distribución normal (3.2.5) y las estimaciones de la longitud del intervalo con distribución t (3.2.12) cuando la varianza verdadera es 1. Se observa que el intervalo normal siempre tiene menor longitud que el intervalo t, especialmente para tamaños de muestras pequeños, pero a medida que n crece, la diferencia es cada vez más pequeña, es decir, el hecho de ignorar el valor verdadero de la varianza no causa pérdida de precisión cuando la muestra es grande.

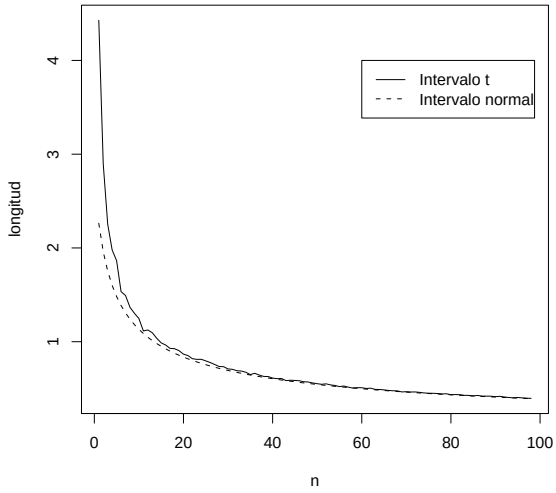

Figura 3.4: *Longitud Estimación de la longitud esperada $E(l)$ del intervalo (3.2.12) para diferentes tamaños de muestra n.*

Bajo el mismo contexto, también podemos comparar el intervalo normal con el intervalo t en términos de la probabilidad de cobertura por medio de estudios de simulación. Se simula 1000 veces $n = 2, \cdots, 100$ datos provenientes de una distribución normal estándar, y para cada conjunto de datos se calculan los intervalos normal y t, y se examina si estos intervalos contienen la media teórica 0. La probabilidad de cobertura real del intervalo normal se calcula como el número de veces que el intervalo contiene a 0 dividido por el número total de iteraciones, 1000. Los resultados de simulación se muestran en la Figura 3.5, donde se observa que la probabilidad de cobertura real de ambos intervalos es muy similar y están cercanos a la probabilidad de cobertura nominal 0.95, aun cuando el tamaño muestral n sea pequeño,

Como vimos anteriormente, el intervalo de confianza para μ se basa en una distribución normal cuando la varianza teórica es conocida y se basa en una distribución t en el otro caso. Por tanto, uno puede cometer el error de afirmar que la distribución de la media muestral $\bar{X}$ es normal si la varianza es conocida, y tiene distribución t si no. El error radica en que la ignorancia acerca del valor de σ no cambia la distribución de los datos y por lo tanto, la distribución de $\bar{X}$ siempre es normal. La que tiene distribución t cuando σ es desconocida es la variable pivote que se basa en $\bar{X}$ y S_{n-1}.

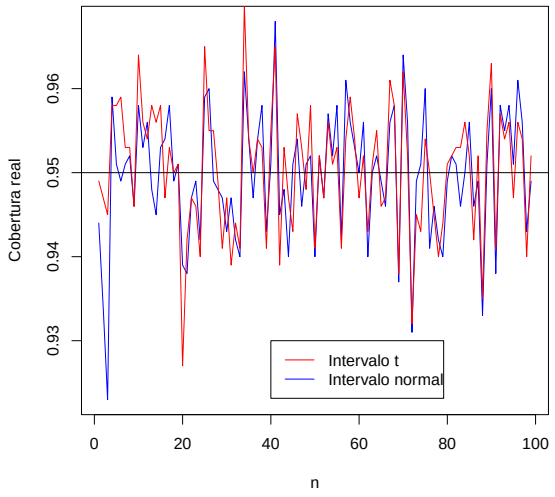

Figura 3.5: *Probabilidad de cobertura real de los intervalos normal y t para diferentes tamaños de muestra n.*

Intervalos de confianza para la varianza σ^2

En esta parte consideramos el problema de estimación por intervalos de confianza de la varianza teórica σ^2 en una distribución normal. Este parámetro es muy importante, en particular, en procesos de producción donde se espera que la varianza sea pequeña para que los productos fabricados sean homogéneos en términos de alguna característica de interés.

Suponga que se tiene $X_1, \cdots, X_n$, una muestra aleatoria proveniente de la distribución $N(\mu, \sigma^2)$. Dado que la distribución normal tiene dos parámetros, la variable pivote para σ^2 depende de si μ es conocido o desconocido, por lo tanto, se consideran los dos casos de forma separada.

Intervalos para σ^2 y σ cuando $\mu = \mu_0$ es conocida

Se ha visto que en este caso, el estimador de máxima verosimilitud de σ^2 es $\sum_{i=1}^n (X_i - \mu_0)^2/n$ y como se ha mencionado anteriormente, se puede encontrar una variable pivote modificando el estimador. Para esto, recordemos que

$$\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sigma^2} \sim \chi^2_n,$$

de donde se observa que la anterior variable depende de σ^2 y su distribución no, de donde concluimos que ésta es una variable pivote para σ^2. Una vez encontrada la

variable pivote, se procede a encontrar percentiles de la distribución de la variable a y b tales que

$$Pr\left(a < \frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{\sigma^2} < b\right) = 1 - \alpha \tag{3.2.14}$$

Recordando que se debe buscar el intervalo para σ^2 que tenga, en lo posible, la menor longitud, entonces despejamos σ^2, se tiene que

$$Pr\left(\frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{b} < \sigma^2 < \frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{a}\right) = 1 - \alpha,$$

cuya longitud dada por

$$l = \sum_{i=1}^{n}(X_i - \mu_0)^2 \left(\frac{1}{a} - \frac{1}{b}\right).$$

De nuevo, l es una variable aleatoria, por lo que se examina su esperanza que está dada por

$$\begin{aligned} E(l) &= \sigma^2 \left(\frac{1}{a} - \frac{1}{b}\right) E\left(\frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{\sigma^2}\right) \\ &= n\sigma^2 \left(\frac{1}{a} - \frac{1}{b}\right). \end{aligned} \tag{3.2.15}$$

Nótese que esta cantidad no depende directamente de $b-a$ como en los intervalos de confianza para μ considerados anteriormente, sino de $1/a - 1/b$, y ésta puede ser grande aun cuando $b-a$ es pequeño, es decir, un intervalo de longitud pequeña para la variable pivote puede conducir a un intervalo de longitud grande para el parámetro σ^2 tal como se ilustra a continuación.

Dado que la distribución χ^2_n es unimodal para $n > 2$, se puede aplicar el Resultado 3.2.1 para encontrar a y b que satisfacen (3.2.14) y que minimizan $b-a$. Estos deben satisfacer que $f(a) = f(b)$ y $a \leq x^* \leq b$ donde $f(\cdot)$ denota la función de densidad de la distribución χ^2_n y x^* es una moda de $f(\cdot)$. El siguiente código de R permite encontrar estos valores a y b.

```
> alpha<-0.05
> n<-10
> x<-seq(0,100,0.01)
> x<-x[-1]
> f<-dchisq(x,n)
> maxi<-which(f==max(f))
> ind<-matrix(NA,maxi,2)
> integrales<-rePr(NA,maxi)
>
> chisq<-function(x){
+ return(dchisq(x,n))
```

```
+ }
> for(i in 1:maxi){
+ ind[i,1]<-i
+ y<-which(abs(f[i]-f[maxi:length(f)])==
min(abs(f[i]-f[maxi:length(f)]))+maxi-1
+ ind[i,2]<-y
+ integrales[i]<-integrate(chisq,x[i],x[y])$value
+ }
>
> val<-which(abs(integrales-(1-0.05))==
+  min(abs(integrales-(1-0.05))))
> x[ind[val,1]]
[1] 2.41
> x[ind[val,2]]
[1] 18.88
```

En el anterior ejemplo, $a = 2.41$ y $b = 18.88$ que corresponden a los percentiles 0.008 y 0.9582 respectivamente, y la longitud del intervalo está dado por 16.47, y $Pr\left(2.41 < \frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\sigma^2} < 18.88\right) = 0.95$. Otra pareja de valores a' y b' pueden ser $a' = \chi^2_{n,\alpha/2}$ y $b' = \chi^2_{n,1-\alpha/2}$, para $n = 10$, $a' = 3.25$ y $b' = 20.48$, dando como resultado $b' - a' = 17.23$ que es mayor que $b - a$. Sin embargo, $1/a' - 1/b' = 0.26$, mientras que $1/a - 1/b = 0.36$. Lo anterior ilustra que el intervalo para σ^2 resultante del intervalo más corto para la variable pivote no es necesariamente el de menor longitud.

La solución al problema de encontrar a y b que satisfacen (3.2.14) y que minimizan $1/a - 1/b$ se puede resolver usando técnicas de minimización. La relación (3.2.14) es equivalente a

$$\int_a^b f(x)dx = 1 - \alpha, \tag{3.2.16}$$

donde $f(\cdot)$ denota la función de densidad de la distribución χ^2_n. Es claro que cuando el valor a cambia, para que la anterior igualdad siga siendo válida b también cambia, por lo tanto, se puede considerar a b como una función de a y lo escribimos como $b = b(a)$. Derivando la cantidad que se desea minimizar $1/a - 1/b$ con respecto a a e igualando a 0, se tiene que $\frac{\partial b}{\partial a} = \frac{b^2}{a^2}$. Ahora derivando ambos lados de (3.2.16) con respecto a a, se tiene que

$$f(b)\frac{\partial b}{\partial a} - f(a) = 0,$$

de donde se tiene que, $a^2 f(a) = b^2 f(b)$. Los valores de a y b que cumplen esta condición serán los que determinan el intervalo de menor longitud para σ^2. El siguiente código de R permite encontrar esos valores a y b para $n = 10$ y $\alpha = 0.05$

```
> n<-10
> alpha<-0.05
```

```
> x<-seq(0,3*n,0.01)
> x<-x[-1]
> f<-dchisq(x,n)
> maxi<-which(f==max(f))
> ind<-matrix(NA,maxi,2)
> integrales<-rePr(NA,maxi)
>
> chisq<-function(x){
+ return(dchisq(x,n))
+ }
>
> for(i in 1:maxi){
+ ind[i,1]<-i
+ aux<-x[-(1:maxi)]^2*f[(maxi+1):length(f)]
+ y<-which(abs(x[i]^2*f[i]-aux)
==min(abs(x[i]^2*f[i]-aux)))+maxi-1
+ ind[i,2]<-y
+ integrales[i]<-integrate(chisq,x[i],x[y])$value
+ }
>
> val<-which(abs(integrales-(1-alpha))==
+  min(abs(integrales-(1-alpha))))
> x[ind[val,1]]
[1] 3.89
> x[ind[val,2]]
[1] 27.24
```

De donde se tiene que $a = 3.89$ y $b = 27.24$, se puede verificar que $3.89^2 f(3.89) = 0.65$ y $27.24^2 f(27.24) = 0.65$. Además $1/3.89 - 1/27.24 = 0.22$, lo cual es menor que las longitudes obtenidas anteriormente. En conclusión, el intervalo con menor longitud usando la variable pivote $\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\sigma^2}$ es

$$IC(\sigma^2) = \left(\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{b}, \frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{a} \right), \tag{3.2.17}$$

donde a y b son tales que $a^2 f(a) = b^2 f(b)$ con f denotando la función de densidad de la distribución χ^2_n. Es claro que encontrar los valores de a y b a veces puede ser complicado, por esta razón, muchas veces se utiliza una alternativa más práctica es solamente tener en cuenta que $Pr\left(a < \frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\sigma^2} < b \right) = 1-\alpha$ e ignorar la longitud. De nuevo existen infitos valores de a y b que satisfacen la anterior igualdad, y por comodidad, se escoge $a = \chi^2_{n,\alpha/2}$ y $b = \chi^2_{n,1-\alpha/2}$. De esta forma, se tiene el siguiente intervalo que es de uso común en la práctica, y que se presenta en la mayoría de los textos de la enseñanza estadística. Lo denotaremos por $IC^*(\sigma^2)$ para hacer una

diferenciación con el intervalo en (3.2.17), y está dado por

$$IC^*(\sigma^2) = \left(\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi^2_{n,1-\frac{\alpha}{2}}}, \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi^2_{n,\frac{\alpha}{2}}} \right) \tag{3.2.18}$$

Teóricamente el intervalo $IC(\sigma^2)$ tiene longitud más corta que el intervalo $IC^*(\sigma^2)$, también se puede calcular la longitud de los dos intervalos y compararlos directamente. Para el intervalo $IC(\sigma^2)$, la longitud esperada está dada por (3.2.15), mientras que para el intervalo $IC^*(\sigma^2)$, la esperanza de la longitud l^* está dada por

$$E(l^*) = n\sigma^2 \left(\frac{1}{\chi^2_{n,\alpha/2}} - \frac{1}{\chi^2_{n,1-\alpha/2}} \right).$$

En la Figura 3.6, se pueden observar los valores de estas dos longitudes esperadas para diferentes valores de n cuando $\sigma^2 = 1$. Nótese que aunque el cómputo de estos dos intervalos depende de la media teórica μ_0, la longitud esperada de ambos intervalos no depende de μ_0. Se puede observar que efectivamente la longitud esperada de IC es más pequeña que la de IC^*; sin embargo, para valores de n grandes, la diferencia es muy pequeña, lo cual se corrobora notando que en ambos intervalos, el tamaño muestral n aparece multiplicando dentro de las longitudes, y por lo tanto, se puede usar el intervalo (3.2.18) sin una pérdida grande de precisión cuando la muestra es grande.

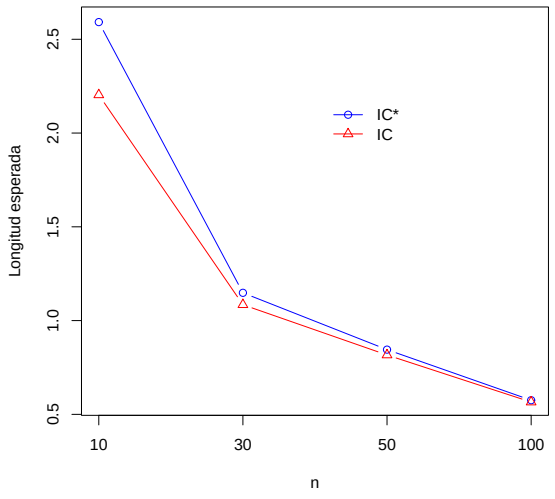

Figura 3.6: *Longitud esperada de los intervalos (3.2.15) y (3.2.16) para σ^2 para diferentes tamaños de muestra.*

Ahora, consideramos los intervalos unilaterales para σ^2 en el mismo escenario cuando $\mu = \mu_0$ es conocido. En primer lugar, para encontrar un intervalo inferior para σ^2, se debe notar que la variable pivote guarda una relación inversamente proporcional con σ^2, entonces se debe buscar un intervalo de la forma $(-\infty, b)$ para la varible pivote, esto es, $Pr\left(\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\sigma^2} < b\right) = 1-\alpha$, de donde se concluye que $b = \chi^2_{n,1-\alpha}$ usando la definición del percentil. Despejando σ^2, se tiene el siguiente intervalo

$$IC(\sigma^2) = \left(\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\chi^2_{n,1-\alpha}}, \infty\right).$$

Análogamente, se puede obtener el intervalo unilateral superior

$$IC(\sigma^2) = \left(-\infty, \frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\chi^2_{n,\alpha}}\right).$$

Teniendo en cuenta que el parámetro σ^2 es siempre positivo, entonces un límite inferior natural es el valor 0, de donde se tiene el siguiente intervalo superior

$$IC(\sigma^2) = \left(0, \frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\chi^2_{n,\alpha}}\right)$$

Ahora, para medir la variación de una población de estudio, la varianza puede ser un poco complicada al momento de la interpretación, puesto que la unidad de la varianza es el cuadrado de la unidad de los datos observados; por ejemplo, para la variable ingreso, si la unidad de medición es pesos, entonces la unidad de la varianza será pesos^2, y esto no solo es complicado para la comprensión de alguien que carece de conocimientos estadísticos, sino para los mismos estadísticos. La solución es utilizar la desviación estándar σ, la cual tiene la misma unidad que los datos observados, y facilita la interpretación; por ejemplo, si la desviación estándar de la variable ingreso es de 100 mil pesos, entonces podemos afirmar que en promedio, los ingresos se difieren entre ellos por un monto de 100 mil pesos, el cual es una interpretación directa y de fácil comprensión.

Dado lo anterior, estamos interesados en encontrar intervalos de confianza para σ, y estos se encuentran directamente usando los intervalos para σ^2 y el Resultado 3.2.2 donde la función g es la función raíz cuadrada, la cual es uno a uno y estrictamente creciente. Como se había visto anteriormente, tenemos dos intervalos bilaterales para σ^2, por lo tanto, podemos encontrar dos intervalos unilaterales para σ, dados por

$$IC(\sigma) = \left(\sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{b}}, \sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{a}}\right), \tag{3.2.19}$$

donde a y b son tales que $a^2 f(a) = b^2 f(b)$ con f la función de densidad de la distribución χ^2_n, estos valores se pueden encontrar con el código en R presentado

anteriormente. Otro intervalo bilateral para σ se puede obtener de (3.2.18) y está dado por

$$IC^*(\sigma) = \left(\sqrt{\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi^2_{n,1-\frac{\alpha}{2}}}}, \sqrt{\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi^2_{n,\frac{\alpha}{2}}}} \right) \tag{3.2.20}$$

Tal como se había visto anteriormente, el intervalo $IC(\sigma^2)$ es más preciso que $IC^*(\sigma^2)$ pero esto no garantiza que $IC(\sigma)$ también sea más preciso que $IC^*(\sigma)$. La comparación teórica de los dos intervalos en términos de la longitud esperada no es trivial y por consiguiente recurrimos a las simulaciones. En la Figura 3.7 se observa las estimaciones de las dos longitudes esperadas basadas en 1000 muestras simuladas de una distribución normal estándar, podemos observar que el intervalo más corto para σ^2 también indujo un intervalo preciso para σ, aunque, una vez más, en muestras grandes, podemos usa el intervalo (3.2.20), el cual se puede computar de manera fácil.

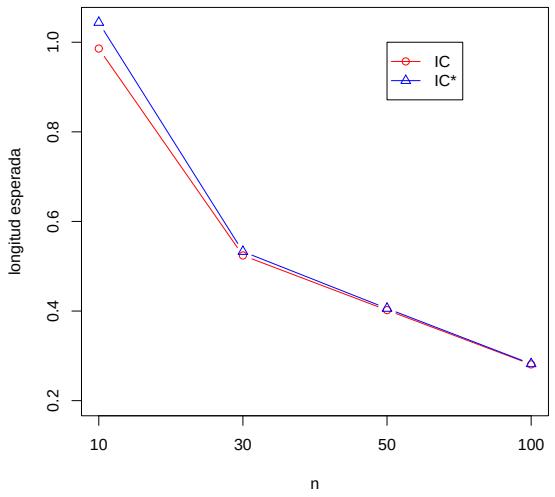

Figura 3.7: *Longitud esperada estimada de los intervalos (3.2.19) y (3.2.20) para σ para diferentes tamaños de muestra.*

Aplicando de nuevo el Resultado 3.2.2, tenemos los siguiente intervalos unilaterales para σ.

$$IC(\sigma) = \left(\sqrt{\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi^2_{n,1-\alpha}}}, \infty \right).$$

y

$$IC(\sigma) = \left(0, \sqrt{\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi^2_{n,\alpha}}} \right)$$

Intervalos para σ^2 y σ cuando μ es desconocida

Cuando la media teórica μ es desconocida, el estimador de máxima verosimilitud de σ^2 es $S_n^2 = \sum_{i=1}^n (X_i - \bar{X})^2/n$ y usando la propiedad

$$\frac{nS_n^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1},$$

se tiene que esta variable depende de σ^2 y no de μ, además su distribución no depende de σ^2 y μ, de donde concluimos que ésta es una variable pivote para σ^2. Una vez más, debemos encontrar valores a y b tales que $Pr\left(a < \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} < b\right) = 1 - \alpha$ para luego encontrar un intervalo bilateral para σ^2. Similar al caso donde μ es conocido, los valores a y b que minimizan la longitud del intervalo son aquellos con $a^2 f(a) = b^2 f(b)$ donde $f(\cdot)$ denota la función de densidad de la distribución χ^2_{n-1}. Con el código de R presentado anteriormente modificando adecuadamente el grado de libertad de la distribución, se pueden encontrar estos valores. De esta forma, se tiene el siguiente intervalo para σ^2

$$\begin{aligned} IC(\sigma^2) &= \left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{b}, \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{a}\right) \\ &= \left(\frac{(n-1)S_{n-1}^2}{b}, \frac{(n-1)S_{n-1}^2}{a}\right) \end{aligned} \tag{3.2.21}$$

Otra alternativa más práctica es escoger a y b como los percentiles $\alpha/2$ y $1-\alpha/2$ de la distribución de la variable pivote, esto es, $a = \chi^2_{n-1,\alpha/2}$ y $b = \chi^2_{n-1,1-\alpha/2}$, y obtenemos el siguiente intervalo para σ^2

$$\begin{aligned} IC^*(\sigma^2) &= \left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) \\ &= \left(\frac{(n-1)S_{n-1}^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S_{n-1}^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) \end{aligned} \tag{3.2.22}$$

Análogo al caso cuando $\mu = \mu_0$ es conocida, se puede ver que la longitud esperada de (3.2.22) es siempre mayor que la de (3.2.21), pero esta diferencia se hace cada vez más pequeña cuando el tamaño muestral es grande, y por consiguiente, podemos utilizar indistintamente los dos intervalos.

Nótese adicionalmente que el intervalo (3.2.22) se basa en el estimador S_{n-1}^2 ampliando a la izquierda y derecha multiplicando por $(n-1)/\chi^2_{n-1,1-\frac{\alpha}{2}}$ y $(n-1)/\chi^2_{n-1,\frac{\alpha}{2}}$ respectivamente, y por consiguiente no es un intervalo simétrico con respecto al estimador.

Se deja como ejercicio el procedimiento para encontrar los siguientes intervalos unilaterales para σ^2 (Ejercicio 3.4),

$$IC(\sigma^2) = \left(\frac{(n-1)S_{n-1}^2}{\chi^2_{n-1,1-\alpha}}, \infty \right), \tag{3.2.23}$$

y

$$IC(\sigma^2) = \left(0, \frac{(n-1)S_{n-1}^2}{\chi^2_{n-1,\alpha}} \right). \tag{3.2.24}$$

Utilizando los anteriores intervalos para σ^2, podemos obtener fácilmente los siguientes intervalos para la desviación estándar σ.

$$IC(\sigma) = \left(\sqrt{\frac{(n-1)S_{n-1}^2}{b}}, \sqrt{\frac{(n-1)S_{n-1}^2}{a}} \right)$$

donde a y b son tales que $a^2 f(a) = b^2 f(b)$ con $f(\cdot)$ denotando la función de densidad de la distribución χ^2_{n-1}. Otro intervalo para σ está dado por

$$IC^*(\sigma) = \left(\sqrt{\frac{(n-1)S_{n-1}^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}}, \sqrt{\frac{(n-1)S_{n-1}^2}{\chi^2_{n-1,\frac{\alpha}{2}}}} \right), \tag{3.2.25}$$

En muestras pequeñas el intervalo $IC(\sigma)$ será mas preciso que $IC^*(\sigma)$, mientras que en muestras grandes, no hay una gran diferencia entre estos dos intervalos en términos de la precisión. Por otro lado, los intervalos unilaterales para σ están dados por

$$IC(\sigma) = \left(\sqrt{\frac{(n-1)S_{n-1}^2}{\chi^2_{n-1,1-\alpha}}}, \infty \right),$$

y

$$IC(\sigma) = \left(0, \sqrt{\frac{(n-1)S_{n-1}^2}{\chi^2_{n-1,\alpha}}} \right).$$

Ilustramos el uso de estos intervalos en el siguiente ejemplo.

Ejemplo 3.2.5. *Retomamos el Ejemplo 2.3.6 donde se considera una línea de producción de láminas de vidrio templado de grosor de 3 cm, se dispone del grosor de 12 láminas producidas por esta línea. La varianza muestral está dada por* $s^2_{n-1} = 0.1068\, cm^2$. *Y los valores de* a *y* b *que satisfacen* $a^2 f(a) = b^2 f(b)$ *con* f *la función de densidad de* χ^2_{n-1} *corresponden a* $a = 4.51$ *y* $b = 28.45$, *y el intervalo (3.2.21) resultante es* $(0.04, 0.26)$, *cuya longitud es de 0.22. Por otro lado, el intervalo (3.2.22) es* $(0.05, 0.31)$ *que tiene como longitud 0.26, y por consiguiente menos preciso que el intervalo (3.2.21).*

Suponga que una línea de producción de vidrio templado de grosor de 3 cm es aceptable si la diferencia promedio de grosor de las láminas producidas no supera a

0.2 cm, esto es σ no debe ser mayor que 0.2 cm. Para saber si la línea en cuestión tiene desviación superior a los 0.2 cm, calculamos el intervalo inferior para σ, el cual está dado por $(0.24, \infty)$ de donde se observa que incluso el límite inferior de σ es mayor que 0.2 cm, y se concluye que la línea de producción tiene una desviación estándar mayor que lo establecido, y por consiguiente, la calidad no es aceptable.

Intervalos de confianza para el coeficiente de variación

Cuando se desea comparar dos poblaciones en términos de la dispersión, si los dos medias teóricas son muy diferentes, no es apropiado usar la varianza o la desviación como medida de dispersión, puesto que éstas dependen de la magnitud de las medias teóricas; otra desventaja de la varianza o la dispersión es que depende de la unidad con que fue medida la variable de estudio. Según eso, si la variable de estudio tiene unidades diferentes en dos poblaciones, por ejemplo, la variable ingreso mensual por familia en Colombia se medirá en pesos colombianos mientras que en Perú se medirá en nuevos soles, las dos desviaciones estarán en unidades de pesos colombianos y soles, respectivamente, y por consiguiente no podrán ser comparados directamente. Una alternativa es el uso del coeficiente de variación definida como $cv = \sigma/\mu$, el cual está libre de unidad de medición y puede ser interpretado como porcentaje. En el capítulo anterior, se vio que bajo normalidad, el estimador de máxima verosimilitud de cv es $S_n/\bar{X}$, y estamos interesados en hallar un intervalo de confianza para cv.

Es claro que si la media teórica o la desviación teórica es conocida, se pueden obtener intervalos para el coeficiente de variación simplemente aplicando el Resultado 3.2.2, por ejemplo, si la desviación teórica es conocida, entonces se tienen el siguiente intervalo bilateral para cv.

$$IC(cv) = \left(\frac{\sigma_0}{\bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}}, \frac{\sigma_0}{\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}} \right).$$

También, si se usa un intervalo t para μ, tenemos el siguiente intervalo para cv

$$IC(cv) = \left(\frac{\sigma_0}{\bar{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{S_{n-1}}{\sqrt{n}}}, \frac{\sigma_0}{\bar{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{S_{n-1}}{\sqrt{n}}} \right).$$

Para muestras grandes, no debe haber una diferencia grande entre los dos intervalos en términos de la longitud. Se deja como ejercicio escribir los intervalos unilaterales cuando σ es conocido y los bilaterales y unilaterales cuando μ es conocido (Ejercicio 3.9).

Cuando ambos parámetros teóricos son desconocidos, es muy difícil utilizar el método de la variable pivote para encontrar un intervalo de confianza para cv, ya que es difícil construir una variable en base de $S_n/\bar{X}$ que dependiera de cv y que a la vez tenga distribución no dependiente de los parámetros. Por esta razón, se proponen tres intervalos intuitivos para cv

Propuesta 1

La primera propuesta es utilizar un intervalo para la desviación estándar σ incorporando el estimador de μ. El intervalo propuesto es como sigue

$$IC_1(cv) = \begin{cases} \left(\dfrac{\sqrt{(n-1)S_{n-1}^2/\chi_{n-1,1-\alpha/2}^2}}{\bar{X}}, \dfrac{\sqrt{(n-1)S_{n-1}^2/\chi_{n-1,\alpha/2}^2}}{\bar{X}} \right), & \text{si } \bar{X} > 0 \\ \left(\dfrac{\sqrt{(n-1)S_{n-1}^2/\chi_{n-1,\alpha/2}^2}}{\bar{X}}, \dfrac{\sqrt{(n-1)S_{n-1}^2/\chi_{n-1,1-\alpha/2}^2}}{\bar{X}} \right), & \text{si } \bar{X} < 0 \end{cases}$$

Propuesta 2

La segunda propuesta se basa en el intervalo t student para μ y además se considera el estimador de σ. El intervalo propuesto es como sigue

$$IC_2(cv) = \left(\frac{S_{n-1}}{\bar{X} + t_{n-1,1-\alpha/2} S_{n-1}/\sqrt{n}}, \frac{S_{n-1}}{\bar{X} - t_{n-1,1-\alpha/2} S_{n-1}/\sqrt{n}} \right).$$

Este intervalo tiene una desventaja en comparación con el intervalo $IC_1(cv)$, puesto que cuando el límite inferior del intervalo t para μ es negativo, pero el límite superior es positivo, el límite superior del intervalo resultante para cv será menor que el límite inferior, y por consiguiente el intervalo carece de utilidad práctica. además que la longitud del intervalo será negativa.

Propuesta 3

La tercera propuesta es utilizar simultáneamente los intervalos bilaterales $IC(\mu)$ y $IC(\sigma)$ para crear un intervalo para el coeficiente de variación, y tenemos el siguiente intervalo

$$IC_3(cv) = \left(\frac{\sqrt{(n-1)S_{n-1}^2/\chi_{n-1,1-\alpha/2}^2}}{\bar{X} + t_{n-1,1-\alpha/2} S_{n-1}/\sqrt{n}}, \frac{\sqrt{(n-1)S_{n-1}^2/\chi_{n-1,\alpha/2}^2}}{\bar{X} - t_{n-1,1-\alpha/2} S_{n-1}/\sqrt{n}} \right).$$

Este intervalo, por su construcción, debe tener una probabilidad de cobertura mayor que los dos anteriores intervalos. Pero al mismo tiempo, se espera que tenga también una longitud mayor. Adicionalmente, este intervalo también puede tener el mismo problema que el intervalo IC_2 en el sentido de que el límite superior puede ser menor que el límite inferior.

Los anteriores tres intervalos fueron construidos usando simplemente la intuición y no mediante algún procedimiento que garantizara que el nivel de confianza sea $100 \times (1-\alpha)\,\%$. Por esta razón, se realiza un estudio de simulación para verificar que la probabilidad de cobertura real sea cercana al $1-\alpha$. Se simularon 1000 muestras provenientes de una distribución normal para valores del coeficiente de variación iguales a 0.2, 0.4, 0.6 y 0.8 [6] y para cada uno de los tres intervalos propuestos se calculó la

[6] La desviación teórica se fijó en 1, y la media teórica varía según el coeficiente de variación.

probabilidad de cobertura real como el número de veces que el intervalo contuvo al coeficiente de variación dividido por 1000, y se calculó la longitud estimada como el promedio de longitud de los intervalos calculados. La probabilidad de cobertura real de los tres intervalos se muestra en la Figura 3.8, donde se observa que el intervalo IC_2 tiene una probabilidad de cobertura muy por debajo del nivel de confianza nominal, mientras que el intervalo IC_1 tiene un desempeño mejor en términos de la probabilidad de cobertura. Finalmente, el intervalo con mayor probabilidad de cobertura es el intervalo IC_3 tal como se sospechaba anteriormente.

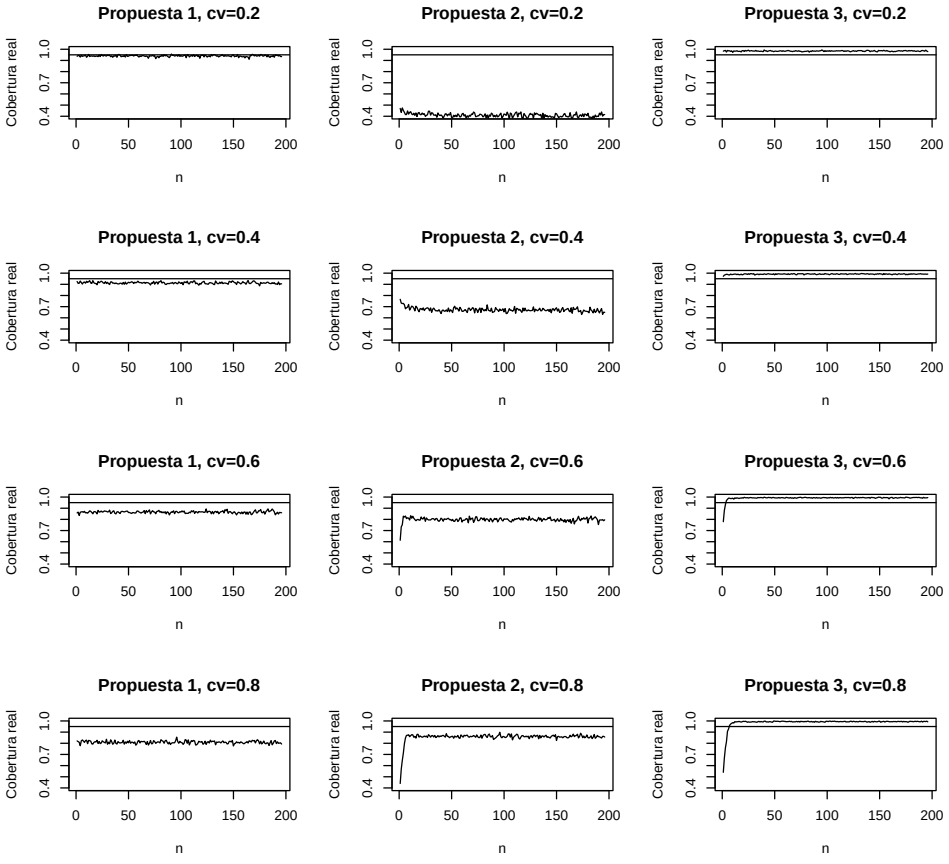

Figura 3.8: *Probabilidades de cobertura para las tres propuestas de intervalo de confianza para el coeficiente de variación.*

Para investigar si la superioridad de los intervalos IC_1 y IC_3 es debido a la gran longitud, se examina en términos de la longitud. Estas longitudes se muestran en la Figura 3.9, en la cual se observa que el intervalo con mayor longitud es IC_3, de donde podemos afirmar que éste tiene la mayor probabilidad de cobertura debido a que es un intervalo muy ancho, y por consiguiente, menos preciso. Por otro lado, el intervalo

IC_1 tiene un comportamiento muy estable en términos de la longitud, pues en primer lugar, la longitud es siempre positiva por la forma como se definió el intervalo, y en segundo lugar, tiene una longitud aceptable entre los tres intervalos propuestos. Como conclusión, se recomienda el intervalo IC_1 para el coeficiente de variación cv en una muestra proveniente de una distribución normal.

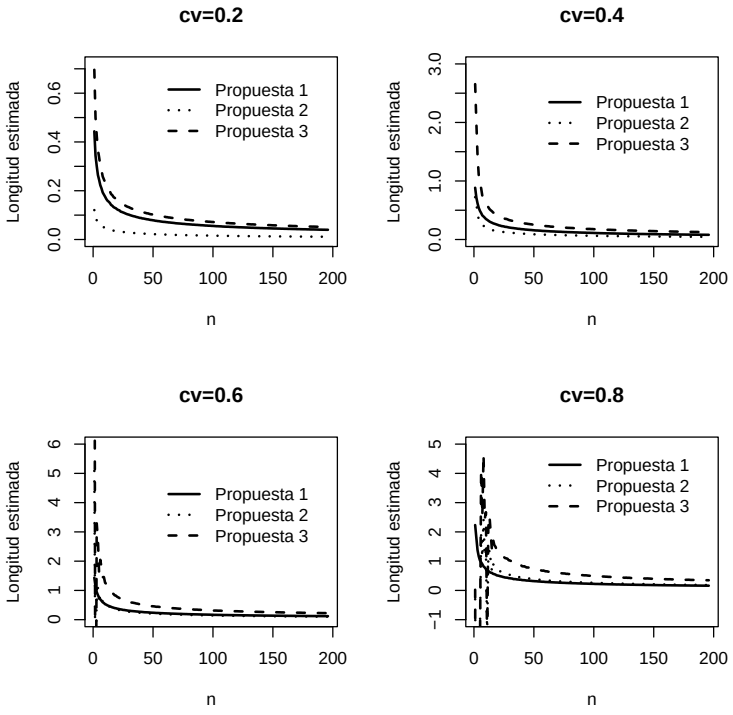

Figura 3.9: *Longitudes estimadas para las tres propuestas de intervalo de confianza para el coeficiente de variación.*

Ejemplo 3.2.6. *Volviendo al problema donde se estudia una variable que se mide en dos unidades diferentes en diferentes poblaciones, suponga que para la variable de estudio ingreso por persona medida en Colombia y Perú, para Colombia, en una muestra de 100 personas, el promedio y la desviación estándar muestral fueron 635000 pesos y 150000 pesos, respectivamente; mientras que en Perú, en una muestra de 70 personas, el promedio y la desviación muestral fueron 935 nuevos soles y 285 nuevos soles, respectivamente. Si se desea analizar estos datos para comparar el ingreso de los colombianos y los peruanos, calculamos el intervalo IC_1 para los dos coeficientes de variación de ambos países. Podemos calcular los dos intervalos usando el siguiente comando en R, y para los datos de Colombia, el intervalo para cv es $(0.21, 0.27)$, mientras que para los datos de Perú, el intervalo para cv es $(0.27, 0.35)$. De donde se puede sospechar que los peruanos son más dispersos en términos del ingreso.*

```
> mc<-635000
> sc<-150000
> mp<-935
> sp<-285
> inf_c<-sqrt((100-1)*(sc^2)/qchisq(0.975,99))/mc
> sup_c<-sqrt((100-1)*(sc^2)/qchisq(0.025,99))/mc
> inf_c
[1] 0.2074032
> sup_c
[1] 0.2744115
> inf_p<-sqrt((100-1)*(sp^2)/qchisq(0.975,99))/mp
> sup_p<-sqrt((100-1)*(sp^2)/qchisq(0.025,99))/mp
> inf_p
[1] 0.2676278
> sup_p
[1] 0.3540935
```

3.2.2 Problemas de dos muestras

En este apartado del libro, se consideran situaciones donde se encuentran dos poblaciones de estudio independientes que pueden ser descritas adecuadamente por distribuciones normales. Por ejemplo, los dos institutos enunciados en el Ejemplo 2.3.12, podemos compararlos en términos del rendimiento obtenido por los respectivos alumnos, esto es, compararlos en términos de las medias teóricas. También podemos compararlos en términos de la homogeneidad de los rendimientos, esto es, compararlos en términos de las varianzas teóricas.

Los supuestos bajo los cuales se desarrollan las teorías en esta parte son: se tienen dos muestras aleatorias de tamaño n_X y n_Y denotados por $X_1, \ldots, X_{n_X}$ y $Y_1, \ldots, Y_{n_Y}$ provenientes de $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$, respectivamente. Además se supone que las dos muestras son independientes, esto es, cualquier conjunto de variables X, es independiente de cualquier conjunto de variables Y. Por consiguiente, cualquier estadística construida en la primera muestra será independiente de cualquier estadística construida en la segunda muestra. A continuación desarrollaremos intervalos de confianza para $(\mu_X - \mu_Y)$ y σ_X^2/σ_Y^2.

Intervalos de confianza para diferencia de medias

Para encontrar un intervalo de confianza para la diferencia de ellos $\mu_X - \mu_Y$, el método de la variable pivote seguirá siendo útil aquí, y para encontrar una variable pivote, se tendrá en cuenta un estimador natural para $\mu_X - \mu_Y$ es $\bar{X} - \bar{Y}$ cuya distribución se da a continuación:

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right). \tag{3.2.26}$$

En problemas de una muestra, cuando se estudió los intervalos de confianza para μ, se vio que estos dependen de si la varianza teórica es conocida o no. En caso afirmativo, el intervalo se basa en una distribución normal estándar, y en caso negativo, una distribución t student. En el contexto de dos muestras, el intervalo para $\mu_X - \mu_Y$ también depende de las varianzas teóricas, σ_X^2 y σ_Y^2, y tenemos los siguientes casos:

σ_X^2 y σ_Y^2 son conocidas.

Teniendo en cuenta (3.2.26), se tiene que

$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} \sim N(0,1),$$

la anterior variable depende del parámetro de interés $\mu_X - \mu_Y$, y su distribución no, de donde se tiene que ésta es una variable pivote para $\mu_X - \mu_Y$. Entonces, siguiendo los lineamientos del método de la variable pivote, se procede a buscar valores a y b con

$$Pr\left(a < \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} < b\right) = 1-\alpha. \tag{3.2.27}$$

Como la función de densidad de la distribución normal es unimodal, entonces por el Resultado 3.2.1, se tiene que los valores de a y b que cumplen (3.2.27) y que minimizan $b-a$ están dados por $a = -z_{1-\alpha/2}$ y $b = z_{1-\alpha/2}$. Para conocer si estos valores también minimizan la longitud del intervalo resultante para μ, despejamos μ en (3.2.27), y tenemos que

$$Pr\left(\bar{X}-\bar{Y}-b\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}} < \mu_X-\mu_Y < \bar{X}-\bar{Y}-a\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}\right) = 1-\alpha, \tag{3.2.28}$$

cuya longitud está dada por $l = (b-a)\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$, y por consiguiente valores de a y b que minimizan $(b-a)$ también minimizan l, y tenemos que el siguiente intervalo bilateral de menor longitud de $100\times(1-\alpha)\,\%$ dado por

$$IC(\mu_X-\mu_Y) \tag{3.2.29}$$

$$= \left(\bar{X}-\bar{Y}-z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}, \bar{X}-\bar{Y}+z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}\right). \tag{3.2.30}$$

Acerca del anterior intervalo, podemos observar, en primer lugar, que éste es un intervalo simétrico, entonces en la práctica, una vez conocido el intervalo, se puede conocer la estimación $\bar{x}-\bar{y}$ como el promedio del límite inferior y superior. Por otro lado, la longitud de este intervalo es una constante y está dada por

$$l = 2z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}},$$

de tal forma que cuando los tamaños de las dos muestras son grandes, se incrementa la precisión del intervalo resultante. Adicionalmente, observe que cuando las dos varianzas teóricas son pequeñas, el intervalo resultante también tendrá una longitud pequeña. Por otro lado, los intervalos unilaterales para $\mu_X - \mu_Y$ están dados por

$$IC(\mu_X - \mu_Y) = \left(-\infty, \bar{X} - \bar{Y} + z_{1-\alpha}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right)$$

y

$$IC(\mu_X - \mu_Y) = \left(\bar{X} - \bar{Y} - z_{1-\alpha}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}, \infty \right)$$

Una consecuencia inmediata al aplicar el Resultado 3.2.2 conduce a los siguientes intervalos de confianza para $\mu_Y - \mu_X$

$$IC(\mu_Y - \mu_X) = \left(\bar{Y} - \bar{X} - z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}, \bar{Y} - \bar{X} + z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right), \tag{3.2.31}$$

$$IC(\mu_Y - \mu_X) = \left(-\infty, \bar{Y} - \bar{X} + z_{1-\alpha}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right)$$

y

$$IC(\mu_Y - \mu_X) = \left(\bar{Y} - \bar{X} - z_{1-\alpha}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}, \infty \right)$$

Los anteriores intervalos no son muy usados en la práctica, puesto que en la mayoría de los casos de estudio, no se conocen los valores de las varianzas teóricas, y el anterior intervalo ya no será aplicable. A continuación estudiamos el caso cuando las varianzas teóricas son desconocidas, pero se pueden asumir iguales.

σ_X^2 y σ_Y^2 son desconocidas, pero iguales.

Cuando las varianzas teóricas son desconocidas, la variable $\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}$ ya no puede ser una variable pivote para $\mu_X - \mu_Y$, puesto que ésta depende de σ_X^2 y σ_Y^2, y estas varianzas son desconocidas. Sin embargo, podemos modificarla para construir una variable pivote. En primer lugar, al suponer que las dos varianzas teóricas son iguales, tenemos que $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ y

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}}} \sim N(0, 1). \tag{3.2.32}$$

Por otro lado, los estimadores insesgados para las varianzas teóricas son $S^2_{n_X-1,X}$ y $S^2_{n_Y-1,Y}$. Y se tienen las siguientes distribuciones

$$\frac{(n_X-1)S^2_{n_X-1,X}}{\sigma^2_X} = \frac{\sum_{i=1}^{n_X}(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2_{n_X-1}$$

y

$$\frac{(n_Y-1)S^2_{n_Y-1,Y}}{\sigma^2_Y} = \frac{\sum_{j=1}^{n_Y}(Y_i-\bar{Y})^2}{\sigma^2} \sim \chi^2_{n_Y-1}.$$

Usando el hecho de que las dos muestras son independientes y el Resultado 1.1.25 se tiene que

$$\frac{\sum_{i=1}^{n_X}(X_i-\bar{X})^2+\sum_{j=1}^{n_Y}(Y_i-\bar{Y})^2}{\sigma^2} \sim \chi^2_{n_x+n_Y-2},$$

esto es

$$\frac{(n_X-1)S^2_{n_X-1,X}+(n_Y-1)S^2_{n_Y-1,Y}}{\sigma^2} \sim \chi^2_{n_x+n_Y-2}. \tag{3.2.33}$$

Usando las propiedades (3.2.32), (3.2.33) y teniendo en cuenta que $\bar{X}-\bar{Y}$ es independiente de $(n_X-1)S^2_{n_X-1,X}+(n_Y-1)S^2_{n_Y-1,Y}$, se tiene que

$$\frac{\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}}}{\sqrt{\frac{(n_X-1)S^2_{n_X-1,X}+(n_Y-1)S^2_{n_Y-1,Y}}{(n_X+n_Y-2)\sigma^2}}} \sim t_{n_X+n_Y-2}. \tag{3.2.34}$$

Considerando que las dos muestras provienen de distribuciones con la misma varianza teórica σ^2, se pueden usar las variables de ambas muestras para estimar la varianza común σ^2. En este caso, tenemos que $\frac{(n_X-1)S^2_{n_X-1,X}+(n_Y-1)S^2_{n_Y-1,Y}}{n_X+n_Y-2}$ es un estimador insesgado para σ^2 (Ejercicio 3.5), a la cual se denomina la varianza combinada en inglés *pooled variance*, denotada por S^2_p. De esta forma (3.2.34) se convierte en

$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{S_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \sim t_{n_X+n_Y-2}. \tag{3.2.35}$$

Se puede verificar fácilmente que la anterior estadística es una variable pivote para el parámetro de interés $\mu_X-\mu_Y$. Siguiendo los mismos pasos en la construcción de $IC(\mu)$ cuando σ^2 es desconocida, se tiene el siguiente intervalo de confianza más preciso de $100\times(1-\alpha)\,\%$ dado por

$$IC(\mu_X-\mu_Y) = \bar{X}-\bar{Y} \pm t_{n_X+n_Y-2,1-\alpha/2}S_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}.$$

Este intervalo es de nuevo simétrico, y la longitud es aleatoria, dada por $l = 2t_{n_X+n_Y-2,1-\alpha/2}S_p\sqrt{\frac{1}{n}_X + \frac{1}{n_Y}}$. Usando la misma variable pivote, se puede encontrar los siguientes intervalos unilaterales para $\mu_X - \mu_Y$

$$IC(\mu_X - \mu_Y) = \left(-\infty, \bar{X} - \bar{Y} + t_{n_X+n_Y-2,1-\alpha}S_p\sqrt{\frac{1}{n}_X + \frac{1}{n_Y}}\right)$$

y

$$IC(\mu_X - \mu_Y) = \left(\bar{X} - \bar{Y} - t_{n_X+n_Y-2,1-\alpha}S_p\sqrt{\frac{1}{n}_X + \frac{1}{n_Y}}, \infty\right)$$

Para el parámetro $\mu_Y - \mu_X$, aplicando el Resultado 3.2.2, tenemos los siguientes intervalos

$$IC(\mu_Y - \mu_X) = \bar{Y} - \bar{X} \pm t_{n_X+n_Y-2,1-\alpha/2}S_p\sqrt{\frac{1}{n}_X + \frac{1}{n_Y}},$$

$$IC(\mu_Y - \mu_X) = \left(-\infty, \bar{Y} - \bar{X} + t_{n_X+n_Y-2,1-\alpha}S_p\sqrt{\frac{1}{n}_X + \frac{1}{n_Y}}\right)$$

y

$$IC(\mu_Y - \mu_X) = \left(\bar{Y} - \bar{X} - t_{n_X+n_Y-2,1-\alpha}S_p\sqrt{\frac{1}{n}_X + \frac{1}{n_Y}}, \infty\right)$$

Ejemplo 3.2.7. *Para los datos del Ejemplo 2.3.12, estamos interesados en comparar los dos institutos en términos del desempeño promedio de los alumnos de los dos institutos. Para poder utilizar los intervalos dados anteriormente, se debe garantizar que las dos varianzas teóricas son iguales, más adelante se estudiarán procedimientos para validar este supuesto, por ahora diremos que el supuesto es válido, puesto que las varianzas muestrales son muy parecidas,* $S_{n_X-1,X} = 8.28$ *y* $S_{n_Y-1,Y} = 8.56$. *Para calcular el intervalo bilateral para la diferencia de los promedios, podemos utilizar los siguientes comandos en R.*

```
> A<-c(75, 87, 83, 73, 74, 88, 88, 74, 64, 92, 73, 87, 91, 83,84)
> B<-c(64, 85, 72, 64, 74, 93, 70, 79, 79, 75, 66, 83 ,74)
> alpha<-0.05
> nx<-length(A)
> ny<-length(B)
> Sp<-sqrt(((nx-1)*var(A)+(ny-1)*var(B))/(nx+ny-2))
> lim.sup<-mean(A)-mean(B)+qt(1-alpha/2,nx+ny-2)*Sp*sqrt(1/nx+1/ny)
> lim.inf<-mean(A)-mean(B)-qt(1-alpha/2,nx+ny-2)*Sp*sqrt(1/nx+1/ny)
> lim.inf
[1] -0.7116975
> lim.sup
[1] 12.38349
```

O equivalentemente usando la función `t.test` *con la opción* `var.equal=T`.

```
> t.test(A,B, var.equal=T)

        Two Sample t-test

data:  A and B
t = 1.8321, df = 26, p-value = 0.07842
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7116975 12.3834924
sample estimates:
mean of x mean of y
 81.06667  75.23077
```

Observe que el intervalo de confianza del 95 % está dado por $(-0.71, 12.38)$*, el cual contiene tanto valores positivos como negativos, entonces si el rendimiento promedio de los alumnos del instituto* A *y* B *se denotan por* μ_X *y* μ_Y*, respectivamente, podemos ver que* $\mu_A - \mu_B$ *puede ser positivo o negativo, indicando que no hay una diferencia significativa entre los alumnos de los dos institutos en términos del rendimiento. Sin embargo, observando más detalladamente el intervalo* $(-0.71, 12.38)$*, podemos ver que la gran mayoría del intervalo se ubica en el eje positivo, de donde surge la inquietud de que el instituto A tiene un desempeño superior al del instituto B. Para verificar si los datos apoyan la afirmación de que* $\mu_A > \mu_B$ *equivalente a* $\mu_A - \mu_B > 0$*, calculamos el intervalo superior en R de la siguiente forma*

```
> lim.sup<-mean(A)-mean(B)+qt(1-alpha,nx+ny-2)*Sp*sqrt(1/nx+1/ny)
> lim.sup
[1] 11.2689
```

de donde se tiene que el intervalo superior del 95 % es $(-\infty, 11.26)$ *y por tanto, los datos muestran evidencias que apoyan la superioridad del instituto A comparado con el instituto B en términos del rendimiento de los alumnos.*

Finalmente, hacemos la anotación de que el cambio del nivel de confianza puede afectar sobre la decisión que se toma con respecto a los parámetros teóricos. Si calculamos un intervalo más preciso, por ejemplo, el intervalo de 90 %, tenemos

```
> alpha<-0.1
> lim.sup<-mean(A)-mean(B)+qt(1-alpha/2,nx+ny-2)*Sp*sqrt(1/nx+1/ny)
> lim.inf<-mean(A)-mean(B)-qt(1-alpha/2,nx+ny-2)*Sp*sqrt(1/nx+1/ny)
> lim.inf
[1] 0.4028956
> lim.sup
[1] 11.2689
```

Observe que el intervalo del 90 % es $(0.20, 11.47)$*, el cual solo contiene valores positivos, indicando que el desempeño del instituto A sí es superior comparado con el instituto B. De lo anterior podemos ver que en algunas situaciones, el valor de* α *puede*

afectar sobre la conclusión acerca de los parámetros teóricos. Discutiremos más sobre estas situaciones en el siguiente capítulo.

Para los datos del ejemplo anterior, se asumió válido el supuesto de que las dos varianzas teóricas son iguales. Se debe analizar los datos para verificar que los datos efectivamente apoyan a esta afirmación, más adelante se estudiará la construcción de intervalos de confianza para el cociente de varianzas, y esto nos permite corroborar o refutar este supuesto.

σ_X^2 y σ_y^2 son desconocidas y diferentes.

En este caso, consideramos de nuevo la distribución de la estadística $\bar{X} - \bar{Y}$ dada en (3.2.26), cuya varianza está dada por $\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$, que puede ser estimada insesgadamente mediante la estadística $S_D^2 = \frac{S_{n_X-1,X}^2}{n_X} + \frac{S_{n_Y-1,Y}^2}{n_Y}$. De esta forma, se tiene un candidato como variable pivote para $\mu_X - \mu_Y$ la siguiente estadística

$$D = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_{n_X-1,X}^2}{n_X} + \frac{S_{n_Y-1,Y}^2}{n_Y}}}. \tag{3.2.36}$$

Es claro que la anterior estadística depende del parámetro de interés $\mu_X - \mu_Y$, y no de los demás parámetros desconocidos σ_X^2 y σ_Y^2. Sin embargo, la distribución de esta estadística depende del cociente σ_X^2/σ_Y^2. Cuando los tamaños muestrales n_X y n_Y son grandes, D tiene distribución normal estándar aproximadamente (Bickel & Doksum 2001). Utilizando este hecho, se pueden construir los siguientes intervalos aproximados para $\mu_X - \mu_Y$.

$$IC(\mu_X - \mu_Y) = \bar{X} - \bar{Y} \pm z_{1-\alpha/2}\sqrt{\frac{S_{n_X-1,X}^2}{n_X} + \frac{S_{n_Y-1,Y}^2}{n_Y}}$$

$$IC(\mu_X - \mu_Y) = \left(-\infty, \bar{X} - \bar{Y} + z_{1-\alpha}\sqrt{\frac{S_{n_X-1,X}^2}{n_X} + \frac{S_{n_Y-1,Y}^2}{n_Y}}\right)$$

y

$$IC(\mu_X - \mu_Y) = \left(\bar{X} - \bar{Y} - z_{1-\alpha}\sqrt{\frac{S_{n_X-1,X}^2}{n_X} + \frac{S_{n_Y-1,Y}^2}{n_Y}}, \infty\right)$$

Por otro lado, para tamaños muestrales pequeños o moderados, se debe hacer uso de la aproximación de Welch (1949) definida como sigue: sea $c = (s_{n_X-1,X}^2/n_X)/s_D^2$, entonces se tiene que una distribución aproximada para D es la distribución t_k con k el entero más cercano a

$$\left[\frac{c^2}{n_X - 1} + \frac{(1-c)^2}{n_Y - 1}\right]^{-1} \tag{3.2.37}$$

De esta forma, tenemos los siguientes intervalos para $\mu_X - \mu_Y$ para muestras pequeñas

$$IC(\mu_X - \mu_Y) = \bar{X} - \bar{Y} \pm t_{k,1-\alpha/2}\sqrt{\frac{S^2_{n_X-1,X}}{n_X} + \frac{S^2_{n_Y-1,Y}}{n_Y}}$$

$$IC(\mu_X - \mu_Y) = \left(-\infty, \bar{X} - \bar{Y} + t_{k,1-\alpha}\sqrt{\frac{S^2_{n_X-1,X}}{n_X} + \frac{S^2_{n_Y-1,Y}}{n_Y}}\right)$$

y

$$IC(\mu_X - \mu_Y) = \left(\bar{X} - \bar{Y} - t_{k,1-\alpha}\sqrt{\frac{S^2_{n_X-1,X}}{n_X} + \frac{S^2_{n_Y-1,Y}}{n_Y}}, \infty\right)$$

El cálculo del intervalo de Welch puede llevarse a cabo usando la instrucción `t.test` en R. Ilustramos el uso de este comando en el siguiente ejemplo.

Ejemplo 3.2.8. *Suponga que se quiere estudiar el efecto de dos dietas diferentes para reducir el nivel de glucosa para pacientes con nivel de glucosa entre 100∼ 130 mg/dl. Los pacientes se sometieron a las dos dietas de forma aleatoria, y después de 2 meses del tratamiento, el nivel de glucosa de estos dos grupos de pacientes fueron: con dieta 1: 105.0, 96.7, 103.9, 110.6, 95.1, 111.2, 93.1, 109.9, 105.8, 95.3, 92.7; con dieta 2: 90.3, 92.1, 96.5, 84.8, 94.0, 86.9, 91.5, 86.1, 94.4, 86.8, 94.4, 91.6. Para calcular un intervalo de confianza para la diferencia del nivel de glucosa con las dos dietas $\mu_1 - \mu_2$, se debe hacer el supuesto acerca de la igualdad de varianza de las dos poblaciones. Aún no disponemos de herramientas que logren tal fin; sin embargo, de los datos muestrales, se puede ver que $S_{n_1-1} = 7.31$ y $S_{n_2-1} = 3.83$ de donde sospechamos que las dos varianzas teóricas pueden ser diferentes. Y considerando que las dos muestras son relativamente pequeñas, se utiliza el intervalo de Welch por medio de los siguientes comandos en R.*

```
> A<-c(105.0, 96.7, 103.9, 110.6,  95.1, 111.2,  93.1, 109.9,
105.8, 95.3, 92.7)
> B<-c(90.3, 92.1, 96.5, 84.8, 94.0, 86.9, 91.5, 86.1, 94.4,
86.8, 94.4, 91.6)
> t.test(A,B)
95 percent confidence interval:
  5.713331 16.229093
 sample estimates:
mean of x mean of y
 101.75455  90.78333
```

Podemos ver que el intervalo bilateral del 95 % para $\mu_1 - \mu_2$ está dado por $(5.71, 16.22)$, y éste contiene solo valores positivos, indicando que el nivel de glucosa promedio de pacientes sometidos a la dieta 1 será mayor que el de los pacientes sometidos a la dieta 2, y se puede concluir que la dieta 2 es más efectiva que la dieta 1.

Intervalos de confianza para cociente de varianzas

En esta parte estudiamos procedimientos para hallar intervalos de confianza para la cociente de dos varianzas teóricas σ_X^2/σ_Y^2 basándonos en dos muestras provenientes de distribuciones normal bajo las especificaciones dadas al principio del capítulo. Estos intervalos nos servirán para verificar si dos varianzas teóricas son iguales o no. Esto es importante porque

- Para hallar intervalos para la diferencia de dos medias teóricas, se necesita tener supuestos acerca de las varianzas teóricas, y dependiendo de si éstas son iguales o no, se emplean diferentes intervalos para la diferencia de medias.
- En algunas prácticas estadísticas, se necesita comparar dos varianzas, y determinar cuál tiene mayor magnitud. Por ejemplo, las líneas de producción industrial deben tener una pequeña variabilidad, para garantizar que los productos sean lo más homogéneo posible en términos de alguna característica.

Nótese que para el propósito de evaluar si las dos varianzas teóricas son iguales, cualquiera de los intervalos $IC(\sigma_X^2/\sigma_Y^2)$ y $IC(\sigma_Y^2/\sigma_X^2)$. Ahora, hemos visto, en el caso de una muestra, que el intervalo de confianza para la varianza depende de si la media teórica es conocida o no. En el caso de dos muestras, también se debe hacer esta distinción y por consiguiente, tenemos los siguientes casos:

μ_X y μ_Y son conocidas

En este caso, los estimadores de máxima verosimilitud de σ_X^2 y σ_Y^2 son $\frac{\sum_{i=1}^{n_X}(X_i-\mu_X)^2}{n_X}$ y $\frac{\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2}{n_Y}$, respectivamente. Y por consiguiente el estimador de máxima verosimilitud de σ_X^2/σ_Y^2 está dado por

$$\frac{\sum_{i=1}^{n_X}(X_i-\mu_X)^2/n_X}{\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2/n_Y}.$$

Aunque la anterior variable claramente no es una variable pivote para σ_X^2/σ_Y^2, podemos modificarla recordando que

$$\frac{\sum_{i=1}^{n_X}(X_i-\mu_X)^2}{\sigma_X^2} \sim \chi_{n_X}^2$$

y

$$\frac{\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2}{\sigma_Y^2} \sim \chi_{n_Y}^2.$$

Usando la independencia de las dos muestras y la Definición 1.1.16, se tiene que

$$\frac{\sum_{i=1}^{n_X}(X_i-\mu_X)^2/(n_X\sigma_X^2)}{\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2/(n_Y\sigma_Y^2)} = \frac{\sigma_Y^2}{\sigma_X^2}\frac{n_Y\sum_{i=1}^{n_X}(X_i-\mu_X)^2}{n_X\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2} \sim F_{n_Y}^{n_X}, \qquad (3.2.38)$$

Podemos ver que la anterior variable es una variable pivote para la cociente de varianzas, y por consiguiente debemos buscar valores a y b tales que

$$Pr\left(a < \frac{\sigma_Y^2}{\sigma_X^2}\frac{n_Y\sum_{i=1}^{n_X}(X_i-\mu_X)^2}{n_X\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2} < b\right) = 1-\alpha. \tag{3.2.39}$$

Los valores óptimos de a y b son los que minimizan la longitud o la longitud esperada del intervalo resultante para la cociente de varianzas dado por

$$IC(\sigma_Y^2/\sigma_X^2) = \left(a\frac{n_X\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2}{n_Y\sum_{i=1}^{n_X}(X_i-\mu_X)^2}, b\frac{n_X\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2}{n_Y\sum_{i=1}^{n_X}(X_i-\mu_X)^2}\right),$$

cuya longitud, como se puede observar claramente, es una variable aleatoria, y por consiguiente calculamos la longitud esperada. Esta está dada por

$$E(l) = (b-a)E\left(\frac{n_X\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2}{n_Y\sum_{i=1}^{n_X}(X_i-\mu_X)^2}\right) = (b-a)\frac{n_X}{n_X-2}\frac{\sigma_Y^2}{\sigma_X^2}$$

para $n_X > 2$. De donde concluimos que los valores de a y b que minimizan $b-a$ también minimizan $E(l)$, y por consiguiente el Resultado 3.2.1 nos indica que los valores óptimos de a y b deben cumplir $f(a) = f(b)$ donde f denota la función de densidad de una distribución $F_{n_X}^{n_Y}$[7]. Anteriormente se presentó un código R que permite encontrar valores de a y b que satisfacen $f(a) = f(b)$ con f la función de densidad de una distribución χ^2, una pequeña modificación de este código nos permite encontrar los valores a y b óptimos. En general para evitar cálculos tediosos, se puede adoptar la solución $a = f_{n_Y,\alpha/2}^{n_X}$ y $b = f_{n_Y,1-\alpha/2}^{n_X}$, que en muestras grandes conduce a intervalos de longitud pequeña.

Adicionalmente, podemos tener el siguiente intervalo para σ_X^2/σ_Y^2 aplicando el Resultado 3.2.2

$$\begin{aligned} IC(\sigma_X^2/\sigma_Y^2) &= \left(\frac{1}{f_{n_Y,1-\alpha/2}^{n_X}}\frac{n_Y\sum_{i=1}^{n_X}(X_i-\mu_X)^2}{n_X\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2}, \frac{1}{f_{n_Y,\alpha/2}^{n_X}}\frac{n_Y\sum_{i=1}^{n_X}(X_i-\mu_X)^2}{n_X\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2}\right) \\ &= \left(f_{n_X,\alpha/2}^{n_Y}\frac{n_Y\sum_{i=1}^{n_X}(X_i-\mu_X)^2}{n_X\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2}, f_{n_X,1-\alpha/2}^{n_Y}\frac{n_Y\sum_{i=1}^{n_X}(X_i-\mu_X)^2}{n_X\sum_{j=1}^{n_Y}(Y_j-\mu_Y)^2}\right). \end{aligned}$$

En muchos de los estudios estadísticos, no se dispone de información auxiliar, o ésta no es del todo confiable, y no se conocen los valores de las medias teóricas. En estos casos, los anteriores intervalos ya no son aplicables, por esta razón, estudiamos los intervalos de confianza para la cociente de varianzas cuando las medias teóricas son desconocidas.

[7] Esto es cierto siempre y cuanto f es unimodal, y esto se cumple cuando ambos grados de libertad son mayores que 3.

μ_X y μ_Y son desconocidas

En este caso, la variable encontrada en (3.2.38) ya no es una variable pivote para la cociente de varianzas, puesto que ésta depende de las medias teóricas desconocidas μ_X y μ_Y. Una propuesta natural es reemplazar μ_X y μ_Y por sus estimadores $\bar{X}$ y $\bar{Y}$, respectivamente. Adicionalmente, recordemos que

$$\frac{\sum_{i=1}^{n_X}(X_i-\bar{X})^2}{\sigma_X^2} \sim \chi^2_{n_X-1}$$

y

$$\frac{\sum_{j=1}^{n_Y}(Y_j-\bar{Y})^2}{\sigma_Y^2} \sim \chi^2_{n_Y-1}.$$

De esta forma tenemos que

$$\frac{\sum_{i=1}^{n_X}(X_i-\bar{X})^2/((n_X-1)\sigma_X^2)}{\sum_{j=1}^{n_Y}(Y_j-\bar{Y})^2/((n_Y-1)\sigma_Y^2)} = \frac{\sigma_Y^2}{\sigma_X^2}\frac{S^2_{n_X-1}}{S^2_{n_Y-1}} \sim F^{n_X-1}_{n_Y-1}, \tag{3.2.40}$$

Y podemos obtener el siguiente intervalo para σ_X^2/σ_Y^2

$$IC(\sigma_X^2/\sigma_Y^2) = \left(f^{n_Y-1}_{n_X-1,\alpha/2}\frac{S^2_{n_X-1}}{S^2_{n_Y-1}}, f^{n_Y-1}_{n_X,1-\alpha/2}\frac{S^2_{n_X-1}}{S^2_{n_Y-1}} \right). \tag{3.2.41}$$

Podemos ver que el anterior intervalo se basa en el estimador $S^2_{n_X-1}/S^2_{n_Y-1}$ para la cociente de varianzas ampliando a la izquierda y derecha multiplicando por los percentiles $f^{n_Y-1}_{n_X-1,\alpha/2}$ y $f^{n_Y-1}_{n_X-1,1-\alpha/2}$ respectivamente, y por consiguiente no es un intervalo simétrico con respecto al estimador.

El cálculo de este intervalo se lleva a cabo en R usando `var.test(stats)`, lo ilustramos en el siguiente ejemplo.

Ejemplo 3.2.9. *En el Ejemplo 3.2.7, se calculó un intervalo de confianza para la diferencia entre rendimientos promedios obtenidos por estudiantes de dos institutos donde se supuso que los dos institutos tienen la misma variación. A continuación se calcula el intervalo (3.2.41) de 95 % para corroborar este afirmación, tenemos*

```
> A<-c(75, 87, 83, 73, 74, 88, 88, 74, 64, 92, 73, 87, 91, 83,84)
> B<-c(64, 85, 72, 64, 74, 93, 70, 79, 79, 75, 66, 83 ,74)
> var.test(A,B)
data:  A and B
95 percent confidence interval:
 0.2918789 2.8544131
sample estimates:
ratio of variances
         0.9358256
```

Observamos que el intervalo del 95 % está dado por $(0.29, 2.85)$*, como este intervalo contiene el valor 1, podemos afirmar que la cociente de varianza* σ_X^2/σ_Y^2 *sí puede tomar el valor 1, es decir, las dos varianzas teóricas sí pueden ser iguales. Y por consiguiente, el intervalo* t *student calculado en el Ejemplo 3.2.7 es válido.*

Ahora, retomamos el Ejemplo 3.2.8, donde se interesaba comparar el nivel de colesterol de dos grupos de pacientes luego de someterlos a dos dietas diferentes. Para estos datos, las dos varianzas muestrales fueron 7.31^2 *y* 3.83^2*. Basado en estas estimaciones puntuales, se consideró adecuado el supuesto de que las varianzas teóricas son diferentes, y por consiguiente se empleó el intervalo de Welch para la diferencia de medias. Aquí calculamos el intervalo de confianza para la cociente de varianzas para verificar que las varianzas teóricas son diferentes, tenemos*

```
> A<-c(105.0, 96.7, 103.9, 110.6, 95.1, 111.2, 93.1, 109.9,
+ 105.8, 95.3, 92.7)
> B<-c(90.3, 92.1, 96.5, 84.8, 94.0, 86.9, 91.5, 86.1, 94.4,
+ 86.8, 94.4, 91.6)
> var.test(A,B)
data:  A and B
95 percent confidence interval:
  1.034110 13.362031
sample estimates:
ratio of variances
          3.645933
```

Podemos ver que el intervalo obtenido $(1.03, 13.36)$ *no contiene el valor 1, indicando que el supuesto de que las dos varianzas teóricas son diferentes es adecuado.*

3.3 Bajo distribuciones diferentes a la normal

En las secciones anteriores, hemos visto que para encontrar un intervalo de confianza para algún parámetro, el método de la variable pivote consiste en encontrar una variable pivote Q, construir un intervalo para Q y finalmente despejar el parámetro de interés. En muestras provenientes de distribuciones normales, el procedimiento se puede llevar a cabo sin mayores complicaciones. Sin embargo, cuando la muestra aleatoria proviene de distribuciones diferentes de la distribución normal, el método de la variable pivote no siempre resulta útil puesto que no siempre es fácil encontrar una variable pivote para el parámetro de interés; más aún, hay casos donde a pesar de disponer de una variable pivote, no se puede despejar el parámetro de interés. Cuando no se puede encontrar la variable pivote, una herramienta que puede resultar útil es el teorema del límite central que nos permite aproximar la distribución del promedio muestral mediante una distribución normal.

Primero presentamos una forma de encontrar variables pivotes para ciertos tipos de parámetros. Estos parámetros se denominan parámetro de escala, y se definen a continuación.

Definición 3.3.1. *Sea X una variable aleatoria con función de densidad de probabilidad $f_X(x)$, la cual depende de un parámetro θ, se dice que θ es un parámetro de escala si la distribución de la variable X/θ o θX no depende de θ.*

De la anterior definición, vemos que para probar que un parámetro es de escala, se debe encontrar la distribución de X/θ o θX. Conociendo la distribución de X, hay tres formas básicas para encontrar esta distribución. Éstas son

- Utilizar directamente la función de distribución o la función de densidad cuando éstas sean fáciles de hallar.
- Cuando la función de distribución y/o de densidad sean de formas complicadas, se puede usar el teorema de transformación que encuentra la función de densidad para una función $g(X)$ mediante el uso de jacobiano, Blanco (2004, p. 256).
- Utilizar la función generadora de momentos[8].

A continuación, se define otra clase de parámetros para la cual es fácil de encontrar una variable de pivote.

Definición 3.3.2. *Sea X una variable aleatoria con función de densidad de probabilidad $f_X(x)$, la cual depende de un parámetro θ, se dice que θ es un parámetro de localización si la distribución de la variable $X+\theta$ o $X-\theta$ no depende de θ.*

Nota: si la función de densidad de una variable depende de más de un parámetro, para verificar que un parámetro específico sea de localización o de escala, los demás parámetros se consideran constantes, como lo ilustra el siguiente ejemplo.

Ejemplo 3.3.1. *Sea X una variable aleatoria con distribución $N(\mu, \sigma^2)$. Se tiene que $X-\mu \sim N(0,\sigma^2)$, y por consiguiente, μ es un parámetro de localización.*

Para un parámetro de localización o de escala, el siguiente resultado nos permite encontrar una variable pivote.

Resultado 3.3.1. *Sea $X_1, \ldots, X_n$ una muestra aleatoria proveniente de una distribución con función de densidad $f(x,\theta)$, y sea T el estimador de máxima verosimilitud de θ, entonces*

- *si θ es parámetro de localización, entonces $T+\theta$ o $T-\theta$ es una variable pivote para θ.*
- *si θ es parámetro de escala, entonces T/θ o θT es una variable pivote para θ.*

[8] Es bien conocido para una variable aleatoria la función generadora de momentos (cuando ésta existe) caracteriza la distribución de la variable, hay ejemplos donde dos variables diferentes pueden tener las funciones características casi idénticas, McCullagh (1994).

3.3.1 Intervalos de confianza con distribución exponencial

Sea X una variable aleatoria con distribución $Exp(\theta)$ con $E(X) = \theta$. Se tiene que $X/\theta \sim Exp(1)$, y por consiguiente, θ es un parámetro de escala. Para ver la distribución de X/θ, se puede hacer uso de la función generadora de momentos, tenemos que

$$M_{X/\theta}(t) = E(e^{tX/\theta}) = M_X(t/\theta) = \frac{1}{1-\theta\frac{t}{\theta}} = \frac{1}{1-\theta},$$

la cual corresponde a la función generadora de momentos de una distribución $Exp(1)$. También se puede encontrar la distribución de X/θ usando directamente la función de distribución como sigue

$$F_{X/\theta}(x) = Pr\left(\frac{X}{\theta} \leq x\right) = Pr(X \leq \theta x) = F_X(\theta x) = 1 - e^{-x}.$$

Y ésta corresponde a la función de distribución de una variable con distribución $Exp(1)$, de donde se concluye que $X\theta \sim Exp(1)$. Y por consiguiente podemos aplicar el Resultado 3.3.1, sea X_1, ..., X_n una muestra aleatoria con distribución $Exp(\theta)$. Se ha visto que θ es un parámetro de escala, por otro lado, el estimador de máxima verosimilitud para θ es $\bar{X}$, por lo tanto, $\bar{X}/\theta$ es una variable pivote para θ. Y nuevamente, se buscan valores a y b con

$$Pr\left(a < \frac{\bar{X}}{\theta} < b\right) = 1 - \alpha. \tag{3.3.1}$$

Claramente, a y b son percentiles de la distribución de la variable $\bar{X}/\theta$, pero surge la dificultad de que esta distribución no es ninguna de las distribuciones comunes en la teoría estadística, y por consiguiente los valores a y b no se pueden hallar directamente.

Sin embargo, usando el Resultado 1.1.17, se tiene que la variable $\sum_{i=1}^{n} X_i \sim Gamma(n, \theta)$. Ahora, usando la función generadora de momentos, se puede ver fácilmente que $\sum_{i=1}^{n} X_i/\theta \sim Gamma(n, 1)$. Por lo tanto, (3.3.1) se convierte en

$$Pr\left(an < \frac{\sum_{i=1}^{n} X_i}{\theta} < bn\right) = 1 - \alpha. \tag{3.3.2}$$

De donde se concluye que an y bn son percentiles de la distribución $Gamma(n, 1)$. El lector puede usar argumentos explicados anteriormente para encontrar los valores de a y b que minimizan la longitud del intervalo resultante para θ. Por simplicidad, tomamos $an = Gamma(n, 1)_{\alpha/2}$ y $bn = Gamma(n, 1)_{1-\alpha/2}$. Finalmente, despejando θ de (3.3.2), se tiene el siguiente intervalo para θ,

$$IC(\theta) = \left(\frac{\sum_{i=1}^{n} X_i}{Gamma(n, 1)_{1-\alpha/2}}, \frac{\sum_{i=1}^{n} X_i}{Gamma(n, 1)_{\alpha/2}}\right). \tag{3.3.3}$$

La longitud de este intervalo está dada por

$$l = \sum_{i=1}^{n} X_i \left(\frac{1}{Gamma(n, 1)_{\alpha/2}} - \frac{1}{Gamma(n, 1)_{1-\alpha/2}}\right)$$

y su valor esperado está dado por

$$E(l) = n\theta \left(\frac{1}{Gamma(n,1)_{\alpha/2}} - \frac{1}{Gamma(n,1)_{1-\alpha/2}} \right). \tag{3.3.4}$$

De lo anterior se observa que mayor sea el parámetro θ, mayor será la longitud del intervalo; por otro lado, no es claro a simple vista cómo es la relación entre la longitud esperada y el tamaño muestral. Más adelante, en la Figura 3.12 se verá que al incrementar el tamaño muestral, el intervalo se hace más preciso.

Ahora, para encontrar los límites unilaterales, se debe notar que la variable pivote $\sum_{i=1}^{n} X_i/\theta$ guarda una relación inversamente proporcional con el parámetro θ. De esta forma, si se busca un intervalo superior para θ, se debe comenzar encontrando un intervalo inferior para la variable pivote; mientras que un intervalo superior para la variable pivote conduce a un intervalo inferior para θ. De esta forma, se obtienen los siguientes intervalos unilaterales para θ.

$$IC(\theta) = \left(\frac{\sum_{i=1}^{n} X_i}{Gamma(n,1)_{1-\alpha}}, \infty \right) \tag{3.3.5}$$

y

$$IC(\theta) = \left(0, \frac{\sum_{i=1}^{n} X_i}{Gamma(n,1)_{\alpha}} \right) \tag{3.3.6}$$

donde el valor 0 en el anterior intervalo es un límite inferior natural para θ pues éste solo puede tomar valores positivos.

Ejemplo 3.3.2. *Para los datos del Ejemplo 2.3.4 donde se dispone de un conjunto de datos correspondientes a tiempo de espera antes de que una llamada sea atendida por el operador, en este ejemplo se ha visto que los datos muestran evidencias de que provienen de una distribución exponencial, y se encontró una estimación de 0.8 minutos para el tiempo promedio de espera. Si estamos interesados en hallar un intervalo de confianza para este promedio, podemos usar el siguiente comando en R*

```
> tiempo<-c(0.13, 0.06, 0.50, 0.41, 1.44, 0.60, 0.22, 1.08,
0.78,0.92, 2.73, 0.83, 0.19, 0.21, 1.75, 0.79, 0.02, 0.05,
2.30,1.03)
> alpha<-0.05
>n<-length(tiempo)
> L.sup<-sum(tiempo)/qgamma(alpha/2,shape=n,scale=1)
> L.inf<-sum(tiempo)/qgamma(1-alpha/2,shape=n,scale=1)
> L.sup
[1] 1.312976
> L.inf
[1] 0.5405979
```

y un intervalo del 95 % para el tiempo promedio de espera es $(0.54, 1.31)$. *Si aumentamos el nivel de confianza al* $98\,\%$, *el intervalo resultante será* $(0.50, 1.45)$,

el cual tiene una longitud mayor que el de $95\,\%$ *confirmando una vez más que al aumentar el nivel de confianza, el intervalo pierde precisión.*

Suponga que ahora se desea saber cuál es el tiempo mínimo que deben esperar los clientes antes de ser atendidos. Para eso se debe encontrar un intervalo de confianza inferior para el tiempo promedio de espera, y el límite inferior se puede calcular con el comando

```
>L.inf<-sum(tiempo)/qgamma(1-alpha,shape=n,scale=1)
> L.inf
[1] 0.5753385
```

de donde se concluye que los clientes que llaman deben esperar por lo menos más de medio minuto antes de ser atendidos por uno de los operadores de la aerolínea.

En algunas prácticas estadísticas, el usuario no tiene en cuenta la distribución de los datos, y para encontrar un intervalo de confianza para la media teórica, simplemente aplica el intervalo t dado en (3.2.12). Para estudiar qué tan buenos son los intervalos obtenidos de esta manera, se realiza el siguiente estudio de simulación. Se simula 1000 veces muestras de tamaño 5, 10, 20, 50, 100, 500, 1000 provenientes de una distribución exponencial con parámetro de valor conocido, digamos igual a 5, y en cada iteración se calcula el intervalo (3.3.3) y el intervalo (3.2.12), y se observa si estos intervalos contienen o no al parámetro verdadero. La probabilidad de cobertura real de un intervalo se calcula como el número de iteraciones donde el intervalo contiene el parámetro verdadero dividido por el número total de iteraciones. El código utilizado es como sigue.

```
> set.seed(123)
>
> n<-c(5,10,20,50,100,500,1000)
> ng<-1000
> theta<-5
> pro.Exp<-pro.t<-matrix(NA)
> aux.Exp<-aux.t<-0
>
> for(i in 1:length(n)){
+ for(j in 1:ng){
+ x<-rgamma(n[i],shape=1,scale=theta)
+ if(t.test(x)$conf.int[1]>theta||t.test(x)$conf.int[2]<theta)
+ {aux.t<-aux.t}
+ if(t.test(x)$conf.int[1]<=theta&&t.test(x)$conf.int[2]>=theta)
+ {aux.t<-aux.t+1}
+ Exp.inf<-sum(x)/qgamma(0.975,shape=n[i],scale=1)
+ Exp.sup<-sum(x)/qgamma(0.025,shape=n[i],scale=1)
+ if(Exp.inf>theta||Exp.sup<theta){aux.Exp<-aux.Exp}
+ if(Exp.inf<=theta&&Exp.sup>=theta){aux.Exp<-aux.Exp+1}
+ }
```

```
+ pro.Exp[i]<-aux.Exp/ng
+ pro.t[i]<-aux.t/ng
+ aux.Exp<-aux.t<-0
+ }

>
> plot(pro.Exp,type="b", ylim=c(0.85,1),xaxt="n",
+ ylab="Probabilida de cobertura",xlab="n")
> lines(pro.t,type="b", pch=2)
> abline(h=0.95)
> axis(1, 1:length(n), n)
> legend(4,0.9,c("Intervalo gamma","Intervalo t"),  lty=c(1,1),
+ pch=c(1,2),bty="n")
```

En la Figura 3.10, se muestran las probabilidades de cobertura para los dos intervalos.

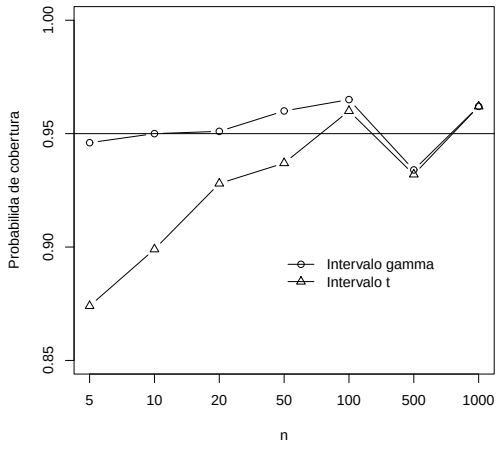

Figura 3.10: *Probabilidades de cobertura para el intervalo gamma y el intervalo t para diferentes tamaños de muestra.*

Se observa que el intervalo gamma siempre tiene la probabilidad de cobertura real muy cercano al nivel nominal del 95 % aún para muestras de tamaño pequeño; por otro lado, el intervalo t tiene probabilidad relativa muy por debajo del 95 % para muestras pequeños, pero a medida que la muestra crece, la diferencia de los dos intervalos ya es muy pequeña. Por lo anterior, podemos observar que ignorar la distribución de los datos y aplicar el intervalo t puede conducir a intervalos no muy confiables cuando la muestra es pequeña. Se deja como ejercicio el diseño de un estudio de simulación para comparar estos dos intervalos en términos de la longitud (Ejercicio 3.15).

Intervalo de confianza en una distribución exponencial usando TLC

En una muestra proveniente de una distribución $Exp(\theta)$, también podemos usar el teorema del límite central para encontrar un intervalo de confianza aproximado. Para eso recordemos que la media y la varianza teórica están dadas por θ y θ^2, respectivamente, entonces tenemos que la variable $\frac{\sqrt{n}(\bar{X}-\theta)}{\theta}$ se distribuye aproximadamente como normal estándar, la cual desempeña la función de la variable pivote. Por consiguiente, tenemos que

$$1-\alpha = Pr\left(a < \frac{\sqrt{n}(\bar{X}-\theta)}{\theta} < b\right)$$

y para encontrar, en lo posible, el intervalo con menor longitud para θ, despejamos θ en la anterior igualdad. Tenemos que

$$1-\alpha = Pr\left(\frac{\sqrt{n}\bar{X}}{b+\sqrt{n}} < \theta < \frac{\sqrt{n}\bar{X}}{a+\sqrt{n}}\right), \tag{3.3.7}$$

cuya longitud está dada por $l = \sqrt{n}\bar{X}(\frac{1}{a+\sqrt{n}} - \frac{1}{b+\sqrt{n}})$, con longitud esperada dada por

$$E(l) = \sqrt{n}\theta\left(\frac{1}{a+\sqrt{n}} - \frac{1}{b+\sqrt{n}}\right) \tag{3.3.8}$$

la cual no depende directamente de la longitud $b-a$ y por consiguiente el uso del Resultado 3.2.1 no arrojará un intervalo de menor longitud para θ. Se puede elaborar un programa computacional similar al del caso de intervalos para σ^2 en una distribución normal. Por ahora, escogemos $a = -z_{1-\alpha/2}$ y $b = z_{1-\alpha/2}$. Y reemplazándolos en (3.3.7) obtenemos el siguiente intervalo aproximado para θ

$$IC(\theta) = \left(\frac{\sqrt{n}\bar{X}}{z_{1-\alpha/2}+\sqrt{n}}, \frac{\sqrt{n}\bar{X}}{-z_{1-\alpha/2}+\sqrt{n}}\right). \tag{3.3.9}$$

Ejemplo 3.3.3. *Siguiendo con el Ejemplo 3.3.2 donde se calculó intervalos exactos basados en una distribución Gamma para los datos del Ejemplo 2.3.4., para calcular los intervalos aproximados basados en la distribución normal estándar, podemos usar el siguiente comando en R*

```
> L.sup<-sqrt(n)*mean(tiempo)/(-qnorm(1-alpha/2)+sqrt(n))
> L.inf<-sqrt(n)*mean(tiempo)/(qnorm(1-alpha/2)+sqrt(n))
> L.sup
[1] 1.42771
> L.inf
[1] 0.5576177
```

y tenemos el intervalo aproximado $(0.56, 1.43)$ *para el tiempo promedio de espera, y podemos ver que éste es más ancho que el intervalo.*

Para averiguar, entre el intervalo exacto encontrado anteriormente y el intervalo exacto dado en (3.3.3), cuál es la mejor opción, realizamos estudios de simulación, para

compararlos en términos de la probabilidad de cobertura real y la longitud esperada. Para compararlos en términos de la probabilidad de cobertura real, se simulan 1000 muestras para $n = 5, \cdots, 200$ provenientes de una distribución $Exp(2)$, y para cada muestra se calcula el intervalo aproximado basado en la distribución normal estándar y el intervalo exacto basado en la distribución Gamma, y se examina si estos intervalos contienen el parámetro teórico $\theta = 2$. La probabilidad de cobertura real de los dos intervalos se calcula como el número de veces que el intervalo contiene a θ dividido por el número total de muestras simuladas para cada valor de n. Los resultados de simulación se muestran en la Figura 3.11, donde se observa que, en general, la probabilidad de cobertura real del intervalo aproximado es siempre mayor que el intervalo exacto y la diferencia es especialmente prominente en muestras de tamaño pequeño. Además, nótese que aún para muestras grandes, el intervalo exacto tiene, casi siempre, la probabilidad de cobertura real inferior a la nominal del $95\,\%$; mientras que la probabilidad de cobertura del intervalo aproximado tiene un comportamiento más estable, oscilando alrededor del valor nominal para muestras grandes. Cabe resaltar que en muestras pequeñas, ambos intervalos tienen una probabilidad de cobertura real muy por debajo de la nominal, situación que no ocurre en casos como intervalos para μ en una distribución normal (Ver Figura 3.5).

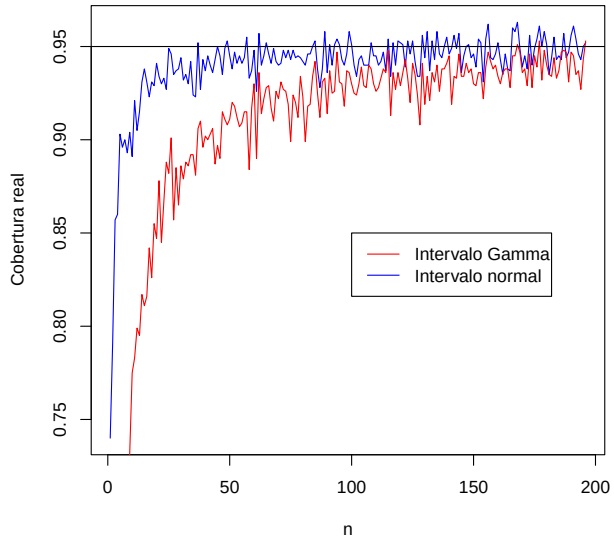

Figura 3.11: *Probabilidades de cobertura para el intervalo aproximado normal y el intervalo exacto gamma para diferentes tamaños de muestra en una distribución* $Exp(2)$.

Ahora, para comparar los dos intervalos en términos de la longitud esperada, podemos calcular estas longitudes. En el caso del intervalo normal, la longitud esperada está dada en (3.3.8); para el intervalo Gamma, la longitud esperada está dada

en (3.3.4). Podemos graficar estas dos longitudes esperadas para diferentes tamaños muestras, esta gráfica se muestra en la Figura 3.12 donde el modelo poblacional es $Exp(1)$. Se observa que en muestras muy pequeñas (de tamaño menor a 10, aproximadamente) el intervalo aproximado puede tener una longitud muy grande con respecto al intervalo exacto, pero esta diferencia se disminuye y desaparece muy rápidamente. Por lo anterior, podemos recomendar el intervalo normal, especialmente cuando el tamaño muestral es grande.

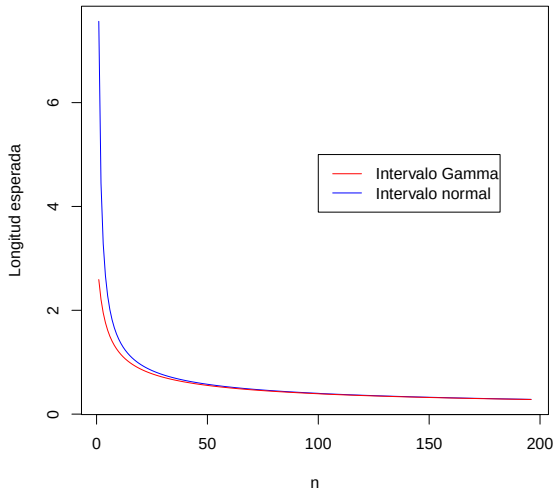

Figura 3.12: *Longitud esperada teórica del intervalo aproximado normal y el intervalo exacto gamma para diferentes tamaños de muestra en una distribución* $Exp(1)$.

En el caso de la distribución exponencial, fue posible calcular las longitudes esperadas teóricas. Cuando esto no es posible, se puede hacer uso de las simulaciones. Podemos calcular la longitud de los dos intervalos en cada una de las 1000 muestras simuladas anteriormente para cada valor fijo de n, y se toma el promedio de estos 1000 valores como una estimación de la longitud esperada. Los resultados se muestran en la Figura 3.13, donde se observa que el comportamiento es muy parecido a los valores teóricos, indicando que el intervalo aproximado puede no ser tan preciso como el intervalo exacto; sin embargo, para muestras moderadamente grandes, no hay diferencias importantes entre los dos intervalos en términos de la precisión, y se recomienda el intervalo aproximado basado en la distribución normal estándar.

3.3.2 Intervalos de confianza con distribución Bernoulli

El procedimiento descrito anteriormente resulta ser muy limitado, puesto que en distribuciones como Poisson, Binomial, el parámetro de interés, λ y p no son de loca-

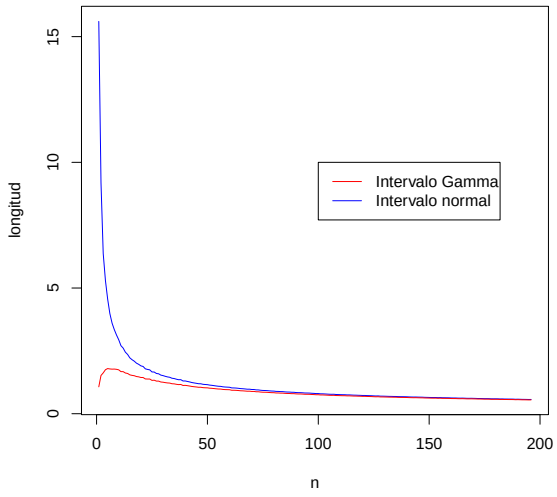

Figura 3.13: *Longitud esperada simulada del intervalo aproximado normal y el intervalo exacto gamma para diferentes tamaños de muestra en una distribución* $Exp(2)$.

lización y tampoco de escala, y por consiguiente no se aplica el procedimiento. Sin embargo, el teorema del límite central nos puede resultar útil en algunos casos.

Consideramos el problema de encontrar un intervalo de confianza para una proporción dadas n observaciones de variables independientes e idénticamente distribuidas provenientes de una distribución $Ber(p)$. Si la muestra se denota por $X_1, \ldots, X_n$, entonces el estimador de máxima verosimilitud es el promedio muestral $\bar{X}$, que se denotará por $\hat{p}$. Por otro lado, la esperanza de $Ber(p)$ es p y la varianza $p(1-p)$, de donde se tiene que

$$\frac{\bar{X} - p}{\sqrt{p(1-p)/n}} \to_D Z \sim N(0,1).$$

Aunque la anterior variable es una variable pivote para p cuando la muestra es grande, y se puede encontrar a y b con $Pr\left(a < \frac{\hat{p} - p}{\sqrt{Pr(1-p)/n}} < b\right) = 1-\alpha$, no se puede despejar el parámetro de interés p. Una solución a este problema es reemplazar p por su estimador $\hat{p}$ en el denominador de la variable pivote, y así construir la variable $\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$. El intervalo para p presentado en la mayoría de los textos estadísticos, el intervalo de Wald, se basa en esta variable, asumiendo que la distribución asintótica sigue siendo la distribución $N(0,1)$. De esta forma, se tiene que

$$Pr\left(-z_{1-\alpha/2} < \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} < z_{1-\alpha/2}\right) = 1-\alpha,$$

despejando el parámetro p, se tiene el siguiente intervalo de confianza para p

$$IC(p) = \left(\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right).$$

A pesar de la simplicidad del cómputo del anterior intervalo, éste tiene varias fallas. En primer lugar, en algunas situaciones el intervalo obtenido puede contener valores fuera del espacio paramétrico de p, $[0, 1]$. Por ejemplo, cuando $n = 30$ que constan de 1 éxito y 29 fracasos, $\hat{p} = 0.033$, se tiene que el límite inferior es -0.0309; mientras que cuando $n = 30$ que constan de 29 éxitos y 1 fracaso, $\hat{p} = 0.967$ el límite superior es 1.0309. En segundo lugar, cuando la proporción muestral toma valor 0 o 1, el límite inferior es igual al límite superior, y la estimación por intervalo de confianza se reduce a la estimación puntual. Adicionalmente, estudios de simulación han mostrado que el intervalo de Wald tiene un mal desempeño, ver por ejemplo Agresti & Coull (1998) y Cepeda, Aguilar, Cervantes, Corrales, Díaz & Rodríguez (2008).

En la literatura estadística reciente, se han desarrollado otros intervalos para p con desempeño muy superior que el de Wald, entre ellos, se encuentra el intervalo de Agresti y Caffo (Agresti & Caffo 2000), que también es muy sencillo de implementar en la práctica. El intervalo de Agresti y Caffo está dado por

$$IC_{AC}(p) = \left(\tilde{p} - z_{1-\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}, \tilde{p} + z_{1-\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}\right),$$

donde $\tilde{n} = n+4$, $\tilde{p} = \tilde{x}/\tilde{n}$ con $\tilde{x} = x+2$. Nótese que el intervalo de Agresti y Caffo consiste simplemente en añadir dos éxitos y dos fracasos a la muestra observada, pero tiene mucho mejores propiedades que el intervalo clásico de Wald, su probabilidad de cobertura real es más alta, y su longitud más corta. Para detalles sobre otros intervalos para p y la comparación entre ellos, consulte Cepeda, Aguilar, Cervantes, Corrales, Díaz & Rodríguez (2008).

Otro problema importante en la práctica es cuando se observan dos muestras independientes $X_1, \ldots, X_{n_1}$ y $Y_1, \ldots, Y_{n_2}$ provenientes de distribuciones $Ber(p_1)$ y $Ber(p_2)$; por ejemplo, tasa de curación de dos medicamentos. En este caso, estamos interesados en hallar intervalos de confianza para la diferencia de las dos proporciones $p_1 - p_2$. Análogo al caso del intervalo para una proporción, el teorema del límite central conduce al intervalo de Wald para dos muestras. En este caso, se tiene en cuenta que

$$\hat{p}_i \sim_{aprox.} N\left(p_i, \frac{p_i(1-p_i)}{n_i}\right),$$

para $i = 1, 2$. Usando el hecho de que las dos muestras son independientes, se tiene que $\hat{p}_1$ y $\hat{p}_2$ también son independientes, y por consiguiente

$$\hat{p}_1 - \hat{p}_2 \sim_{aprox.} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

Estandarizando $\hat{p}_1 - \hat{p}_2$ conduce al intervalo de Wald para $p_1 - p_2$, dado por:

$$IC(p_1 - p_2) = \hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Desafortunadamente, este intervalo también tiene propiedades no deseables además de malos desempeños, al igual que el intervalo de Wald para una proporción tal como se muestra en estudios de simulación en Zhang, Gutiérrez & Cepeda (2009). Aquí se presenta el intervalo de Agresti y Caffo para dos muestras, dado por

$$IC_{AC}(p) = \tilde{p}_1 - \tilde{p_2} \pm z_{1-\alpha/2}\sqrt{V(\tilde{p}_1, \tilde{n}_1) + V(\tilde{p}_2, \tilde{n}_2)},$$

con

$$V(\tilde{p}_i, \tilde{n}_i) = \frac{1}{\tilde{n}_i}\left[\tilde{p}_i - \tilde{p}_i\frac{n_i}{\tilde{n}_i} + \frac{1}{2\tilde{n}_i}\right],$$

$\tilde{n}_i = n_i + 2$ para $i = 1, 2$, $\tilde{p}_1 = (\bar{X} + 1)/\tilde{n}_1$ y $\tilde{p}_2 = \bar{Y} + 1/\tilde{n}_2$, esto es, $\tilde{p}_i$ se calcula añadiendo un éxito y un fracaso en la i-ésima muestra, con $i = 1, 2$.

Otro intervalo para $p_1 - p_2$ fácil de calcular es el intervalo de Newcombe, (Newcombe 1998). El cómputo de este intervalo consiste en, primero, resolver para p_i la ecuación $|\hat{p_i} - p_i| = z_{1-\alpha/2}$, y denotamos las dos soluciones como l_i y u_i con $l_i < u_i$, para $i = 1, 2$. El intervalo de Newcombe está dado por

$$IC_{Newcombe}(p_1 - p_2) = \hat{p_1} - \hat{p_2} \pm z_{1-\alpha/2}\sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}}.$$

En Zhang, Gutiérrez & Cepeda (2009) se observa que el comportamiento de los intervalos Agresti y Caffo y el intervalo de Newcombe son muy similares, ambos muy superiores comparando con el de Wald. El intervalo de Newcombe puede tener un comportamiento levemente mejor que el de Agresti y Caffo.

3.3.3 Intervalos de confianza con distribución Poisson

Supongamos que $X_1, \cdots, X_n$ son variables aleatorias con distribución $Pois(\lambda)$ que constituyen una muestra aleatoria. Garwood (1936) propone calcular el intervalo de confianza para λ usando la estadística $Y = \sum_{i=1}^{n} X_i$ cuya distribución es $Pois(n\lambda)$. En una muestra observada donde Y toma el valor y_0, el intervalo propuesto se calcula como $(\lambda_1,\ \lambda_2)$ donde los límites λ_1 y λ_2 satisfacen

$$P(Y \leq y_0) = \sum_{i=0}^{y_0} \frac{e^{-n\lambda_1}(n\lambda_1)^i}{i!} = \frac{\alpha}{2} \tag{3.3.10}$$

y

$$P(Y \geq y_0) = \sum_{i=y_0}^{\infty} \frac{e^{-n\lambda_2}(n\lambda_2)^i}{i!} = \frac{\alpha}{2} \tag{3.3.11}$$

Para encontrar el valor de λ_1 que satisface (3.3.10), hacemos el uso del Resultado 1.1.15, donde liga la distribución Poisson con la distribución χ^2. Tenemos que

$$\frac{\alpha}{2} = P(Y \leq y_0) = 1 - P(X \leq 2n\lambda_1)$$

con $X \sim \chi^2_{2(y_0+1)}$, de donde $P(X \leq 2n\lambda_1) = 1 - \alpha/2$, y de esta forma $2n\lambda_1 = \chi^2_{2(y_0+1),1-\alpha/2}$, y tenemos que

$$\lambda_1 = \frac{1}{2n}\chi^2_{2(y_0+1),1-\alpha/2}.$$

De forma análoga, tenemos que

$$\frac{\alpha}{2} = P(Y \geq y_0) = P(X \leq 2n\lambda_2)$$

con $X \sim \chi^2_{2y_0}$, de donde tenemos que $\lambda_2 = \frac{1}{2n}\chi^2_{2y_0,\alpha/2}$. Y, finalmente un intervalo de confianza de $1-\alpha$ para λ está dado por

$$IC(\lambda) = \left(\frac{1}{2n}\chi^2_{2y_0,\alpha/2},\ \frac{1}{2n}\chi^2_{2(y_0+1),1-\alpha/2}\right).$$

En el caso de que y_0 se toma $\chi^2_{2y_0,\alpha/2} = 0$, y el límite inferior del anterior intervalo será 0.

Ejemplo 3.3.4. *Retomando los datos del Ejemplo 2.3.2, que corresponden al número de muertes violentas en 15 barrios de una ciudad en un determinado mes. En este contexto, $n = 15$, el número total de muertes violentas en estos 15 barrios es de 47, es decir, $y_0 = 47$. Si queremos calcular un intervalo del 90 % para el número promedio de muertos en un barrio, tenemos que $\alpha = 0.1$, y los percentiles $\chi^2_{2y_0,\alpha/2}$ y $\chi^2_{2(y_0+1),1-\alpha/2}$ toman valores de 72.64 y 119.87 respectivamente, conduciendo al intervalo $(2.42,\ 3.99)$ para el número promedio de muertes violentas en un barrio de esta ciudad.*

Más aún, podemos tener una estimación por intervalo de confianza para la probabilidad de que en mes no ocurra ninguna muerte en un barrio determinado. Esta probabilidad está dada por $e^{-\lambda}$, la cual es una función decreciente de λ, así un intervalo de confianza para esta probabilidad se puede calcular como $(e^{-3.99},\ e^{-2.42}) = (1.85\,\%,\ 8.89\,\%)$. Análogamente podemos calcular un intervalo de confianza para la probabilidad de que en un mes ocurran menos de dos muertes violentas en un barrio. Esta probabilidad está dada por $e^{-\lambda} + \lambda e^{-\lambda}$, que también es una función decreciente de λ. De esta forma, un intervalo de confianza para esta probabilidad se puede calcular como $(e^{-3.99} + 3.99e^{-3.99},\ e^{-2.42} + 2.42e^{-2.42}) = (9.23\,\%,\ 30.41\,\%)$.

También, teniendo en cuenta que la ciudad está conformada por 63 barrios, podemos calcular un intervalo de confianza para el número promedio de muertes violentas en la ciudad como $(2.42 * 63,\ 3.99 * 63) = (152.46,\ 251,37)$.

3.4 Ejercicios

3.1 Complete los pasos para hallar el intervalo unilateral inferior (3.2.9).

3.2 Para los datos de la Tabla 2.3 (en el Ejercicio 2.14 se vio que la distribución normal es apropiada para cada uno de los dos grupos de datos).

(a) Calcula un intervalo bilateral de 95 % para el kilómetro promedio recorrido por los automóviles de la marca A.

(b) Si se restringe que el margen de error máximo permitido es de 3 kilómetros, ¿cómo haría para determinar el tamaño muestral requerido?

(c) Encuentra un límite inferior para el kilómetro promedio recorrido por los automóviles de la marca A y la marca B de manera separada.

(d) Si los fabricantes de la marca A y marca B afirman que los automóviles en promedio recorren más de 45 y 48 kilómetros por galón, respectivamente. ¿Qué se puede decir acerca de estas dos afirmaciones usando los resultados de la parte (c)?

(e) Si las especificaciones técnicas de los automóviles de la marca A establecen que la diferencia promedio entre los automóviles en términos del número de kilómetros recorridos por galón de gasolina no puede ser mayor a 5 kilómetros, mediante el cálculo de un intervalo de confianza, discuta si esta especificación se está cumpliendo o no.

3.3 Para los mismos datos del punto anterior,

(a) Calcula el intervalo de confianza (3.2.21) para la varianza teórica de los automóviles de la marca A

(b) Calcula el intervalo de confianza de longitud más corta para la varianza teórica de los automóviles de la marca A modificando levemente el código de la página 169.

(c) Usando los intervalos de las partes (a) y (b), calcula dos intervalos de confianza para la desviación estándar. ¿Cuál intervalo tiene menor longitud?

(d) Calcula un intervalo de confianza para el coeficiente de variación para las dos marcas.

3.4 Complete los pasos para hallar los intervalos unilaterales para σ^2 dados en (3.2.23) y (3.2.24).

3.5 Dadas dos muestras aleatorias independientes $X_1, \cdots, X_{n_X}$ y $Y_1, \cdots, Y_{n_Y}$ provenientes de $N(\mu_X, \sigma^2)$ y $N(\mu_Y, \sigma^2)$ respectivamente, demuestre que

$$S_p^2 = \frac{(n_X - 1)S_{n_X-1,X}^2 + (n_Y - 1)S_{n_Y-1,Y}^2}{n_X + n_Y - 2}$$

es un estimador insesgado para la varianza común σ^2.

3.6 Dadas dos muestras aleatorias independientes $X_1, \cdots, X_{n_X}$ y $Y_1, \cdots, Y_{n_Y}$ provenientes de $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$ respectivamente, complete los pasos para encontrar el intervalo de confianza unilateral de forma $(-\infty, T)$ para $\mu_X - \mu_Y$ dado en el texto cuando

(a) σ_X^2 y σ_Y^2 son conocidas;

(b) σ_X^2 y σ_Y^2 son desconocidas, pero iguales.

3.7 Dadas dos muestras aleatorias independientes $X_1, \cdots, X_{n_X}$ y $Y_1, \cdots, Y_{n_Y}$ provenientes de $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$ respectivamente, con $\sigma_X^2 \neq \sigma_Y^2$ y desconocidas, complete los pasos para hallar los intervalos unilaterales y bilaterales para $\mu_X - \mu_Y$ cuando

(a) Las muestras son grandes

(b) Las muestras son moderadas o pequeñas.

3.8 Dadas dos muestras aleatorias independientes $X_1, \cdots, X_{n_X}$ y $Y_1, \cdots, Y_{n_Y}$ provenientes de $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$ respectivamente, construya un intervalo de confianza unilateral de forma (T, ∞) para σ_X^2/σ_Y^2 cuando

(a) μ_X y μ_Y son desconocidos;

(b) μ_X es conocido y μ_Y no es conocido.

3.9 En una muestra aleatoria proveniente de $N(\mu, \sigma^2)$

(a) Cuando $\mu = \mu_0$ escriba los posibles intervalos bilaterales y unilaterales para el coeficiente de variación cv.

(b) Cuando $\sigma = \sigma_0$ escriba los posibles intervalos unilaterales para el coeficiente de variación cv.

3.10 Un ganadero desea aumentar la producción lechera diaria de sus vacas, y decide probar un nuevo concentrado. Para verificar la efectividad del nuevo concentrado, el ganadero separa 35 vacas, de las cuales 15 son alimentadas con el concentrado actual y las restantes con el concentrado nuevo. Después de tres semanas de alimentación, él toma nota de la producción lechera. Para las vacas alimentadas con el concentrado actual, los resultados fueron (en litros): 16.4, 18.9, 15.7, 20.2, 16.8, 19.4, 14.7, 17.8, 19.5, 16.8, 18.4, 14.6, 20.7, 21.1, 17.3, y para las vacas alimentadas con el concentrado nuevo, los resultados fueron: 19.4, 18.1, 21.0, 20.4, 20.5, 17.4, 19.6, 18.4, 21.4, 19.2, 15.7, 22.8, 21.6, 17.2, 18.4, 19.4, 20.5, 23.6, 18.4, 18.3. ¿Debe el ganadero aceptar el concentrado nuevo o seguir con el concentrado actual? Construya un intervalo de confianza apropiado para contestar esta pregunta.

3.11 Demuestre que en la distribución $Ber(p)$, el parámetro p no es un parámetro de localización y tampoco es de escala.

3.12 Repita el punto anterior para el parámetro θ en la distribución $Pois(\theta)$.

3.13 En una muestra aleatoria proveniente de $Exp(\theta)$, complete los pasos para encontrar los intervalos (3.3.5) y (3.3.6).

3.14 En una muestra aleatoria proveniente de $Exp(\theta)$, encuentre los intervalos unilaterales para θ usando el teorema del límite central.

3.15 Diseñe un estudio de simulación para comparar los intervalos (3.3.3) y (3.3.9) en términos de la longitud. Muestre los resultados en una gráfica e interprete.

3.16 Considerando la propagación de una enfermedad respiratoria, suponga que las variables el tiempo que demora entre el contacto con una persona enferma y la manifestación de las síntomas pueden ser descritas con una distribución exponencial. Para conocer más acerca de esta distribución y consecuentemente conocer más acerca de esta enfermedad, se registra esta variable en varios pacientes, estos datos (medidos en horas) son 27, 6, 9, 4, 50, 15, 39, 1, 7, 14, 13, 5, 11, 3, 13, 70, 7, 37 y 17.

(a) Elabora un QQ plot para verificar que la distribución exponencial es apropiada para estos datos.

(b) Encuentra un intervalo del 95 % para la media teórica. ¿Cómo se interpreta este intervalo?

(c) Encuentra un intervalo de confianza del 95 % para el porcentaje de pacientes que demoran menos de 12 horas en manifestar los síntomas contando desde el momento de contacto.

3.17 El vendedor de seguros que hace visitas a posibles clientes masculinos para ofrecer plan médico para mascotas obtuvo 4 ventas exitosas en 15 visitas, usando el intervalo de Wald.

(a) Encuentra un intervalo de confianza para la probabilidad de tener éxito en una visita.

(b) Encuentra un intervalo de confianza para la probabilidad de tener dos éxitos en dos visitas.

(c) Encuentra un límite superior para la probabilidad de tener éxito en una visita.

(d) Repite la parte (a) y (b) usando el intervalo de Agresti y Caffo y el intervalo score definido en Cepeda, Aguilar, Cervantes, Corrales, Díaz & Rodríguez (2008).

3.18 En el punto anterior, el mismo vendedor en 20 visitas a clientas femeninas consiguió 9 ventas exitosas, mediante el cálculo de un intervalo de confianza apropiado para determinar si la probabilidad de lograr una venta exitosa con las mujeres es diferente que con los hombres.

Capítulo 4

Pruebas de hipótesis

El tema de las pruebas de hipótesis se difiere con la estimación puntual y la estimación por intervalo de confianza en el sentido de que éstos proveen valor o un conjunto de valores específicos que el parámetro de interés puede tomar, mientras que una prueba de hipótesis trata de verificar si una cierta afirmación acerca del parámetro puede considerarse como válida basándose en una muestra observada. Por consiguiente, una prueba de hipótesis es muy útil en situaciones donde no es de mucho interés el valor (estimado) del parámetro, sino la validez de la afirmación en cuestión. A continuación se presenta un ejemplo donde resulta útil el uso de pruebas de hipótesis.

Para el propósito de importación de ciertas motocicletas, la entidad ambiental del país importador necesita verificar que el nivel de contaminantes producidos por estas motocicletas cumple con las normas del país. En particular la emisión de monóxido de carbono (CO), denotada por μ, no debe superar a $5.5g/Km$. En este caso sólo se necesita verificar si la afirmación $\mu \leq 5.5g/Km$ puede considerarse como válida, mientras que la estimación puntual de μ no es de gran interés, como se verá más adelante.

4.1 Conceptos preliminares

Dada una muestra aleatoria $X_1, \ldots, X_n$ provenientes de una distribución con función de densidad $f(x, \theta)$, donde el espacio paramétrico del parámetro θ se denota por $\mathbf{\Theta}$. Un sistema de hipótesis está conformado por dos afirmaciones acerca de θ de la siguiente forma

$$H_0 : \ \theta \in \mathbf{\Theta}_0 \quad vs. \quad H_1 : \ \theta \in \mathbf{\Theta}_1 \tag{4.1.1}$$

donde $\mathbf{\Theta}_0, \mathbf{\Theta}_1$ son subconjuntos de $\mathbf{\Theta}$ y se denominan espacio paramétrico nulo y espacio paramétrico alterno, respectivamente. La única condición que se debe cumplir es $\mathbf{\Theta}_0 \cap \mathbf{\Theta}_1 = \emptyset$. H_0 se denomina la hipótesis nula, y H_1 la hipótesis alterna. El procedimiento del prueba de hipótesis consiste en decidir cuál de las dos hipótesis puede ser aceptada como válida, basado en una muestra aleatoria observada. Para eso,

se necesita, en primer lugar, encontrar una estadística de prueba; y posteriormente se debe establecer una regla de decisión que nos indica para qué valores de la estadística de prueba se debe rechazar H_0 y aceptar H_1.

La búsqueda de una regla de decisión para un sistema de hipótesis se lleva a cabo teniendo en cuenta que en el procedimiento de la prueba de hipótesis, se puede cometer dos tipos de errores:

- Rechazar H_0 cuando ésta es verdadera. Este error se denomina error tipo I.
- No rechazar H_0 cuando ésta es falsa. Este error se denomina error tipo II.

Es claro que ambos errores conducen a decisiones erróneas y éstas pueden causar pérdidas económicas y demás daños y perjuicios; por consiguiente, una buena regla de decisión debe garantizar que la probabilidad de cometer tanto el error tipo I como el error tipo II sea pequeña. Sin embargo, no es fácil, por no decir imposible, minimizar las dos probabilidades simultáneamente. Una solución a este problema es plantear el sistema (4.1.1) de tal forma que el error tipo I sea menos grave que el error tipo II, y garantizar únicamente que la probabilidad de cometer error tipo I sea pequeña. Una falla evidente de este procedimiento es que no se controla la magnitud del error tipo II, y la probabilidad de cometer este error puede ser realmente grande en algunos casos. Lo anterior ha sido siempre una de las críticas que existen en la literatura hacia el procedimiento de pruebas de hipótesis. Para implementar el anterior procedimiento, se necesita determinar, para cada sistema de hipótesis, cuál de los dos errores es más grave. En muchas situaciones prácticas se necesita la ayuda de expertos en el tema. Considere el ejemplo de motocicletas enunciado al comienzo de este capítulo.

Ejemplo 4.1.1. *Suponga que la emisión de CO de cierto tipo de motocicletas* Scooter *no debe superar a 5.5g/Km. La entidad ambiental responsable selecciona un número determinado de estas motocicletas para efectuar pruebas correspondientes. Si se denota la emisión de CO de estas motocicletas por μ, entonces existen dos hipótesis acerca de μ: $\mu \leq 5.5$ y $\mu > 5.5$. Basada en los resultados de la muestra, la entidad debe decidir cuál de las dos hipótesis aceptar. Si el sistema de hipótesis planteado es*

$$H_0 : \ \mu \leq 5.5 \quad vs. \quad H_1 : \ \mu > 5.5 \tag{4.1.2}$$

entonces tenemos que

- *El error tipo I implica rechazar $\mu \leq 5.5$ cuando realmente $\mu \leq 5.5$, esto es, motocicletas que están emitiendo una cantidad permitida de CO no pasan la prueba. Y esto puede causar una pérdida económica para la empresa que fabrica y la empresa que importa estas motocicletas. Y esto puede causar despido de empleados, y en el peor de los casos, la quiebra de las dos empresas.*
- *El error tipo II implica aceptar $\mu \leq 5.5$ cuando en la realidad $\mu > 5.5$, esto implica que motocicletas que emiten una gran cantidad de CO pasan la prueba y pueden ser importadas. Y esto causará inevitablemente la contaminación excesiva del medio ambiente.*

Si el experto del tema considera más grave la pérdida económica de las empresas que la contaminación del medio ambiente, se puede mantener el sistema (4.1.2) planteado anteriormente; mientras que si la prioridad es conservar el medio ambiente, el sistema que se plantea debe ser

$$H_0: \ \mu > 5.5 \quad vs. \quad H_1: \ \mu \leq 5.5,$$

el cual, como se verá más adelante, tiene la misma regla de decisión que el sistema

$$H_0: \ \mu \geq 5.5 \quad vs. \quad H_1: \ \mu < 5.5$$

En este ejemplo, considerando el contexto del problema, el espacio paramétrico de μ *es* $\boldsymbol{\Theta} = (0, \infty)$*, y en el sistema (4.1.2),* $\boldsymbol{\Theta}_0 = (0, 5.5]$ *y* $\boldsymbol{\Theta}_1 = (5.5, \infty)$.

En el ejemplo anterior, la unión de los conjuntos $\boldsymbol{\Theta}_0$ y $\boldsymbol{\Theta}_1$ conforma espacio paramétrico completo $\boldsymbol{\Theta}$, todos los sistemas de hipótesis tienen que cumplir con esta condición. Suponga que la empresa que fabrica las motocicletas del ejemplo anterior diseña un dispositivo para disminuir la emisión del gas CO. Si las motocicletas sin el dispositivo emiten $3.7g/Km$, entonces para probar la eficiencia del dispositivo, el sistema de hipótesis que se plantea será

$$H_0: \ \mu = 3.7 \quad vs. \quad H_1: \ \mu < 3.7.$$

Y en este caso, $\boldsymbol{\Theta}_0 = \{3.7\}$, $\boldsymbol{\Theta}_1 = (0, 3.7)$ y la unión de estos conjuntos no conforman el espacio paramétrico completo.

Para el desarrollo de la teoría básica concerniente al tema de pruebas de hipótesis primero se estudian muestras provenientes de la distribución normal y posteriormente se consideran muestras provenientes de distribuciones diferentes a la normal.

4.2 Una muestra bajo normalidad

Como supuesto general para esta parte del libro, suponga que se dispone de una muestra aleatoria $X_1, \cdots, X_n$ provenientes de una distribución $N(\mu, \sigma^2)$. Estudiaremos por separado los sistemas de hipótesis para μ y σ^2.

4.2.1 Pruebas de hipótesis para la media teórica

En esta sección, estudiamos las pruebas de hipótesis acerca de la media teórica μ en una distribución $N(\mu, \sigma^2)$, es decir, basado en una muestra aleatoria $X_1, \cdots, X_n$ con distribución teórica $N(\mu, \sigma^2)$, estamos interesados en encontrar una regla de decisión para el sistema

$$H_0: \ \mu \in \boldsymbol{\Theta}_0 \quad vs. \quad H_1: \ \mu \in \boldsymbol{\Theta}_1$$

El procedimiento para encontrar una regla de decisión depende de si el valor de la varianza teórica es conocido o no[1]; también plantearemos diferentes sistemas de hipótesis según $\boldsymbol{\Theta}_0$ y $\boldsymbol{\Theta}_1$ toma diferentes formas.

[1] Análogo a los intervalos de confianza para μ que también depende de si σ^2 es conocida o no.

$H_0: \mu = \mu_0$ **vs.** $H_1: \mu \neq \mu_0$ **con** σ^2 **conocida**

Este sistema de hipótesis es apto para cuando hay un valor supuesto para la media teórica y se quiere verificar la validez de este supuesto. Por ejemplo, muchas especificaciones técnicas que encontramos en los empaques de productos como eléctricos, farmaceúticos, de ferretería, entre otros, son valores que se suponen válidos para estos productos, en el caso de que haya sospecha de que el valor propuesto μ_0 ya no es válido para la población, y tampoco hay sospecha de que el valor verdadero de la media teórica esté mayor o menor que μ_0. Podemos utilizar la hipótesis

$$H_0: \mu = \mu_0 \quad vs. \quad H_1: \mu \neq \mu_0 \tag{4.2.1}$$

Necesitamos encontrar una regla de decisión en términos de alguna estadística de prueba que nos permita juzgar el sistema para cada posible muestra observada. Hay diferentes formas de encontrar una regla de decisión, ilustramos el razonamiento de uno de estos métodos con una situación práctica.

Suponga que la longitud de los clavos producidos en cierta fábrica debe ser de 5cm, para verificar que los productos terminados en una línea de producción satisfacen con este requisito, se seleccionan aleatoriamente 50 clavos, se mide la longitud de cada uno de estos y se obtiene la longitud promedio de estos 50 clavos. Es claro que el promedio muestral de estos 50 clavos es una estimación de la media teórica; de esta forma, si este promedio muestral es de 7cm, claramente se concluye que la línea de producción no cumple con el requisito, ya que intuitivamente 7cm está muy lejos del valor especificado de 5cm. Se tiene la misma conclusión si este promedio es de 3cm, por ejemplo. En otras palabras, la simple lógica sugiere que se debe rechazar la hipótesis de que la longitud promedio de los clavos producidos por la línea de producción es de 5cm si el promedio muestral está muy alejado de 5cm.

Escribiendo formalmente la idea anterior, se rechaza H_0 en (4.2.1) cuando el valor de $\bar{x}$ está muy alejado de μ_0, y podemos describirlo como

$$\text{Rechazar } H_0 \text{ cuando } |\bar{x} - \mu_0| > K \text{ para algún } K > 0.$$

El uso del valor absoluto se debe a que matemáticamente la distancia entre dos valores a y b se mide con $|a-b|$. La anterior afirmación es la regla de decisión para el sistema (4.2.1) y necesitamos encontrar el valor de K para completarla. Para eso, podemos hacer uso de la definición del error tipo I, pero en primer lugar, se debe determinar cuál es la máxima magnitud permitida de la probabilidad de cometer este tipo de errores. Este límite superior se denotará por α y se conoce como el tamaño de la prueba o el nivel de significación. Teóricamente, el valor de α puede ser cualquier valor entre 0 y 1, pero como ésta sirve para restringuir la probabilidad de cometer error tipo I, se escogen valores pequeños para α. Usualmente en la práctica se usan valores como 0.01, 0.02, 0.05, 0.10. De esta forma, tenemos que

$$\alpha \geq Pr(\text{Rechazar} H_0 \mid H_0 \text{ es verdadera})$$

Ahora, una regla de decisión se define en términos de una estadística, entonces para calcular la anterior probabilidad, se necesita conocer la distribución de la estadística

suponiendo cierta H_0. Esta distribución se conoce como la distribución nula, y depende de H_0. Dependiente de la forma de H_0, tenemos los dos siguientes casos

- Cuando H_0 es de la forma $\theta = \theta_0$ para algún valor específico θ_0, la hipótesis se conoce como **hipótesis simple**, el espacio paramétrico nulo toma la forma $\mathbf{\Theta}_0 = \{\theta_0\}$. En este caso, la distribución nula de la estadística de prueba está determinada de manera única, y tomamos

$$\alpha = Pr(\text{Rechazar}H_0 \mid H_0 \text{ es verdadera}).$$

- Cuando H_0 es tal que el espacio paramétrico nulo contiene más de un valor para el parámetro se denomina **hipótesis compuesta**, y la distribución nula de la estadística de prueba consiste en una familia de distribuciones, y por consiguiente $Pr(\text{Rechazar}H_0 \mid H_0 \text{ es verdadera})$ puede tomar más de un valor. En este caso, tomamos

$$\alpha = \sup_{\theta \in \mathbf{\Theta}_0} Pr(\text{Rechazar}H_0 \mid H_0 \text{ es verdadera}).$$

En el sistema de hipótesis (4.2.1), H_0 es una hipótesis simple; por consiguiente, tenemos que

$$\begin{aligned}
\alpha &= Pr(\text{Rechazar}H_0 \mid H_0 \text{ es verdadera}) \\
&= Pr(|\bar{X} - \mu_0| > K \mid H_0 \text{ es verdadera}) \\
&= Pr(\bar{X} - \mu_0 > K \mid H_0 \text{ es verdadera}) + Pr(\bar{X} - \mu_0 < -K \mid H_0 \text{ es verdadera})
\end{aligned}$$

Nótese que cuando H_0 es verdadera, tenemos que $\mu = \mu_0$, entonces la distribución nula de $\bar{X}$ es $N(\mu_0, \sigma^2/n)$. Usando esta distribución nula de $\bar{X}$, tenemos que

$$\begin{aligned}
\alpha &= Pr\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{K}{\sigma/\sqrt{n}}\right) + Pr\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -\frac{K}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(\frac{K}{\sigma/\sqrt{n}}\right) + \Phi\left(-\frac{K}{\sigma/\sqrt{n}}\right) \\
&= 2\Phi\left(-\frac{K}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

de donde se concluye que $-\frac{K}{\sigma/\sqrt{n}} = z_{\alpha/2}$, de donde $K = -z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$. Con el valor de K encontrado, tenemos la siguiente regla de decisión para el sistema (4.2.1):

$$\text{Rechazar } H_0 \text{ cuando } |\bar{X} - \mu_0| > z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Se puede ver fácilmente que la anterior regla de decisión es equivalente a

$$\text{Rechazar } H_0 \text{ cuando } |\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}| > z_{1-\alpha/2}.$$

O equivalente a

Rechazar H_0 cuando $\bar{X} > \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ ó $\bar{X} < \mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$.

En conclusión, para el anterior sistema de hipótesis, existen tres reglas de decisión equivalentes:

(a) $|\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}| > z_{1-\alpha/2}$ donde la estadística de prueba es $\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}$

(b) $|\bar{X} - \mu_0| > z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ donde la estadística de prueba es $\bar{X} - \mu_0$

(c) $\bar{X} > \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ ó $\bar{X} < \mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ donde la estadística de prueba es $\bar{X}$.

Las anteriores reglas de decisión se basan en una estadística de prueba, en la práctica cuando se observa un conjunto de datos $x_1, \cdots, x_n$, se calcula el valor de la estadística y se verifica si se cumple cualquiera de las anteriores reglas de decisión. Presentamos una ilustración en el siguiente ejemplo.

Ejemplo 4.2.1. *Retomando el ejercicio 6 del capítulo 2. Una máquina de llenado de botellas debe estar programada para efectuar un llenado de 350ml, y se quiere conocer el funcionamiento real de la máquina. Para eso se extrajeron aleatoriamente 20 botellas llenadas por la máquina y se midió el contenido de la botella, los resultados fueron: 355, 350, 340, 345, 354, 358, 350, 343, 349, 346, 351, 358, 342, 350, 356, 345, 349, 356, 354, 346. Una simple gráfica QQ nos ilustra que se puede suponer que los datos provienen de una distribución normal. Suponga adicionalmente que la desviación estándar es igual a 5ml, si se desea evaluar la calidad de la máquina de llenado, el sistema de hipótesis es*

$$H_0: \ \mu = 350 \quad vs. \quad H_1: \ \mu \neq 350,$$

y supongamos que se probará el sistema con un nivel de significación igual a 5 %.

Como se comentó anteriormente, la regla de decisión puede ser cualquiera de las tres dadas anteriormente, puesto que las tres son equivalentes y conducen a la misma decisión. Si usamos la regla de decisión (b), se calcula $|\bar{x} - 350|$ que es igual a 0.15, y por el otro lado se calcula $z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ que es igual a 1.75, de donde se concluye que $|\bar{x} - 350|$ no es mayor a $z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$, lo cual conduce a la decisión de no rechazar o aceptar H_0, y se puede afirmar que la máquina efectivamente realiza un llenado de 350ml.

Región de rechazo

Otro concepto importante en las pruebas de hipótesis es la región de rechazo asociada a una regla de decisión, y se define como el conjunto conformado por todos los valores de la estadística de prueba que conducen a la decisión de rechazar H_0. En cada una de las tres anteriores reglas de decisión (a), (b) y (c), la estadística de prueba es diferente, y para cada una de ellas, podemos obtener la correspondiente región de rechazo:

- Para la estadística $\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}$ de la regla de decisión (a), la región de rechazo es $\{c \in \mathbb{R} : c > z_{1-\alpha/2}$ ó $c < -z_{1-\alpha/2}\}$. Podemos ilustrar esta región de rechazo en la Figura 4.1, junto con la distribución nula de la estadística de prueba que corresponde a la distribución normal estándar.

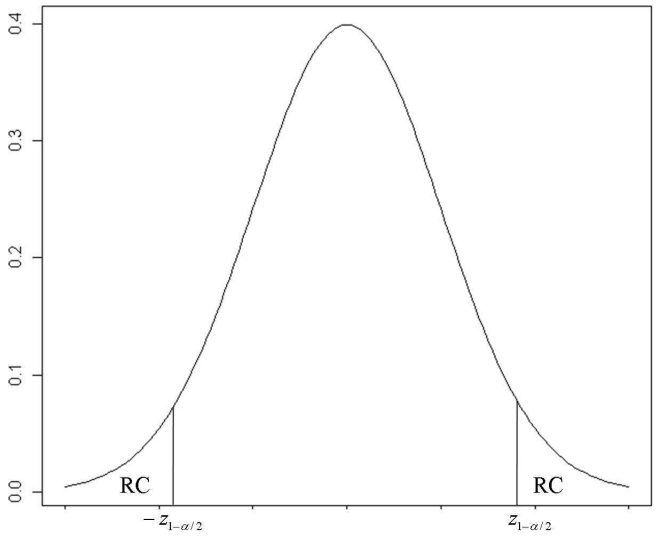

Figura 4.1: *Ilustración de la región de rechazo de la regla de decisión (a).*

- Ahora para la estadística $\bar{X}$ de la regla de decisión (c), la región de rechazo es $\{c \in \mathbb{R} : c > \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ ó $c < \mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$. Ahora nótese que la distribución bajo la hipótesis nula de $\bar{X}$ es $N(\mu_0, \sigma^2/n)$. Ilustramos la región de rechazo en la Figura 4.2 suponiendo que $\sigma = 2$ y $n = 20$.

En resumen, para un sistema de hipótesis dado, puede haber varias reglas de decisión, y asociadas a ellas, varias estadísticas de pruebas, y para cada una de las estadísticas de prueba, se tiene la respectiva región de rechazo. Por lo tanto, cuando se refiere al término región de rechazo, siempre debe especificar cuál es la estadística de prueba correspondiente.

p valor

Una vez dado el concepto de región de rechazo, ahora estudiamos un concepto de fundamental importancia en las pruebas de hipótesis: el denominado p valor. Para introducir el concepto, considérese de nuevo el sistema de hipótesis

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0,$$

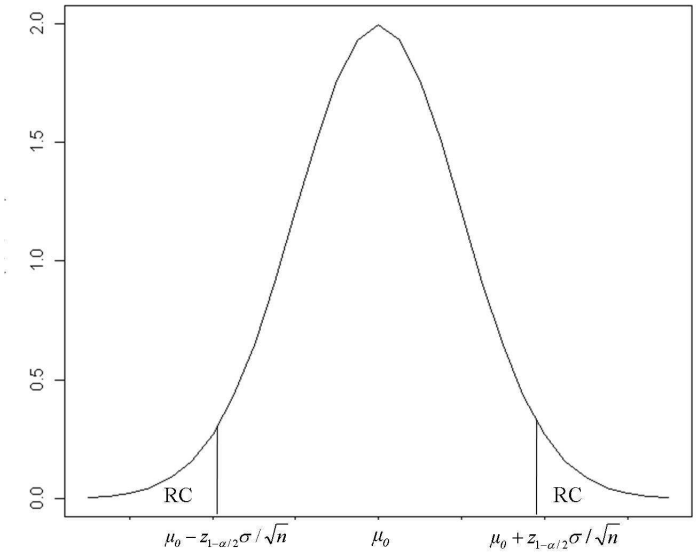

Figura 4.2: *Ilustración del región de rechazo de la regla de decisión (c).*

basándose en una muestra con distribución normal, supongamos que se usa la regla de decisión (a) dada anteriormente, de manera que cuando se observa la realización de una muestra $x_1, \cdots, x_n$, se calcula el valor de la estadística de prueba, en este caso, $\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}$. Si el valor de la estadística está dentro de la región de rechazo $\{c \in \mathbb{R} : c > z_{1-\alpha/2}$ ó $c < -z_{1-\alpha/2}\}$, entonces se rechaza H_0, de lo contrario se acepta H_0.

Retomando el Ejemplo 4.2.1, el valor de la estadística $\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}$ en la muestra observada es de -0.1341. Si el nivel de significación es de 5 %, entonces $z_{1-\alpha/2} = 1.96$, y claramente el valor -0.1341 no se encuentra dentro de la región de rechazo. Nótese que la decisión tomada es la misma que en el Ejemplo 4.2.1; sin embargo, tanto el razonamiento del Ejemplo 4.2.1 como el uso de la región de rechazo tiene una desventaja que radica en que si el usuario decide usar otro nivel de significación, el procedimiento debe realizarse de nuevo y eso no es eficiente computacionalmente. Es por esta razón que los programas o software estadísticos siempre calculan el p valor que nos permite realizar la prueba de hipótesis para diferentes niveles de significación.

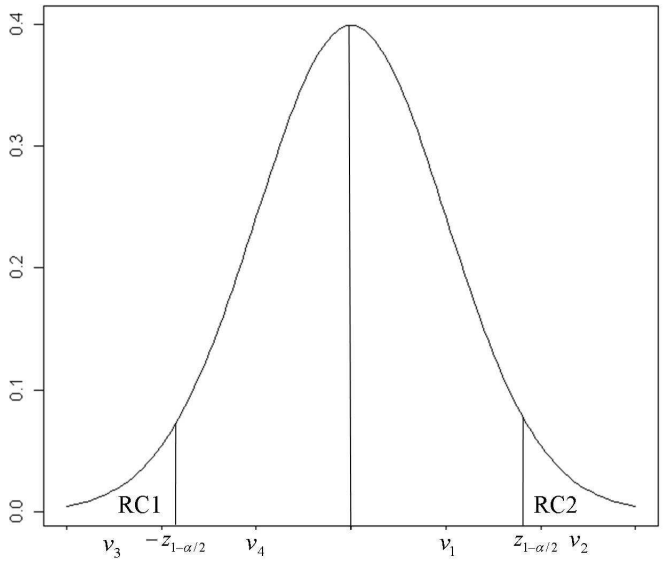

Figura 4.3: *Ilustración del región de rechazo y p valor de la regla de decisión (a).*

El concepto del p valor está íntimamente ligado con el de región de rechazo. Observe la Figura 4.3, donde se muestra la región de rechazo si se utiliza la regla de decisión (a). Dada la forma de la distribución nula de la estadística de prueba y la región de rechazo, los valores que puede tomar la estadística de prueba se pueden dividir en cuatro grupos: (1) entre 0 y $z_{1-\alpha/2}$, (2) mayor que $z_{1-\alpha/2}$, (3) entre 0 y $-z_{1-\alpha/2}$ y finalmente, (4) menor que $-z_{1-\alpha/2}$. De esta forma, si en una muestra observada el valor de la estadística de prueba, digamos v_1, es como el caso (1) mostrado en la gráfica, es claro que no se rechaza H_0. Nótese que en este caso, el área hacia la derecha de v_1 es mayor que el área de RC2, el cual es $\alpha/2$. Ahora si el valor de la estadística v_2 es como el caso (2), entonces pertenece a la región de rechazo, que es equivalente al hecho de que el área hacia la derecha de v_2 es menor que $\alpha/2$. Lo anterior sugiere establecer una nueva regla de decisión:

Rechazar H_0 si el área a la derecha del valor de la estadística es menor a $\alpha/2$

que es equivalente a

Rechazar H_0 si el área a la derecha del valor de la estadística multiplicado por 2 es menor a α.

Ahora, la anterior regla de decisión es correcta si el valor de la estadística es mayor que 0; en el caso contrario, el análisis es diferente. Suponga que el valor de la estadística es igual a v_3, es claro que el área hacia la derecha de v_3 es mayor a $\alpha/2$; de hecho, es mayor a 0.5. Sin embargo, éste pertenece a la región de rechazo. El análisis correcto, cuando el valor de la estadística es menor que 0, es observar el área hacia la izquierda. Entonces para el valor v_3, el área hacia la izquierda es menor a $\alpha/2$ y se rechaza H_0; para el valor v_4, el área hacia la izquierda es mayor a $\alpha/2$, y no se rechaza H_0.

En conclusión, denotando el valor de la estadística de prueba $\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}$ como v, podemos establecer la siguiente regla de decisión:

Rechazar H_0 si

$$\begin{cases} \text{el área a la derecha de } v \text{ multiplicada por 2 es menor a } \alpha, & \text{para } v > 0 \ ; \\ \text{el área a la izquierda de } v \text{ multiplicada por 2 es menor a } \alpha, & \text{para } \ v < 0. \end{cases}$$

Ahora, recordando que un área bajo la curva de una función de densidad se puede interpretar como una probabilidad, la anterior regla de decisión se convierte en:

$$\text{Rechazar } H_0 \text{ si} \begin{cases} 2Pr\left(\dfrac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} > v\right) < \alpha, & \text{para } v > 0 \ ; \\ 2Pr\left(\dfrac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} < v\right) < \alpha, & \text{para } \ v < 0. \end{cases}$$

Y el p valor asociado al valor de la estadística v se define como

$$p \text{ valor} = \begin{cases} 2Pr\left(\dfrac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} > v\right), & \text{para } v > 0 \ ; \\ 2Pr\left(\dfrac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} < v\right), & \text{para } \ v < 0. \end{cases}$$

O equivalentemente

$$p \text{ valor} = \begin{cases} 2Pr(Z > v), & \text{para } v > 0 \ ; \\ 2Pr(Z < v), & \text{para } \ v < 0. \end{cases}$$

donde Z denota una variable aleatoria con distribución normal estándar. Dada la anterior definición del p valor podemos reescribir la regla de decisión en término del p valor:

Rechazar H_0 si el p valor es menor al nivel de significación α.

Ilustremos el cálculo y el uso del p valor con el problema descrito en el Ejemplo 4.2.1. usando la regla de decisión (a). El valor de la estadística $\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}$ es -0.1341, el cual es menor que 0, por lo tanto, el p valor se define como

$$2Pr\left(\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} < -0.1341\right) = 0.8933.$$

Entonces si el nivel de significación es de 5 %, no se rechaza H_0, pues el p valor es mayor a 5 %. Ahora, si se cambia el nivel de significación a 10 %, no es necesario volver a realizar cálculos numéricos para tomar la decisión, pues simplemente comparamos el p valor con el 10 %, y la decisión sigue siendo "No rechazar H_0". Por esta razón, en la práctica, el uso de p valor puede ser más sencillo que los otros procedimientos, y los software estadísticos, en la mayoría de los casos, simplemente arrojan el valor de

la estadística de prueba y el p valor, y los usuarios pueden comparar el p valor con el nivel de significación deseado.

Es difícil dar una definición formal de p valor, puesto que su cálculo depende en primer lugar del planteamiento del sistema de hipótesis, y en segundo lugar depende de la estadística de prueba que se utiliza. Una definición muy popular de p valor, pero errónea, afirma que el p valor es la probabilidad de que H_0 es verdadera. De esta manera, si el p valor es menor que el nivel de significación α, implica que la probabilidad de que H_0 sea verdadera es pequeña, conduciendo a la decisión de rechazar H_0. A pesar de que el anterior razonamiento funciona, no es correcto, puesto que para H_0, solo hay dos posibilidades, o es verdadera o es falsa, así que la probabilidad de que H_0 sea verdadera es o bien 0 o bien 1, mientras que p valor puede tomar cualquier valor entre 0 y 1.

Función de potencia

Para un sistema de hipótesis, puede haber más de una regla de decisión. En este caso, necesitamos comparar estas reglas de decisión en términos de la probabilidad de cometer los dos tipos de errores, para lo cual, definimos la función de potencia.

Definición 4.2.1. *La función de potencia de una regla de decisión para un sistema de hipótesis es una función del parámetro θ definida como*

$$\beta(\theta) = Pr(\textit{Rechazar } H_0) \tag{4.2.2}$$

Nótese que $1-\beta(\theta) = Pr(\text{Aceptar } H_0)$, y podemos asociar la función de potencia con los dos errores que se pueden cometer en una prueba de hipótesis de la siguiente forma

- Si la hipótesis H_0 es verdadera, esto es, si $\theta \in \mathbf{\Theta}_0$, entonces al rechazar H_0, se está cometiendo el error tipo I, y la máxima probabilidad de cometer el error tipo I es el nivel de significación α. De esta forma $\beta(\theta) \leq \alpha$
- Si la hipótesis H_0 es falsa, esto es, si $\theta \in \mathbf{\Theta}_0^c$ (no necesariamente $\theta \in \mathbf{\Theta}_1$), entonces aceptar H_0 equivale a cometer el error tipo II. Y si denotamos la probabilidad de cometer error tipo II como β, tenemos que $1-\beta(\theta) = \beta$.

Y en conclusión, la función de potencia queda expresada como

$$\beta(\theta) = \begin{cases} \text{Probabilidad de cometer error tipo I} \leq \alpha & \text{si } \theta \in \mathbf{\Theta}_0 \\ 1-\beta & \text{si } \theta \in \mathbf{\Theta}_0^c \end{cases} \tag{4.2.3}$$

y podemos ver claramente que la función de potencia es una función del parámetro θ. Ahora miremos cómo se calcula la función de potencia para la regla de decisión encontrada anteriormente para el sistema de hipótesis

$$H_0: \ \mu = \mu_0 \quad vs. \quad H_1: \ \mu \neq \mu_0,$$

Tenemos que

$$\begin{aligned}
\beta(\mu) &= Pr(\text{Rechazar } H_0) \\
&= Pr\left(\bar{X} > \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ ó } \bar{X} < \mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\
&= Pr\left(\bar{X} > \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) + Pr\left(\bar{X} < \mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\
&= Pr\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right) + Pr\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{\mu_0-\mu}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right) \\
&= 1 - \Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right) + \Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right) \qquad (4.2.4)
\end{aligned}$$

donde Φ denota la función de distribución de una variable normal estándar. Podemos observar las siguientes propiedades de la función $\beta(\mu)$:

- Para $\mu \in \boldsymbol{\Theta}_0$, esto es, cuando $\mu = \mu_0$, $\beta(\mu) = 1-\Phi(z_{1-\alpha/2})+\Phi(-z_{1-\alpha/2}) = \alpha$, y coincide con el primer caso de (4.2.3).
- Para $\mu \in \boldsymbol{\Theta}_1$, esto es, cuando $\mu \neq \mu_0$, entre más alejado esté μ de μ_0, el término $\frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2}$ se aproxima a $\frac{\mu_0-\mu}{\sigma/\sqrt{n}} - z_{1-\alpha/2}$ y en consecuencia $\beta(\mu)$ se acerca al valor 1. Dado (4.2.3), esto indica que cuando μ es muy diferente de μ_0, la probabilidad de cometer error tipo II es muy pequeña, es decir, la regla de decisión será capaz de reconocer hipótesis nulas falsas.

  También, cuando el tamaño de muestra n es grande, $\frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2}$ y $\frac{\mu_0-\mu}{\sigma/\sqrt{n}} - z_{1-\alpha/2}$ se hace grande, y ambos $\Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right)$ y $\Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right)$ se acercan a 1 y la función $\beta(\mu)$ también se acerca a 1, esto es, entre mayor sea la muestra, una hipótesis nula falsa es más fácil de detectar.
- Es claro que la función de potencia depende del nivel de significación α que se interpreta como la probabilidad de cometer error tipo I, y se determina de antemano. Al principio se comentó que no se pueden minimizar simultáneamente las probabilidades de cometer error tipo I y error tipo II, entonces se espera que la función de potencia tome valores pequeños cuando α es pequeño y al observar la forma de (4.2.4), confirmamos lo dicho anteriormente.

Los anteriores comentarios se pueden ilustrar graficando la función de potencia para diferentes tamaños de muestra. Utilizamos el siguiente código, y la gráfica resultante se muestra en la Figura 4.4.

```
> po_norm<-function(mu,n,mu0,alpha,sigma){
+ 1-pnorm((mu0-mu)*sqrt(n)/sigma+qnorm(1-alpha/2))+
+ pnorm((mu0-mu)*sqrt(n)/sigma-qnorm(1-alpha/2))
+ }
> alpha<-0.05
> sigma<-1
```

```
> mu0<-2
> n1<-10
> n2<-30
> n3<-50
>
> plot(function(x) po_norm(x,n1,mu0,alpha,sigma),0,4,type="l",
+ xlab="mu",ylab="función de potencia")
> curve(po_norm(x,n2,mu0,alpha,sigma),0,4,lty=2,add=T)
> curve(po_norm(x,n3,mu0,alpha,sigma),0,4,lty=3,add=T)
> legend(3,0.4,c("n=10","n=30","n=50"),lty=c(1,2,3))
> windows()
>
> alpha1<-0.03
> alpha2<-0.05
> alpha3<-0.1
> sigma<-1
> mu0<-2
> n<-20
>
> plot(function(x) po_norm(x,n,mu0,alpha1,sigma),0,4,type="l",
+ xlab="mu",ylab="función de potencia")
> curve(po_norm(x,n,mu0,alpha2,sigma),0,4,lty=2,add=T)
> curve(po_norm(x,n,mu0,alpha3,sigma),0,4,lty=3,add=T)
> legend(2.6,0.3,c("alpha=0.03","alpha=0.05","alpha=0.1"),lty=c(1,2,3))
```

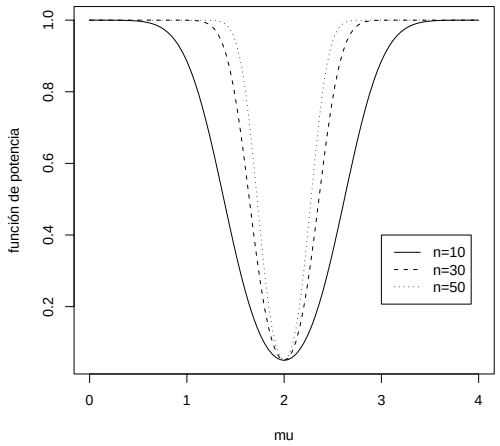

Figura 4.4: *Función de potencia (4.2.4) para diferentes tamaños de muestra con* $\mu_0 = 2$, $\sigma = 1$ *y* $\alpha = 0.05$

Por otro lado, también podemos graficar la función de potencia para diferentes niveles de significación, para corroborar que entre mayor sea α, mayor es la potencia de la prueba. Una modificación del anterior código arroja la gráfica en la Figura 4.5.

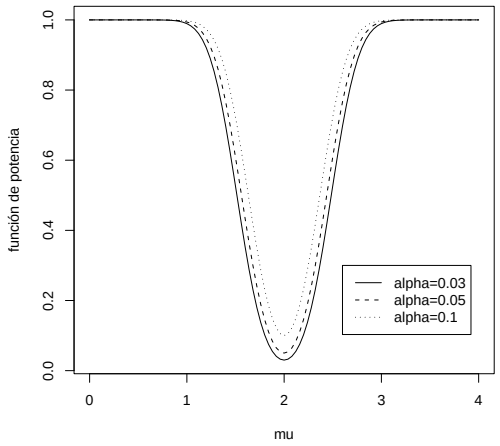

Figura 4.5: *Función de potencia (4.2.4) para diferentes niveles de significación con* $\mu_0 = 2$, $\sigma = 1$ *y* $n = 20$.

Aceptar $H_0: \ \mu = \mu_0$ **no significa que** μ **tome el valor** μ_0

Los procedimientos vistos hasta el momento nos ayudan a tomar una decisión acerca de un sistema de hipótesis, y es muy común pensar que al aceptar la hipótesis nula $\mu = \mu_0$, ya tenemos la certeza de que en la media teórica μ toma el valor de μ_0, lo cual no es correcto, puesto que el hecho de que no se rechaza $\mu = \mu_0$ solo indica que al asumir el valor μ_0 para μ, no se presenta ninguna discordancia con lo observado en la muestra, mas puede haber otros valores ≪apropiados≫ para μ.

Además, en la toma de decisiones acerca de H_0, hay posibilidad de cometer error tipo II, esto es, aceptar $H_0: \ \mu = \mu_0$ aún cuando el valor verdadero de μ no sea μ_0. Para ilustrarlo, se realiza un estudio de simulación donde el valor de μ es 3, pero se plantean varias hipótesis nulas $\mu = \mu_0$ con $\mu_0 = 2.9,\ 2.95,\ 3,\ 3.05$, etc. Para cada uno de estos valores de μ_0 se simulan 1000 muestras de tamaño 30 de la distribución $N(3, 1)$, y para cada una de las muestras simuladas se aplica la regla de decisión rechazar H_0 si $|\sqrt{n}(\bar{x} - \mu_0)|/\sigma > z_{1-\alpha/2}$, y se calcula cuántas veces se acepta H_0. Los resultados se presentan en la Figura 4.6, donde se observa que entre más cercano esté μ_0 del μ verdadero, más fácil es aceptar $H_0: \ \mu = \mu_0$, es decir, hay mayor probabilidad de cometer error tipo II. Esto es lógico puesto que si μ_0 es muy cercano a μ hay poca diferencia entre una muestra proveniente de $N(\mu_0, \sigma^2)$ y una proveniente de $N(\mu, \sigma^2)$, y por consiguiente es difícil detectar la falsedad de la afirmación $\mu = \mu_0$.

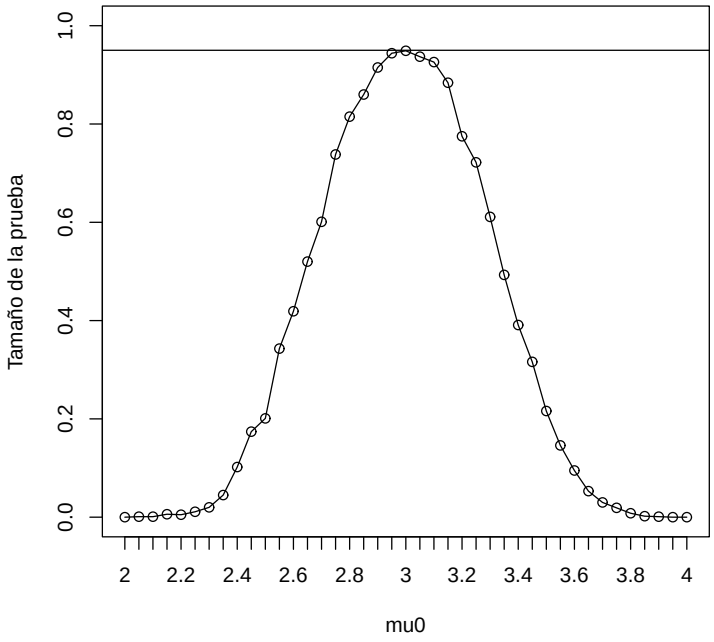

Figura 4.6: *Ilustración del región de rechazo y p valor de la regla de decisión (a).*

En el Ejemplo 4.2.1 sobre el problema de llenado de botella, se trató de probar si el llenado promedio es de 350 ml, basado en una muestra. Se llegó a la conclusión de que se puede asumir que $H_0: \ \mu = 350ml$ es verdadera. Sin embargo, según las anotaciones anteriores, existe la posibilidad de que μ no tome el valor de 350 ml, y a pesar de esto, se llega a la conclusión de aceptar H_0.

Entonces al observar la decisión de aceptar H_0, no estamos diciendo que tenemos la seguridad de que $\mu = 350\ ml$, sino que la diferencia entre μ y $350\ ml$ es muy pequeña y por consiguiente no se detecta con los datos. Ahora, en la práctica, tenemos que estar conscientes de que el valor a probar μ_0 no necesariamente es el único valor que es aceptable para el experto del tema en cuestión. Por ejemplo, se supone que la máquina debe efectuar un llenado de 350ml, pero si la máquina efectúa un llenado de 348ml, también puede ser aceptable para la compañía.

Los estudiantes deben tener claro que la ciencia de estadística se trata de incertidumbres, y no hay verdades absolutas como en la ciencia de matemática; por lo tanto, al aceptar $\mu = \mu_0$, lo único que podemos concluir es que los datos no manifiestan ningún conflicto con esta afirmación. De hecho se puede ver que para los datos del Ejemplo 4.2.1, si planteamos el sistema de hipótesis como $H_0: \ \mu = 349\ ml$ vs. $H_1: \ \mu \neq 349\ ml$ también se llega a la decisión de aceptar H_0.

$H_0: \ \mu \leq (=)\mu_0$ **vs.** $H_1: \ \mu > \mu_0$ **con** σ^2 **conocida**

Ahora, consideramos un sistema de hipótesis donde H_0 es compuesta y la búsqueda de una regla de decisión es un poco diferente que cuando H_0 es simple. Supongamos ahora que la hipótesis que se desea verificar es que la media teórica no supere a cierto valor μ_0, y el sistema que se considera es el siguiente

$$H_0: \ \mu \leq \mu_0 \quad vs. \quad H_1: \ \mu > \mu_0, \tag{4.2.5}$$

donde el espacio paramétrico nulo está dado por $\boldsymbol{\Theta}_0 = (\infty, \mu_0]$. Para encontrar una regla de decisión, es natural pensar en rechazar H_0 y aceptar $H_1: \ \mu > \mu_0$ cuando en una muestra observada la estimación de μ, $\bar{x}$ es muy grande comparada con μ_0. Esto conduce a la siguiente regla de decisión:

$$\text{Rechazar } H_0 \text{ si } \bar{X} - \mu_0 > K \tag{4.2.6}$$

para algún $K > 0$. Para encontrar el valor de K, se hace uso de la definición del error tipo I, y restringiendose a que la probabilidad de cometer este tipo de error sea a lo más α. Considerando que H_0 es una hipótesis compleja, tenemos que

$$\begin{aligned} \alpha &= \sup_{\mu \in \boldsymbol{\Theta}_0} Pr(\text{Rechazar } H_0 | \ H_0 \text{ es verdadera}) \\ &= \sup_{\mu \in \boldsymbol{\Theta}_0} Pr(\bar{X} - \mu_0 > K \mid H_0 \text{ es verdadera}) \end{aligned}$$

Cuando H_0 es verdadera, la media teórica μ es menor o igual a μ_0, por lo tanto la distribución nula de $\bar{X}$ es $N(\mu, \sigma^2/n)$, donde $\mu \leq \mu_0$, estandarizando la variable $\bar{X}$, tenemos que

$$\begin{aligned} \alpha &= \sup_{\mu \in \boldsymbol{\Theta}_0} Pr\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{K + \mu_0 - \mu}{\sigma/\sqrt{n}} \right) \\ &= \sup_{\mu \in \boldsymbol{\Theta}_0} \left(1 - \Phi\left(\frac{K + \mu_0 - \mu}{\sigma/\sqrt{n}} \right) \right) \end{aligned}$$

Para encontrar el valor de la constante K, necesitamos ver cuál es el valor de μ en $\boldsymbol{\Theta}_0$ que hace máximo a $1 - \Phi\left(\frac{K+\mu_0-\mu}{\sigma/\sqrt{n}}\right)$. Es claro que $1 - \Phi\left(\frac{K+\mu_0-\mu}{\sigma/\sqrt{n}}\right)$ es una función creciente de μ, es decir, un valor grande de μ hace grande a $1 - \Phi\left(\frac{K+\mu_0-\mu}{\sigma/\sqrt{n}}\right)$, pero la hipótesis nula asegura que $\mu \leq \mu_0$, entonces podemos ver que el valor de μ que maximiza a $1 - \Phi\left(\frac{K+\mu_0-\mu}{\sigma/\sqrt{n}}\right)$ bajo H_0 es $\mu = \mu_0$. Por consiguiente, tenemos que $\alpha = 1 - \Phi(\sqrt{n}K/\sigma)$, y se encuentra que $K = z_{1-\alpha}\sigma/\sqrt{n}$. Y la regla de decisión para (4.2.5) es

$$\text{Rechazar } H_0 \text{ si } \bar{X} - \mu_0 > z_{1-\alpha}\tfrac{\sigma}{\sqrt{n}}.$$

En otras situaciones puede haber un valor supuesto para la media teórica μ_0 pero se sospecha que éste no es adecuado, sino que la media teórica es mayor a μ_0. Por ejemplo, la fábrica de clavos del Ejemplo 4.2.1 despacha un pedido de clavos de 5cm, pero recibe devolución de este pedido puesto que el cliente manifiesta que los clavos de este pedido son más largos que los de costumbre. En estos casos, el sistema de hipótesis de interés puede ser de la forma

$$H_0 : \ \mu = \mu_0 \quad vs. \quad H_1 : \ \mu > \mu_0 \tag{4.2.7}$$

En este caso, la hipótesis alterna es la misma que el sistema (4.2.5), y como una regla de decisión establece cuándo se debe rechazar la hipótesis nula, o equivalentemente, cuándo aceptar la hipótesis alterna, entonces **dos sistemas que difieren en la hipótesis nula, pero concuerdan en la hipótesis alterna tienen la misma regla de decisión.** Por consiguiente, la regla de decisión para (4.2.7) es

$$\text{Rechazar } H_0 \text{ si } \bar{X} - \mu_0 > z_{1-\alpha}\frac{\sigma}{\sqrt{n}}.$$

O equivalentemente

$$\text{Rechazar } H_0 \text{ si } \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} > z_{1-\alpha}.$$

Región de rechazo

Y es claro que para la anterior regla de decisión, la estadística de prueba es $\dfrac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$ y el región de rechazo asociado es $\{c \in \mathbb{R} : c > z_{1-\alpha}\}$.

p valor

Ahora consideramos la forma de calcular el p valor para la anterior regla de decisión para los sistemas (4.2.5) y (4.2.7). En primer lugar, observemos la forma de la región de rechazo correspondiente ilustrada en la Figura 4.7. Suponga que el valor de la estadística $\dfrac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$ observada en una muestra es el valor v (ver Figura 4.7), entonces la decisión de aceptar o rechazar H_0 depende del área bajo curva hacia la derecha de v, esto es, $Pr\left(\dfrac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} > v\right)$. Si esta área es mayor que el área de la región de rechazo α, entonces v no cae en la región de rechazo, y por consiguiente, se acepta H_0; por otro lado, si el área bajo curva hacia la derecha de v es menor que α, se rechaza H_0.

De esta forma, el p valor para el sistema de hipótesis (4.2.5) o (4.2.7) ligado al valor observado de la estadístca v se define como

$$p \text{ valor} = Pr\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} > v\right)$$

o equivalentemente

$$p \text{ valor} = Pr(Z > v) = 1 - \Phi(v) \tag{4.2.8}$$

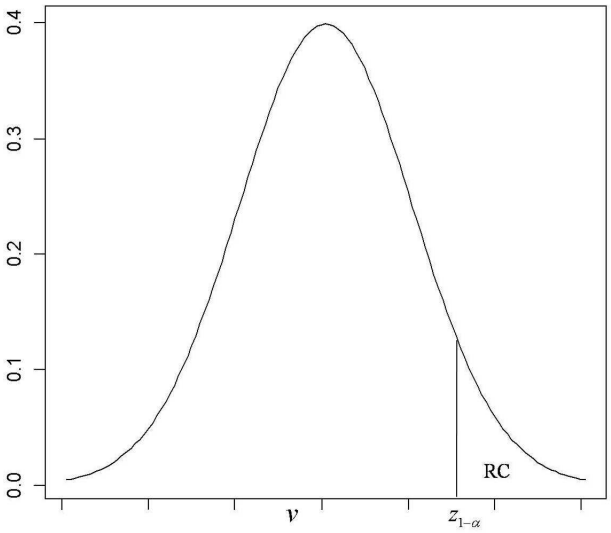

Figura 4.7: *Ilustración del región de rechazo de la regla del sistema (4.2.5) y (4.2.7).*

Y rechazamos la hipótesis nula si el p valor es pequeño comparado con el nivel de significación.

Ilustramos una aplicación de lo anterior en el siguiente ejemplo.

Ejemplo 4.2.2. *Retomamos el Ejemplo 4.1.1, donde se desea conocer el comportamiento de un tipo de motocicleta en términos de la emisión del gas CO. Si el sistema planteado es*

$$H_0: \ \mu \leq 5.5 \quad vs. \quad H_1: \ \mu > 5.5 \tag{4.2.9}$$

y en una muestra de 10 motocicletas, mediante el uso de QQ plot, se verificó que la distribución normal describe bien el comportamiento de los datos [2]*. Se observa un promedio de emisión de CO del* $7.2g/Km$ *y suponga que en promedio las motocicletas se difieren en* $1.2g/Km$ *en términos de emisión de CO.*

Para observar si los datos muestrales muestran evidencia en contra de H_0*, calculamos el valor de la estadística de prueba* $\dfrac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}$*, la cual tiene un valor de 4.48, y es claramente mayor que el percentil* $z_{1-\alpha}$ *para todos los valores comunes de* α *en la práctica, y esto revela que los datos muestran una fuerte evidencia en contra de que la emisión de CO no supere a* $5.5g/Km$*.*

Por otro lado, la decisión se puede tomar en base del p *valor que se calcula según (4.2.8). Y tenemos que*

$$p \text{ valor} = 1 - \Phi(4.48) = 3.73e-06$$

[2] Dado que el tamaño muestral es pequeño, no se utiliza el histograma para verificar la distribución teórica.

De donde podemos anotar que los datos observados están en contra de la hipótesis nula, ya que p valor es extremadamente pequeño.

Función de potencia

Finalmente, estudiamos la función de potencia para la regla de decisión encontrada para los sistemas (4.2.5) y (4.2.7). Tenemos que

$$\begin{aligned}\beta(\mu) &= Pr(\text{Rechazar } H_0)\\ &= Pr(\bar{X} - \mu_0 > z_{1-\alpha}\frac{\sigma}{\sqrt{n}})\\ &= Pr\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{1-\alpha}\right)\\ &= 1 - \Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) \qquad (4.2.10)\end{aligned}$$

Se puede verificar que los comentarios para la relación entre esta función de potencia con el tamaño muestral y el nivel de significación son los mismos que los del sistema $\mu = \mu_0\ vs.\ \mu \neq \mu_0$. Esto es, la potencia de la prueba incrementa a medida que (1) incrementa el tamaño muestral (2) aumenta el nivel de significación α. En las Figuras 4.8 y 4.9 se grafican (4.2.10) para diferentes tamaños de muestra y diferentes niveles de significación.

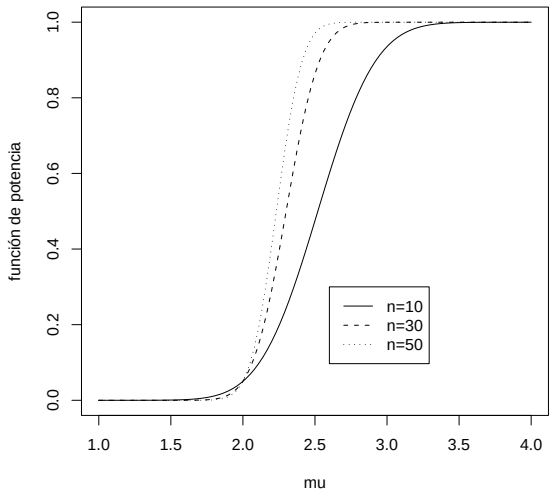

Figura 4.8: *Función de potencia (4.2.10) para diferentes tamaños de muestra con $\mu_0 = 2$, $\sigma = 1$ y $\alpha = 0.05$.*

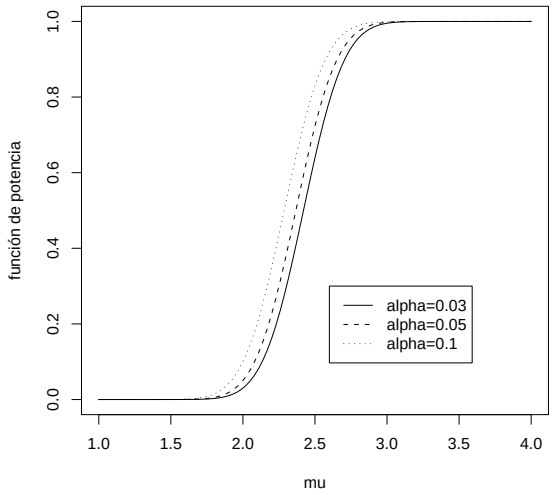

Figura 4.9: *Función de potencia (4.2.10) para diferentes niveles de significación con* $\mu_0 = 2$, $\sigma = 1$ *y* $n = 20$.

En la práctica se debe tener en cuenta que cuando se acepta $H_0 : \ \mu \leq \mu_0$ todavía es posible que el μ verdadero sea mayor a μ_0, es decir, siempre se debe tener en cuenta que estamos propensos a cometer el error tipo II, y hasta ahora, no se ha puesto un límite a la probabilidad de cometer este tipo de errores. Dadas las discusiones previas acerca de la función de potencia, tenemos que esta probabilidad se puede calcular como $\beta = 1 - \beta(\mu) = \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha}\right)$ con $\mu > \mu_0$. La forma de estas probabilidades se muestra en la Figura 4.10 (con $\mu_0 = 2$), donde podemos observar que entre más cercano esté μ de μ_0 mayor será la probabilidad de aceptar una hipótesis nula falsa.

$H_0 : \ \mu \geq (=)\mu_0$ **vs.** $H_1 : \ \mu < \mu_0$ **con** σ^2 **conocida**

Para el sistema

$$H_0 : \ \mu = \mu_0 \quad vs. \quad H_1 : \ \mu < \mu_0$$

o

$$H_0 : \ \mu \geq \mu_0 \quad vs. \quad H_1 : \ \mu < \mu_0$$

Se deja como ejercicio completar el procedimiento para obtener la siguiente regla de decisión de nivel de significación α (Ejercicio 4.2)

$$\text{Rechazar } H_0 \text{ si } \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} < -z_{1-\alpha}. \tag{4.2.11}$$

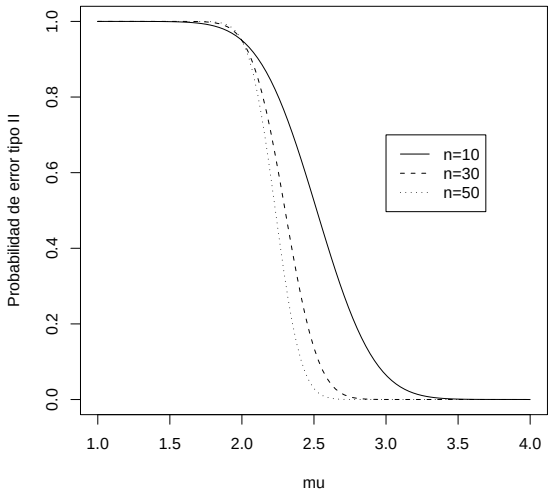

Figura 4.10: *Probabilidad de cometer error tipo II de la regla de decisión encontrada para sistemas (4.2.5) y (4.2.7) con* $\mu_0 = 2$.

Y asociada a la anterior regla de decisión si el valor de la estadística de prueba $\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}$ se denota por v, entonces el p valor está dado por

$$p \text{ valor} = Pr(Z < v) = \Phi(v) \tag{4.2.12}$$

y como de costumbre se rechaza H_0 si el p valor es menor que el nivel de significación α.

También se deja como ejercicio el procedimiento para encontrar la función de potencia dada por (Ejercicio 4.2)

$$\beta(\mu) = \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{1-\alpha}\right).$$

Prueba de razón de verosimilitud

Otra forma de encontrar una regla de decisión para las anteriores sistemas es el método de la prueba de razón de verosimilitud que se describe a continuación.

Definición 4.2.2. *Sea* $X_1, \cdots, X_n$ *una muestra aleatoria con función de densidad* $f(x_i, \theta)$*, donde* θ *es el parámetro desconocido. Suponga que se quiere probar el siguiente sistema de hipótesis:*

$$H_0 : \ \theta = \theta_0 \quad vs. \quad H_1 : \ \theta = \theta_1,$$

se denomina la prueba de razón de verosimilitud a la prueba con regla de decisión dada por

$$\textit{Rechazar } H_0 \textit{ si } \lambda > K \textit{ para algún constante } K,$$

donde λ es la razón de verosimilitud dada por

$$\lambda = \frac{L(\theta_1, x_1, \cdots, x_n)}{L(\theta_0, x_1, \cdots, x_n)} = \frac{\prod_{i=1}^{n} f(x_i, \theta_1)}{\prod_{i=1}^{n} f(x_i, \theta_0)}, \tag{4.2.13}$$

Para entender mejor la anterior definición, recordemos que la función de verosimilitud $L(\theta, x_1, \cdots, x_n)$ se puede interpretar como la probabilidad de observar valores $x_1, \cdots, x_n$ cuando el valor del parámetro es θ. Así que si $L(\theta_1, x_1, \cdots, x_n)$ es mucho más grande que $L(\theta_0, x_1, \cdots, x_n)$, podemos concluir que el valor θ_1 es más creíble que θ_0 para el parámetro, puesto que los valores observados son $x_1, \cdots, x_n$ y por consiguiente, la probabilidad de observar estos valores debe ser grande. En conclusión, cuando λ es grande, los datos muestran evidencias a favor de θ_1, y se rechaza el valor de θ_0.

Una forma equivalente pero más sencilla de la prueba de razón de verosimilitud es rechazar H_0 cuando $\ln \lambda > \ln K$, puesto que la función logarítmica es una función estrictamente creciente, y valores grandes de λ conducen a valores grandes de $\ln \lambda$ y viceversa. De esta manera, una regla de decisión equivalente es

$$\text{Rechazar } H_0 \text{ si } \sum_{i=1}^{n} (\ln f(x_i, \theta_1) - \ln f(x_i, \theta_0)) > K^* \text{ para alguna constante } K^*.$$

La metodología de prueba de razón de verosimilitud sirve, aparentemente, para sistemas de hipótesis donde tanto la hipótesis nula como la alterna son igualdad. Es claro que muchos sistemas de hipótesis no son de esta forma, pero generalmente se pueden escribir en forma de igualdades, como lo indica en el Tabla 4.1. Ahora, aunque un sistema de hipótesis de forma de igualdad versus desigualdad se puede escribir como un sistema conformado por dos igualdades, generalmente no es posible aplicar el método de la prueba de razón de verosimilitud, sino la prueba generalizada de razón de verosimilitud que se expondrá más adelante.

$\theta = \theta_0 \ vs. \ \theta < \theta_0$	$\theta = \theta_0 \ vs. \ \theta = \theta_1$ con $\theta_1 < \theta_0$
$\theta = \theta_0 \ vs. \ \theta > \theta_0$	$\theta = \theta_0 \ vs. \ \theta = \theta_1$ con $\theta_1 > \theta_0$
$\theta \leq \theta^* \ vs. \ \theta > \theta^*$	$\theta = \theta_0 \ \ vs. \ \theta = \theta_1$ con $\theta_0 \leq \theta^*$ y $\theta_1 > \theta^*$
$\theta \geq \theta^* \ vs. \ \theta < \theta^*$	$\theta = \theta_0 \ vs. \ \theta = \theta_1$ con $\theta_0 \geq \theta^*$ y $\theta_1 < \theta^*$
$\theta = \theta_0 \ vs. \ \theta \neq \theta_0$	$\theta = \theta_0 \ vs. \ \theta = \theta_1$ con $\theta_1 \neq \theta_0$.

Tabla 4.1: *Sistemas de hipótesis equivalentes.*

Retomamos el sistema (4.2.7) en el siguiente ejemplo para ilustrar el uso de la prueba de razón de verosimilitud.

Ejemplo 4.2.3. *Sea $X_1, \cdots, X_n$ una muestra aleatoria proveniente de $N(\mu, \sigma^2)$ con $\sigma^2 = \sigma_0^2$ conocida, consideramos el siguiente sistema de hipótesis*

$$H_0 : \ \mu = \mu_0 \quad vs. \quad H_1 : \ \mu > \mu_0. \tag{4.2.14}$$

De acuerdo con la Tabla 4.1, el anterior sistema es equivalente a

$$H_0 : \ \mu = \mu_0 \quad vs. \quad H_1 : \ \mu = \mu_1 \tag{4.2.15}$$

con $\mu_1 > \mu_0$. De esta manera, se puede establecer una regla de decisión usando la prueba de razón de verosimilitud. Tenemos que

$$\begin{aligned}\sum_{i=1}^{n}(\ln f(x_i,\theta_1) - \ln f(x_i,\theta_0)) &= \sum_{i=1}^{n}\left(\frac{1}{2}\frac{(x_i-\mu_0)^2}{\sigma_0^2} - \frac{1}{2}\frac{(x_i-\mu_1)^2}{\sigma_0^2}\right)\\ &= \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\left((x_i-\mu_0)^2 - (x_i-\mu_1)^2\right)\\ &= \frac{1}{\sigma_0^2}(\mu_1-\mu_0)\sum_{i=1}^{n}x_i + \frac{n(\mu_0^2-\mu_1^2)}{2\sigma_0^2}.\end{aligned}$$

De manera que la regla de decisión de la prueba de razón de verosimilitud está dada por:

Rechazar H_0 si $\frac{1}{\sigma_0^2}(\mu_1-\mu_0)\sum_{i=1}^{n}x_i + \frac{n(\mu_0^2-\mu_1^2)}{2\sigma_0^2} > K^*$ para alguna constante K^*.

La anterior regla de decisión, sin duda, tiene una forma poco amigable. Para lograr una regla de decisión equivalente pero más simple, se observa que en la anterior regla de decisión la única cantidad aleatoria que varía de muestra a muestra es la realización de la estadística $\sum_{i=1}^{n}x_i$, entonces se despeja para este valor, y se tiene que

$$\text{Rechazar } H_0 \text{ si } \sum_{i=1}^{n}x_i > \frac{\sigma_0^2}{\mu_1-\mu_0}(K^* - \frac{n(\mu_0^2-\mu_1^2)}{2\sigma_0^2}) = K_1,$$

o también se puede escribir la anterior regla de decisión como:

$$\text{Rechazar } H_0 \text{ si } \bar{x} > K_1/n = K.$$

Una vez establecida la regla de decisión, el siguiente paso es encontrar el valor de la constante involucrada K. Para eso se procede de la manera corriente, recurriendo a la definición del error tipo I, y se despeja el valor de K de la igualdad

$$\alpha = Pr(\text{cometer error tipo I}),$$

de donde se tiene que $K = \frac{\sigma_0}{\sqrt{n}}z_{1-\alpha} + \mu_0$, y de esta manera, se completa la regla de decisión:

$$\text{Rechazar } H_0 \text{ si } \bar{x} > \frac{\sigma_0}{\sqrt{n}}z_{1-\alpha} + \mu_0.$$

O equivalentemente

$$\textit{Rechazar } H_0 \textit{ si } \frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma_0} > z_{1-\alpha}.$$

El lector puede darse cuenta de que la anterior regla de decisión coincide con la hallada anteriormente, pero con mucho más operaciones algebraicas, entonces ¿para qué utilizar la prueba de razón de verosimilitud?, ¿qué ventajas tiene ésta? Hay por lo menos las dos siguientes razones: en primer lugar, la prueba de razón de verosimilitud es una herramienta estándar que puede ser utilizada en muchas áreas de la estadística donde no es fácil proponer una regla de decisión; en segundo lugar, anteriormente se comentó que aunque matemáticamente se puede calcular la función de potencia asociada a una prueba, en la práctica no se puede cuantificar la potencia, entonces surge la pregunta de si puede existir otra regla de decisión que tenga mayor potencia comparada con las reglas encontradas anteriormente. El lemma de Neyman Pearson responde esta pregunta y afirma que la regla de decisión encontrada usando el método de la prueba de razón de verosimilitud es la más potente, es decir, la prueba de razón de verosimilitud es la prueba que tiene menor probabilidad de cometer error tipo II.

Resultado 4.2.1. *(Lema de Neyman-Pearson) Sea $X_1, \cdots, X_n$ una muestra aleatoria con función de densidad $f(x_i, \theta)$, donde θ es el parámetro desconocido. Suponga que se quiere probar el siguiente sistema de hipótesis:*

$$H_0: \ \theta = \theta_0 \quad vs. \quad H_1: \ \theta = \theta_1,$$

entonces la prueba de razón de verosimilitud: Rechazar H_0 si $\lambda > K$ *es la prueba más potente de tamaño α si $\beta(\theta_0) = \alpha$.*

Para verificar que la regla de decisión Rechazar H_0 si $\dfrac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma_0} > z_{1-\alpha}$ para (4.2.7) la más potente de tamaño α, se debe verificar que $\beta(\mu_0) = \alpha$. Ahora, la función de potencia está dada en (4.2.10), de donde $\beta(\mu_0) = 1 - \Phi(1-\alpha) = \alpha$.

Se debe hacer énfasis en que la anterior prueba es la más potente de tamaño α, puesto que al aumentar el nivel de significación, la potencia también aumenta, por lo tanto pruebas con un alto nivel de significación pueden ser más potentes que la anterior prueba de razón de verosimilitud.

También hacemos aclaración de que el hecho de que una prueba sea la más potente se refiere a que la función de potencia $\beta(\theta)$ es no inferior a la función de potencia de cualquier otra prueba para todo $\theta \in \mathbf{\Theta}_1$, no para algún θ particular.

Ahora, como se mencionaba antes, aunque algunos sistemas de hipótesis como (4.2.1) que se desean probar pueden ser expresados como

$$H_0: \ \theta = \theta_0 \quad vs. \quad H_1: \ \theta = \theta_1,$$

no siempre se puede encontrar una regla de decisión usando el método de la razón de verosimilitud. En estos casos, se puede aplicar el método de la razón generalizada de verosimilitud, que se describe a continuación.

Definición 4.2.3. *Sea $X_1, \cdots, X_n$ una muestra aleatoria con función de densidad $f(x_i, \theta)$, donde θ es el parámetro desconocido. Suponga que se quiere probar el siguiente sistema de hipótesis:*

$$H_0 : \ \theta \in \mathbf{\Theta}_0 \quad vs. \quad H_1 : \ \theta \in \mathbf{\Theta_1},$$

con $\mathbf{\Theta}_0 \cup \mathbf{\Theta}_1 \subseteq \mathbf{\Theta}$, el espacio paramétrico de θ, y $\mathbf{\Theta}_0 \cap \mathbf{\Theta}_0 = \emptyset$, entonces la prueba con regla de decisión dada por

$$\textit{Rechazar } H_0 \textit{ si } \lambda > K \textit{ para algún constante } K,$$

donde λ es la razón generalizada de verosimilitud dada por

$$\lambda = \frac{\sup_{\mathbf{\Theta}_0 \cup \mathbf{\Theta}_1} L(\theta, x_1, \cdots, x_n)}{\sup_{\mathbf{\Theta}_0} L(\theta, x_1, \cdots, x_n)}, \tag{4.2.16}$$

donde $\sup_A L(\theta, x_1, \cdots, x_n)$ se denota el valor máximo que puede tomar la función de verosimilitud L en el conjunto A.

La lógica de esta prueba es similar a la prueba de razón de verosimilitud, cuando λ es muy grande, $\sup_{\mathbf{\Theta}_0 \cup \mathbf{\Theta}_1} L(\theta, x_1, \cdots, x_n)$ es mayor que $\sup_{\mathbf{\Theta}_0} L(\theta, x_1, \cdots, x_n)$, implicando que el valor máximo de $L(\theta, x_1, \cdots, x_n)$ en $\mathbf{\Theta}_1$ es mayor que el valor máximo en $\mathbf{\Theta}_0$. Esto es, es mas creíble que θ toma valor en $\mathbf{\Theta}_1$ que en $\mathbf{\Theta}_0$, y por consiguiente se rechaza H_0.

Por otro lado, cuando $\mathbf{\Theta}_1 = \mathbf{\Theta}^c$, se tiene que $\mathbf{\Theta}_0 \cup \mathbf{\Theta}_1 = \mathbf{\Theta}$, y dado que $\mathbf{\Theta}$ es el espacio paramétrico completo de θ, entonces el numerador de λ en (4.2.16) se convierte simplemente en la función de verosimilitud evaluada en el estimador MV de θ, esto es, $L(\hat{\theta}_{MV}, x_1, \cdots, x_n)$.

Anteriormente se mencionó que cuando el sistema de hipótesis que se desea probar es de la forma de igualdad frente a desigualdad, no siempre se puede aplicar la prueba de razón de verosimilitud. Veamos en el siguiente ejemplo que la prueba de razón generalizada de verosimilitud puede resultar útil.

Ejemplo 4.2.4. *Dada una muestra aleatoria $X_1, \cdots, X_n$ proveniente de una distribución $N(\mu, \sigma^2)$, usaremos la prueba de razón generalizada de verosimilitud para encontrar una regla de decisión para el sistema*

$$H_0 : \ \mu = \mu_0 \quad vs. \quad H_1 : \ \mu \neq \mu_0.$$

En primer lugar, nótese que $\mathbf{\Theta}_0$ es el conjunto cuyo único valor es μ_0, esto es, $\mathbf{\Theta}_0 = \{\mu_0\}$, y $\mathbf{\Theta}_1 = \mathbf{\Theta}_0^c$, así que el numerador de λ es la función de verosimilitud evaluada en $\hat{\mu}_{MV}$.

Ahora, para calcular la razón generalizada de verosimilitud λ, recordemos, en primer lugar, que la función de verosimilitud está dada por

$$L(\mu, x_1, \cdots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}.$$

Ahora, la estimación MV de μ es el promedio muestral $\bar{x}$, entonces el numerador de λ se convierte en

$$L(\hat{\mu}_{MV}, x_1, \cdots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}.$$

Y por otro lado, $\boldsymbol{\Theta}_0 = \{\mu_0\}$, entonces el máximo valor de L en $\boldsymbol{\Theta}_0$ es L evaluado en μ_0, entonces

$$L(\mu_0, x_1, \cdots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu_0)^2 \right\}.$$

De donde se tiene que la razón generalizada de verosimilitud está dada por

$$\begin{aligned} \lambda &= \frac{(2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2\}}{(2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu_0)^2\}} \\ &= \exp\left\{ \frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (x_i - \mu_0)^2 - \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] \right\}. \end{aligned}$$

Usando la monotocidad de la función logarítmica y eliminando constantes de la desigualdad, se tiene la regla de decisión:

$$\textit{Rechazar } H_0 \textit{ si } \textstyle\sum_{i=1}^{n} [(x_i - \mu_0)^2 - (x_i - \bar{x})^2] > K^*$$

para alguna constante K^*. Ahora, es fácil ver que $\sum_{i=1}^{n} [(x_i - \mu_0)^2 - (x_i - \bar{x})^2] = n(\bar{x} - \mu_0)^2$, entonces se rechaza H_0 cuando $(\bar{x} - \mu_0)^2 > K_1$, la cual es equivalente a $|\bar{x} - \mu_0| > K$ encontrada al principio de este capítulo. Finalmente, se encuentra el valor para K siguiendo a los pasos expuestos anteriormente.

Aunque en muchos casos, como en el ejemplo anterior, la regla de decisión encontrada al utilizar la prueba de razón (generalizada) de verosimilitud es equivalente a la encontrada simplemente analizando el sistema de hipótesis y usando el sentido común. La prueba de razón de verosimilitud ofrece una metodología estándar para una amplia gama de sistemas de hipótesis, ésta es particularmente útil cuando la distribución de donde proviene la muestra es diferente que la distribución normal y/o el sentido común no da ninguna pista sobre cómo debe ser la regla de decisión.

$H_0: \ \mu = \mu_0$ **vs.** $H_1: \ \mu \neq \mu_0$ **con** σ^2 **desconocida**

En muchos casos, por falta de información acerca de la población objetiva, la varianza teórica no es conocida; en este caso, para el sistema de hipótesis

$$H_0: \ \mu = \mu_0 \quad vs. \quad H_1: \ \mu \neq \mu_0$$

las reglas de decisión dadas para el caso cuando σ^2 es conocida ya no serán válidas, aunque el procedimiento para obtener las reglas de decisión son las mismas. En primer

lugar, dado que el sistema de hipótesis es el mismo, la regla de decisión sigue siendo: Rechazar H_0 cuando $|\bar{x} - \mu_0| > K$ para alguna constante positivo K. Al restringir la magnitud de cometer el error tipo I a ser igual al nivel de significación α, se tiene

$$\alpha = Pr(|\bar{X} - \mu_0| > K)$$

cuando $\mu = \mu_0$. Para encontrar el valor de K en la anterior ecuación, se multiplica $\frac{\sqrt{n}}{S_{n-1}}$, de donde

$$\begin{aligned}\alpha =& Pr\left(\frac{\sqrt{n}|\bar{X} - \mu_0|}{S_{n-1}} > \frac{\sqrt{n}K}{S_{n-1}}\right) \\ &= Pr\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{S_{n-1}} > \frac{\sqrt{n}K}{S_{n-1}}\right) + Pr\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{S_{n-1}} < -\frac{\sqrt{n}K}{S_{n-1}}\right).\end{aligned}$$

Ahora bajo la hipótesis nulo $\mu = \mu_0$, se tiene que la distribución de $\frac{\sqrt{n}(\bar{X}-\mu_0)}{S_{n-1}}$ es la distribución t_{n-1}, que es simétrica con respecto a 0. Por lo tanto, las dos probabilidades en la ecuación anterior son iguales, entonces se tiene

$$\frac{\alpha}{2} = Pr\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{S_{n-1}} > \frac{\sqrt{n}K}{S_{n-1}}\right),$$

de donde $\frac{\sqrt{n}K}{S_{n-1}} = t_{n-1,1-\alpha/2}$, de esta manera se tiene que $K = t_{n-1,1-\alpha/2}\frac{S_{n-1}}{\sqrt{n}}$. De esta manera, se completa la regla de decisión, y tiene forma:

$$\text{Rechazar } H_0 \text{ si } |\bar{x} - \mu_0| > t_{n-1,1-\alpha/2}\frac{s_{n-1}}{\sqrt{n}},$$

de manera equivalente se tiene la siguiente regla de decisión:

$$\text{Rechazar } H_0 \text{ si } \frac{\sqrt{n}(\bar{x} - \mu_0)}{s_{n-1}} > t_{n-1,1-\alpha/2} \text{o } \frac{\sqrt{n}(\bar{x} - \mu_0)}{s_{n-1}} < -t_{n-1,1-\alpha/2}, \quad (4.2.17)$$

donde la estadística de prueba es $\frac{\sqrt{n}(\bar{X}-\mu_0)}{S_{n-1}}$, y la región de rechazo está dada por $\{c \in \mathbb{R} : c > t_{n-1,1-\alpha/2} \text{ ó } c < -t_{n-1,1-\alpha/2}\}$. Esta prueba es conocida como la prueba t a dos colas, en la Figura 4.11, se muestra esta región de rechazo.

p valor

También podemos calcular el p valor para esta prueba. El razonamiento es análogo al caso cuando σ^2 es conocido, lo único que es diferente en el caso cuando σ^2 es desconocido es que la estadística de prueba es $\frac{\sqrt{n}(\bar{X}-\mu_0)}{S_{n-1}}$ cuya distribución nula es t_{n-1}; de esta forma, el p valor para la prueba t de dos colas está dado por:

$$p \text{ valor} = \begin{cases} 2Pr\left(\dfrac{\sqrt{n}(\bar{X} - \mu_0)}{S_{n-1}} > v\right), & \text{para } v > 0 \text{ ;} \\ 2Pr\left(\dfrac{\sqrt{n}(\bar{X} - \mu_0)}{S_{n-1}} < v\right), & \text{para } \ v < 0. \end{cases}$$

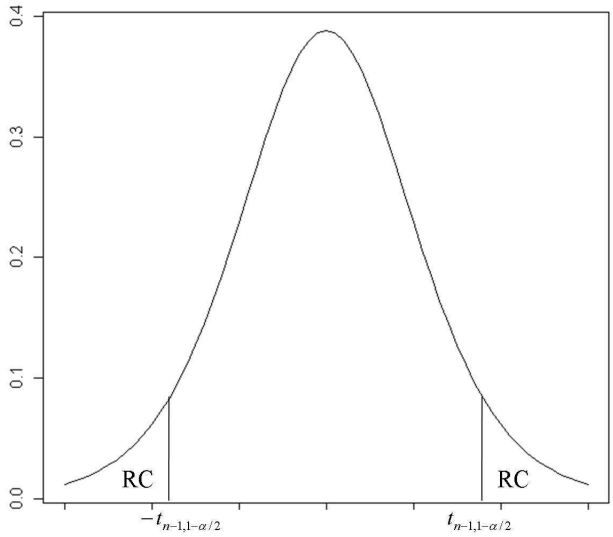

Figura 4.11: *Ilustración del región de rechazo de la prueba t a dos colas.*

o equivalentemente

$$p \text{ valor} = \begin{cases} 2Pr(T_{n-1} > v), & \text{para } v > 0 \text{ ;} \\ 2Pr(T_{n-1} < v), & \text{para } v < 0. \end{cases}$$

donde T_{n-1} denota una variable aleatoria con distribución t_{n-1}.

Volviendo al Ejemplo 4.2.1, suponga que el valor de la desviación estándar es desconocida; en este caso, el sistema de hipótesis sigue siendo

$$H_0 : \ \mu = 350 \quad vs. \quad H_1 : \ \mu \neq 350,$$

y la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{\sqrt{n}(\bar{x} - \mu_0)}{s_{n-1}} > t_{n-1,1-\alpha/2} \text{ ó } \frac{\sqrt{n}(\bar{x} - \mu_0)}{s_{n-1}} < -t_{n-1,1-\alpha/2}$$

donde se supone que $\alpha = 5\,\%$. Para la muestra observada de tamaño 20, se tiene que $\bar{x} = 349.85$ y $s_{n-1} = 5.4$, entonces $\frac{\sqrt{n}(\bar{x}-\mu_0)}{s_{n-1}} = -0.124$, y $t_{n-1,1-\alpha/2} = 2.093$, de donde se observa que el valor de la estadística de prueba no se encuentra dentro de la región de rechazo, entonces se puede concluir que la máquina sí lleva a cabo un llenado de 350ml.

Por otro lado, calculamos el p valor. El valor de la estadística de prueba -0.124 es negativo, de manera que

$$p \text{ valor} = 2Pr(T_{n-1} < -0.124),$$

con la ayuda de una tabla estadística de la distribución t o la función `pt` del software R, se tiene que el p valor es 0.9026, el cual es mayor a cualquier nivel de significación α usado en la práctica, de donde se concluye no se rechaza H_0.

El anterior procedimiento también se puede llevar a cabo usando el comando `t.test`, el uso y el resultado arrojado es como sigue

```
> x<-c(355, 350, 340, 345, 354, 358, 350, 343, 349, 346, 351, 358,
+ 342, 350, 356, 345, 349, 356, 354, 346)
> t.test(x,mu=350)

        One Sample t-test

data:  x
t = -0.1242, df = 19, p-value = 0.9025
alternative hypothesis: true mean is not equal to 350
95 percent confidence interval:
 347.3216 352.3784
sample estimates:
mean of x
   349.85
```

Podemos ver que los resultados son iguales a los obtenidos manualmente. Adicionalmente, vemos que un intervalo del 95 % para la media teórica es (347.3,352.4), y el valor 350 se encuentra dentro de este intervalo, conduciéndonos a la misma conclusión de aceptar H_0.

Función de potencia

Para encontrar la función de potencia de la anterior regla de decisión, procedemos de forma similar al caso cuando σ^2 es desconocida. Tenemos que

$$\begin{aligned}\beta(\mu) &= Pr(\text{Rechazar } H_0)\\ &= Pr\left(\frac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}} > t_{n-1,1-\alpha/2}\right) + Pr\left(\frac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}} < t_{n-1,1-\alpha/2}\right)\end{aligned}$$

Ahora, $\bar{X} \sim N(\mu, \sigma^2/n)$ de donde $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N\left(\frac{\mu-\mu_0}{\sigma/\sqrt{n}}, 1\right)$, y por otro lado, $\frac{(n-1)S^2_{n-1}}{\sigma^2} \sim \chi^2_{n-1}$. Utilizando la independencia entre $\bar{X}$ y S^2_{n-1}, tenemos que

$$\frac{\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{S^2_{n-1}}{\sigma^2}}} = \frac{\bar{X}-\mu_0}{S_{n-1}/\sqrt{n}} \sim t^{nc}_{n-1,\delta}.$$

con $\delta = \frac{\mu-\mu_0}{\sigma/\sqrt{n}}$. Y de allí podemos calcular la función de potencia como

$$\beta(\mu) = 1 - F_{t^{nc}_{n,\delta}}(t_{n-1,1-\alpha/2}) + F_{t^{nc}_{n,\delta}}(t_{n-1,\alpha/2}) \tag{4.2.18}$$

donde $F_{t^{nc}_{n-1,\delta}}(\cdot)$ denota la función de distribución de la distribución $t^{nc}_{n-1;\delta}$. Para reproducir gráficas similares a las de las Figuras 4.4 y 4.5, tenemos el siguiente código

```
> po_t_nocen<-function(mu,n,mu0,alpha,sigma){
+ delta<-(mu-mu0)*sqrt(n)/sigma
+ 1-pt(qt(1-alpha/2,n-1),df=n-1,ncp=delta)+pt(qt(alpha/2,n-1),
+ df=n-1,ncp=delta)
+ }
>
> alpha<-0.05
> sigma<-1
> mu0<-2
> n1<-10
> n2<-30
> n3<-50
>
>
> mu<-mu0+seq(-2,2,0.05)
>
> plot(po_t_nocen(mu,n1,mu0,alpha,sigma),type="l",xaxt="n",xlab="mu",
+ ylab="función de potencia")
> axis(1, 1:length(mu), mu)
> lines(po_t_nocen(mu,n2,mu0,alpha,sigma),lty=2)
> lines(po_t_nocen(mu,n3,mu0,alpha,sigma),lty=3)
> legend(60,0.4,c("n=10","n=30","n=50"),lty=c(1,2,3))

> windows()
> alpha1<-0.03
> alpha2<-0.05
> alpha3<-0.1
> n<-20
>
> plot(po_t_nocen(mu,n,mu0,alpha1,sigma),type="l",xaxt="n",xlab="mu",
+ ylab="función de potencia")
> axis(1, 1:length(mu), mu)
> lines(po_t_nocen(mu,n,mu0,alpha2,sigma),lty=2)
> lines(po_t_nocen(mu,n,mu0,alpha3,sigma),lty=3)
> legend(55,0.3,c("alpha=0.03","alpha=0.05","alpha=0.1"),lty=c(1,2,3))
```

En las Figuras 4.12 y 4.13 se muestra la función (4.2.18) variando el tamaño de muestral y el nivel de significación. Podemos ver que las figuras son muy similares que las de la función de potencia (4.2.4) debido a la similitud entre la distribución t y la distribución normal, y como consecuencia, concluimos que la potencia aumenta cuando se tiene una muestra grande y también cuando el nivel de significación α es grande.

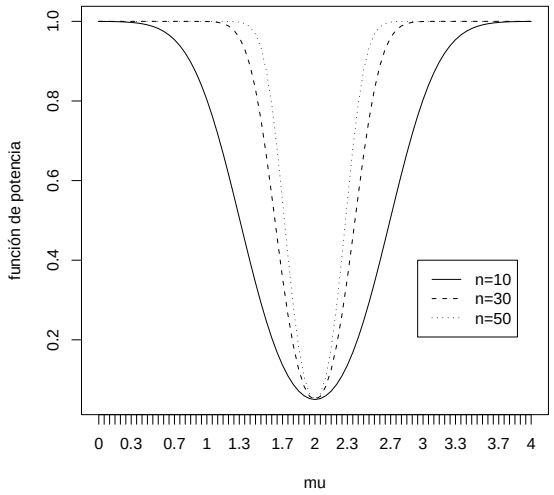

Figura 4.12: *Función de potencia (4.2.18) con* $\alpha = 0.05$, $\sigma = 1$ *y diferentes tamaños de muestra.*

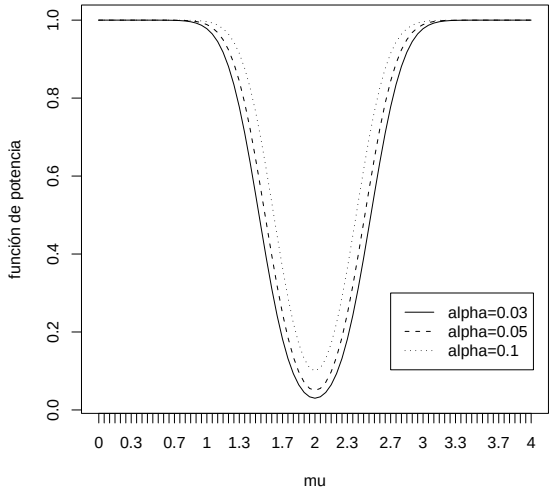

Figura 4.13: *Función de potencia (4.2.18) para diferentes niveles de significación con* $\mu_0 = 2$, $\sigma = 1$ *y* $n = 20$.

$H_0: \ \mu \leq (=)\mu_0$ **vs.** $H_1: \ \mu > \mu_0$ **con** σ^2 **desconocida**

Si el sistema de interés es

$$H_0: \ \mu = \mu_0 \quad vs. \quad H_1: \ \mu > \mu_0.$$

o

$$H_0: \ \mu \leq \mu_0 \quad vs. \quad H_1: \ \mu > \mu_0.$$

Se vio anteriormente que para el caso cuando σ^2 es conocido, la regla de decisión encontrada es

$$\text{Rechazar } H_0 \text{ si } \bar{x} > K.$$

y el procedimiento para encontrar esta regla de decisión no está sujeto al supuesto de que la varianza es conocida. Entonces cuando no se tiene este supuesto, se obtiene la misma regla de decisión. Y el hecho de que la varianza es desconocida conlleva a que al encontrar el valor de la constante K, la distribución usada será la distribución t_{n-1}. Y se tiene que

$$\text{Rechazar } H_0 \text{ si } \frac{\sqrt{n}(\bar{x} - \mu_0)}{s_{n-1}} > t_{n-1,1-\alpha}.$$

p valor

Se puede encontrar la forma de calcular el p valor cuando la estadística de prueba $\bar{X}$ toma un valor v, análogo al caso cuando σ^2 es conocida. Se deja como ejercicio ver que (Ejercicio 4.3)

$$p \text{ valor} = Pr\left(\frac{\sqrt{n}(\bar{X} - \mu)}{S_{n-1}} > v\right) = Pr(T_{n-1} > v),$$

y se rechaza H_0 cuando el p valor es menor que el nivel de significación α.

Función de potencia

Es fácil ver que la función de potencia está dada por

$$\beta(\mu) = 1 - F_{t^{nc}_{n-1,\delta}}\left(t_{n-1,1-\alpha}\right)$$

donde $\delta = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$.

Dualidad entre $IC(\mu)$ y pruebas de hipótesis

En el capítulo anterior, se mencionó que los intervalos de confianza para μ son útiles para tomar decisiones sobre hipótesis como $\mu = \mu_0$, $\mu \leq \mu_0$ o $\mu \geq \mu_0$. Veamos que las decisiones tomadas usando los intervalos de confianza concuerdan con las vistas en el presente capítulo.

En el caso de sistema de hipótesis

$$H_0: \ \mu = \mu_0 \quad vs. \quad H_1: \ \mu \neq \mu_0,$$

con σ^2 desconocida, se vio que un intervalo de confianza para μ es

$$(\bar{x} - t_{n-1,1-\alpha/2} s_{n-1}/\sqrt{n}, \bar{x} + t_{n-1,1-\alpha/2} s_{n-1}/\sqrt{n}).$$

Y se rechaza H_0 si el intervalo no contiene a μ_0, esto es, si

$$\mu_0 < \bar{x} - t_{n-1,1-\alpha/2} s_{n-1}/\sqrt{n} \ \text{ o } \ \mu > \bar{x} + t_{n-1,1-\alpha/2} s_{n-1}/\sqrt{n}$$

que es equivalente a

$$t_{n-1,1-\alpha/2} < \sqrt{n}(\bar{x} - \mu_0)/s_{n-1} \ \text{ o } \ \sqrt{n}(\bar{x} - \mu_0)/s_{n-1} < -t_{n-1,1-\alpha/2},$$

la cual es la regla de decisión encontrada anteriormente, ver (4.2.17).

Por otro lado, si el sistema de interés es

$$H_0: \ \mu \leq \mu_0 \quad vs. \quad H_1: \ \mu > \mu_0,$$

entonces el intervalo útil es $(\bar{x} - z_{1-\alpha}\sigma/\sqrt{n}, \infty)$, y rechaza H_0, si

$$\bar{x} - z_{1-\alpha}\sigma/\sqrt{n} > \mu_0,$$

equivalente a

$$\sqrt{n}(\bar{x} - \mu_0)/\sigma > z_{1-\alpha},$$

que es la regla de decisión encontrada anteriormente, ver (4.2.6)

Aunque en los anteriores, el uso de los intervalos de confianza es equivalente a los procedimientos de prueba de hipótesis, el uso de los intervalos de confianza es muy limitado, pues no funcionan para sistemas donde $\boldsymbol{\Theta}_1 \neq \boldsymbol{\Theta}_0^c$. Por consiguiente, no funcionan para sistemas como

$$H_0: \ \mu = \mu_0 \quad vs. \quad H_1: \ \mu > \mu_0.$$

Otra observación importante acerca de los prueba de hipótesis es con respecto al nivel de significación α. En la práctica, el valor usado para α es generalmente 0.02, 0.05 o 0.1. Como se ha visto anteriormente, ésta es la probabilidad de cometer el error tipo I, pero en muchos informes estadísticos, se presenta a α como la magnitud

del error, sin mencionar el error tipo II. Y esto causa la falsa impresión de que entre más pequeño sea α, más confiable es el resultado del procedimiento. Lo que ocurre realmente al escoger un valor de α pequeño, como 0.02 o 0.01, es que el área de la región de rechazo también disminuye, pues ésta es igual a α. De esta forma, es más difícil rechazar H_0. De hecho, existen situaciones donde al disminuir el valor de α, la decisión puede cambiar de rechazar al no rechazar.

Considera el sistema

$$H_0 : \ \mu = 0 \quad vs. \quad H_1 : \ \mu \neq 0,$$

con σ^2 desconocido. Se ha visto anteriormente que la regla de decisión es: rechazar H_0 si $\frac{\sqrt{n}(\bar{x}-\mu_0)}{s} > t_{n-1,1-\alpha/2}$ ó $\frac{\sqrt{n}(\bar{x}-\mu_0)}{s} < -t_{n-1,1-\alpha/2}$. Suponga que en una muestra de 20 observaciones, $\bar{x} = 0.5$, y $s^2_{n-1} = 1$, entonces $\sqrt{n}\bar{x}/s_{n-1} = 2.24$. Ahora, si el nivel de significación $\alpha = 0.05$, entonces $t_{n-1,1-\alpha/2} = 2.09$, y llegamos a la conclusión de rechazar H_0; por otro lado, si $\alpha = 0.02$, entonces $t_{n-1,1-\alpha/2} = 2.54$, y llegamos a la conclusión de no rechazar H_0, y observamos cómo los resultados cambian al cambiar el nivel de significación.

Finalmente, volvemos a tomar el tema del planteamiento de un sistema de hipótesis. Cuando se plantea un sistema como

$$H_0 : \ \theta \in \mathbf{\Theta}_0 \quad vs. \quad H_1 : \ \theta \in \mathbf{\Theta}_1.$$

La decisión se toma en base a una muestra observada, y la muestra debe tener suficiente evidencia en contra de H_0 para llegar a la decisión de rechazar H_0. De hecho, si revisamos los sistemas de hipótesis vistos anteriormente, podemos observar que el área del región de rechazo es de tan solo α. En algunas situaciones, una muestra observada puede no mostrar suficiente evidencia en contra de H_0, ni en contra de H_1. En estas situaciones, el planteamiento del sistema de hipótesis es crucial en la toma correcta de decisiones.

Considera el sistema

$$H_0 : \ \mu \geq 5 \quad vs. \quad H_1 : \ \mu < 5,$$

con σ^2 desconocido. Suponga que el nivel de significación es $\alpha = 0.05$. Se ha visto anteriormente que la regla de decisión es: rechazar H_0 si $\sqrt{n}(\bar{x}-5)/s_{n-1} < t_{n-1,\alpha} = -1.73$. Suponga que en una muestra de 20 observaciones, $\bar{x} = 5.2$, y $s^2_{n-1} = 1$, entonces $\sqrt{n}(\bar{x}-5)/s_{n-1} = 0.89$, y no se rechaza H_0, es decir, podemos aceptar $\mu \geq 5$.

Ahora cambiamos los hipótesis en el anterior sistema, y consideramos el siguiente sistema

$$H_0 : \ \mu \leq 5 \quad vs. \quad H_1 : \ \mu > 5,$$

se puede ver que la regla de decisión en este caso es: rechazar H_0 si $\sqrt{n}(\bar{x}-5)/s_{n-1} > t_{n-1,1-\alpha} = 1.73$, para la misma muestra observada, $\sqrt{n}(\bar{x}-5)/s_{n-1} = 0.89$, y llegamos a la conclusión de no rechazar H_0, es decir, aceptamos $\mu \leq 5$.

Obsérvese que dos personas pueden llegar a conclusiones totalmente diferentes utilizando los mismos datos, inclusive utilizando ambos los correctos procedimientos estadísticos. Situaciones como ésta forman parte de las críticas que existen hacia los procedimientos de pruebas de hipótesis. El problema radica en el planteamiento de la hipótesis, debemos recordar que en el procedimiento de buscar reglas de decisión, sólo se está teniendo en cuenta la magnitud del error tipo I, sin considerar el error tipo II; por lo tanto, el usuario de las técnicas de prueba de hipótesis debe asociarse con el experto del tema específico, y plantear el sistema donde el error tipo I es menos grave que el error tipo II.

Una observación interesante acerca de las pruebas de hipótesis es que el procedimiento de éste es similar al procedimiento matemático utilizado para hacer una demostración por contradicción, miremos por qué. Desde nuestros primeros contactos con la estadística nos hemos dado cuenta de que ésta está muy ligada a los fundamentos matemáticos (alguien dijo que la matemática es la esclava de todas las ciencias). También sabemos que una herramienta fundamental de la estadística es, sin duda alguna, la prueba de hipótesis; mientras que para los matemáticos la demostración por contradicción es indispensable en muchos desarrollos teóricos. Bien, para la sorpresa de muchos, hay una similitud increíble entre estos métodos.

Antes que todo recordemos cómo se demuestra la veracidad de una proposición, P, mediante contradicción: se supone que la proposición, P, es falsa, luego se observa si con este supuesto se puede llegar a alguna contradicción. Si así sucede, podemos concluir que el supuesto anterior es falso; es decir, que la falsedad de P es falsa, y por consiguiente se concluye lo que se quería demostrar: que P es verdadera.

Ahora pensemos en el procedimiento de la prueba de una hipótesis H_0: primero se supone que H_0 es verdadero. Bajo este supuesto, se observa si el valor muestral de la estadística de prueba pertenece a la región de rechazo que equivale a la contradicción, lo cual sucede con una probabilidad α que, por lo general, es muy pequeña. Si este evento sucede, concluimos que H_0 es falso. Se puede ver la similitud entre estos dos procedimientos teniendo en cuenta que en el primer paso se asume un supuesto en ambos casos. En el segundo paso, podemos ver que una contradicción equivale, en el caso de prueba de hipótesis, a que un evento, con probabilidad de ocurrencia muy pequeña, suceda. Este evento es: el valor de la estadística pertenece a la región de rechazo. En el tercer paso, si se llega a la contradicción se concluye que el supuesto planteado en el primer paso es falso; es decir, se rechaza H_0.

4.2.2 Pruebas de hipótesis acerca de la varianza teórica

$H_0:\ \sigma^2 = \sigma_0^2$ **vs.** $H_1:\ \sigma^2 \neq \sigma_0^2$. **con** μ **conocido**

Consideramos el siguiente sistema de hipótesis en una muestra aleatoria $X_1, \ldots, X_n$ proveniente de una distribución $N(\mu, \sigma^2)$ con μ conocida.

$$H_0:\ \sigma^2 = \sigma_0^2 \quad vs. \quad H_1:\ \sigma^2 \neq \sigma_0^2. \tag{4.2.19}$$

El estimador MV de σ^2 está dado por $\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$. Considerando la

forma del sistema de hipótesis, se puede plantear una regla inicial de decisión: rechazar H_0 cuando $\hat{\sigma}^2_{MV}$ es muy grande comparado con σ^2_0 o muy pequeño comparado con σ^2_0. Esta idea puede ser formalizada como

$$\text{Rechazar } H_0 \text{ cuando } \hat{\sigma}^2_{MV} - \sigma^2_0 > K_1 \text{ o } \hat{\sigma}^2_{MV} - \sigma^2_0 < K_2$$

para constantes $K_1 > 0$ y $K_2 < 0$. Para completar la regla de decisión, es necesario encontrar los valores de K_1 y K_2; para ello, recurrimos nuevamente a la definición del error tipo I, y al limitar a la probabilidad de cometer este error a ser igual a α, tenemos

$$\begin{aligned}\alpha &= Pr(\hat{\sigma}^2_{MV} - \sigma^2_0 > K_1) + Pr(\hat{\sigma}^2_{MV} - \sigma^2_0 < K_2) \quad \text{Cuando } H_0 \text{ es cierta} \\ &= Pr\left(\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2_0} > \frac{nK_1}{\sigma^2_0} + n\right) + Pr\left(\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2_0} < \frac{nK_2}{\sigma^2_0} + n\right).\end{aligned}$$

Recordando que $\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2_0} \sim \chi^2_n$ bajo la hipótesis nula de $\sigma^2 = \sigma^2_0$, se tiene que $\frac{nK_1}{\sigma^2_0} + n$ y $\frac{nK_2}{\sigma^2_0} + n$ son percentiles de la distribución χ^2_n, y existen muchos percentiles que satisfacen la anterior igualdad. Los percentiles que se usan con más frecuencia son $\frac{nK_1}{\sigma^2_0} + n = \chi^2_{n,1-\alpha/2}$ y $\frac{nK_2}{\sigma^2_0} + n = \chi^2_{n,\alpha}$. De donde se tiene que $K_1 = \frac{\chi^2_{n,1-\alpha/2} - n}{n\sigma^2_0}$ y $K_2 = \frac{\chi^2_{n,\alpha/2} - n}{n\sigma^2_0}$. De esta forma, tenemos la siguiente regla de decisión para el sistema (4.2.19)

$$\text{Rechazar } H_0 \text{ cuando } \hat{\sigma}^2_{MV} - \sigma^2_0 > \frac{\chi^2_{n,1-\alpha/2} - n}{n\sigma^2_0} \text{ o } \hat{\sigma}^2_{MV} - \sigma^2_0 < \frac{\chi^2_{n,\alpha/2} - n}{n\sigma^2_0}.$$

Simples operaciones algebraicas nos conducen a la siguiente regla de decisión equivalente

$$\text{Rechazar } H_0 \text{ cuando } \frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2_0} > \chi^2_{n,1-\alpha/2} \text{ o } \frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2_0} < \chi^2_{n,\alpha/2}.$$

p valor

Para obtener la forma de calcular el p valor para el sistema (4.2.19) ligado a la anterior regla de decisión, se tiene en cuenta que la estadística de prueba es $\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2_0}$. En una muestra observada, el valor que toma la estadística puede ser mayor o menor al percentil 0.5 de la distribución χ^2_n, como lo ilustra en la Figura 4.14. Puede tomar valores como v_1 o como v_2. Cuando el valor de la estadística v es mayor a $\chi^2_{n,0.5}$, podemos observar que v cae en la región de rechazo cuando la probabilidad a la derecha es menor a $\alpha/2$, el área del región de rechazo a la derecha, esto es, cuando $Pr(\chi^2_n > v) < \alpha/2$, con χ^2_n una variable aleatoria con distribución χ^2_n; por otro lado, cuando el valor de la estadística v es menor a $\chi^2_{n,0.5}$, se rechaza H_0 cuando $Pr(\chi^2_n < v) < \alpha/2$. En conclusión

$$\text{Rechazar } H_0 \text{ si } \begin{cases} 2Pr(\chi^2_n > v) < \alpha, & \text{para } v > \chi^2_{n,0.5} \text{ ;} \\ 2Pr(\chi^2_n < v) < \alpha, & \text{para } v < \chi^2_{n,0.5}. \end{cases}$$

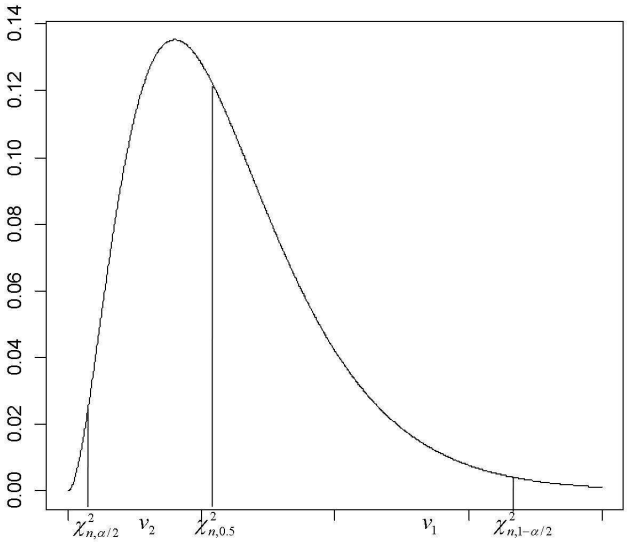

Figura 4.14: *Ilustración del p valor para la hipótesis (4.2.19)*

De esta forma, se define el p valor como

$$p \text{ valor} = \begin{cases} 2Pr(\chi^2_n > v), & \text{para } v > \chi^2_{n,0.5} \text{ ;} \\ 2Pr(\chi^2_n < v), & \text{para } v < \chi^2_{n,0.5}. \end{cases}$$

y se rechaza H_0 cuando el p valor es menor que el nivel de significación α.

Función de potencia

De la definición de la función de potencia tenemos que

$$\begin{aligned} \beta(\sigma^2) &= Pr\left(\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma_0^2} > \chi^2_{n,1-\alpha/2}\right) + Pr\left(\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma_0^2} < \chi^2_{n,\alpha/2}\right) \\ &= Pr\left(\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2} > \frac{\sigma_0^2}{\sigma^2}\chi^2_{n,1-\alpha/2}\right) + Pr\left(\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2} < \frac{\sigma_0^2}{\sigma^2}\chi^2_{n,\alpha/2}\right) \\ &= 1 - F_{\chi^2_n}\left(\frac{\sigma_0^2}{\sigma^2}\chi^2_{n,1-\alpha/2}\right) + F_{\chi^2_n}\left(\frac{\sigma_0^2}{\sigma^2}\chi^2_{n,\alpha/2}\right) \end{aligned} \tag{4.2.20}$$

donde $F_{\chi^2_n}(\cdot)$ es la función de distribución de uan distribución χ^2_n. Nótese que en primer lugar, esta función se define para valores positivos, y además no depende de la media teórica μ. Observando simplemente la forma de esta función, no es claro si al aumentar el tamaño muestral la potencia incrementa o no, pero podemos graficar la función $\beta(\sigma^2)$ para diferentes tamaños muestrales y cómo afectan estos en la potencia de la prueba.

```
> po_sigma<-function(sigma,n,sigma0,alpha){
+ 1-pchisq(sigma0/sigma*qchisq(1-alpha/2,n),n)+
+ pchisq(sigma0/sigma*qchisq(alpha/2,n),n)   }
>
> alpha<-0.05
> sigma0<-10
> n1<-10
> n2<-30
> n3<-50
>
> sigma<-sigma0+seq(-10,30,0.1)
>
> plot(function(x) po_sigma(x,n1,sigma0,alpha),0,40,type="l",
+ xlab="sigma^2",ylab="función de potencia")
> curve(po_sigma(x,n2,sigma0,alpha),0,40,,lty=2,add=T)
> curve(po_sigma(x,n3,sigma0,alpha),0,40,,lty=3,add=T)
> legend(30,0.4,c("n=10","n=30","n=50"),lty=c(1,2,3))
```

En la Figura 4.15 se muestra esta función de potencia para diferentes tamaños de muestra y se observa que la potencia aumenta al incrementarse el tamaño muestral. Además observe que $\beta(\sigma^2)$ no es simétrica con respecto a σ^2, esto implica que si se tienen dos muestras aleatorias, la primera con varianza teórica $\sigma_0^2+\delta$ y la segunda con $\sigma_0^2-\delta$ para algún $0<\delta<\sigma_0^2$, y se juzga el sistema $H_0:\ \sigma^2=\sigma_0^2$ vs. $H_1:\ \sigma^2\neq\sigma_0^2$, es menos probable aceptar H_0 en la segunda muestra que en la primera. En otras palabras, cuando en la práctica se acepta $H_0:\ \sigma^2=\sigma_0^2$, es posible que en la población σ^2 sea diferente que σ_0^2 y lo que ilustran las gráficas de (4.2.20) es que es más probable que en la población ocurra $\sigma^2>\sigma_0^2$ que ocurra $\sigma^2<\sigma_0^2$.

$H_0:\ \sigma^2\leq(=)\sigma_0^2$ **vs.** $H_1:\ \sigma^2>\sigma_0^2$. **con** μ **conocido**

Ahora, consideramos el siguiente sistema de hipótesis

$$H_0:\ \sigma^2=\sigma_0^2 \quad vs. \quad H_1:\ \sigma^2>\sigma_0^2. \tag{4.2.21}$$

que tiene la misma regla de decisión que el sistema

$$H_0:\ \sigma^2\leq\sigma_0^2 \quad vs. \quad H_1:\ \sigma^2>\sigma_0^2.$$

Usaremos la prueba de la razón de verosimilitud para encontrar una regla de decisión. Para eso, primero transformamos el anterior sistema en el siguiente

$$H_0:\ \sigma^2=\sigma_0^2 \quad vs. \quad H_1:\ \sigma^2=\sigma_1^2, \tag{4.2.22}$$

con $\sigma_1^2>\sigma_0^2$.

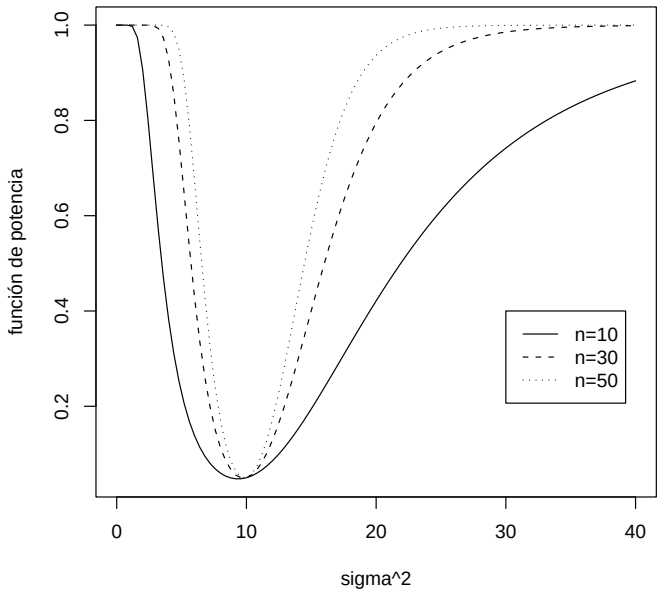

Figura 4.15: *Función de potencia (4.2.20) para diferentes tamaños de muestra con* $\alpha = 0.05$ *y* $\sigma_0^2 = 10$.

La función de verosimilitud en una muestra proveniente de la distribución normal está dada por

$$L(\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right\}.$$

Entonces la razón de verosimilitudes está dada por

$$\lambda = \frac{(\sigma_1^2)^{-n/2}}{(\sigma_0^2)^{-n/2}} \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_0^2}\right)\sum_{i=1}^{n}(X_i-\mu)^2\right\},$$

y H_0 se rechaza para valores grandes de λ. Ahora, en la expresión de λ, la parte aleatoria que toma valores diferentes en muestras diferentes es $\sum_{i=1}^{n}(X_i-\mu)^2$, y obsérvese que la expresión $\frac{(\sigma_1^2)^{-n/2}}{(\sigma_0^2)^{-n/2}}$ y $-\frac{1}{2}\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_0^2}\right)$ son ambos positivos, entonces cuando $\sum_{i=1}^{n}(X_i-\mu)^2$ toma valores grandes, λ también lo hace. De esta forma podemos afirmar que se rechaza H_0 para valores grandes de $\sum_{i=1}^{n}(X_i-\mu)^2$. Esto es

$$\text{Rechazar } H_0 \text{ cuando } \sum_{i=1}^{n}(X_i-\mu)^2 > K$$

para algún $K > 0$. De nuevo, para encontrar el valor de K, se utiliza la definición del error tipo I, de donde

$$\begin{aligned}\alpha &= Pr\left(\sum_{i=1}^{n}(X_i-\mu)^2 > K\right) \quad \text{Cuando } H_0 \text{ es cierta} \\ &= Pr\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2} > \frac{K}{\sigma_0^2}\right).\end{aligned}$$

Cuando H_0 es cierta, $\dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2} \sim \chi_n^2$, entonces la anterior expresión indica que $\dfrac{K}{\sigma_0^2} = \chi^2_{n,1-\alpha}$. De esta forma se tiene la siguiente regla de decisión

$$\text{Rechazar } H_0 \text{ cuando } \textstyle\sum_{i=1}^{n}(X_i-\mu)^2 > \chi^2_{n,1-\alpha}\sigma_0^2,$$

o equivalentemente

$$\text{Rechazar } H_0 \text{ cuando } \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2} > \chi^2_{n,1-\alpha}.$$

Región de rechazo y p valor

En la anterior regla de decisión la región de rechazo asociada con la estadística de prueba $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2}$ está dada por $\{c :\in \mathbb{R} :\ c > \chi^2_{n,1-\alpha}\}$. Y asociado a esta región de rechazo el p valor se calcula como

$$p \text{ valor} = Pr(\chi_n^2 > v),$$

donde v es el valor observado de la estadística de prueba $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2}$.

Función de potencia

Teniendo en cuenta la definición de la función de potencia, tenemos que

$$\beta(\sigma^2) = 1 - F_{\chi_n^2}\left(\frac{\sigma_0^2}{\sigma^2}\chi^2_{n,1-\alpha}\right). \tag{4.2.23}$$

Ilustramos la forma de esta función en la Figura 4.16 con $\sigma_0^2 = 10$, $\alpha = 0.05$ para diferentes tamaños de muestra.

$H_0 :\ \sigma^2 \geq (=)\sigma_0^2$ **vs.** $H_1 :\ \sigma^2 < \sigma_0^2$. **con** μ **conocido**

El procedimiento para encontrar una regla de decisión para este sistema es similar a lo desarrollado anteriormente y se deja como ejercicio (Ejercicio 4.7).

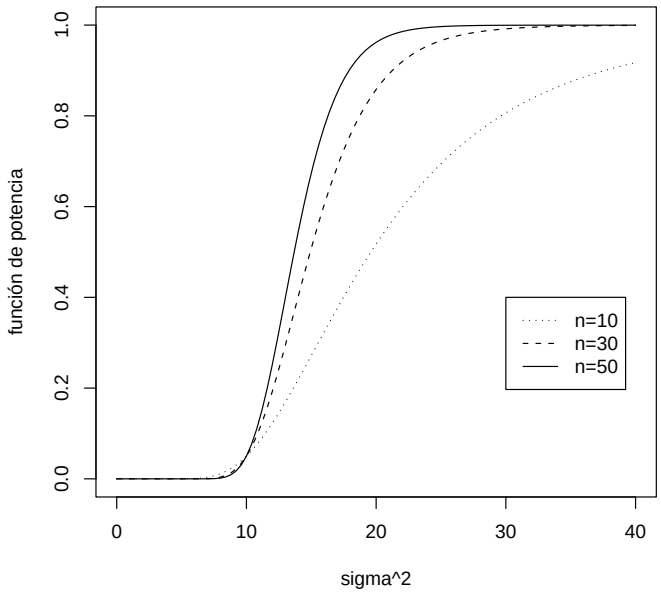

Figura 4.16: *Función de potencia (4.2.23) para diferentes tamaños de muestra con* $\alpha = 0.05$ *y* $\sigma_0^2 = 10$.

$H_0: \ \sigma^2 = \sigma_0^2$ **vs.** $H_1: \ \sigma^2 \neq \sigma_0^2$. **con** μ **desconocido**

Cuando la media teórica μ es desconocida, el estimador de máxima verosimilitud para σ^2 es $\hat{\sigma^2}_{MV} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, y al adoptar el procedimiento presentado al principio de la sección 4.2.2 para el caso cuando μ es conocido, se puede encontrar que para el sistema

$$H_0: \ \sigma^2 = \sigma_0^2 \quad vs. \quad H_1: \ \sigma^2 \neq \sigma_0^2,$$

una regla de decisión es

$$\text{Rechazar } H_0 \text{ cuando } \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha/2} \text{ o}$$
$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma_0^2} < \chi^2_{n-1,\alpha/2}.$$

Y se encuentra, similarmente, el p valor dado por

$$p \text{ valor} = \begin{cases} 2Pr(\chi^2_{n-1} > v), & \text{para } v > \chi^2_{n-1,0.5} \ ; \\ 2Pr(\chi^2_{n-1} < v), & \text{para } \ v < \chi^2_{n-1,0.5}. \end{cases}$$

donde χ^2_{n-1} denota una variable aleatoria con distribución χ^2_{n-1} y v es el valor observado de la estadística de prueba $\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\sigma_0^2}$.

También, es fácil verificar que la función de potencia está dada por

$$\beta(\sigma^2) = 1 - F_{\chi^2_{n-1}}\left(\frac{\sigma_0^2}{\sigma^2}\chi^2_{n-1,1-\alpha/2}\right) + F_{\chi^2_{n-1}}\left(\frac{\sigma_0^2}{\sigma^2}\chi^2_{n-1,\alpha/2}\right) \tag{4.2.24}$$

El lector puede ver que la anterior función de potencia es muy similar a (4.2.20) cuando μ es conocido excepto que el grado de libertad de la distribución χ^2 se cambia de n a $n-1$, por consiguiente la forma de la función de potencia (4.2.24) tamién es muy similar a la de (4.2.20) presentada en la Figura 4.15.

$H_0:\ \sigma^2 \leq (=)\sigma_0^2$ **vs.** $H_1:\ \sigma^2 > \sigma_0^2$. **con** μ **desconocido**

Para el sistema $H_0:\ \sigma^2 = \sigma_0^2$ vs. $H_1:\ \sigma^2 > \sigma_0^2$ hacemos uso de la prueba de razón de verosimilitudes escribiendo al sistema como

$$H_0:\ \sigma^2 = \sigma_0^2 \quad vs. \quad H_1:\ \sigma^2 = \sigma_1^2,$$

con $\sigma_1^2 > \sigma_0^2$. Al replicar el procedimiento de la prueba de razón de verosimilitud para el caso cuando μ es conocido, se encuentra la misma estadística de prueba $\sum_{i=1}^n (X_i-\mu)^2$, que no se puede calcular cuando μ es desconocido. La solución a este problema es reemplazar μ por su estimador $\bar{X}$. De esta forma, se tiene que la razón de verosimilitud está dada por

$$\lambda = \frac{(\sigma_1^2)^{-n/2}}{(\sigma_0^2)^{-n/2}} \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_0^2}\right)\sum_{i=1}^n (X_i-\bar{X})^2\right\}.$$

Usando $\sigma_1^2 > \sigma_0^2$, se concluye que H_0 se rechaza cuando $\sum_{i=1}^n (X_i-\bar{X})^2 > K$. Finalmente, usando la propiedad

$$\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\sigma_0^2} \sim \chi^2_{n-1},$$

bajo H_0, se tiene que $K = \sigma_0^2\chi^2_{n-1,1-\alpha}$, y la regla de decisión está dada por

$$\text{Rechazar } H_0 \text{ cuando } \frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha}.$$

Y el p valor asociado está dado por

$$p \text{ valor} = Pr(\chi^2_{n-1} > v)$$

donde v es el valor observado de la estadística de prueba $\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\sigma_0^2}$.

$H_0:\ \sigma^2 \geq (=)\sigma_0^2$ **vs.** $H_1:\ \sigma^2 < \sigma_0^2$. **con** μ **desconocido**

El procedimiento para encontrar una regla de decisión para este sistema es similar a lo desarrollado anteriormente y se deja como ejercicio (Ejercicio 4.7).

4.3 Dos muestras

Ahora consideramos el problema de dos muestras, donde se encuentran dos poblaciones que tienen una característica común de interés, y se desea comparar las dos poblaciones con base en muestras de estas poblaciones. Suponga que tienen dos muestras aleatorias independientes de tamaño n_X y n_Y denotadas por $X_1, \ldots, X_{n_X}$ y $Y_1, \ldots, Y_{n_Y}$ provenientes de $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$, respectivamente.

4.3.1 Comparación entre dos medias

En primer lugar, consideramos el problema de comparar las dos poblaciones en términos de las medias teóricas. Por ejemplo, en la industria, cuando se dispone de dos maquinarias o dos líneas de producción que realizan la misma labor, se quiere comparar las dos maquinarias en términos de alguna variable de interés como calidad de productos o eficiencia de producción.

El lector recuerda que en problemas de una muestra cuando se juzgan hipótesis acerca de la media teórica μ, la regla de decisión depende de si la varianza teórica es conocida o no. En el problema de dos muestras tenemos dos varianzas teóricaes σ_X^2 y σ_Y^2, y la regla de decisión para sistemas de hipótesis acerca de las medias teóricas también depende de si estas dos varianzas son conocidas o no.

$H_0:\ \mu_X^2 = \mu_Y^2$ **vs.** $H_1:\ \mu_X^2 \neq \mu_Y^2$, **con** σ_X^2 **y** σ_Y^2 **conocidas**

En primer lugar, suponemos que las varianzas teóricas, σ_X^2 y σ_Y^2 son conocidas. Usaremos la prueba de razón generalizada de verosimilitudes para encontrar una regla de decisión para el sistema

$$H_0:\ \mu_X^2 = \mu_Y^2 \quad vs.\ H_1:\ \mu_X^2 \neq \mu_Y^2 \tag{4.3.1}$$

En primer lugar, dado que las dos muestras son independientes, se tiene que la función de verosimilitud de las dos muestras es simplemente el producto de las dos funciones de verosimilitudes. Por lo tanto ésta está dada por

$$L(x_1, \ldots, x_{n_X}, y_1, \ldots, y_{n_Y}) = (2\pi\sigma_X^2)^{-n_X/2} \exp\left\{-\frac{1}{2\sigma_X^2}\sum_{i=1}^{n_X}(X_i - \mu_X)^2\right\}$$
$$(2\pi\sigma_Y^2)^{-n_Y/2} \exp\left\{-\frac{1}{2\sigma_Y^2}\sum_{i=1}^{n_Y}(Y_i - \mu_Y)^2\right\}$$

Dado que en el sistema considerado, $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_0^c$, se tiene que el numerador de la razón generalizada de verosimilitudes λ está dado por la función de verosimilitud evaluada en los estimadores MV de μ_X y μ_Y. Esto es,

$$L(\hat{\mu}_{X,MV}, \hat{\mu}_{Y,MV}) = (2\pi\sigma_X^2)^{-n_X/2} \exp\left\{-\frac{1}{2\sigma_X^2}\sum_{i=1}^{n_X}(X_i - \bar{X})^2\right\} \\ (2\pi\sigma_Y^2)^{-n_Y/2} \exp\left\{-\frac{1}{2\sigma_Y^2}\sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2\right\}.$$

Por otro lado, bajo la hipótesis nula H_0, $\mu_X = \mu_Y$, entonces las dos muestras provienen de distribución con la misma media teórica, μ. Se ha visto en (2.3.6) que el estimador MV de la media teórica común está dado por

$$\hat{\mu}_{MV} = \frac{\sigma_Y^2 n_X \bar{X} + \sigma_X^2 n_Y \bar{Y}}{n_X\sigma_Y^2 + n_Y\sigma_X^2}. \tag{4.3.2}$$

Y de esta forma, el denominador de λ está dado por

$$L(\hat{\mu}_{MV}) = (2\pi\sigma_X^2)^{-n_X/2} \exp\left\{-\frac{1}{2\sigma_X^2}\sum_{i=1}^{n_X}(X_i - \hat{\mu}_{MV})^2\right\} \\ (2\pi\sigma_Y^2)^{-n_Y/2} \exp\left\{-\frac{1}{2\sigma_Y^2}\sum_{j=1}^{n_Y}(Y_j - \hat{\mu}_{MV})^2\right\}.$$

En conclusión, $\ln\lambda$ está dado por

$$\ln\lambda = -\frac{1}{2\sigma_X^2}\left\{\sum_{i=1}^{n_X}\left[(X_i - \bar{X})^2 - (X_i - \hat{\mu}_{MV})^2\right]\right\} - \frac{1}{2\sigma_Y^2} \\ \left\{\sum_{j=1}^{n_Y}\left[(Y_j - \bar{Y})^2 - (Y_j - \hat{\mu}_{MV})^2\right]\right\},$$

donde

$$\sum_{i=1}^{n_X}\left[(X_i - \bar{X})^2 - (X_i - \hat{\mu}_{MV})^2\right] = \sum_{i=1}^{n_X}(2X_i - \bar{X} - \hat{\mu}_{MV})(\hat{\mu}_{MV} - \bar{X}) \\ = -n_X(\bar{X} - \hat{\mu}_{MV})^2.$$

Simples operaciones algebraicas muestran que

$$\bar{X} - \hat{\mu}_{MV} = \frac{\sigma_X^2 n_Y(\bar{X} - \bar{Y})}{n_X\sigma_Y^2 + n_Y\sigma_X^2}, \tag{4.3.3}$$

de donde

$$\sum_{i=1}^{n_X}\left[(X_i-\bar{X})^2-(X_i-\hat{\mu}_{MV})^2\right]=-\frac{n_X n_Y^2\sigma_X^4(\bar{X}-\bar{Y})^2}{(n_X\sigma_Y^2+n_Y\sigma_X^2)^2}.$$

Análogamente, se tiene que

$$\sum_{j=1}^{n_Y}\left[(Y_i-\bar{Y})^2-(Y_i-\hat{\mu}_{MV})^2\right]=-\frac{n_X^2 n_Y\sigma_Y^4(\bar{X}-\bar{Y})^2}{(n_X\sigma_Y^2+n_Y\sigma_X^2)^2}.$$

De donde, tenemos que

$$\ln\lambda=\frac{n_X n_Y}{2(n_X\sigma_Y^2+n_Y\sigma_X^2)^2}(\bar{X}-\bar{Y})^2.$$

Se debe rechazar H_0 para valores grandes de $\ln\lambda$, la cual es equivalente a rechazar H_0 para valores grandes de $(\bar{X}-\bar{Y})^2$ o $|\bar{X}-\bar{Y}|$. Y tenemos la siguiente regla de decisión

$$\text{Rechazar } H_0 \text{ cuando } |\bar{X}-\bar{Y}|>K, \text{ para algún } K>0.$$

Para encontrar el valor de K, tenemos que

$$\begin{aligned}\alpha &= Pr(|\bar{X}-\bar{Y}|>K)\\ &= Pr(\bar{X}-\bar{Y}>K)+Pr(\bar{X}-\bar{Y}<-K),\end{aligned}\tag{4.3.4}$$

suponiendo que H_0 es cierta. En este caso,

$$\bar{X}-\bar{Y}\sim N\left(0,\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}\right),$$

y de (4.3.4), se concluye que $Pr(\bar{X}-\bar{Y}>K)=\alpha/2$, y se encuentra que $K=z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$. Y tenemos la regla de decisión

$$\text{Rechazar } H_0 \text{ cuando } |\bar{X}-\bar{Y}|>z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$$

o equivalentemente

$$\text{Rechazar } H_0 \text{ cuando } \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}>z_{1-\alpha/2} \text{ o } \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}<-z_{1-\alpha/2}.$$

Nótese que la anterior regla de decisión es equivalente al uso del intervalo de confianza (3.2.31). Puesto que con el uso del intervalo, se rechaza $\mu_X=\mu_Y$ cuando el valor 0 no pertenece al intervalo, esto es, $0<\bar{X}-\bar{Y}-z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$ o $0>\bar{X}-\bar{Y}+z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$, lo cual es equivalente a la regla de decisión encontrada anteriormente.

p valor

La forma de calcular el p valor es similar a lo discutido para el sistema $\mu = \mu_0$ vs. $\mu \neq \mu_0$, cuando la varianza teórica es conocida. Y se encuentra que

$$p \text{ valor} = \begin{cases} 2Pr(Z > v), & \text{para } v > 0 \ ; \\ 2Pr(Z < v), & \text{para } \ v < 0. \end{cases}$$

donde v es el valor observado de la estadística de prueba $(\bar{X} - \bar{Y})/\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$ y Z denota una variable aleatoria con distribución normal estándar.

Función de potencia

Para encontrar la función de potencia tenemos que

$$\begin{aligned} & \beta(\mu_X, \mu_Y) && (4.3.5) \\ &= Pr(\text{Rechazar } H_0) \\ &= Pr\left(\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} > z_{1-\alpha/2} \right) + Pr\left(\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} < -z_{1-\alpha/2} \right) \\ &= Pr\left(\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} > z_{1-\alpha/2} + \frac{(\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \right) && (4.3.6) \\ &\quad + Pr\left(\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} < -z_{1-\alpha/2} + \frac{(\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \right) \\ &= 1 - \Phi\left(z_{1-\alpha/2} + \frac{(\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \right) + \Phi\left(-z_{1-\alpha/2} + \frac{(\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \right) && (4.3.7) \end{aligned}$$

Observemos que la función $\beta(\mu_X, \mu_Y)$ depende, inicialmente, de dos parámetros μ_X y μ_Y, por lo tanto es una función de dos argumentos, y la gráfica debe ser como la mostrada en la Figura 4.17.

Sin embargo, en (4.3.5) podemos ver que la función de potencia depende de μ_X y μ_Y únicamente a través de $\mu_X - \mu_Y$, de hecho la gráfica en Figura 4.17 es simétrica con respecto al plano $\{(\mu_X, \mu_Y) : \ \mu_X = \mu_Y\}$. De esta forma, podemos visualizar esta función de potencia solo variando los valores de $\mu_X - \mu_Y$. El siguiente código calcula y grafica la función (4.3.5) para diferentes tamaños de muestra, y el resultado se observa en la Figura 4.18

```
> pote_2_norm<-function(dif,nx,ny,alpha,sigmaX,sigmaY){
+ 1-pnorm(dif/sqrt((sigmaX/nx)+(sigmaY/ny))+qnorm(1-alpha/2))+
+ pnorm(dif/sqrt((sigmaX/nx)+(sigmaY/ny))-qnorm(1-alpha/2)) }
```

```
> alpha<-0.05
> sigmaX<-1
> sigmaY<-4
> n1<-10
> n2<-30
> n3<-50

> plot(function(x) pote_2_norm(x,n1,n1,alpha,sigmaX,sigmaY),-4,4,
+ type="l",xlab="mu_X-mu_Y",ylab="función de potencia")
> curve(pote_2_norm(x,n2,n2,alpha,sigmaX,sigmaY),-4,4,lty=2,add=T)
> curve(pote_2_norm(x,n3,n3,alpha,sigmaX,sigmaY),-4,4,lty=3,add=T)
> legend(1.8,0.4,c("nx=ny=10","nx=ny=30","nx=ny=50"),lty=c(1,2,3))
```

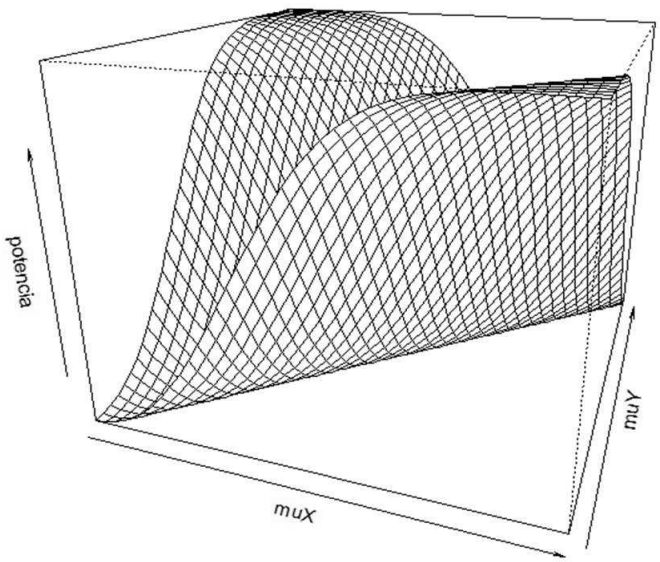

Figura 4.17: *Función de potencia (4.2.23) para con* $\alpha = 0.05$, $\sigma_X^2 = 1$, $\sigma_Y^2 = 4$ *y* $n_X = n_Y = 10$.

Observamos que la fórmula de la anterior función de potencia se asemeja a la del sistema (4.2.1) con varianza teórica conocida, y por consiguiente es de esperar que la forma de las dos funciones de potencia también sean similares, lo cual se confirma observando la Figura 4.18.

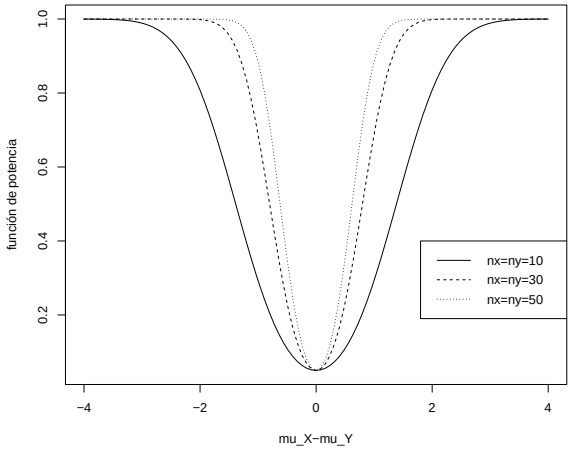

Figura 4.18: *Función de potencia (4.3.5) para diferentes tamaños de muestra con* $\alpha = 0.05$ *y* $\sigma_0^2 = 10$.

σ_X^2 y σ_Y^2 son desconocidas, pero iguales.

Suponga que $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ es desconocida, se debe reemplazarlas por el estimador MV de la varianza común σ^2. Como se mencionaba en el capítulo 2, cuando las dos muestras provienen de distribuciones con la misma varianza teórica σ^2, se puede usar las variables de las dos muestras para estimar la varianza común σ^2, en este caso, y se tiene que

$$\hat{\sigma}_{MV}^2 = \frac{(n_X - 1)S_{n_X-1,X}^2 + (n_Y - 1)S_{n_Y-1,Y}^2}{n_X + n_Y}.$$

De esta forma tenemos que el numerador de la razón generalizada de verosimilitudes está dado por

$$\begin{aligned} & L(\hat{\mu}_{X,MV}, \hat{\mu}_{Y,MV}, \hat{\sigma}_{MV}^2) \\ &= \left(2\pi\hat{\sigma}_{MV}^2\right)^{-(n_X+n_Y)/2} \exp\left\{-\frac{1}{2\hat{\sigma}_{MV}^2}\left[\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + \sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2\right]\right\} \\ &= \left(2\pi\hat{\sigma}_{MV}^2\right)^{-(n_X+n_Y)/2} \exp\left\{-\frac{n_X + n_Y}{2}\right\}. \end{aligned} \tag{4.3.8}$$

Por otro lado, suponiendo H_0 verdadera, $\mu_X = \mu_Y = \mu$, entonces las dos muestras provienen de la misma distribución, y en este caso el estimador MV de μ está dado por (2.3.5), esto es

$$\hat{\mu}_{MV} = \frac{\sum_{i=1}^{n_X} X_i + \sum_{j=1}^{n_Y} Y_j}{n_X + n_Y},$$

y

$$\hat{\sigma}^2_{0,MV} = \frac{\sum_{i=1}^{n_X}(X_i - \hat{\mu}_{MV})^2 + \sum_{j=1}^{n_Y}(Y_j - \hat{\mu}_{MV})^2}{n_X + n_Y}.$$

De donde el denominador de λ está dado por

$$\begin{aligned} & L(\hat{\mu}_{MV}, \hat{\sigma^2}_{0,MV}) \\ &= \left(2\pi\hat{\sigma}^2_{0,MV}\right)^{-(n_X+n_Y)/2} \exp\left\{-\frac{1}{2\hat{\sigma}^2_{0,MV}}\left[\sum_{i=1}^{n_X}(X_i - \hat{\mu}_{MV})^2 + \sum_{j=1}^{n_Y}(Y_j - \hat{\mu}_{MV})^2\right]\right\} \qquad (4.3.9) \\ &= \left(2\pi\hat{\sigma}^2_{0,MV}\right)^{-(n_X+n_Y)/2} \exp\left\{-\frac{n_X + n_Y}{2}\right\}. \qquad (4.3.10) \end{aligned}$$

Dado lo anterior, podemos tener que

$$\lambda = \left(\frac{\hat{\sigma}^2_{MV}}{\hat{\sigma}^2_{0,MV}}\right)^{-(n_X+n_Y)/2},$$

y podemos concluir que se rechaza H_0 para valores grandes de la estadística $\frac{\hat{\sigma}^2_{0,MV}}{\hat{\sigma}^2_{MV}}$, donde

$$\begin{aligned} \frac{\hat{\sigma}^2_{0,MV}}{\hat{\sigma}^2_{MV}} &= \frac{\sum_{i=1}^{n_X}(X_i - \hat{\mu}_{MV})^2 + \sum_{j=1}^{n_Y}(Y_j - \hat{\mu}_{MV})^2}{\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + \sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2} \\ &= \frac{\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + n_X(\bar{X} - \hat{\mu}_{MV})^2 + \sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2 + n_Y(\bar{Y} - \hat{\mu}_{MV})^2}{\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + \sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2} \\ &= 1 + \frac{n_X(\bar{X} - \hat{\mu}_{MV})^2 + n_Y(\bar{Y} - \hat{\mu}_{MV})^2}{\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + \sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2}. \qquad (4.3.11) \end{aligned}$$

Ahora, simples operaciones algebraicas muestran que

$$n_X(\bar{X} - \hat{\mu}_{MV})^2 = \frac{n_X n_Y^2(\bar{X} - \bar{Y})^2}{(n_X + n_Y)^2},$$

y

$$n_Y(\bar{Y} - \hat{\mu}_{MV})^2 = \frac{n_Y n_X^2(\bar{X} - \bar{Y})^2}{(n_X + n_Y)^2}.$$

De esta forma, reemplazando en (4.3.11), tenemos que

$$\frac{\hat{\sigma}^2_{0,MV}}{\hat{\sigma}^2_{MV}} = 1 + \frac{n_X n_Y}{n_X + n_Y}\frac{(\bar{X} - \bar{Y})^2}{\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + \sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2}$$

Entonces la regla de decisión para el sistema de interés es

$$\text{Rechazar } H_0 \text{ cuando } \frac{(\bar{X}-\bar{Y})^2}{\sum_{i=1}^{n_X}(X_i-\bar{X})^2+\sum_{j=1}^{n_Y}(Y_j-\bar{Y})^2} > K,$$

para algún $K > 0$. La cual es equivalente a

$$\text{Rechazar } H_0 \text{ cuando } \frac{\bar{X}-\bar{Y}}{\sqrt{\sum_{i=1}^{n_X}(X_i-\bar{X})^2+\sum_{j=1}^{n_Y}(Y_j-\bar{Y})^2}} > K_1 \text{ o}$$
$$\frac{\bar{X}-\bar{Y}}{\sqrt{\sum_{i=1}^{n_X}(X_i-\bar{X})^2+\sum_{j=1}^{n_Y}(Y_j-\bar{Y})^2}} < -K_1$$

para algún $K_1 > 0$.

Para encontrar el valor de K_1, se debe conocer la distribución de la estadística de prueba bajo $H_0: \mu_X = \mu_Y$, aunque esta distribución no es ninguna de las comunes, recordamos (3.2.34), y tenemos que bajo H_0

$$\frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \sim t_{n_X+n_Y-2},$$

con $S_p^2 = \frac{\sum_{i=1}^{n_X}(X_i-\bar{X})^2+\sum_{j=1}^{n_Y}(Y_j-\bar{Y})^2}{n_X+n_Y-2}$. Usando esta distribución, podemos modificar la regla de decisión encontrada anteriormente y así

$$\text{Rechazar } H_0 \text{ cuando } \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} > K_2 \text{ o } \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} < -K_2,$$

de donde usando la definición de error tipo I, se tiene que $K_2 = t_{n_X+n_Y-2,1-\alpha/2}$, y completamos la regla de decisión. Se deja como ejercicio encontrar la fórmula del p valor y la función de potencia (Ejercicio 4.9).

Ejemplo 4.3.1. *En el Ejemplo 2.3.12, se planteó el problema de comparar dos institutos A y B de capacitación en términos de calificación obtenida por sus alumnos, y se verificó que la distribución normal parece ser apropiada para describir estos datos. Si μ_A y μ_B denotan la calificación promedio de los estudiantes de los dos institutos, entonces para ver si hay diferencia significativa entre los dos institutos, planteamos el sistema $H_0: \mu_A = \mu_B$ vs. $H_1: \mu_A \neq \mu_B$; para ello, necesitamos conocer si las varianzas teóricas pueden considerarse iguales o no. Las estimaciones de las desviaciones estándares son 8.276185 y 8.55525, respectivamente, podemos ver que son bastante similares, lo cual puede ser un indicio de que las varianzas teóricas son iguales.*

Dado lo anterior, aplicamos la prueba t asumiendo igualdad entre las varianzas teóricas,

```
> A<-c(75, 87, 83, 73, 74, 88, 88, 74, 64, 92, 73, 87, 91, 83,84)
> B<-c(64, 85, 72, 64, 74, 93, 70, 79, 79, 75, 66, 83 ,74)
> t.test(A,B, var.equal=T)
```

```
        Two Sample t-test

data:  A and B
t = 1.8321, df = 26, p-value = 0.07842
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7116975 12.3834924
sample estimates:
mean of x mean of y
 81.06667  75.23077
```

Observamos que el p-valor de esta prueba es de 0.07842, entonces con un nivel de significación de 5 % no se rechaza H_0, pero si se cambia el nivel de significación de 10 % sí se rechaza H_0. El nivel de significación α es el límite superior para el error de rechazar una hipótesis verdadera, entonces al tomar un α pequeño, es más difícil rechazar H_0; de hecho, el área de la región de rechazo es más pequeña, y es más difícil que el valor de la estadística se sitúe dentro de la región de rechazo.

σ_X^2 y σ_Y^2 son desconocidos y diferentes

Cuando las varianzas de las dos poblaciones son desconocidas y además diferentes, tenemos la misma situación considerada en la sección 3.1.2, donde se introdujo la estadística D dada por (3.2.36), cuya distribución es t_k, con k dado por (3.2.37). Y la regla de decisión queda determinada como

$$\text{Rechazar } H_0 \text{ cuando } D > t_{k,1-\alpha/2} \text{ o } D < -t_{k,1-\alpha/2}.$$

4.3.2 Comparación entre dos varianzas

En la anterior sesión, se vio que para juzgar un sistema de hipótesis acerca de $\mu_X - \mu_Y$ es necesario conocer la estructura de las dos poblaciones en términos de las varianzas. Cuando éstas no son conocidas, hay que determinar si se puede asumir que sean iguales, y dependiendo de esto, se aplica la regla de decisión apropiada. Por lo anterior, es necesario considerar el siguiente sistema de hipótesis

$$H_0 : \ \sigma_X^2 = \sigma_Y^2 \quad vs. \quad H_1 : \ \sigma_X^2 \neq \sigma_Y^2, \tag{4.3.12}$$

Teniendo en cuenta que los estimadores de máxima verosimilitud de σ_X^2 y σ_Y^2 son $\frac{1}{n_X}\sum_{i=1}^{n_X}(X_i - \bar{X})^2$ y $\frac{1}{n_Y}\sum_{j=1}^{n_Y}(Y_i - \bar{Y})^2$, respectivamente, podemos proponer la siguiente regla de decisión

$$\text{Rechazar } H_0 \text{ cuando } \frac{\frac{1}{n_X}\sum_{i=1}^{n_X}(X_i-\bar{X})^2}{\frac{1}{n_Y}\sum_{j=1}^{n_Y}(Y_i-\bar{Y})^2} > K_1 \text{ o } \frac{\frac{1}{n_X}\sum_{i=1}^{n_X}(X_i-\bar{X})^2}{\frac{1}{n_Y}\sum_{j=1}^{n_Y}(Y_i-\bar{Y})^2} < K_2$$

para constantes $K_1 > 1$ y $K_2 < 1$. Para encontrar los valores de K_1 y K_2, recordemos, en primer lugar, la siguiente distribución

$$\frac{\sigma_Y^2 S_X^2}{\sigma_X^2 S_Y^2} = \frac{\sigma_Y^2 (n_Y - 1)\sum_{i=1}^{n_X}(X_i - \bar{X})^2}{\sigma_X^2 (n_X - 1)\sum_{j=1}^{n_Y}(Y_i - \bar{Y})^2} \sim F_{n_Y-1}^{n_X-1},$$

que, bajo H_0, se convierte en

$$\frac{S_{X,n_X-1}^2}{S_{Y,n_Y-1}^2} = \frac{(n_Y - 1)\sum_{i=1}^{n_X}(X_i - \bar{X})^2}{(n_X - 1)\sum_{j=1}^{n_Y}(Y_i - \bar{Y})^2} \sim F_{n_Y-1}^{n_X-1}, \tag{4.3.13}$$

Y usando la definición del error tipo I, tenemos que

$$\begin{aligned}\alpha &= Pr\left(\frac{n_Y \sum_{i=1}^{n_X}(X_i - \bar{X})^2}{n_X \sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2} > K_1\right) + Pr\left(\frac{n_Y \sum_{i=1}^{n_X}(X_i - \bar{X})^2}{n_X \sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2} < K_2\right) \\ &= Pr\left(\frac{S_{X,n_X-1}^2}{S_{Y,n_Y-1}^2} > \frac{(n_Y - 1)n_X K_1}{(n_X - 1)n_Y}\right) + Pr\left(\frac{S_{X,n_X-1}^2}{S_{Y,n_Y-1}^2} < \frac{(n_Y - 1)n_X K_2}{(n_X - 1)n_Y}\right)\end{aligned}$$

asumiendo que H_0 es verdadera. Recurriendo a (4.3.13), se tiene que $\frac{(n_Y-1)n_X K_1}{(n_X-1)n_Y}$ y $\frac{(n_Y-1)n_X K_2}{(n_X-1)n_Y}$ son percentiles de la distribución $F_{n_Y-1}^{n_X-1}$. Por facilidad, podemos escoger $\frac{(n_Y-1)n_X K_1}{(n_X-1)n_Y} = f_{n_y-1,1-\alpha/2}^{n_X-1}$ y $\frac{(n_Y-1)n_X K_2}{(n_X-1)n_Y} = f_{n_y-1,\alpha/2}^{n_X-1}$. De esta forma, se tiene la siguiente regla de decisión:

$$\text{Rechazar } H_0 \text{ cuando } \frac{\frac{1}{n_X}\sum_{i=1}^{n_X}(X_i - \bar{X})^2}{\frac{1}{n_Y}\sum_{j=1}^{n_Y}(Y_i - \bar{Y})^2} > \frac{(n_X - 1)n_Y f_{n_y-1,1-\alpha/2}^{n_X-1}}{(n_Y - 1)n_X} \text{ o}$$
$$\frac{\frac{1}{n_X}\sum_{i=1}^{n_X}(X_i - \bar{X})^2}{\frac{1}{n_Y}\sum_{j=1}^{n_Y}(Y_i - \bar{Y})^2} > \frac{(n_X - 1)n_Y f_{n_y-1,\alpha/2}^{n_X-1}}{(n_Y - 1)n_X},$$

la cual es equivalente a

$$\text{Rechazar } H_0 \text{ cuando } \frac{S_{X,n_X-1}^2}{S_{Y,n_Y-1}^2} > f_{n_y-1,1-\alpha/2}^{n_X-1} \text{ ó } \frac{S_{X,n_X-1}^2}{S_{Y,n_Y-1}^2} > f_{n_y-1,\alpha/2}^{n_X-1}.$$

En R, la función `var.test` lleva a cabo el procedimiento.

p valor

Ahora, consideramos el p valor asociado a la anterior regla de decisión, cuya estadística de prueba es $\frac{S_{X,n_X-1}^2}{S_{Y,n_Y-1}^2}$. Suponga que se observa una muestra aleatoria, el valor que toma la estadística puede ser mayor o menor que el percentil 0.5 de la distribución $f_{n_y-1}^{n_X-1}$ como lo ilustra la Figura 4.19. Si el valor de la estadística es mayor que $f_{n_y-1,0.5}^{n_X-1}$ (el valor v_1 en la figura), se rechaza H_0 cuando $Pr(F_{n_Y-1}^{n_X-1} > v_1) < \alpha/2$, donde $F_{n_Y-1}^{n_X-1}$

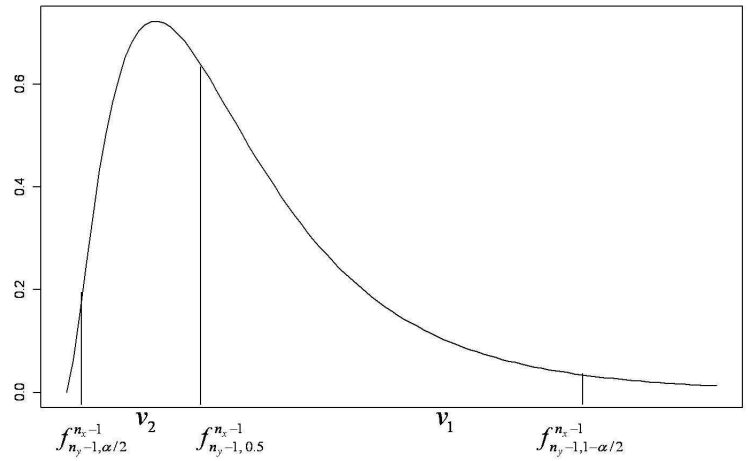

Figura 4.19: *Ilustración del p valor para la hipótesis (4.3.12).*

denota una variable aleatoria con distribución f^{nX-1}_{nY-1}; por otro lado, si el valor de la estadística es menor que $f^{nX-1}_{n_y-1,0.5}$, (el valor v_2 en la figura), se rechaza H_0 cuando $Pr(F^{nX-1}_{nY-1} < v_1) < \alpha/2$

Con lo anterior, podemos obtener la siguiente forma de calcular el p valor.

$$p \text{ valor} = \begin{cases} 2Pr(F^{nX-1}_{nY-1} > v), & \text{para } v > f^{nX-1}_{n_y-1,0.5} \text{ ;} \\ 2Pr(F^{nX-1}_{nY-1} < v), & \text{para } v < f^{nX-1}_{n_y-1,0.5}. \end{cases}$$

El siguiente código de R nos permite calcular el p valor para el sistema de hipótesis (4.3.12) para dos muestras.

```
> p.val<-function(x,y){
+ nx<-length(x)
+ ny<-length(y)
+ if(var(x)/var(y)>=qf(0.5,nx-1,ny-1)){
+ p.val<-2*(1-pf(var(x)/var(y),nx-1,ny-1))
+ }
+ if(var(x)/var(y)<qf(0.5,nx-1,ny-1)){
+ p.val<-2*(pf(var(x)/var(y),nx-1,ny-1))}
+ p.val
+ }
```

Ejemplo 4.3.2. *En el Ejemplo 4.3.1, donde se compararon dos institutos de capacitación en términos de calificación obtenida por sus alumnos, y para juzgar el sistema $H_0 : \mu_A = \mu_B$ vs. $H_1 : \mu_A \neq \mu_B$, se hizo el supuesto de que las varianzas teóricas son iguales teniendo en cuenta las estimaciones muestrales. Ahora, efectuamos la prueba F para verificar la validez de este supuesto usando la anterior función para calcular el p-valor.*

```
> A<-c(75, 87, 83, 73, 74, 88, 88, 74, 64, 92, 73, 87, 91, 83,84)
> B<-c(64, 85, 72, 64, 74, 93, 70, 79, 79, 75, 66, 83 ,74)
> p.val(A,B)
[1] 0.8954233
```

Observamos que el p-valor es grande comparado con cualquier valor común en la práctica de α, de donde podemos afirmar que $\sigma_A^2 = \sigma_B^2$ es un supuesto razonable para los datos.

En R, la función `var.test` *lleva a cabo esta prueba F, y también calcula un intervalo de confianza para la cociente de varianzas σ_A^2/σ_B^2.*

```
> var.test(A,B)

        F test to compare two variances

data:  A and B
F = 0.9358, num df = 14, denom df = 12, p-value = 0.8954
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2918789 2.8544131
sample estimates:
ratio of variances
         0.9358256
```

Función de potencia

La función de potencia para la anterior regla de decisión puede ser calculada como

$$\begin{aligned}
&\beta(\sigma_X^2, \sigma_Y^2) \\
&= Pr\left(\frac{S_{X,n_X-1}^2}{S_{Y,n_Y-1}^2} > f_{n_y-1,1-\alpha/2}^{n_X-1}\right) + Pr\left(\frac{S_{X,n_X-1}^2}{S_{Y,n_Y-1}^2} > f_{n_y-1,\alpha/2}^{n_X-1}\right) \\
&= Pr\left(\frac{\sigma_Y^2 S_{X,n_X-1}^2}{\sigma_X^2 S_{Y,n_Y-1}^2} > \frac{\sigma_X^2}{\sigma_Y^2} f_{n_y-1,1-\alpha/2}^{n_X-1}\right) + Pr\left(\sigma_Y^2 \frac{S_{X,n_X-1}^2}{\sigma_X^2 S_{Y,n_Y-1}^2} > \frac{\sigma_X^2}{\sigma_Y^2} f_{n_y-1,\alpha/2}^{n_X-1}\right) \\
&= 1 - F_{n_X-1,n_Y-1}\left(\frac{\sigma_X^2}{\sigma_Y^2} f_{n_y-1,1-\alpha/2}^{n_X-1}\right) + F_{n_X-1,n_Y-1}\left(\frac{\sigma_X^2}{\sigma_Y^2} f_{n_y-1,\alpha/2}^{n_X-1}\right) \quad (4.3.14)
\end{aligned}$$

donde $F_{n_X-1,n_Y-1}(\cdot)$ denota la función de distribución de la distribución $F_{n_Y-1}^{n_X-1}$. Nótese que inicialmente la anterior función de potencia también es una función de dos argumentos que depende de σ_X^2 y σ_Y^2; sin embargo, $\beta(\sigma_X^2, \sigma_Y^2)$ depende de estos parámetros solo a través de la cociente σ_X^2/σ_Y^2, por lo tanto, graficamos esta función en función de σ_X^2/σ_Y^2.

El siguiente código grafica $\beta(\sigma_X^2, \sigma_Y^2)$ para diferentes tamaños de muestra.

```
> pote_sigX_sigY<-function(cociente,nx,ny,alpha){
+ 1-pf(cociente*qf(1-alpha/2,nx-1,ny-1),nx-1,ny-1)+
+ pf(cociente*qf(alpha/2,nx-1,ny-1),nx-1,ny-1)
+ }
> alpha<-0.05
> n1<-10
> n2<-30
> n3<-50
>
> plot(function(x) pote_sigX_sigY(x,n3,n3,alpha),0,10,type="l",
+ xlab="sig_X/sig_Y",ylab="función de potencia")
> curve(pote_sigX_sigY(x,n2,n2,alpha),0,10,lty=2,add=T)
> curve(pote_sigX_sigY(x,n1,n1,alpha),0,10,lty=3,add=T)
> legend(6,0.4,c("nx=ny=10","nx=ny=30","nx=ny=50"),lty=c(3,2,1))
```

En la Figura 4.20 se muestra esta función de potencia para diferentes tamaños de muestra con $n_X = n_Y$ y se observa un comportamiento similar a la función de potencia del sistema $H_0: \ \sigma^2 = \sigma_0^2$ vs. $H_1: \ \sigma^2 \neq \sigma_0^2$.

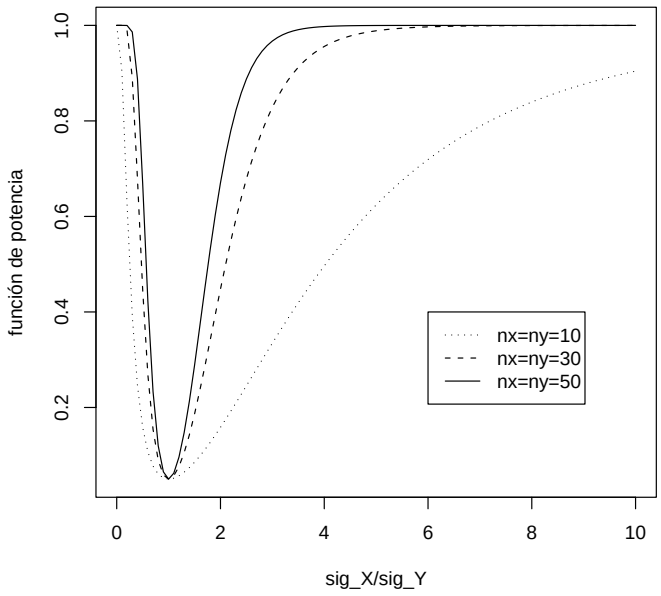

Figura 4.20: *Función de potencia (4.3.14) para diferentes tamaños de muestra con* $\alpha = 0.05$.

4.4 k muestras

Suponga que se dispone de k muestras independientes, donde la i ésima muestra de tamaño n_i se denota por X_1^i, ..., $X_{n_i}^i$. Suponga además que las muestras provienen de distribución $N(\mu_i, \sigma_i^2)$ para $i = 1, \ldots, k$.

4.4.1 Igualdad de k medias

El sistema de interés es

$$H_0: \ \mu_1 = \ldots = \mu_k \quad vs. \quad H_1: \text{ existen por lo menos dos medias diferentes.} \tag{4.4.1}$$

Lo anterior es una generalización del problema de dos muestras estudiado anteriormente, donde se vio que dependiendo de las varianzas teóricas, la regla de decisión cambia según si éstas son conocidas o no. En el caso de k muestras, hay k varianzas teóricas, y puede haber un gran número de casos que dificultan el desarrollo teórico respectivo. Por esta razón, suponemos que las k varianzas teóricas son iguales, esto es, $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2$. Bajo este supuesto, la función de verosimilitud de las k muestras está dada por

$$L(\mu_1, \ldots, \mu_k, \sigma^2) = (2\pi\sigma^2)^{-\sum_{i=1}^k n_i/2} \exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^k \sum_{j=1}^{n_i} (X_j^i - \mu_i)^2 \right] \right\} \tag{4.4.2}$$

Y desarrollamos la prueba de razón generalizada de verosimilitudes como sigue.

En primer lugar, los estimadores de máxima verosimilitud de μ_1, ..., μ_k y la varianza común σ^2 están dados por

$$\hat{\mu}_{i,MV} = \bar{X}^i \tag{4.4.3}$$

para $i = 1, \ldots, k$, y

$$\hat{\sigma}_{MV}^2 = \frac{n_1 S_{1,n_1}^2 + \ldots + n_k S_{k,n_k}^2}{n_1 + \ldots + n_k} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (X_j^i - \bar{X}^i)^2}{\sum_{i=1}^k n_i} \tag{4.4.4}$$

donde $\bar{X}^i$ y S_{i,n_i}^2 denotan el promedio muestral y la varianza muestral (dividiendo sobre n_i) de la i ésima muestra.

De esta forma, el numerador de la razón generalizada de verosimilitudes está dado por

$$L(\hat{\mu}_{1,MV},\ldots,\hat{\mu}_{k,MV},\hat{\sigma}^2_{MV}) = (2\pi\hat{\sigma}^2_{MV})^{-\sum_{i=1}^k n_i/2}\exp\left\{-\frac{\sum_{i=1}^k n_i}{2}\right\},$$

similar a la expresión (4.3.8) obtenida para el caso de dos muestras.

Ahora, bajo H_0, las k medias teóricas son iguales, y las k muestras provienen de una misma distribución $N(\mu,\sigma^2)$. En este caso el estimador de máxima verosimilitud de la media común μ está dado por el promedio de las k muestras, esto es,

$$\hat{\mu}_{0,MV} = \frac{n_1\bar{X}^1+\ldots+n_k\bar{X}^k}{n_1+\ldots+n_k} = \frac{\sum_{i=1}^k n_i\bar{X}^i}{\sum_{i=1}^k n_i} = \frac{\sum_{i=1}^k\sum_{j=1}^{n_i} X^i_j}{\sum_{i=1}^k n_i}, \tag{4.4.5}$$

y el estimador de la varianza común σ^2 está dado por

$$\hat{\sigma}^2_{0,MV} = \frac{\sum_{i=1}^k\sum_{j=1}^{n_i}(X^i_j-\hat{\mu}_{0,MV})^2}{\sum_{i=1}^k n_i}, \tag{4.4.6}$$

De esta forma, el denominador de la razón generalizada de verosimilitudes está dado por

$$L(\hat{\mu}_{MV},\hat{\sigma}^2_{0,MV}) = (2\pi\hat{\sigma}^2_{0,MV})^{-\sum_{i=1}^k n_i/2}\exp\left\{-\frac{\sum_{i=1}^k n_i}{2}\right\}.$$

De esta forma, tenemos que la razón generalizada de verosimilitudes está dada por

$$\lambda = \left(\frac{\hat{\sigma}^2_{MV}}{\hat{\sigma}^2_{0,MV}}\right)^{-\sum_{i=1}^k n_i/2}$$

y podemos concluir que se rechaza H_0 para valores grandes de la estadística $\frac{\hat{\sigma}^2_{0,MV}}{\hat{\sigma}^2_{MV}}$, con

$$\begin{aligned}\frac{\hat{\sigma}^2_{0,MV}}{\hat{\sigma}^2_{MV}} &= \frac{\sum_{i=1}^k\sum_{j=1}^{n_i}(X^i_j-\hat{\mu}_{0,MV})^2}{\sum_{i=1}^k\sum_{j=1}^{n_i}(X^i_j-\bar{X}^i)^2}\\ &= \frac{\sum_{i=1}^k\sum_{j=1}^{n_i}(X^i_j-\bar{X}^i)^2+\sum_{i=1}^k n_i(\bar{X}^i-\hat{\mu}_{0,MV})^2}{\sum_{i=1}^k\sum_{j=1}^{n_i}(X^i_j-\bar{X}^i)^2}\\ &= 1+\frac{\sum_{i=1}^k n_i(\bar{X}^i-\hat{\mu}_{0,MV})^2}{\sum_{i=1}^k\sum_{j=1}^{n_i}(X^i_j-\bar{X}^i)^2}.\end{aligned}$$

Y podemos afirmar que se debe rechazar H_0 cuando

$$\frac{\sum_{i=1}^k n_i(\bar{X}^i-\hat{\mu}_{0,MV})^2}{\sum_{i=1}^k\sum_{j=1}^{n_i}(X^i_j-\bar{X}^i)^2} > K, \tag{4.4.7}$$

para algún $K > 0$. De nuevo, para encontrar el valor de K, se debe conocer la distribución de la estadística de prueba bajo la hipótesis nula. Para eso, en primer lugar tengamos en cuenta que la distribución de $\sum_{j=1}^{n_i}(X_j^i - \bar{X}^i)^2/\sigma^2$ es la distribución $\chi^2_{n_i-1}$, y usando la independencia de las k muestras, se tiene que

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(X_j^i - \bar{X}^i)^2}{\sigma^2} \sim \chi^2_{\sum_{i=1}^k n_i - k}. \tag{4.4.8}$$

Ahora, consideramos el numerador de la estadística de prueba en (4.4.7), tenemos que

$$\begin{aligned}\sum_{i=1}^{k}\frac{n_i(\bar{X}^i-\mu)^2}{\sigma^2} &= \sum_{i=1}^{k}\frac{(\sqrt{n_i}\bar{X}^i - \sqrt{n_i}\hat{\mu}_{0,MV} + \sqrt{n_i}\hat{\mu}_{0,MV} - \sqrt{n_i}\mu)^2}{\sigma^2}\\ &= \sum_{i=1}^{k}\frac{(\sqrt{n_i}\bar{X}^i - \sqrt{n_i}\hat{\mu}_{0,MV})^2}{\sigma^2} + \sum_{i=1}^{k}\frac{(\sqrt{n_i}\hat{\mu}_{0,MV} - \sqrt{n_i}\mu)^2}{\sigma^2}\\ &= \sum_{i=1}^{k}\frac{(\sqrt{n_i}\bar{X}^i - \sqrt{n_i}\hat{\mu}_{0,MV})^2}{\sigma^2} + \frac{\sum_{i=1}^k n_i(\hat{\mu}_{0,MV}-\mu)^2}{\sigma^2}\end{aligned} \tag{4.4.9}$$

Bajo la hipótesis nula, se tiene que $\bar{X}^i \sim N(\mu, \sigma^2/n_i)$, de donde se tiene que $n_i(\bar{X}^i-\mu)^2/\sigma^2 \sim \chi^2_1$, y usando la independencia de las k muestras, se tiene que bajo la hipótesis nula

$$\sum_{i=1}^{k}\frac{n_i(\bar{X}^i-\mu)^2}{\sigma^2} \sim \chi^2_k.$$

Por otro lado, bajo H_0, $\hat{\mu}_{0,MV}$ es el promedio de las k muestras provenientes de una distribución $N(\mu,\sigma^2)$, entonces se tiene que $\hat{\mu}_{0,MV} \sim N(\mu, \sigma^2/\sum_{i=1}^k n_i)$, entonces

$$\frac{\hat{\mu}_{0,MV}-\mu}{\sqrt{\dfrac{\sigma^2}{\sum_{i=1}^k n_i}}} \sim N(0,1), \tag{4.4.10}$$

de donde

$$\frac{\sum_{i=1}^k n_i(\hat{\mu}_{0,MV}-\mu)^2}{\sigma^2} \sim \chi^2_1. \tag{4.4.11}$$

Usando las distribuciones (4.4.10), (4.4.11), y la identidad (4.4.9), podemos concluir que bajo H_0,

$$\sum_{i=1}^{k}\frac{(\sqrt{n_i}\bar{X}^i - \sqrt{n_i}\hat{\mu}_{0,MV})^2}{\sigma^2} = \sum_{i=1}^{k}\frac{n_i(\bar{X}^i - \hat{\mu}_{0,MV})^2}{\sigma^2} \sim \chi^2_{k-1}. \tag{4.4.12}$$

Usando las distribuciones (4.4.8) (4.4.12) y la independencia de las dos estadísticas, se tiene que

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{X}^i - \hat{\mu}_{0,MV})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_j^i - \bar{X}^i)^2/(\sum_{i=1}^{k} n_i - k)} \sim f^{k-1}_{\sum_{i=1}^{k} n_i - k}.$$

Ahora, retomando la regla de decisión encontrada anteriormente (4.4.7), usando nuevamente la definición del error tipo I, se tiene que $K = f^{k-1}_{\sum_{i=1}^{k} n_i - k, 1-\alpha}$, y la regla de decisión finalmente está dada por

$$\text{Rechazar } H_0 \text{ cuando } F = \frac{\sum_{i=1}^{k} n_i(\bar{X}^i - \hat{\mu}_{0,MV})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_j^i - \bar{X}^i)^2/(\sum_{i=1}^{k} n_i - k)} > f^{k-1}_{\sum_{i=1}^{k} n_i - k, 1-\alpha}.$$

Dada la anterior regla de decisión, el p-valor se calcula como $p-\text{valor} = Pr(F > v)$ donde $F \sim f^{k-1}_{\sum_{i=1}^{k} n_i - k}$ y v denota el valor observado de la estadística F.

Ejemplo 4.4.1. *Suponga que se quiere comparar tres marcas de carros con respecto al rendimiento en términos de la distancia recorrida por galón de tres marcas de automóviles en referencia a especificaciones similares bajo circunstancias similares con respecto a la carretera, clima y demás condiciones controlables por los técnicos y expertos automovilísticos. Los datos se muestran en la Tabla 4.2 y para ver si hay diferencia entre las tres marcas de automóviles, probamos el sistema $H_0: \mu_A = \mu_B = \mu_C$ frente a la alternativa de que por lo menos dos medias son diferentes. Las estimaciones de las desviaciones estándares son 2.3576, 1.663 y 3.466, respectivamente. Por ahora, asumimos que las varianzas teóricas son iguales, y más adelante comprabaremos la validez de este supuesto usando una prueba estadística.*

	Distancia recorrida (en Km) por galón
Marca A	39.4, 41.1, 39.5, 40.0, 43.7, 46.0, 43.5, 42.1
Marca B	42.7, 39.2, 41.2, 40.7, 37.4, 40.0, 40.7
Marca C	52.6, 49.4, 49.4, 46.4, 51.2, 49.2, 55.0, 53.6, 55.7, 57.4

Tabla 4.2: *Datos usados en el Ejemplo 4.4.1: kilometro recorrido por un galón de gasolina en tres marcas de automóviles.*

Los siguientes comandos de R calculan el valor de estadística F, el percentil $f^{k-1}_{\sum_{i=1}^{k} n_i - k, 1-\alpha}$ y el p-valor.

```
> A<-c(39.4, 41.1, 39.5, 40.0, 43.7, 46.0, 43.5, 42.1)
> B<-c(42.7, 39.2, 41.2, 40.7, 37.4, 40.0, 40.7)
> C<-c(52.6, 49.4, 49.4, 46.4, 51.2, 49.2, 55.0, 53.6, 55.7, 57.4)
> k<-3
> alpha<-0.05
> n1<-length(A)
> n2<-length(B)
> n3<-length(C)
```

```
> mu.comun<-mean(c(A,B,C))
> mu1<-mean(A)
> mu2<-mean(B)
> mu3<-mean(C)
> f1<-(n1*(mu1-mu.comun)^2+n2*(mu2-mu.comun)^2+n3*(mu3-mu.comun)^2)/(k-1)
> f2<-(var(A)*(n1-1)+var(B)*(n2-1)+var(C)*(n3-1))/(n1+n2+n3-k)
> estad<-f1/f2
> p.val<-pf(estad,k-1,n1+n2+n3-k,lower.tail=F)
> estad
[1] 48.10022
> qf(1-alpha,k-1,n1+n2+n3-k)
[1] 3.443357
> p.val
[1] 9.286228e-09
```

Podemos ver que el p-valor es pequeño si utilizamos un nivel de significación del 5 %, de donde concluimos que el rendimiento de las tres marcas de carros no es igual. Pero eso no implica que las tres medias μ_A, μ_B y μ_C son todas diferentes entre ellos, sino que por lo menos hay una media diferente. Entonces para detectar cuál marca de automóviles tine un rendimiento sustancialmente diferente, debemos probar por separado las hipótesis $\mu_A = \mu_B$, $\mu_A = \mu_C$ y $\mu_B = \mu_C$. Para eso, usamos `t.test` *para cada una de estas hipótesis. Los resultados se encuentran en la Tabla 4.3, y allí observamos que las marcas A y B tienen rendimientos similares, mientras que la marca C es muy diferente de las marcas A y B.*

Hipótesis	Estimaciones		Estadística	p-valor
$\mu_A = \mu_B$	$\hat{\mu}_A = 41.91$	$\hat{\mu}_B = 40.27$	1.5346	0.1489
$\mu_A = \mu_C$	$\hat{\mu}_A = 41.91$	$\hat{\mu}_C = 51.99$	-7.0082	2.953e-06
$\mu_B = \mu_C$	$\hat{\mu}_B = 40.27$	$\hat{\mu}_C = 51.99$	-8.2465	5.914e-07

Tabla 4.3: *Prueba de igualdad de dos medias del Ejemplo 4.4.1.*

4.4.2 Igualdad de varianzas

El sistema de interés es

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2 \quad vs. \quad H_1 : \text{existen por lo menos dos varianzas diferentes.} \tag{4.4.13}$$

Nuevamente, hacemos uso de la prueba de la razón generalizada de verosimilitudes. La función de verosimilitud de las k muestras está dada por (4.4.2). Bajo la hipótesis nula, $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2$, se ha visto anteriormente que los estimadores de máxima verosimilitud de σ^2, μ_1, ..., μ_k están dados por (4.4.3) y (4.4.4). Por otro lado, los estimadores de máxima verosimilitud de μ_i y σ_i^2 están dados por las respectivas medias y varianzas muestrales (dividiendo por $n_i - 1$) de la i-ésima muestra para $i = 1, \ldots, k$.

De esta forma, la razón generalizada de verosimilitudes está dada por

$$\lambda = \frac{\prod_{i=1}^{n}(\hat{\sigma}^2_{i,MV})^{-n_i/2}}{(\hat{\sigma}^2_{MV})^{-\sum_{i=1}^{k} n_i/2}},$$

y se tiene que bajo H_0, $2\ln\lambda$ se distribuye aproximadamente como χ^2_{k-1} (Bickel & Doksum 2001, p. 394). Y por consiguiente, se rechaza H_0 cuando $\sum_{i=1}^{k} n_i \ln\hat{\sigma}^2_{MV} - \sum_{i=1}^{k} n_i \ln\hat{\sigma}^2_{i,MV} > \chi^2_{k-1,1-\alpha}$.

Bartlett (1937) hizo una modificación a la anterior estadística de prueba con el fin de que la distribución de la estadística de prueba se acercara más a la distribución χ^2_{k-1}. La modificación de Bartlett consiste en reemplazar los estimadores de máxima verosimilitud por los estimadores insesgados, reemplazar n_i por $n_i - 1$ y dividir la estadística por la constante c dada por

$$c = 1 + \left(\frac{1}{3(k-1)}\right)\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{k} n_i - k}\right).$$

La estadística de prueba queda entonces expresada como

$$A = \frac{1}{c}\left\{(\sum_{i=1}^{k} n_i - k)\ln S^2 - \sum_{i=1}^{k}(n_i - 1)\ln S_i^2\right\}$$

donde S_i^2 es el estimador insesgado de σ_i^2 en la i-ésima muestra, y $S^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2/(\sum_{i=1}^{k} n_i - k)$. Y se rechaza H_0 cuando $A > \chi^2_{k-1,1-\alpha}$.

Ejemplo 4.4.2. *En el Ejemplo 4.4.1 se probó la hipótesis de igualdad de tres medias teóricas para comparar tres marcas de automóviles en términos de la distancia recorrida por un galón de gasolina. En ese ejemplo se hizo la suposición de que las tres varianzas teóricas son iguales. A continuación, verificamos la validez de este supuesto usando la teoría desarrollada anteriormente y el siguiente código.*

```
> A<-c(39.4, 41.1, 39.5, 40.0, 43.7, 46.0, 43.5, 42.1)
> B<-c(42.7, 39.2, 41.2, 40.7, 37.4, 40.0, 40.7)
> C<-c(52.6, 49.4, 49.4, 46.4, 51.2, 49.2, 55.0, 53.6, 55.7, 57.4)
> k<-3
> alpha<-0.05
> n1<-length(A)
> n2<-length(B)
> n3<-length(C)
> S1<-var(A)
> S2<-var(B)
> S3<-var(C)
> S<-(S1*(n1-1)+S2*(n2-1)+S3*(n3-1))/(n1+n2+n3-k)
>
```

```
> cons.c<-1+(1/(n1-1)+1/(n2-1)+1/(n3-1)-1/(n1+n2+n3-k))/(3*(k-1))
> estad<-(log(S)*(n1+n2+n3-k)-(n1-1)*log(S1)-(n2-1)*log(S2)-
+ (n3-1)*log(S3))/cons.c
> p.val<-pchisq(estad,k-1,lower.tail=F)
> estad
[1] 3.443527
> qchisq(1-alpha,k-1)
[1] 5.991465
> p.val
[1] 0.1787507
```

De donde vemos que p-valor sugiere que el supuesto de igualdad de varianzas es adecuado con base en los datos observados. Esta anterior prueba también se puede llevar a cabo usando la función `bartlett.test`*, pero se debe crear un vector de indicadores para etiquetar cada dato según a qué marca corresponde. El uso de esta función se ilustra a continuación.*

```
> ind<-c(rep("A",n1),rep("B",n2),rep("C",n3))
> ind
 [1] "A" "A" "A" "A" "A" "A" "A" "A" "B" "B" "B" "B" "B"
[14] "B" "B" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C"
> dato<-c(A,B,C)
> bartlett.test(dato,ind)

        Bartlett test of homogeneity of variances

data:  dato and ind
Bartlett's K-squared = 3.4435, df = 2, p-value = 0.1788
```

4.5 Muestras provenientes de la distribución Bernoulli y binomial

4.5.1 Una muestra

Suponga que la muestra aleatoria $X_1, \ldots, X_n$ constituye una muestra aleatoria proveniente de la distribución Bernoulli con probabilidad de éxito p. Esta distribución es muy útil en ciencias como la medicina donde p puede ser como probabilidad de contagiar alguna enfermedad, o probabilidad de éxito en cirugías complicadas como por ejemplo, trasplante de corazón; también en la investigación de mercados, para estudiar la preferencia de los clientes entre dos marcas.

En general, planteamos el sistema de hipótesis acerca de la probabilidad p

$$H_0 : \ p = p_0 \quad vs. \quad H_1 : \ p \neq p_0, \tag{4.5.1}$$

Una de las prueba más sencillas hace uso del teorema límite central, que en el caso de una variable X con distribución $Bin(n, p)$, se tiene que

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \rightarrow_D N(0,1),$$

de donde

$$Z^2 = \frac{(X - np_0)^2}{np_0(1-p_0)} \sim_{aprox} \chi^2_1,$$

bajo la hipótesis $H_0 : \ p = p_0$. Podemos ver que la estadística Z^2 mide la ≪diferencia≫ entre la variable X y su esperanza bajo H_0. De esta forma, un valor grande de Z^2 indica que X está muy lejos de su esperanza suponiendo H_0 cierta, y esto nos conduce a la decisión de rechazar H_0. Dado lo anterior, podemos rechazar H_0 si el valor de Z^2 es mayor que el percentil $\chi^2_{1,1-\alpha}$, y el p-valor se puede calcular como $Pr(Z^2 > v)$ donde v es el valor observado de la estadística Z^2.

Ejemplo 4.5.1. *Suponga que se quiere conocer la probabilidad de que una clienta prefiera la marca de shampoo ≪LIZ≫ ante otras marcas de shampoo. Si en 30 clientas que compraron shampoo, 7 compraron la marca ≪LIZ≫, entonces qué podemos concluir acerca de la sospecha de que la probabilidad de que una clienta prefiera la marca de shampoo ≪LIZ≫ sea de 0.5. En este caso nuestro sistema de interés es* $H_0 : \ p = 0.5$ *vs.* $H_1 : \ p \neq 0.5$. *El cómputo de la estadística* Z^2 *y el* p *valor se puede calcular como*

```
> x<-7
> n<-30
> alpha<-0.05
> Z2<-(x-n*p)^2/(n*p*(1-p))
> Z2
[1] 8.533333
> p.val<-pchisq( 8.533333,1,lower.tail=F)
> p.val
[1] 0.003487006
```

De donde observamos que el p *valor es muy pequeño, indicando que los datos sugieren que 0.5 no es un valor aceptable para la probabilidad* p. *El anterior cálculo también se puede realizar usando la instrucción* `prop.test` *de la siguiente forma. Nótese que el valor de la estadística y el* p *valor coinciden con los obtenidos anteriormente. Sin embargo, la instrucción* `prop.test` *nos provee una estimación por intervalo de confianza.*

```
> prop.test(7,30,0.5,correct=F)

        1-sample proportions test without continuity correction

data:  7 out of 30, null probability 0.5
X-squared = 8.5333, df = 1, p-value = 0.003487
```

```
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.1179239 0.4092833
sample estimates:
        p
0.2333333
```

En el anterior ejemplo cuando se utilizó la función `prop.test` se agregó la opción de `correct=F`; al no poner esta opción, se agrega la corrección de continuidad de Yates, debido a que al utilizar el teorema límite central, estamos aproximando una distribución discreta con una distribución normal. Con esta corrección de continuidad, la estadística de prueba se calcula como

$$Z_c^2 = \begin{cases} \dfrac{(X - np_0)^2}{np_0(1-p_0)} & \text{si } x = np_0 \\ \dfrac{(X - 0.5 - np_0)^2}{np_0(1-p_0)} & \text{si } x > np_0 \\ \dfrac{(X + 0.5 - np_0)^2}{np_0(1-p_0)} & \text{si } x < np_0 \end{cases}$$

Podemos ver que la corrección de continuidad de Yates consiste en acercar el valor observado x 0.5 unidad hacia el valor esperado bajo la hipótesis nula np_0, en caso de que x difiera de éste. Con esta corrección, podemos ver que, excepto $x = np_0$, el valor de Z_c^2 siempre será menor que el de Z^2, y como se rechaza H_0 para valores grandes de estas estadísticas, podemos afirmar que al utilizar la corrección de Yates es más difícil rechazar H_0, es decir, se requieren evidencias realmente fuertes en los datos para poder rechazar H_0. Si aplicamos esta corrección de continuidad a los datos del Ejemplo 4.5.1, tenemos que

```
> prop.test(7,30,0.5)

        1-sample proportions test with continuity correction

data:  7 out of 30, null probability 0.5
X-squared = 7.5, df = 1, p-value = 0.00617
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.1063502 0.4270023
sample estimates:
        p
0.2333333
```

Podemos ver que efectivamente la estadística de prueba ahora es más pequeña, además el intervalo de confianza ahora es más ancho, de donde también podemos ver que con la corrección de continuidad es más fácil aceptar H_0.

Otra prueba para el sistema (4.5.1) es utilizar directamente la distribución de la variable X y no recurrir a aproximaciones como el teorema límite central. Sin embargo,

el desarrollo de esta prueba es más fácil de entender partiendo desde el punto de vista del p valor que iniciando una búsqueda de una regla de decisión. Ilustramos este proceso dentro del contexto del Ejemplo 4.5.1.

Un razonamiento natural es calcular la estimación de esta probabilidad que en este caso es $\hat{p} = 7/30 = 0.233$ que es diferente del valor 0.7. Sin embargo, una variable con distribución $Bin(30, 0.5)$ también puede tomar valor de 7, que fue lo observado en la muestra, en vez de tomar el valor esperado $30 * 0.5 = 15$. ¿Pero qué tan probable es que eso ocurra? Si $X \sim Bin(30, 0.5)$, entonces podemos calcular la probabilidad de que el valor de X difiera de su valor esperado por lo menos $15 - 7 = 8$, esto es

$$\begin{aligned} Pr(|X - 15| \geq 8) &= Pr(X \geq 23) + Pr(X \leq 7) \\ &= 0.005222879, \end{aligned}$$

lo cual es una probabilidad bastante pequeña, indicando que si $p = 0.5$, es muy improbable observar $X = 7$ en la muestra. Lo anterior calcula realmente la probabilidad de que se observen valores más extremos que el observado, y esta es precisamente la definición o la interpretación que se da al p-valor, y usando el p-valor podemos tomar una decisión acerca de $p = p_0$. Este proceso se conoce como la prueba binomial exacta y en R se lleva a cabo usando el comando `binom.test`. Para los datos de preferencia del shampoo ≪LIZ≫, la instrucción y el correspondiente resultado es:

```
> binom.test(7,30,0.5)

        Exact binomial test

data:  7 and 30
number of successes = 7, number of trials = 30, p-value = 0.005223
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.09933786 0.42283652
sample estimates:
probability of success
             0.2333333
```

El p-valor calculado en R es 0.005223, y podemos ver que es el mismo resultado obtenido anteriormente calculando a mano.

Ahora, generalizamos la forma de calcular el p valor para el sistema de hipótesis (4.5.1). Suponga que se observa la realización x de una variable aleatoria X con distribución $Bin(n, p)$. Según la ilustración al principio del capítulo, el p valor se calcula como la probabilidad de que la variable X se difiera de su esperanza bajo H_0 por más de la diferencia observada en la muestra. Recordando que la esperanza de X bajo H_0 es np_0, tenemos que

$$p - \text{valor} = Pr(|X - np_0| \geq |np_0 - x|) = \begin{cases} Pr(|X - np_0| \geq np_0 - x) & \text{si } np_0 > x \\ Pr(|X - np_0| \geq x - np_0) & \text{si } np_0 < x \\ 1 & \text{si } np_0 = x \end{cases}$$

De allí, el p valor se calcula según los siguientes casos

- Si $np_0 > x$, entonces

$$\begin{aligned} p-\text{valor} &= Pr(|X - np_0| \geq np_0 - x) \\ &= Pr(X - np_0 \geq np_0 - x) + Pr(X - np_0 \leq x - np_0) \\ &= Pr(X \geq 2np_0 - x) + Pr(X \leq x) \\ &= 1 - Pr(X < 2np_0 - x) + Pr(X \leq x) \end{aligned}$$

  donde $[\cdot]$ denota el operador parte entera. Podemos calcular este p valor en R como `1-pbinom(ceiling(2*n*p0-x)-1,n,p0)+pbinom(x,n,p0)`

- Si $np_0 < x$, entonces

$$\begin{aligned} p-\text{valor} &= Pr(|X - np_0| \geq x - np_0) \\ &= Pr(X - np_0 \geq x - np_0) + Pr(X - np_0 \leq np_0 - x) \\ &= Pr(X \geq x) + Pr(X \leq 2np_0 - x) \\ &= 1 - Pr(X < x) + Pr(X \leq 2np_0 - x) \\ &= 1 - Pr(X \leq x - 1) + Pr(X \leq 2np_0 - x) \end{aligned}$$

  donde $[\cdot]$ denota el operador parte entera. Podemos calcular este p valor en R como `1-pbinom(x-1,n,p0)+pbinom(2*n*p0-x,n,p0)`

- Si $np_0 = x$, entonces claramente $p-\text{valor} = Pr(|X - np_0| \geq 0) = 1$. Obsérvese que de allí, cuando la estimación puntual de p dé igual al valor especificado de H_0, se acepta H_0, lo cual es muy lógico.

Para sistemas de tipo

$$H_0 : \ p = p_0 \quad vs. \quad H_1 : \ p > p_0,$$

o

$$H_0 : \ p \leq p_0 \quad vs. \quad H_1 : \ p > p_0.$$

Es claro que se rechaza H_0 para valores grandes de la estadística X, y por consiguiente podemos calcular el p valor como $Pr(X \geq x)$ bajo H_0, es decir, con $X \sim Binom(n, p_0)$. También se puede emplear la función `binom.test` con la opción `greater`.

Por otro lado, para sistemas como

$$H_0 : \ p = p_0 \quad vs. \quad H_1 : \ p < p_0, \tag{4.5.2}$$

o

$$H_0 : \ p \geq p_0 \quad vs. \quad H_1 : \ p < p_0, \tag{4.5.3}$$

Se rechaza H_0 para valores pequeños de X, y el p valor se calcula como $Pr(X \leq x)$ con $X \sim Binom(n, p_0)$. En R se debe utilizar la opción `less` en la función `binom.test`.

Ejemplo 4.5.2. *Suponga que con las nuevas tecnologías en la ciencia médica se cree que la probabilidad de éxito de una cirugía de trasplante de corazón es mayor a 0.7 y suponga que en 15 cirugías de este tipo 11 fueron exitosas. Para ver qué tan acorde es esta hipótesis con los datos observados, tenemos el sistema*

$$H_0: \ p \geq 0.7 \quad vs. \quad H_1: \ p < 0.7, \tag{4.5.4}$$

Al utilizar el comando `binom.test`*, tenemos que*

```
> binom.test(11,15,0.7,"less")

        Exact binomial test

data:  11 and 15
number of successes = 11, number of trials = 15, p-value = 0.7031
alternative hypothesis: true probability of success is less than 0.7
95 percent confidence interval:
 0.0000000 0.9033417
sample estimates:
probability of success
             0.7333333
```

Del anterior resultado, observamos que la estimación muestral es $11/15 = 0.73$*, y el* p*-valor es de 0.70 indicando que los datos están acordes con la hipótesis de* $p \geq 0.7$.

Otra prueba que podemos derivar para el sistema (4.5.1) es la prueba de razón generalizada de verosimilitudes donde teniendo en cuenta que en el sistema (4.5.1) $\boldsymbol{\Theta}_0 \bigcup \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}$ la estadística λ se calcula como

$$\lambda = \frac{L(\hat{p}_{MV})}{L(p_0)} = \frac{\bar{x}^{\sum_{i=1}^n x_i}(1-\bar{x}^{n-\sum_{i=1}^n x_i})}{p_0^{\sum_{i=1}^n x_i}(1-p_0^{n-\sum_{i=1}^n x_i})}$$

De acuerdo al razonamiento de la prueba de razón de verosimilitud, se rechaza H_0 para valores grandes de λ y para establecer una regla de decisión explícita, se necesita la distribución nula de λ y por la forma de la anterior expresión, encontrar su distribución nula puede ser muy difícil. Sin embargo, existe un resultado asintótico muy poderoso con respecto a la distribución nula de $2\ln\lambda$ que lo enunciamos a continuación, y lo vamos a utilizar en adelante en repetidas ocasiones.

Resultado 4.5.1. *Sea* $X_1, \cdots, X_n$ *una muestra aleatoria con función de densidad* $f(x_i, \theta)$*, para el sistema de hipótesis* $H_0: \ \theta \in \boldsymbol{\Theta}_0$ *vs.* $H_1: \ \theta \notin \boldsymbol{\Theta}_0$*, la estadística* $2\ln\lambda$ *converge a una distribución* $\chi^2_{v_1-v_0}$ *bajo* H_0*, donde* v_0 *y* v_1 *son números de parámetros libres en los espacios* $\boldsymbol{\Theta}_0$ *y* $\boldsymbol{\Theta}$*, respectivamente.*

Utilizando el anterior resultado, tenemos que

$$2\ln\lambda = 2\left(\sum_{i=1}^{n} x_i \ln\bar{x} + (n-\sum_{i=1}^{n} x_i)\ln(1-\bar{x}) - \sum_{i=1}^{n} x_i \ln p_0 - (n-\sum_{i=1}^{n} x_i)\ln(1-p_0)\right)$$
$$= 2\left(\sum_{i=1}^{n} x_i \ln\frac{\bar{x}}{p_0} + (n-\sum_{i=1}^{n} x_i)\ln\frac{1-\bar{x}}{1-p_0}\right).$$

Teniendo en cuenta que bajo H_0, $2\ln\lambda$ se distribuye como $\chi^2_{v_1-v_0}$ que en el caso del sistema (4.5.1), tenemos que $v_1 = 1$ y $v_0 = 0$, y de esta forma $2\ln\lambda \sim_{asym} \chi^2_1$. Y la regla de decisión será rechazar H_0 si $2\ln\lambda > \chi^2_{1,1-\alpha}$ y teniendo en cuenta esta regla de decisión, el p-valor se puede calcular como $1-F_{\chi^2_1}(v)$ donde v denota el valor observado de la estadística $2\ln\lambda$.

Ejemplo 4.5.3. *Para los datos del Ejemplo 4.5.1 acerca de preferencia de una marca de shampoo, podemos utilizar la siguiente función en R para calcular la estimación puntual, el valor de la estadística $2\ln\lambda$ y el correspondiente p-valor.*

```
> binom<-function(n,x,p0){
+ estimacion<-x/n
+ estad<-2*(x*log(estimacion/p0)+(n-x)*log((1-estimacion)/(1-p0)))
+ p.val<-pchisq(estad,1,lower.tail=F)
+ list(estimacion=estimacion,estadistica=estad,p.val=p.val)
+ }
>
> binom(30,7,0.5)
$estimacion
[1] 0.2333333

$estadistica
[1] 8.992464

$p.val
[1] 0.002710952
```

Podemos ver que el p-valor es bastante pequeño, indicando que H_0 no es una hipótesis adecuada de acuerdo con los datos observados, lo cual coincide con los resultados obtenidos con las pruebas anteriores.

Finalmente, realizamos un estudio de simulación para comparar las tres pruebas mencionadas anteriormente, con el fin de examinar qué tan buenas son estas tres pruebas y así aceptar una hipótesis nula verdadera y rechazar una hipótesis falsa. Simulamos 10000 muestras provenientes de una distribución $Bin(n, p_0)$ para diferentes valores de $n = 5, 15, 30, 50, 100$ y $p = 0.1, 0.3, 0.5, 0.7, 0.9$. En cada muestra simulada se aplican las tres pruebas para la hipótesis $p = p_0$, y se calcula el tamaño de cada prueba como el número de veces que se rechaza $p = p_0$ dividido por 10000. Los resultados se muestran en la Tabla 4.4.

	Prueba asintótica normal sin corrección				
	$n = 5$	$n = 15$	$n = 30$	$n = 50$	$n = 100$
$p_0 = 0.1$	0.0827	0.0530	0.0265	0.0310	0.0642
$p_0 = 0.3$	0.0313	0.0865	0.0696	0.0439	0.0669
$p_0 = 0.5$	0.0618	0.0331	0.0401	0.0666	0.0599
$p_0 = 0.7$	0.0331	0.0849	0.0683	0.0429	0.0625
$p_0 = 0.9$	0.0793	0.0519	0.0248	0.0278	0.0636
	Prueba asintótica normal con corrección				
$p_0 = 0.1$	0.0072	0.0128	0.0265	0.0310	0.0284
$p_0 = 0.3$	0.0018	0.0196	0.0255	0.0439	0.0391
$p_0 = 0.5$	0.0000	0.0331	0.0401	0.0318	0.0361
$p_0 = 0.7$	0.0027	0.0224	0.0274	0.0429	0.0380
$p_0 = 0.9$	0.0081	0.0128	0.0248	0.0278	0.0270
	Prueba exacta binomial				
$p_0 = 0.1$	0.0072	0.0128	0.0265	0.0310	0.0443
$p_0 = 0.3$	0.0313	0.0196	0.0462	0.0439	0.0522
$p_0 = 0.5$	0.0000	0.0331	0.0401	0.0318	0.0361
$p_0 = 0.7$	0.0331	0.0224	0.0468	0.0429	0.0500
$p_0 = 0.9$	0.0081	0.0128	0.0248	0.0278	0.0435

Tabla 4.4: *Comparación de tamaños de prueba para la prueba exacta binomial, la prueba asintótica normal sin corrección y la prueba asintótica normal con corrección bajo una distribución binomial.*

	Prueba asintótica normal sin corrección				
	$n = 5$	$n = 15$	$n = 30$	$n = 50$	$n = 100$
$p = 0.1$	0.5926	0.9470	0.9995	1.0000	1.0000
$p = 0.3$	0.1710	0.2902	0.5873	0.8608	0.9876
$p = 0.7$	0.1739	0.2984	0.5905	0.8591	0.9881
$p = 0.9$	0.5949	0.9454	0.9995	1.0000	1.0000
	Prueba asintótica normal con corrección				
$p = 0.1$	0	0.9470	0.9995	1.0000	1.0000
$p = 0.3$	0	0.2902	0.5873	0.7842	0.9789
$p = 0.7$	0	0.2984	0.5905	0.7822	0.9797
$p = 0.9$	0	0.9454	0.9995	1.0000	1.0000
	Prueba exacta binomial				
$p = 0.1$	0	0.9470	0.9995	1.0000	1.0000
$p = 0.3$	0	0.2902	0.5873	0.7842	0.9789
$p = 0.7$	0	0.2984	0.5905	0.7822	0.9797
$p = 0.9$	0	0.9454	0.9995	1.0000	1.0000

Tabla 4.5: *Comparación de potencia de prueba para la prueba exacta binomial, la prueba asintótica normal sin corrección y la prueba asintótica normal con corrección bajo una distribución binomial.*

Podemos ver que tal como se comentó anteriormente, la prueba normal con corrección tiende a aceptar más fácil H_0, por consiguiente siempre tiene un tamaño menor que la prueba normal sin corrección. Por otro lado, no parece haber una diferencia marcada entre la prueba binomial y las pruebas asintóticas normales.

Con respecto a la potencia de estas pruebas, simulamos 10000 muestras de una distribución $Bin(n,p)$ con $p = 0.1, 0.3, 0.7, 0.9$ y en cada muestra se aplican las tres pruebas para el hipótesis $p = 0.5$ el cual no corresponde al valor teórico de p. Estimamos la potencia de estas tres pruebas como el número de veces que se rechaza $p = 0.5$ dividido por 10000, y los resultados se muestran en la Tabla 4.5.

Asimismo, es posible observar que en muestras pequeñas con $n = 5$, la prueba normal con corrección y la prueba binomial tienen potencia igual a 0, es decir, en ninguna de las 10000 muestras simuladas, se llegó a rechazar la hipótesis nula $p = 0.5$, mientras que la prueba normal sin corrección tiene una potencia significativamente mayor. En muestras más grandes la potencia de estas tres pruebas puede coincidir, aunque la prueba normal sin corrección siempre tuvo una potencia mayor o igual que las otras dos pruebas. De donde podemos concluir que estas simulaciones mostraron que la prueba normal sin corrección es mejor que las otras dos.

4.5.2 Dos muestras

En esta parte, consideramos el problema de comparar dos muestras Bernoulli en términos de las probabilidades de éxito, o equivalentemente, tenemos X que denota el número de éxitos en n_X ensayos donde cada ensayo tiene como probabilidad de éxito p_1 y Y el número de éxitos en n_Y ensayos con p_2 denotando la probabilidad de éxito en esta población. Ejemplo de ello es la probabilidad de que un cliente compre una determinada marca de arroz con el empaque actual y la probabilidad de que un cliente compre esta marca de arroz con un nuevo empaque; también en la medicina, la probabilidad de cura de una enfermedad con el medicamento A y la misma probabilidad de cura con el medicamento B. En estos casos, el sistema de hipótesis de interés puede ser

$$H_0 : \ p_1 = p_2 \quad vs. \quad H_1 : \ p_1 \neq p_2, \tag{4.5.5}$$

En primer lugar, es claro que los parámetros p_1 y p_2 se pueden estimar con los porcentajes de éxito en las dos muestras, esto es, $\hat{p}_1 = X/n_X$ y $\hat{p}_2 = Y/n_Y$. Para encontrar una regla de decisión para el sistema (4.5.5), un primer acercamiento es utilizar el teorema de límite central, en el tema de intervalo de confianza para la diferencia de dos proporciones se había hecho uso de este teorema, y se encontró que

$$\hat{p}_1 - \hat{p}_2 \sim_{aprox} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_X} + \frac{p_2(1-p_2)}{n_Y}\right)$$

cuya distribución nula bajo $H_0 : \ p_1 = p_2 = p$ es

$$\hat{p}_1 - \hat{p}_2 \sim_{aprox} N\left(0, p(1-p)\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\right)$$

de donde

$$\frac{(\hat{p}_1 - \hat{p}_2)^2}{p(1-p)\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)} \sim_{aprox} \chi^2_1$$

En la anterior expresión, la probabilidad común p no es conocida, por consiguiente, para establecer una regla de decisión usando la anterior estadística, se debe obtener un estimador de p. Se deja como ejercicio verificar que el estimador de máxima verosimilitud de esta probabilidad común p es (Ejercicio 4.15)

$$\hat{p}_{MV} = \frac{X+Y}{n_X + n_Y}. \tag{4.5.6}$$

De esta forma, comparamos el valor de la estadística

$$Z^2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}_{MV}(1-\hat{p}_{MV})\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}$$

con el percentil $\chi^2_{1,1-\alpha}$ para tomar una decisión para el sistema (4.5.5) con un nivel de significación de α. Y el p-valor de esta prueba se calcula como $p - \text{valor} = Pr(Z^2 > v)$ donde $Z^2 \sim \chi^2_1$ y v denota el valor observado de la estadística Z^2.

En R, la prueba anterior se lleva a cabo mediante el comando `prop.test` que nos fue útil para el problema de una muestra. En Gutiérrez & Zhang (2009), se planteó un problema de prueba de cambio de empaque en la investigación de mercados, lo presentamos a continuación.

Ejemplo 4.5.4. *Supongamos que una empresa desea cambiar el empaque y la forma de presentación de un producto particular que está regularmente posicionado en el mercado. Para evaluar el impacto de la nueva presentación en la intención de compra del producto, el gerente de marketing planea una prueba de empaque por medio de la recolección de información en una sesión de grupo (focus group). La prueba fue realizada en 98 consumidores, donde a cada uno de ellos se le pregunta sobre la preferencia entre el empaque nuevo y el actual, en términos de la intención de compra, y los resultados de la prueba de empaque se muestran en la Tabla 4.6.*

Empaque	Compra	No compra	Total
Nuevo	32	31	63
Actual	11	24	35

Tabla 4.6: *Datos de la prueba de empaque del Ejemplo 4.5.4.*

Para conocer si el nuevo empaque tiene un efecto significativo sobre la preferencia de los consumidores comparado con el empaque actual, estamos interesados en comparar la probabilidad de compra con el empaque nuevo p_1 y la probabilidad de compra con el empaque actual p_2. En primer lugar, podemos ver que de los 63 consumidores a quienes se les preguntó sobre la preferencia del empaque nuevo, 32 de ellos se mostraron favorables; mientras que de los 35 consumidores a quienes se

les preguntó la preferencia del empaque actual, 11 se mostraron favorables. De allí, podemos ver que la estimación de las probabilidades están dadas por $\hat{p}_1 = 0.508$ *y* $\hat{p}_2 = 11/35 = 0.314$.

```
> nx <- 63 ; x <- 32
> ny <- 35 ; y <- 11
> prop.test(c(x,y),c(nx,ny))
    2-sample test for equality of proportions with continuity correction

data:  c(x, y) out of c(nx, ny)
X-squared = 2.6852, df = 1, p-value = 0.1013
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.02578597  0.41308755
sample estimates:
   prop 1    prop 2
0.5079365 0.3142857
```

De esta manera, para un nivel de significación del 5 %, no se rechaza la hipótesis de igualdad de proporciones. En otras palabras, no se encuentra evidencia de que el cambio al empaque nuevo tenga algún efecto sobre la decisión de compra comparado con el empaque actual. También podemos ver que el intervalo de confianza del 95 % para $p_1 - p_2$ *contiene el valor 0, conduciendo a la misma conclusión.*

También notemos que con la anterior instrucción, se incluye la corrección de continuidad de Yates de manera semejante a lo expuesto en el caso de una muestra, si se desea excluir esta corrección se debe usar la opción `correct=F`.

Existe otro enfoque basado en una tabla de contingencia que estudia dos variables cualitativas, y los datos registrados corresponden a conteos y éstos se ubican en celdas en tablas como los datos del Ejemplo 4.5.4, donde la hipótesis de que la probabilidad de compra con el empaque nuevo sea igual a la probabilidad de compra con el empaque actual es equivalente a la hipótesis de que la intención de compra es independiente de la presentación del empaque, y en términos de una tabla de contingencia, equivale a probar la independencia entre las filas y las columnas.

Retomando el sistema de interés

$$H_0 : \ p_1 = p_2 \quad vs. \quad H_1 : \ p_1 \neq p_2,$$

Si $p_1 = p_2$, entonces se espera que en la muestra $\hat{p}_1 \approx \hat{p}_2$ y $1 - \hat{p}_1 \approx 1 - \hat{p}_2$, y por consiguiente $\frac{\hat{p}_1(1-\hat{p}_1)}{\hat{p}_2(1-\hat{p}_2)} \approx 1$. Y esta cociente se define como la razón de *odds* muestral[3]. Entre más se difiere esta razón de odds, más evidencia hay en los datos en desfavor de la hipótesis $p_1 = p_2$. Fisher desarrolló la prueba usando este enfoque. En el apéndice expondremos los temas relacionados con el análisis de tabla de contingencias usando esta prueba y una prueba adicionalmente de χ^2, también muy usada.

[3]Este nombre viene de la palabra en inglés *Odds ratio*, que ocasionalmente en español se traduce como cociente de probabilidades ó razón de ventaja.

Ejemplo 4.5.5. *Los datos mostrados en la Tabla 4.7 fueron la base de una demanda en un caso de discriminación racial en 1980 entre los solicitantes de empleo en una fábrica de placas metálicas (Carlin & Louis 1996).*

Raza	Admitido	Rechazado	Total
Blanca	41	39	80
Negra	14	30	44
Total	55	69	124

Tabla 4.7: *Datos de la discriminación racial del Ejemplo 4.5.5.*

Mediante el análisis estadístico de estos datos se debe responder a la siguiente pregunta: ¿existe evidencia de discriminación racial? O equivalentemente, ¿la aceptación de una solicitud de empleo depende de la raza del solicitante? Para aplicar la prueba exacta de Fisher o la prueba χ^2*, podemos utilizar las funciones* `fisher.test` *y* `chisq.test` *que ilustramos a continuación.*

```
> racis<-matrix(c(41,39,14,30),2,2)
> fisher.test(racis)

        Fisher's Exact Test for Count Data

data:  racis
p-value = 0.04025
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.9796526 5.2952825
sample estimates:
odds ratio
  2.237988
> chisq.test(racis)

        Pearson's Chi-squared test with Yates' continuity correction

data:  racis
X-squared = 3.5913, df = 1, p-value = 0.05808
```

Podemos observar que con la prueba exacta de Fisher la razón de odds tomó el valor de 2.238, el cual se puede sospechar que es muy diferente del valor 1; por otro lado, el p*-valor es de 0.04, indicando evidencias en desfavor de* $p_1 = p_2$ *con un nivel de significación de 5 %. Señalamos que al observar el intervalo de confianza para la razón de odds poblacional* $\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$*, podemos ver que éste contiene el valor 1, lo cual produce una decisión contraria al observar el* p*-valor.*

Por otro lado, observando los resultados de la prueba χ^2*, advertimos que con el mismo nivel de significación de 5 %, se llega a la conclusión de aceptar la hipótesis* $H_0 : \ p_1 = p_2$.

4.6 Muestras provenientes de una distribución Poisson

4.6.1 Una muestra

En la práctica es común encontrar datos que corresponden a resultados de conteos y posiblemente pueden ser descritos por medio de la distribución Poisson. En el Ejemplo 2.3.2, se planteó la situación donde se investiga el nivel de violencia en una determinada ciudad por medio de datos que denotan el número de muertes violentos que ocurren mensualmente en distintos barrios de la ciudad y que por la naturaleza del problema, una distribución Poisson puede ser apropiada.

Suponga que el sistema de interés es

$$H_0: \ \lambda = \lambda_0 \quad vs. \quad H_1: \ \lambda \neq \lambda_0.$$

Utilizamos la prueba generalizada de razón de verosimilitud definida, como en el anterior sistema, en $\boldsymbol{\Theta}_0 \bigcup \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}$, entonces

$$\lambda = \frac{L(\hat{\lambda}_{MV})}{L(\lambda_0)} = \exp\left\{ n\lambda_0 - \sum_{i=1}^{n} x_i \right\} \left(\frac{\bar{x}}{\lambda_0} \right)^{\sum_{i=1}^{n} x_i}$$

Y tenemos que

$$\begin{aligned} 2\ln\lambda &= 2\left(\ln L(\hat{\lambda}_{MV}) - \ln L(\lambda_0) \right) \\ &= 2\left(-n\bar{x} + \sum_{i=1}^{n} x_i \ln\bar{x} + n\lambda_0 - \sum_{i=1}^{n} \ln\lambda_0 \right) \\ &= 2\left(n\bar{x}\ln\frac{\bar{x}}{\lambda_0} + n(\lambda_0 - \bar{x}) \right) \end{aligned}$$

Dado el razonamiento en la formulación de la prueba de razón de verosimilitud, valores grandes de $2\ln\lambda$ muestran la gran evidencia que tienen los datos en contra de H_0, y de allí podemos establecer dos reglas de decisión.

1. Anteriormente se ha establecido la distribución asintótica de $2\ln\lambda$ bajo H_0 que es $\chi^2_{v_1 - v_0}$ que en este caso $v_1 = 1$ y $v_0 = 0$. Así, podemos encontrar la regla de decisión: rechazar H_0 si $2\ln\lambda > \chi^2_{1,1-\alpha}$.

2. Por otro lado, observando la forma de la estadística $2\ln\lambda$, vemos que ésta compara el estimador de λ, $\bar{X}$ con el valor especificado por H_0, λ_0. Y esta estadística depende de los valores de la muestra sólo mendiante $\bar{X}$, así que podemos pensar en encontrar la relación que existe entre $\bar{X}$ y $2\ln\lambda$, puesto que si $2\ln\lambda$ es una función creciente de $\bar{X}$ entonces valores grandes de $\bar{X}$ conducen a valores grandes de $2\ln\lambda$ y por consiguiente, se rechaza H_0 para

valores grandes de $\bar{X}$. Para ver dónde $2\ln\lambda$ es una función creciente/decreciente de $\bar{X}$, derivamos a $2\ln\lambda$ con respecto a $\bar{X}$. Tenemos que

$$\frac{\partial 2\ln\lambda}{\partial\bar{X}} = 2n\ln\frac{\bar{X}}{\lambda_0}$$

En la anterior derivada, si $\bar{X}$ es muy grande, entonces la anterior derivada será positiva mientras que la derivada será negativa si los valores de $\bar{X}$ son pequeños. Esto indica que $2\ln\lambda$ es función creciente de $\bar{X}$ para valores grandes de $\bar{X}$ y función decreciente para valores pequeños de $\bar{X}$. En conclusión, cuando $\bar{X}$ toma valores muy grandes o muy pequeños, $2\ln\lambda$ toma valores grandes. Por consiguiente, tenemos la regla de decisión de rechazar H_0 si $\bar{X} > K_1$, o $\bar{X} < K_2$ para algunos valores K_1 y K_2.

Al utilizar la restricción de que el máximo error de tipo I permitido es α y la distribución nula de $n\bar{X}$ dada por $Pois(\lambda_0)$, podemos encontrar que $K_1 = Pois(\lambda_0, 1-\alpha/2)/n$ y $K_2 = Pois(\lambda_0, \alpha/2)/n$, respectivamente, y así tenemos la regla de decisión: rechazar H_0 si $\bar{X} > Pois(\lambda_0, 1-\alpha/2)/n$ o $\bar{X} < Pois(\lambda_0, \alpha/2)/n$.

De lo anterior vemos que existen por lo menos dos pruebas distintas para el sistema de hipótesis considerado bajo una distribución Poisson: una basada en la distribución exacta de $n\bar{X}$, y la otra basada en la distribución asintótica de la estadística $2\ln\lambda$. Dado que la distribución asintótica es válida cuando $n \to \infty$, que en la práctica se refleja en las muestras grandes, es interesante conocer cuál prueba es mejor en el sentido en que los errores tipo I ocurren con frecuencia inferior a α y los errores tipo II no ocurren con frecuencia. Un sencillo ejercicio de simulación nos puede dar pistas acerca del funcionamiento de estas dos pruebas. A continuación describimos cómo se pueden realizar estas simulaciones.

1. Para comparar las dos pruebas con respecto al error tipo I, simulamos 10000 muestras de tamaño n provenientes de una distribución $Pois(\lambda_0)$, y aplicamos las dos pruebas a cada una de las muestras simuladas para probar el sistema $H_0: \lambda = \lambda_0$ vs. $H_1: \lambda \neq \lambda_0$, es decir, ya conocemos de antemano que H_0 es verdadero, entonces de las 1000 muestras simuladas, el número de veces que una prueba rechaza H_0 debe ser pequeño, no superior a α. Realizamos el procedimiento para $n = 5, 15, 30, 50, 100$ y $\lambda_0 = 2, 5, 15, 25$, y los resultados se muestran en la Tabla 4.8, donde cada entrada es el porcentaje de muestras donde erróneamente rechaza H_0, es decir, una aproximación del error tipo I.

   Observamos que con la prueba exacta casi siempre se tiene menor riesgo de rechazar una hipótesis falsa, aunque la diferencia entre las pruebas es muy reducida, inclusive para muestras muy pequeñas.

2. Para comparar las dos pruebas con respecto a la potencia, es decir, la capacidad de rechazar una hipótesis falsa, el procedimiento de simulación es similar al caso anterior, pero probamos el sistema $H_0: \lambda = \lambda_0$ vs. $H_1: \lambda \neq \lambda_0$ en muestras provenientes de distribución $Pois(\lambda)$ con $\lambda \neq \lambda_0$, es decir, ya conocemos de

	Prueba exacta				
	$n = 5$	$n = 15$	$n = 30$	$n = 50$	$n = 100$
$\lambda_0 = 2$	0.0238	0.0428	0.0408	0.0456	0.0487
$\lambda_0 = 5$	0.0454	0.0516	0.0468	0.0487	0.0471
$\lambda_0 = 15$	0.0483	0.0458	0.0515	0.0478	0.0481
$\lambda_0 = 25$	0.0498	0.0475	0.0482	0.0481	0.0489
	Prueba asintótica				
$\lambda_0 = 2$	0.0533	0.0428	0.0534	0.0521	0.0487
$\lambda_0 = 5$	0.0454	0.0516	0.0520	0.0518	0.0471
$\lambda_0 = 15$	0.0483	0.0494	0.0533	0.0478	0.0481
$\lambda_0 = 25$	0.0498	0.0475	0.0482	0.0519	0.0502

Tabla 4.8: *Comparación de tamaños de prueba para la prueba exacta y la prueba asintótica bajo una distribución Poisson.*

antemano que H_0 es falsa. De esta forma, una buena prueba será aquella que rechaza más veces H_0 en las 10000 muestras simuladas. Las simulaciones se realizaron para $n = 5, 15, 30, 50, 100$, $\lambda_0 = 20$ y $\lambda = 13, 15, 17, 19, 21, 23, 25, 27$. Los resultados se muestran en la Tabla 4.9, donde cada entrada es el porcentaje de muestras donde correctamente rechaza H_0, es decir, una aproximación de la potencia de la prueba.

	Prueba exacta				
	$n = 5$	$n = 15$	$n = 30$	$n = 50$	$n = 100$
$\lambda = 13$	0.9689	1.0000	1.0000	1.0000	1.0000
$\lambda = 15$	0.7364	0.9965	1.0000	1.0000	1.0000
$\lambda = 17$	0.3202	0.7699	0.9643	0.9989	1.0000
$\lambda = 19$	0.0725	0.1429	0.2208	0.3439	0.6063
$\lambda = 21$	0.0763	0.1434	0.2279	0.3470	0.5933
$\lambda = 23$	0.2929	0.7113	0.9489	0.9954	1.0000
$\lambda = 25$	0.6452	0.9842	0.9999	1.0000	1.0000
$\lambda = 27$	0.9013	0.9997	1.0000	1.0000	1.0000
	Prueba asintótica				
$\lambda = 13$	0.9761	1.0000	1.0000	1.0000	1.0000
$\lambda = 15$	0.7707	0.9965	1.0000	1.0000	1.0000
$\lambda = 17$	0.3563	0.7699	0.9681	0.9990	1.0000
$\lambda = 19$	0.0880	0.1429	0.2335	0.3570	0.6063
$\lambda = 21$	0.0792	0.1434	0.2280	0.3471	0.5933
$\lambda = 23$	0.2929	0.7113	0.9489	0.9954	1.0000
$\lambda = 25$	0.6452	0.9842	0.9999	1.0000	1.0000
$\lambda = 27$	0.9013	0.9997	1.0000	1.0000	1.0000

Tabla 4.9: *Comparación de potencia de prueba para la prueba exacta y la prueba asintótica bajo una distribución Poisson.*

Observamos en primer lugar que para cada n fijo, entre más se aleja λ, el

parámetro verdadero del valor especificado por H_0, mayor potencia tiene cada prueba. En segundo lugar, observamos que la potencia de la prueba asintótica, aún en muestras pequeñas, es más alta que la prueba exacta, aunque las diferencias no son significativas.

De los anteriores comentarios, podemos concluir que el desempeño de la prueba asintótica y la prueba exacta son muy similares, y por consiguiente en la práctica se puede optar por cualquiera de estas dos, aunque la prueba asintótica puede ser más fácil, por lo menos al momento de calcular el p valor.

Presentamos una aplicación en el siguiente ejemplo.

Ejemplo 4.6.1. *En el Ejemplo 2.3.2, con el fin de conocer el nivel de violencia de una ciudad, se disponen los datos que corresponden al número de muertes violentas de 15 de los 63 barrios: éstos son 1, 1, 5, 5, 2, 3, 3, 6, 4, 3, 2, 3, 2, 3 y 4. Supongamos que se cree que mensualmente suceden en promedio 70 muertes violentas en toda la ciudad. ¿Cómo podemos confirmar o refutar esta creencia? En primer lugar, nótese que los datos que tenemos a la mano corresponden a observaciones no sobre la ciudad, sino sobre barrios que son parte de la ciudad*[4]*. Entonces realmente la inferencia se puede hacer sobre el número promedio de muertes violentas λ_b por barrio en esta ciudad, pero el parámetro de interés no es al nivel de barrios sino en toda la ciudad λ_c; sin embargo, por las anotaciones hechas en el Ejemplo 2.3.2, se puede establecer que $\lambda_c = 63 * \lambda_b$, por lo tanto la hipótesis $\lambda_c = 70$ equivale a $\lambda_b = 70/63 \approx 1.11$. De esta forma, ya podemos utilizar directamente los datos a la mano. Usando la prueba asintótica, encontramos que la estadística $2 \ln \lambda \approx 36.78$, y el p valor están dados por $F_{\chi^2_1}(36.78) = 1.31e{-}09$ mostrando una fuerte discordancia entre los datos observados y la hipótesis $\lambda_b = 70/63 \approx 1.11$ ó $\lambda_c = 70$. Se deja como ejercicio la aplicación de la prueba exacta.*

4.6.2 Dos muestras

$X_1, \cdots, X_{n_X}$ provenientes de una distribución $Pois(\lambda_X)$ y $Y_1, \cdots, Y_{n_Y}$ provenientes de una distribución $Pois(\lambda_Y)$

$$H_0 : \ \lambda_X = \lambda_Y \quad vs. \quad H_1 : \ \lambda_X \neq \lambda_Y \tag{4.6.1}$$

Para el anterior sistema, es natural pensar que los datos muestran evidencia en contra de H_0 si las estimaciones de λ_X y λ_Y son muy diferentes, esto es, si $|\bar{X} - \bar{Y}|$ es muy grande. A pesar de que este razonamiento es completamente lógico y válido, no es posible o es muy complicado encontrar la distribución nula exacta[5] de $\bar{X} - \bar{Y}$.

[4] Si se hace observación directa sobre la ciudad, las observaciones se harán en diferentes meses, y en este caso las observaciones ya no constituyen una muestra aleatoria por ser observaciones hechas a través del tiempo.

[5] Aunque se puede utilizar el teorema del límite central para encontrar que bajo $\lambda_X = \lambda_Y = \lambda$ $\bar{X} - \bar{Y} \sim_{asym} N\left(0, \lambda\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\right)$, pero la varianza de esta distribución es desconocida y por consiguiente tampoco nos es útil para encontrar una regla de decisión.

Por lo tanto, hacemos uso de la prueba generalizada de razón de verosimilitud. Para calcular esta estadística, tengamos presente que

$$\ln L = -n_X\lambda_X - n_Y\lambda_Y + \sum_{i=1}^{n_X} X_i \ln \lambda_X + \sum_{j=1}^{n_Y} Y_j \ln \lambda_Y - \ln \prod_{i=1}^{n_X} X_i - \ln \prod_{j=1}^{n_Y} Y_i$$

Los estimadores de máxima verosimilitud λ_X y λ_Y son $\bar{X}$ y $\bar{Y}$, respectivamente; por otro lado, bajo $H_0: \ \lambda_X = \lambda_Y = \lambda$, el estimador de máxima verosimilitud de λ es $(\sum_{i=1}^{n_X} X_i + \sum_{j=1}^{n_Y} Y_j)/(n_X + n_Y)$, es decir, el promedio de los datos de ambas muestras. Dado lo anterior, la estadística $2\ln\lambda$ se puede calcular como

$$2\ln\lambda = 2\left(n_X\bar{X}\ln\bar{X} + n_Y\bar{Y}\ln\bar{Y} - (n_X\bar{X} + n_Y\bar{Y})\ln\frac{n_X\bar{X} + n_Y\bar{Y}}{n_X + n_Y}\right)$$

Para encontrar la distribución nula de $2\ln\lambda$ observamos que en primer lugar, hay dos parámetros teóricos en el espacio paramétrico completo, mientras que bajo H_0, los dos parámetros son iguales, y por consiguiente se reducen a un solo parámetro. De esta forma, tenemos que $v_1 = 2$ y $v_0 = 1$, y por consiguiente

$$2\ln\lambda \sim_{asym} \chi^2_1.$$

Y se rechaza H_0 si $2\ln\lambda > \chi^2_{1,1-\alpha}$. También podemos calcular el correspondiente p valor como

$$p \text{ valor} = 1 - F_{\chi^2_1}(v)$$

donde $F_{\chi^2_1}$ es la función de distribución de una distribución χ^2_1 y v denota el valor que toma la estadística $2\ln\lambda$.

Ejemplo 4.6.2. *Una empresa de telefonía móvil (A) lanza una promoción limitada de precios especiales en equipos para atraer clientes, y la directa competencia (B) de esta empresa quiere saber si la promoción de la compañía A surtió algún efecto significativo sobre la venta de los equipos*[6]*. Para eso la empresa B observa el número de clientes que realizan compras en 10 puntos de ventas de la empresa A en un día determinado durante la vigencia de la promoción, y después de la promoción. Suponga que los datos que se registraron son los de la Tabla 4.10.*

	Punto de venta									
	1	2	3	4	5	6	7	8	9	10
Durante la promoción	36	31	28	41	35	52	25	31	32	34
Después de la promoción	31	41	29	20	29	35	31	29	33	30

Tabla 4.10: *Datos del Ejemplo 4.6.2.*

[6]Nótese que la empresa A tiene pleno conocimiento sobre el efecto de la promoción puesto que tiene los registros sobre el volumen de ventas, pero la competencia B no tiene esa información y por lo tanto debe recurrir a otros medios para investigar el fenómeno.

Dada la naturaleza del problema, existen dos tipos de comportamientos que pueden considerarse como dos poblaciones, una concerniente al número de ventas durante la promoción, y la otra, después de la promoción. Los datos registrados corresponden al número de ventas diarias y pueden considerarse como realización de variables tipo Poisson. De esta forma, los datos de la primera muestra corresponden al número de ventas durante la promoción en diez puntos de venta, y los datos de la segunda muestra, después de la promoción. Si denotamos el número promedio de ventas diarias en un punto de venta durante y después de la promoción como λ_1 y λ_2, respectivamente, entonces al considerar el sistema de hipótesis

$$H_0: \ \lambda_1 = \lambda_2 \quad vs. \quad H_1: \ \lambda_1 \neq \lambda_2$$

podríamos hacernos una idea acerca de si la promoción ha mejorado significativamente las ventas.

El siguiente código en R computa las estimaciones de λ_1 y λ_2, la estadística $2 \ln \lambda$ y el p valor.

```
> pois_2<-function(x,y){
+ nx<-length(x)
+ ny<-length(y)
+ est.X<-mean(x)
+ est.Y<-mean(y)
+ l1<-sum(x)*log(est.X)+sum(y)*log(est.Y)
+ l2<-(sum(x)+sum(y))*log((sum(x)+sum(y))/(nx+ny))
+ estad<-2*(l1-l2)
+ p<-pchisq(estad,1,lower.tail = F)
+ list(estima.X=est.X,estima.Y=est.Y,estadistica=estad,p.valor=p)
+ }
>
> Durante<-c(36, 31 ,28, 41 ,35 ,52, 25, 31, 32, 34)
> Despues<-c(31 ,41 ,29 ,20, 29, 35 ,31, 29, 33 ,30)
> pois_2(Durante,Despues)
$estima.X
[1] 34.5

$estima.Y
[1] 30.8

$estadistica
[1] 2.097601

$p.valor
[1] 0.1475305
```

De los resultados vemos que las estimaciones de λ_1 y λ_2 son bastante similares, y la prueba de razón de verosimilitudes indica que la diferencia entre estas estimaciones

no es significativa. De donde los directivos de la compañía B pueden concluir que la promoción lanzada por la compañía A no obtuvo un efecto significativo en el volumen de ventas. Es decir, probablemente la promoción no tuvo en cuenta las necesidades que tienen los usuarios de la telefonía móvil y/o los aspectos de interés que llaman la atención de los clientes. De allí, la compañía B puede lanzar una promoción que corrige estas deficiencias, atraer más clientes y tomar la delantera en el mercado.

4.7 Muestras provenientes de la distribución exponencial

4.7.1 Una muestra

Suponga que tenemos $X_1, \cdots, X_n$, una muestra proveniente de una distribución $Exp(\theta)$. Y estamos interesados en probar el sistema de hipótesis

$$H_0: \ \theta = \theta_0 \quad vs. \quad H_1: \ \theta \neq \theta_0 \tag{4.7.1}$$

Hay por lo menos tres formas de encontrar una regla de decisión para este sistema. Las dos primeras teniendo en cuenta la dualidad entre los intervalos de confianza y la prueba de sistema de hipótesis, y la tercera usando la prueba generalizada de razón de verosimilitud.

1. Anteriormente se encontró el siguiente intervalo bilateral exacto para θ dado por

$$IC(\theta) = \left(\frac{\sum_{i=1}^n X_i}{Gamma(n,1)_{1-\alpha/2}}, \frac{\sum_{i=1}^n X_i}{Gamma(n,1)_{\alpha/2}} \right).$$

   Utilizando el anterior intervalo, podemos rechazar $H_0: \ \theta = \theta_0$ si θ_0 no se encuentra dentro del intervalo, es decir:

   Rechazar H_0 si $\sum_{i=1}^n X_i > \theta_0 Gamma(n,1)_{1-\alpha/2}$ ó si $\sum_{i=1}^n X_i < \theta_0 Gamma(n,1)_{\alpha/2}$. El p valor asociado a esta prueba se puede calcular como

$$p \text{ valor} = \begin{cases} 2(1 - F_{G(n,\theta_0)}(v)) & \text{si } v > Gamma(n,\theta_0)_{0.5} \\ 2F_{G(n,\theta_0)}(v) & \text{si } v < Gamma(n,\theta_0)_{0.5} \end{cases}$$

   donde v es el valor observado de la estadística $\sum_{i=1}^n X_i$ y $F_{G(n,\theta_0)}(\cdot)$ denota la función de distribución de la distribución $Gamma(n,\theta_0)$.

2. Usando el mismo argumento del punto anterior y el intervalo aproximado de θ dado por

$$IC(\theta) = \left(\frac{\sqrt{n}\bar{X}}{z_{1-\alpha/2} + \sqrt{n}}, \frac{\sqrt{n}\bar{X}}{-z_{1-\alpha/2} + \sqrt{n}} \right).$$

   Podemos obtener la prueba:

Rechazar H_0 si $\bar{X} > \frac{\theta_0}{\sqrt{n}} z_{1-\alpha/2} + \theta_0$ ó si $\bar{X} < -\frac{\theta_0}{\sqrt{n}} z_{1-\alpha/2} + \theta_0$. El p valor asociado a esta prueba se puede calcular como

$$p \text{ valor} = \begin{cases} 2(1-\Phi(v)) & \text{si } v \geq 0 \\ 2\Phi(v) & \text{si } v < 0 \end{cases}$$

donde v es el valor observado de la estadística $\sqrt{n}(\bar{X} - \theta_0)/\theta_0$.

3. Aplicando la prueba generalizada de razón de verosimilitudes, es fácil verificar que

$$2\ln\lambda = 2n\left(\frac{\bar{X}}{\theta_0} - \log\frac{\bar{X}}{\theta_0} - 1\right) \tag{4.7.2}$$

cuya distribución nula asintótica es χ^2_1, y por consiguiente tenemos la decisión de

Rechazar H_0 si $2\ln\lambda > \chi^2_{1,1-\alpha}$.

Dado que hay tres diferentes pruebas para el mismo sistema de hipótesis, debemos compararlas en términos del tamaño y la potencia.

Primero comprobamos que el tamaño de las pruebas sea cercano al nivel de significación nominal α, y lo realizamos mediante simulaciones. El procedimiento es similar a las comparaciones realizadas bajo la distribución Poisson, es decir, se simulan 10000 muestras provenientes de una distribución $Exp(\theta_0)$ y en cada muestra simulada, se juzga la hipótesis $H_0: \theta = \theta_0$, y el tamaño de la prueba es el número de veces que se rechaza erróneamente H_0 divido por 10000. Los resultados de estas simulaciones para diferentes tamaños de muestra y diferentes valores de θ_0 se encuentran en la Tabla 4.11, donde podemos observar que las tres pruebas, incluyendo las pruebas asintóticas con pequeños tamaños de muestra, tienen buen desempeño en términos del tamaño, ya que el porcentaje de error tipo I cometido es similar al nivel de signficación α.

Con respecto a la potencia, para la prueba exacta, podemos hallar la función de potencia como

$$\begin{aligned}\beta_1(\theta) &= Pr\left(\sum_{i=1}^n X_i > \theta_0 Gamma(n,1)_{1-\alpha/2}\right) + Pr\left(\sum_{i=1}^n X_i < \theta_0 Gamma(n,1)_{\alpha/2}\right) \\ &= 1 - F_{G(n,\theta)}(\theta_0 Gamma(n,1)_{1-\alpha/2}) + F_{G(n,\theta)}(\theta_0 Gamma(n,1)_{\alpha/2})\end{aligned}$$

puesto que $\sum_{i=1}^n X_i \sim Gamma(n,\theta)$, y $F_{G(n,\theta)}(\cdot)$ denota la función de distribución de la distribución $Gamma(n,\theta)$.

De manera análoga y teniendo en cuenta que $\frac{\sqrt{n}(\bar{X}-\theta)}{\theta} \sim_{asym} N(0,1)$ por el teorema límite central, podemos ver que la potencia de la prueba asintótica basada en la distribución normal en el numeral 2 es

$$\beta_2(\theta) = 1 - \Phi\left(\frac{\theta_0}{\theta} z_{1-\alpha/2} + \sqrt{n}(\frac{\theta_0}{\theta} - 1)\right) + \Phi\left(-\frac{\theta_0}{\theta} z_{1-\alpha/2} + \sqrt{n}(\frac{\theta_0}{\theta} - 1)\right). \tag{4.7.3}$$

	Prueba exacta				
	$n = 5$	$n = 15$	$n = 30$	$n = 50$	$n = 100$
$\theta_0 = 2$	0.0490	0.0483	0.0499	0.0488	0.0503
$\theta_0 = 5$	0.0495	0.0526	0.0473	0.0473	0.0470
$\theta_0 = 15$	0.0491	0.0491	0.0477	0.0493	0.0491
$\theta_0 = 25$	0.0523	0.0502	0.0474	0.0500	0.0456
	Prueba asintótica normal				
$\theta_0 = 2$	0.0409	0.0454	0.0466	0.0462	0.0508
$\theta_0 = 5$	0.0414	0.0479	0.0463	0.0471	0.0460
$\theta_0 = 15$	0.0408	0.0459	0.0464	0.0486	0.0484
$\theta_0 = 25$	0.0462	0.0480	0.0474	0.0510	0.0448
	Prueba asintótica χ^2				
$\theta_0 = 2$	0.0536	0.0494	0.0501	0.0484	0.0499
$\theta_0 = 5$	0.0554	0.0523	0.0480	0.0489	0.0474
$\theta_0 = 15$	0.0519	0.0516	0.0484	0.0513	0.0495
$\theta_0 = 25$	0.0547	0.0495	0.0489	0.0501	0.0463

Tabla 4.11: *Comparación de tamaños de prueba para la prueba exacta y las dos pruebas asintóticas bajo una distribución Exponencial.*

Finalmente, debemos hallar la función de potencia de la prueba de razón de verosimilitud cuya regla de decisión se basa en una distribución χ^2. Sin embargo, no es nada fácil hallar esta función y por cuestión de simplicidad usamos simulaciones para aproximar esta función de potencia, similar a las simulaciones de potencia realizadas previamente bajo una distribución Poisson. En la Figura 4.21, graficamos conjuntamente las funciones de potencia $\beta_1(\theta)$, $\beta_2(\theta)$ y las aproximaciones de la función de potencia de la prueba asintótica de verosimilitud basada en la distribución χ^2 con $n = 30$.

```
> set.seed(1234)
> alpha<-0.05
> theta0<-20
> theta<-theta0+seq(-7,7,2)
> n<-30
> n.sim<-10000
>
> ## Potencia de la prueba exacta gamma
> pote_gamma<-function(theta,theta0,n,alpha){
+ 1-pgamma(theta0*qgamma(1-alpha/2,n,1),n,1/theta)+
+ pgamma(theta0*qgamma(alpha/2,n,1),n,1/theta)
+ }
> # Potencia de la prueba asintótica normal
> pote_norm<-function(theta,theta0,n,alpha){
+ 1-pnorm((theta0/theta)*(qnorm(1-alpha/2)+sqrt(n))-sqrt(n))+
+ pnorm((theta0/theta)*(-qnorm(1-alpha/2)+sqrt(n))-sqrt(n))
+ }
```

```
> ## Potencia de la prueba asintótica Ji-cuadrado
> pote_chi<-function(theta,theta0,n,alpha){
+ aux<-0
+ for(k in 1:n.sim){
+ muestra<-rexp(n,1/theta)
+ ba<-mean(muestra)
+ LRT<-2*n*(ba/theta0-log(ba/theta0)-1)
+ if(LRT>qchisq(1-alpha,1)){aux<-aux+1}
+ }
+ aux/n.sim
+ }
> ##
>
> pote_chis<-pote_gammas<-pote_norms<-matrix(NA)
> theta<-seq(5,50,0.1)
> for(i in 1:length(theta)){
+ pote_chis[i]<-pote_chi(theta[i],theta0,n,alpha)
+ pote_gammas[i]<-pote_gamma(theta[i],theta0,n,alpha)
+ pote_norms[i]<-pote_norm(theta[i],theta0,n,alpha)
+ }
>
> plot(pote_chis,type="l",xaxt="n",xlab="theta",ylab="Potencia")
> axis(1,1:length(theta),theta)
> lines(pote_norms,lty=2)
> lines(pote_gammas,lty=3)
> legend(220,0.3,c("Prueba exacta", "Prueba asintótica normal",
+ "Prueba asintótica chi2"),lty=c(3,2,1),bty="n")
```

Observamos que, en primer lugar, el compotamiento de la prueba exacta basada en la distribución Gamma y la prueba asintótica basada en la distribución χ^2 son muy similares, mientras que la prueba asintótica basada en la distribución normal difiere levemente, en el sentido de que tiene mayor capacidad para detectar una hipótesis falsa del tipo $\theta = \theta_0$ cuando θ_0 es menor que el parámetro verdadero, pero cuando el valor especificado por H_0 es mayor que el parámetro verdadero, esta prueba tiene menor potencia que las restantes dos.

Ejemplo 4.7.1. *En el Ejemplo 2.3.4, se planteó el problema de monitorear llamadas contestadas por operadores de una aerolínea con el fin de evitar pérdida de clientes potenciales debido a largos tiempos de espera en línea. Los minutos transcurridos en 20 llamadas antes de ser contestadas fueron 0.13, 0.06, 0.50, 0.41, 1.44, 0.60, 0.22, 1.08, 0.78, 0.92, 2.73, 0.83, 0.19, 0.21, 1.75, 0.79, 0.02, 0.05, 2.30 y 1.03, y en el Ejemplo 2.3.4 se hizo la anotación de que la distribución exponencial describe adecuadamente los datos. Suponga que la hipótesis que sostiene el supervisor de estos operadores es que en promedio se necesita medio minuto para que una llamada sea contestada. Para ver la validez de esta hipótesis usando los datos muestrales, planteamos el sistema*

$$H_0 : \ \theta = 0.5 \quad vs. \quad H_1 : \ \theta \neq 0.5$$

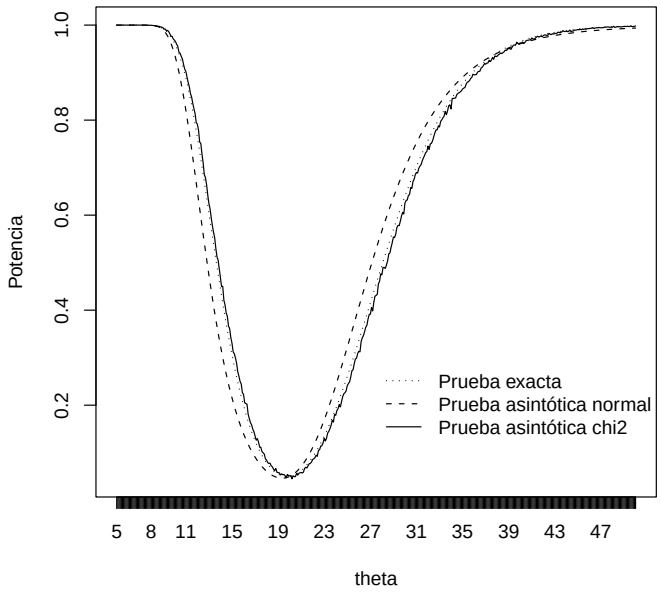

Figura 4.21: *Función de potencia de las tres pruebas para la media teórica bajo la distribución exponencial con* $\alpha = 0.05$ *y* $n = 30$.

donde θ *denota el tiempo promedio (en minutos) de espera. El siguiente código efectúa cualquiera de las tres pruebas descritas anteriormente dependiendo de la opción que el usuario le dé.*

```
>     exp_1<-function(x,theta0,type = c("gamma","normal", "Ji")){
+
+     n<-length(x)
+     est<-mean(x)
+
+     if(type=="gamma"){
+       prueba<-"Prueba exacta gamma"
+       if(sum(x)>qgamma(0.5,n,1/theta0)){
+       p<-2*pgamma(sum(x),n,1/theta0,lower.tail=F)        }
+       else{p<-2*pgamma(sum(x),n,1/theta0)   }
+     }
+
+     if(type=="normal"){
+       prueba<-"Prueba asintótica normal"
+       v<-sqrt(n)*(est-theta0)/theta0
```

```
+       if(v<0){
+       p<-2*pnorm(v)        }
+       else{p<-2*pnorm(v,lower.tail=F)   }
+     }
+
+     if(type=="Ji"){
+       prueba<-"Prueba asintótica Ji-cuadrado"
+       lambda<-2*n*((est/theta0)-log(est/theta0)-1)
+       p<-pchisq(lambda,1,lower.tail=F)
+     }
+
+     list(tipo=prueba,estimación=est,p.valor=p)
+     }
```

Para los datos de este ejemplo, la estimación del parámetro θ *es* $\bar{x} = 0.802$*, el cual es superior al valor especificado en* H_0*, de donde podríamos sospechar que posiblemente el valor de* θ *también sea mayor de* θ_0*, y por consiguiente utilizar la prueba asintótica basada en la distribución normal ya que en la Figura 4.21 se observó que ésta es la más potente en este caso. Por lo tanto utilizamos la función* `exp_1` *de la siguiente manera*

```
>     tiempo<-c(0.13, 0.06, 0.50, 0.41, 1.44, 0.60, 0.22,
+    1.08, 0.78, 0.92, 2.73, 0.83, 0.19, 0.21, 1.75, 0.79,
+    0.02, 0.05, 2.30, 1.03)
>     exp_1(tiempo,0.5,type="normal")
$tipo
[1] "Prueba asintótica normal"

$estimación
[1] 0.802

$p.valor
[1] 0.006909599
```

El p *valor muestra que la hipótesis* $\theta = 0.5$ *no es apoyada por los datos observados, y por consiguiente se rechaza* H_0*.*

4.7.2 Dos muestras

$X_1, \cdots, X_{n_X}$ provenientes de una distribución $Exp(\theta_X)$ y $Y_1, \cdots, Y_{n_Y}$ provenientes de una distribución $Exp(\theta_Y)$.

$$H_0 : \ \theta_X = \theta_Y \quad vs. \quad H_1 : \ \theta_X \neq \theta_Y \tag{4.7.4}$$

Se deja como ejercicio verificar con el método de la prueba generalizada de verosimilitudes, tenemos que (Ejercicio 4.24)

$$2\ln\lambda = 2\left((n_X+n_Y)\ln\frac{n_X\bar{X}+n_Y\bar{Y}}{n_X+n_Y} - n_X\ln\bar{X} - n_Y\ln\bar{Y}\right) \tag{4.7.5}$$

cuya distribución nula asintótica es χ^2_1 y se rechaza H_0 cuando $2\ln\lambda > \chi^2_{1,1-\alpha}$.

Ejemplo 4.7.2. *En el problema de control del tiempo de espera de llamadas que entran a una aerolínea, ahora, según si las llamadas salen del capital, Bogotá, o de ciudades diferentes a Bogotá, los números marcados son diferentes. Para llamadas salientes de Bogotá, los clientes llaman a un número fijo de 7 dígitos, mientras que los clientes que llaman de otra ciudad marcan un número del tipo 018000 de 12 dígitos. Dado lo anterior, es claro que el sistema de operación es diferente, y por consiguiente podemos formular la inquietud de si esta diferencia puede causar que el tiempo de espera de llamadas salientes de Bogotá y las salientes de otra ciudad sea sustancialmente diferente.*

En este contexto, al denotar el tiempo promedio de espera de estos dos tipos de llamadas por θ_X y θ_Y, respectivamente, el sistema de hipótesis de interés es de la forma (4.7.4). Dadas dos muestras de las dos poblaciones, para aplicar la prueba de razón de verosimilitud, solo necesitamos conocer los medios muestrales $\bar{X}$ y $\bar{Y}$, además de verificar que la distribución exponencial es apropiada para describir ambas muestras, éste se puede corroborar con una sencilla QQ plot para la distribución exponencial. Si en dos muestras de tamaño 54 y 38, las estimaciones muestrales fueron $\bar{x} = 0.83$ minutos y $\bar{y} = 0.79$ minutos, entonces la estadística $2\ln\lambda = 0.054$, el cual es claramente más pequeño que los percentiles $\chi^2_{1,1-\alpha}$ para los valores comunes de α en la práctica, y podemos concluir que con base en las muestras, $\theta_X = \theta_Y$ puede ser una hipótesis apropiada para las dos poblaciones, es decir, el tiempo de espera no es influenciado por la ciudad de donde provienen las llamadas.

4.8 Acerca del p-valor

4.8.1 Diversos puntos de vistas acerca del p-valor

El libro ≪The cult of statistical significance≫ de Ziliak y McCloskey (2008) tiene un buen punto acerca de los tópicos de la inferencia estadística y se basa en una crítica científica a la mala costumbre de los estadísticos en la prueba de hipótesis. Los autores se preguntan ¿por qué las decisiones científicas están restringidas a un espacio discreto binario $\{0,1\}$ inducido por una regla de decisión? Los autores del libro sugieren que tendría más sentido científico que las decisiones estuvieran sujetas a una función de pérdida continua en el intervalo $(0,1)$.

Tiene sentido, más aún cuando a la hora de realizar contrastes sea cual sea la rama de aplicación (econometría, mercadeo, epidemiología, ciencia política, etc.), siempre se utiliza la misma regla de decisión que Fisher impuso hace varias décadas: si el valor p es menor que 0.05, entonces rechace la hipótesis. Pero la verdad que todos sabemos, y

a veces no queremos aceptar, es otra. A continuación se presenta un ejemplo detallado adaptado de las primeras páginas del libro.

Imagínese que usted y su pequeño niño de cuatro años caminan por una de las aceras de la ciudad. Se detienen en una esquina y compran un perro caliente (hot dog). El vendedor del carrito de perros lo atiende muy amablemente y le da justo lo que usted pidió. El semáforo se va a poner en rojo pero usted se atreve a cruzar la calle. Situación número uno: cuando va a llegar a la otra acera, usted se da cuenta que el vendedor olvidó colocar mostaza en su perro. Si usted y su hijo se atreven a devolverse y cruzar la calle esquivando carros, motos y tractomulas, existe una probabilidad, digamos 0.95, de que logren tener la mostaza en su perro caliente sin que haya ocurrido ningún accidente. Situación número dos: cuando usted va a llegar a la otra acera, usted se da cuenta que olvido a su hijo y cuando voltea su mirada, el niño está intentando cruzar la calle. Inmediatamente usted se devuelve esquivando carros, motos y tractomulas. Existe una probabilidad de 0.95 de que usted alcance a su hijo y llegue a la otra acera de la calle sano y salvo.

Dos situaciones con dos premios distintos, la mostaza o su hijo, y con la misma probabilidad. La significación estadística ignora esta diferencia puesto que las dos decisiones son iguales en cuanto a la probabilidad de "éxito". Ambas variables NIÑO y MOSTAZA son significativas si $p < 0.05$ y la conclusión sería: existen dos razones, que son igualmente importantes, para cruzar la calle.

Es claro que lo anterior es un punto muy bueno. Los métodos estadísticos deben tener validez teórica desde el punto de vista del usuario. De esta manera, en epidemiología, economía, contaduría, sociología o en marketing, los métodos estadísticos tienen validez siempre y cuando sirvan para apoyar la teoría desarrollada en estas áreas del conocimiento. No le digan a un gerente de mercadeo que la variable satisfacción del consumidor no entra en el modelo de regresión porque el beta no resultó significativo. En estos aspectos, la ciencia estadística debe ser vista como una herramienta. Ahora, como sucede con toda herramienta, es necesario adecuarla al terreno y afinarla de tal manera que se convierta en una herramienta indispensable en manos del experto y no en una más de las herramientas a las cuales se puede acceder. En esta ocasión, le tocó el turno al análisis de correspondencias. Específicamente al muy conocido y bien ponderado mapa perceptual, resultado de este análisis.

Jim Berger ha diseñado un software que demuestra que las interpretaciones usuales acerca de los p-valores pueden ser erradas. Al respecto, John Cook hace una lista de cinco autores que tienen puntos de vista muy críticos acerca de la práctica e interpretación usual del estadístico con respecto al procedimiento de las pruebas de hipótesis.

Andrew Gelman afirma que en la realidad, la hipótesis nula es siempre falsa. ¿Es el tratamiento A igual de efectivo al tratamiento B? Seguramente no. Está claro que antes de la realización de un experimento deben existir algunas diferencias que se pueden manifestar con un número suficiente de datos.

Según Jim Berger, un p-valor pequeño implica que los datos recolectados son inverosímiles bajo la hipótesis nula. Sin embargo, también pueden serlo bajo la hipótesis alternativa. Las comparaciones de las hipótesis deberían estar condicionadas a la realización de los datos.

Stephen Ziliak y Deirdra McCloskey indican que la significación estadística no es lo mismo que la significación científica. La cuestión más importante para la ciencia es el tamaño de un efecto y no si existe o no tal efecto.

Para William Gosset, el error estadístico es sólo uno de los componentes del error real y quizás sea un componente pequeño.

John Ioannidis, por último, señala que los p-valores pequeños no implican una probabilidad pequeña de que la hipótesis nula sea incorrecta. En una revisión de estudios médicos se encontró que el 74 % de los estudios con p-valores menores que 0.05 llegaba a conclusiones erróneas.

Algun extremista diría que la herramienta de las pruebas de hipótesis y de sus respectivos p-valores es una mala herramienta. Nuestro punto de vista es que cuando se entiende que un p-valor es una variable aleatoria, entonces las conclusiones y por consiguiente la toma de decisiones se hacen con más cuidado. Sin embargo, existe otra herramienta estadística que puede ser usada como complemento a los p-valores. Se trata de los factores de Bayes que son la razón entre las probabilidades a posteriori de las dos hipótesis, dada la realización de los datos. Según John Cook, los factores de Bayes no tienen las debilidades de las pruebas de hipótesis, especialmente las que señalan los criticismos de Jim Berger y John Ioannidis.

4.8.2 p valores aleatorios

En esta época de avances computacionales, una lección de intervalos de confianza incluye, además de teoría, simulaciones que tienden a enfatizar el carácter aleatorio de los límites de los intervalos de confianza: un parámetro se fija y el 95 % de los intervalos construidos en la simulación lo cubren. Pero qué pasa con la enseñanza de otros conceptos fundamentales de la inferencia estadística. En esta entrada vamos a enfocarnos en una metodología alternativa en la enseñanza del p valor.

La respuesta que muchos usuarios de la estadística, no estadísticos, encuentran frente a la pregunta ¿qué es un p valor? ¿es un p valor la probabilidad de que la hipótesis nula (H_o) sea cierta?.

La anterior respuesta es, además de pragmática y utilitarista, falsa. Lo cierto es que, técnicamente, la definición de p valor es la siguiente: un p valor es la probabilidad, calculada al asumir que Ho es cierta, de que la estadística de prueba tome valores tan o más extremos que los calculados con la muestra actual.

Ahora, dado que las estadísticas de prueba se construyen para cuantificar las desviaciones de la hipótesis nula con los datos actuales, entonces rechazamos Ho cuando el p valor es pequeño porque si éste es pequeño entonces los datos actuales proveen una fuerte evidencia en contra de Ho. En otras palabras, el hecho de que el p valor sea grande hace que Ho sea difícil de rechazar; por tanto, es casi intuitivo, pero no valido, tomar al p valor como una medida de soporte en contra (o a favor) del rechazo de Ho.

Sin embargo, esta presentación estándar esconde la aleatoriedad del p valor. El p valor es una estadística, por tanto es aleatorio y no puede ser interpretado como una medida de soporte. Se sugiere, siguiendo los lineamentos de Murdoch (2008), que la

enseñanza de este importante concepto siga una metodología alternativa, basada en simulaciones, totalmente diferente a lo que hasta ahora se está realizando. Con un simple ejemplo es posible que el estudiante entienda que un p valor es una cantidad aleatoria condicionada a las realizaciones de las variables aleatorias de la muestra y, por consiguiente, será posible liberarnos de las definiciones incorrectas que pueden guiar a malinterpretaciones en el campo aplicado.

Considere una prueba t, basada en una muestra aleatoria de tamaño n y con distribución $N(\mu, 1)$, apoyada en el siguiente sistema de hipótesis

$H_0: \ \mu = 0$ vs. $H_a: \ \mu \neq 0$.

Es claro que la estadística de prueba sigue una distribución t-student con (n-1) grados de libertad. Para presentar los resultados, es conveniente empezar con H_0 : $\mu = 0$.

Bajo la hipótesis nula, el histograma de los p valores toma la forma de una distribución plana y uniforme sobre el intervalo [0, 1]. Para enfatizar el punto de que un p valor no es la probabilidad de que Ho sea cierto, el instructor sólo necesita explicar este histograma, en donde claramente Ho es cierta; sin embargo, el sp valor está uniformemente distribuido entre cero y uno. Bajo la hipótesis alternativa, la distribución de los p valores no es uniforme (ver Figura 4.22). Es claro que el chance de obtener p valores menores al nivel de significación será más alto bajo la hipótesis alterna que bajo la hipótesis nula y ese efecto es más claro a medida que mu incrementa su valor.

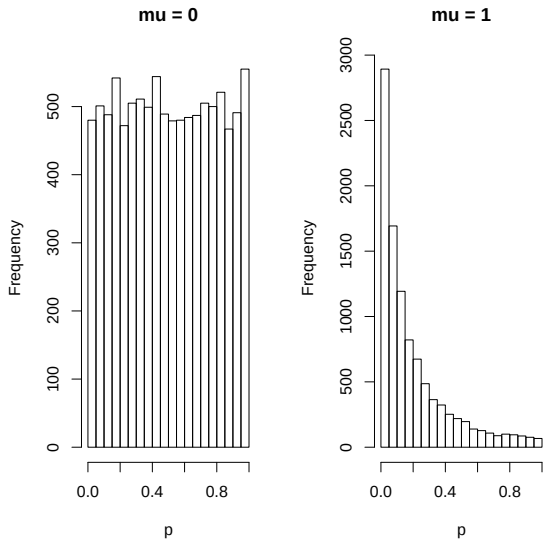

Figura 4.22: p *valores aleatorios*

Ahora, consideramos el sistema de la forma

$H_0: \ \mu = 0$ vs. $H_a: \ \mu < 0$.

Bajo H_a, la distribución de los p valores sobre el intervalo [0, 1] no será uniforme y tenderá al valor uno (ver Figura 4.23). De esta forma, queda claro que la distribución de los p valores no está determinada por el sistema de hipótesis sino por los parámetros.

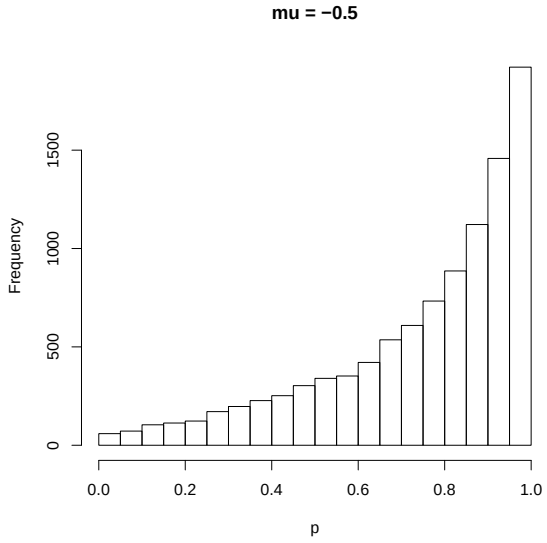

Figura 4.23: p *valores aleatorios*

El programa en R de la simulación de los p valores que generaron las anteriores gráficas se encuentra a continuación.

```
set.seed(7654321,kind=NULL)
nsim<-10000
p<-rep(NA,nsim)
par(mfrow=c(1,2))
#################### H0 mu=0 ####################
for(m in 1:nsim){
n<-4
mu<-0.5
x<-rnorm(n,mu,1)
xbar<-mean(x)
s<-sqrt(var(x))
T<-abs((xbar-mu)/(s/sqrt(n)))
p[m]<-2*(1-pt(T,n-1))
#-----------------------
if(floor(m/1000)==m/1000) print(m)
}
hist(p,main=c("mu = 0"))
#################### Ha mu=1 ####################
```

```
for(m in 1:nsim){
n<-4
mu<-0
x<-rnorm(n,mu,1)
xbar<-mean(x)
s<-sqrt(var(x))
T<-abs((xbar-1)/(s/sqrt(n)))
p[m]<-2*(1-pt(T,n-1))
#-----------------------
if(floor(m/1000)==m/1000) print(m)
}
hist(p,main=c("mu = 1"))
#################### H0 mu<0 ######################
for(m in 1:nsim){
n<-4
mu<--0.5
x<-rnorm(n,mu,1)
xbar<-mean(x)
s<-sqrt(var(x))
T<-(xbar-0)/(s/sqrt(n))
p[m]<-1-pt(T,n-1)
#-----------------------
if(floor(m/1000)==m/1000) print(m)
}
hist(p,main=c("mu = -0.5"))
```

Nota bibliográfica

Erin Leahey, en un reciente artículo, escribe acerca del uso del nivel de significación en pruebas estadísticas, el valor 0.05 y el sistema de tres estrellas que se han convertido en métodos legítimos y dominantes en la mayoría de las investigaciones de tipo social. De acuerdo a Erin, el sistema de hipótesis merece una estrella cuando el p-valor es menor de 0.05, dos estrellas si el p-valor es menor de 0.01 y tres estrellas si el p-valor es menor de 0.001. Erin atribuye el primer uso del nivel de significación 0.05 a Ronald Fisher en su libro publicado en 1935, *Diseño de experimentos*. También nota que otras formas de pruebas de significación eran muy populares en la década de 1930, cuando cerca del 40 % de los artículos publicados en ASR y AJS aplicaban sólo una técnica de prueba de significación.

El famoso 0.05, que nos da de comer a la mayoría de nosotros, fue muy usado desde 1930 hasta 1950, pero declinó hasta 1970. Sin embargo, volvió a revivir hasta nuestra época. Actualmente, cerca del 80 % de los artículos publicados en ASR y AJS emplean ambos procedimientos (nivel de significación y estrellas). El sistema de tres estrellas emergió en la década de 1950, pero se volvió muy popular sólo después de 1970. Un porcentaje cercano al 40 % de artículos publicados en los anteriores journals utiliza la metodología de las tres estrellas.

¿Qué es lo que cuenta en la difusión de tales prácticas? Erin da varios argumentos para responder a esta pregunta. Por ejemplo, ella concluye que los factores institucionales como inversión en investigación y computadores, entrenamiento a nivel de postgrado y la preferencia del editor del journal pueden ser algunos de los factores más importantes en la difusión de tales prácticas. En una conclusión muy interesante, ella encontró que los egresados de Harvard tenían un efecto negativo significativo al adoptar tales prácticas estadísticas.

Por supuesto, este estudio está limitado a la muestra que tomó Erin y no puede ser generalizado. Sin embargo, es una lectura divertida. Si alguien está interesado en los elementos históricos de cómo las prácticas estadísticas fueron introducidas y comenzaron a legitimarse en la investigación social, Camic y Xie (1994) es un muy buen punto de partida.

4.8.3 El p valor no es una medida de soporte

Schervish (1996) afirma que los p-valores están siendo usados por los usuarios de la estadística como medidas de soporte (además de algunas otras malinterpretaciones) cuando éstos precisamente se caracterizan por carecer de consistencia como medidas de la evidencia a favor de un conjunto de hipótesis. Al respecto, es plausible pensar que si es posible obtener evidencia de que cierto animal es un oso, entonces debe existir también evidencia para afirmar que ese animal es un mamífero. Nótese que en el ejemplo anterior existen dos hipótesis: la primera hace referencia a que el animal es un oso y la segunda a que el animal es un mamífero y, por supuesto, la primera está contenida en la segunda. Ahora, utilizar los p-valores como una medida de soporte a favor de la evidencia de la segunda hipótesis puede ser una muy mala idea.

Una medida de soporte debería satisfacer la siguiente propiedad (muy útil en el contexto de comparaciones múltiples):

> Si una hipótesis H_1 implica una hipótesis H_2, entonces una medida de soporte es coherente si el rechazo de H_2 siempre implica el rechazo de H_1

En otras palabras:

> Si una hipótesis H_1 implica otra H_2, entonces la evidencia a favor de H_2 debe ser al menos tan grande como la evidencia en favor de H_1

Teniendo en cuenta este criterio, se sigue que el p-valor es una pésima medida de soporte. Schervish lo explica con el siguiente ejemplo: suponga que se observa la realización de una variable aleatoria con distribución normal de varianza una y media desconocida. Sea $H_1 : \mu \in (-0.5, 0.5)$ y sea $H_2 : \mu \in (-0.82, 0.52)$, claramente el espacio paramétrico de H_1 está contenido en H_2 y, por consiguiente, H_1 implica H_2.

Schervish (1996), citando el libro de Lehmann (1986), establece la regla de decisión para sistemas del tipo $H_0 : \mu \in (\mu_1, \mu_2)$ vs. $H_0 : \mu \notin (\mu_1, \mu_2)$ basado en una observación como rechazar H_0 si $|x - 0.5(\mu_1 + \mu_2)| > c$ donde c es una constante que

satisface $\Phi(0.5(\mu_1-\mu_2)-c)+\Phi(0.5(\mu_2-\mu_1)-c)=\alpha$. El p-valor de esta regla de decisión se puede calcular como

$$p-\text{valor}=\begin{cases}\Phi(x-\mu_1)+\Phi(x-\mu_2) & \text{si } x<0.5(\mu_1+\mu_2)\\ \Phi(\mu_1-x)+\Phi(\mu_2-x) & \text{si } x\geq 0.5(\mu_1+\mu_2)\end{cases}$$

Ahora, si la observación correspondió a $x=2.18$, entonces para probar $H_1:\mu\in(-0.5,0.5)$, $\mu_1=-0.5$, $\mu_2=0.5$ y $x>0.5(\mu_1+\mu_2)$, por consiguiente el p-valor se calcula como $\Phi(-0.5-2.18)+\Phi(0.5-2.18)=0.05016$. Por otro lado, si se desea probar $H_2:\mu\in(-0.82,0.52)$, entonces $\mu_1=-0.82$, $\mu_2=0.52$ t de nuevo $x>0.5(\mu_1+\mu_2)$, y el p-valor se calcula como $\Phi(-0.82-2.18)+\Phi(0.52-2.18)=0.0498$. Lo anterior implica que, tomando el p-valor como medida de soporte, existe más evidencia a favor de H_1 que a favor de H_2, lo cual es contradictorio con el sentido común. Más aún, si el nivel de significación es de 0.05, la regla de decisión implicaría que debemos rechazar H_2 y aceptar H_1. En otras palabras, la media de la distribución puede estar entre (-0.5, 0.5), pero de ninguna manera puede estar entre (-0.82, 0.52), lo cual es muy contradictorio.

Lo anterior nos ilustra que en la práctica, cuando tenemos varios sistemas de hipótesis, cada uno debe ser tratado de manera individual. Es errado tomar decisiones acerca de un sistema usando como base la decisión tomada acerca de otro sistema o lo que puede llamar la transitividad, tal como se ilustró anteriormente.

4.8.4 Acerca de la igualdad en la hipótesis nula

Para terminar el tema de prueba de hipótesis, discutimos la posibilidad de incluir el signo igual en la hipótesis nula. Suponga que se dispone de una muestra aleatoria de tamaño n denotada por $X_1,\cdots,X_n$ y proviene de una distribución $N(\mu,\sigma_0^2)$ con μ desconocido y σ_0^2 conocido. Y supongamos que el sistema de hipótesis de interés es

$$H_0:\ \mu>\mu_0 \quad vs.\ H_1:\ \mu\leq\mu_0, \tag{4.8.1}$$

donde la igualdad $\mu=\mu_0$ está incluida dentro del espacio paramétrico especificado por la hipótesis alterna, H_1. En la teoría estadística, no se establece con claridad si se pueden plantear sistemas de esta forma. Jortiz & Zhang (2010) encontró que en el sistema (4.8.1), se puede encontrar una regla de decisión sin mayores complicaciones, pero a pesar de esto, al incluir la igualdad en H_1, se conducirá a una contradicción con el estimador de máxima verosimilitud, y por consiguiente se recomienda poner la igualdad en la hipótesis nula H_0.

La hipótesis alterna del sistema (4.8.1) especifica espacios paramétricos que incluyen valores menores o iguales a μ_0, entonces es natural pensar en rechazar H_0 cuando $\bar{X}<K$ para alguna constante K. Ahora, para determinar el valor de K y completar la regla de decisión, recurrimos a la definición del tamaño de una prueba dada por

$$\alpha=sup\{P(\text{rechazar } H_0)\} \text{ cuando } H_0 \text{ es verdadera.} \tag{4.8.2}$$

En el sistema específico (4.8.1), la anterior definición se convierte en

$$\alpha = sup\left\{P(\bar{X} < K)\right\} \text{ cuando } H_0 \text{ es verdadera.} \tag{4.8.3}$$

Para determinar el valor de K es necesario encontrar la distribución nula de $\bar{X}$, esto es, la distribución de $\bar{X}$ cuando H_0 es verdadera. Ahora, la hipótesis nula $\mu > \mu_0$ es equivalente a $\mu = \mu^*$ con $\mu^* > \mu_0$. De esta forma, tenemos que

$$\frac{\sqrt{n}(\bar{X} - \mu^*)}{\sigma_0} \sim \ N(0,1). \tag{4.8.4}$$

Por lo tanto, la definición (4.8.3) se convierte en

$$\alpha = sup\left\{\mu^* > \mu_0 : \ P\left(\frac{\sqrt{n}(\bar{X} - \mu^*)}{\sigma_0} < \frac{\sqrt{n}(K - \mu^*)}{\sigma_0}\right)\right\} \tag{4.8.5}$$

$$= sup\left\{\mu^* > \mu_0 : \ \Phi(\frac{\sqrt{n}(K - \mu^*)}{\sigma_0})\right\}, \tag{4.8.6}$$

donde $\Phi(\cdot)$ denota la función de distribución correspondiente a la distribución normal estándar. Como $\Phi(\frac{\sqrt{n}(K - \mu^*)}{\sigma_0})$ es una función decreciente de μ^*, entonces el supremo del conjunto $\left\{\mu^* > \mu_0 : \ \Phi(\frac{\sqrt{n}(K - \mu^*)}{\sigma_0})\right\}$ se da cuando $\mu^* = \mu_0$. Y en este caso, se tiene que $\alpha = \Phi(\frac{\sqrt{n}(K - \mu_0)}{\sigma_0})$, y podemos obtener el valor de K como

$$K = \mu_0 + \Phi^{-1}(\alpha)\frac{\sigma_0}{\sqrt{n}} = \mu_0 + z_\alpha \frac{\sigma_0}{\sqrt{n}}, \tag{4.8.7}$$

y así, encontramos a la regla de decisión: rechazar H_0 si $\bar{X} < \mu_0 + z_\alpha \frac{\sigma_0}{\sqrt{n}}$, equivalente a rechazar H_0 si $\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma_0} < z_\alpha$ y el tamaño de la prueba es α. Nótese que

- La regla de decisión encontrada anteriormente coincide con la regla de decisión del sistema de hipótesis $H_0 : \ \mu \geq \mu_0 \quad vs. \ H_1 : \ \mu < \mu_0$. Es decir, el hecho de que H_1 incluye la igualdad no afectó matemáticamente en la regla de decisión.
- La razón por la que la inclusión de la igualdad en H_1 no influyó en la regla de decisión es el hecho de que el tamaño de una prueba se define con término del supremo tal como se definió en (4.8.3). Si en lugar del supremo se usara el máximo, el anterior procedimiento ya no será válido, y no será posible matemáticamente encontrar una regla de decisión de tamaño α.

Ahora, veamos que esta regla de decisión puede conducir a una contradicción con el estimador de máxima verosimilitud en algunos casos.

Dada la regla de decisión, el p-valor se calcula como $\Phi(z)$ con z el valor que toma la estadística de prueba $\sqrt{n}(\bar{X} - \mu_0)/\sigma_0$. Suponga que en una muestra observada, el estimador de máxima verosimilitud dio exactamente el valor μ_0, es decir, $\bar{x} = \mu_0$. En este caso, $z = 0$, y el p-valor será $\Phi(0) = 0.5$ y nos lleva a la decisión de aceptar $H_0 : \ mu > \mu_0$ el cual no incluye el valor de μ_0. Esto es una clara contradicción, puesto que se rechaza el valor de $\bar{x}$ como posible valor de μ.

4.9 Ejercicios

4.1 Para las siguientes situaciones, plantee un sistema de hipótesis donde el error tipo I es más grave que el error tipo II.

(a) El gerente de una empresa de ventas desea conocer el rendimiento del empleado Pérez, y el índice que el gerente tiene en cuenta es el porcentaje de ventas exitosas p, y el gerente cree que un porcentaje inferior a los 20 % es indicio de un desempeño pobre. Desde el punto de vista del gerente ¿cómo se puede plantear el sistema de hipótesis?

(b) En un departamento del país se modifica la fecha para pagar los impuestos prediales, y la entidad está interesada en conocer si la población está enterada de ese cambio, y además la entidad considera que con las publicidades realizadas, más del 80 % de la población conoce del cambio, ¿cómo se puede plantear el sistema?

4.2 Sea $X_1, \cdots, X_n$ proveniente de una población $N(\mu, \sigma^2)$ donde $\sigma^2 = \sigma_0^2$ es conocida, para el siguiente sistema de hipótesis

$$H_0 : \ \mu \geq \mu_0 \quad vs. \quad H_1 : \ \mu < \mu_0,$$

Desarrolle una regla de decisión para el anterior sistema. Además, especifique la estadística de prueba, dibuje la región de rechazo y obtenga la fórmula del p valor y la función de potencia. ¿El uso de cuál intervalo de confianza es equivalente al uso de la regla de decisión encontrada?

4.3 Repetir el anterior ejercicio con el sistema de hipótesis

$$H_0 : \ \mu \leq \mu_0 \quad vs. \quad H_1 : \ \mu > \mu_0$$

con σ^2 desconocida.

4.4 Sea $X_1, \cdots, X_n$, proveniente de una población $N(\mu, \sigma^2)$ donde σ^2 es desconocida. Encuentre una regla de decisión para el siguiente sistema de hipótesis:

$$H_0 : \ \mu = \mu_0 \quad vs. \quad H_1 : \ \mu < \mu_0.$$

(a) Sin usar la prueba de razón de verosimilitud.

(b) Usando la prueba de razón de verosimilitud.

4.5 Un tipo de bombillo industrial debe tener la vida útil promedio superior a 12 mil horas para poder entrar al mercado colombiano. Para verificar que cumplen con este requisito, se seleccionó 20 bombillos y los resultados del laboratorio indican que la vida útil de estos 20 bombillos es repectivamente (en miles de horas): 11.4, 12.1, 12.5, 13.1, 12.6, 11.9, 12.4, 13.1, 14.0, 11.9, 13.1, 12.8, 12.6, 13.2, 12.4, 11.6, 13.0, 12.4, 12.6, 12.5.

(a) Encuentre una distribución apropiada para los datos.

(b) Plantee un sistema de hipótesis adecuado y aplique el procedimiento adecuado (calculando el p valor) para decidir si los bombillos pueden o no entrar al mercado colombiano.

4.6 Demuestre las expresiones (4.2.23) y (4.2.24).

4.7 Sea $X_1, \cdots, X_n$ proveniente de una población $N(\mu, \sigma^2)$, encuentre una regla de decisión, p valor y la función de potencia para el sistema de hipótesis

$$H_0: \ \sigma^2 = \sigma_0^2 \quad vs. \quad H_1: \ \sigma^2 < \sigma_0^2.$$

cuando

(a) μ es conocido.

(b) μ es desconocido.

4.8 Para los datos del Ejercicio 4.4, plantee un sistema de hipótesis para corroborar o refutar la hipótesis de que la vida útil entre los bombillos se difiere a lo más por 500 horas.

4.9 Escriba la fórmula del p valor y la función de potencia para el sistema $H_0: \ \mu_X = \mu_Y$ vs. $H_0: \ \mu_X \neq \mu_Y$ cuando las varianzas son iguales pero desconocidas. Utilice la fórmula encontrada para calcular el p valor para los datos del Ejemplo 4.3.1, y compare con el p valor obtenido en el ejemplo con el comando `t.test`.

4.10 Sea $X_1, \cdots, X_n$ y $Y_1, \cdots, Y_m$ dos muestras aleatorias independientes provenientes de $N(\mu_X, \sigma_X^2)$ $N(\mu_Y, \sigma_Y^2)$ respectivamente. Encuentre reglas de decisión para los siguientes sistemas de hipótesis.

(a) $H_0: \ \mu_X - \mu_Y \leq \mu_0 \quad vs. \quad H_1: \ \mu_X - \mu_Y > \mu_0$ con $\sigma_X^2 = \sigma_Y^2$ desconocidas.

(b) $H_0: \ \sigma_X^2 = \sigma_Y^2 \quad vs. \quad H_1: \ \sigma_X^2 \neq \sigma_Y^2$ con μ_X y μ_Y conocidos.

4.11 Un ganadero desea aumentar la producción lechera diaria de sus vacas, y decide probar un nuevo concentrado. Para verificar la efectividad del nuevo concentrado, el ganadero separa 35 vacas, de las cuales 15 son alimentadas con el concentrado actual y las restantes con el concentrado nuevo. Después de tres semanas de alimentación, él toma nota de la producción lechera. Para las vacas alimentadas con el concentrado actual, los resultados fueron (en litros): 16.4, 18.9, 15.7, 20.2, 16.8, 19.4, 14.7, 17.8, 19.5, 16.8, 18.4, 14.6, 20.7, 21.1, 17.3, y para las vacas alimentadas con el concentrado nuevo, los resultados fueron: 19.4, 18.1, 21.0, 20.4, 20.5, 17.4, 19.6, 18.4, 21.4, 19.2, 15.7, 22.8, 21.6, 17.2, 18.4, 19.4, 20.5, 23.6, 18.4, 18.3. Contesta las siguientes preguntas usando técnicas de prueba de hipótesis

(a) Verifique si los datos pueden ser descritos de manera adecuada con la distribución normal.

(b) ¿Los dos tipos de concentrados son iguales de efectivos? Si la respuesta es negativa, ¿cuál es más efectivo?

(c) ¿Cuál concentrado produce resultados más homogéneos?

4.12 Verifique las expresiones (4.4.3), (4.4.4), (4.4.5) y (4.4.6).

4.13 Suponga que para aumentar la producción de la cosecha arrocera hay tres tipos de abonos. Para conocer cuál de los tres tipos de abonos es el más eficiente, se divide una finca de 100 hectáreas con cultivos de arroz en 4 grupos de 25 hectáreas cada uno, y a 3 de 4 grupos se les aplican los 3 abonos, y al grupo restante no se le aplica ningún abono. Para cada una de las 100 hectáreas se mide la producción arrocera (en toneladas), los datos se muestran en la Tabla 4.12.

Abono	Producción (en toneladas)
A	3.0, 3.1, 2.6, 2.6, 2.1, 2.9, 2.5, 3.5, 3.3, 3.6, 3.9, 3.8, 1.8 2.8, 3.7, 3.8, 3.1, 2.8, 2.4, 3.0, 2.6, 2.9, 3.9, 4.2, 3.4
B	3.3, 3.6, 4.1, 2.6, 3.1, 2.9, 3.0, 2.7, 2.7, 3.4, 2.7, 3.3, 3.1 2.8, 3.4, 3.3, 3.1, 2.9, 2.8, 2.9, 2.3, 3.0, 3.2, 3.1, 3.1
C	5.1, 5.1, 5.5, 4.9, 5.1, 5.6, 5.0, 5.5, 4.7, 5.7, 5.3, 4.5, 4.4 5.1, 5.1, 5.1, 5.1, 4.9, 4.6, 4.9, 4.9, 4.5, 4.4, 5.4, 4.6
D	5.8, 5.8, 5.6, 5.3, 6.1, 5.7, 6.0, 5.6, 6.1, 5.9, 5.1, 5.3, 5.0 5.6, 5.3, 5.5, 4.6, 5.4, 6.0, 6.0, 5.6, 5.7, 5.2, 5.9 , 5.6
Ninguno	3.1, 2.9, 2.9, 3.1, 3.4, 2.5, 3.3, 2.7, 3.2, 1.9, 2.5, 2.9 3.1, 2.2, 2.4, 2.9, 3.2, 2.5, 2.2, 3.0, 2.7, 4.0, 2.7, 3.3, 3.6

Tabla 4.12: *Datos del Ejercicio 4.13.*

(a) Lleve a cabo un procedimiento de prueba de hipótesis para ver si hay diferencias significativas en términos de producción entre la aplicación y no de abonos.

(b) Lleve a cabo un procedimiento de prueba de hipótesis para ver si hay diferencias significativas entre los diferentes tipos de abonos.

(c) Si hay diferencia entre los diferentes tipos de abonos, diga cuáles abonos son similares y cuáles son diferentes.

4.14 Sea $X_1, \cdots, X_n$ proveniente de una población $Ber(p)$, encuentre una regla de decisión para el sistema de hipótesis usando la razón de verosimilitud:

$$H_0 : \ p = p_0 \quad vs. \quad H_1 : \ p > p_0,$$

4.15 Demuestre la expresión (4.5.6).

4.16 Repita el Ejemplo 4.5.1 usando la prueba exacta. Compare el resultado obtenido con el de la prueba asintótica.

4.17 Una semana antes de comenzar las elecciones presidenciales, a los estudiantes mayores de 18 años de una universidad privada se les pregunta: ¿usted va a votar en estas elecciones?. Se observó que entre los 312 estudiantes entrevistados, 198 respondieron sí a la pregunta. Suponga que el parámetro de interés es el porcentaje de estudiantes que votarán este 30 de mayo.

(a) ¿Se puede afirmar que más de la mitad de los estudiantes de esta universidad votarán en las elecciones?

(b) Suponga que en una universidad pública, se realizó el mismo estudio, y de los 571 entrevistados, 420 dijeron que sí. ¿Se puede afirmar que el porcentaje de los estudiantes que votarán es el mismo en las dos universidades?

4.18 El director de un hospital, con el fin de mejorar la atención en la sección de urgencias, necesita conocer acerca del número de clientes que llegan en una hora a urgencias. Si durante dos semanas se observa el número de pacientes durante una misma hora, y los resultados son 5, 6, 3, 6, 7, 3, 5, 4, 8, 1, 5, 4, 1 y 6.

(a) ¿Qué distribución parece ser apropiada para estos datos?

(b) Si la hipótesis que maneja el hospital antes del estudio es que en promedio llegan 5 pacientes por hora, ¿qué conclusión se puede obtener acerca de esta hipótesis basándose en los datos observados?

4.19 En el contexto planteado en el ejercicio anterior, en época de invierno, más pacientes recurren a urgencias. Si los datos del ejemplo anterior denotan el número de pacientes en días normales, y en épocas de invierno, los datos registrados en dos semanas fueron 8, 6, 13, 8, 7, 8, 3, 10, 8, 7, 5, 9, 7 y 8. ¿Se puede afirmar que el número de clientes que llegan a urgencias en una hora es diferente en invierno?

4.20 Sea $X_1, \cdots, X_n$ proveniente de una población $Pois(\theta)$, encuentre una regla de decisión para el sistema de hipótesis:

$$H_0: \ \theta \geq \theta_0 \quad vs. \quad H_1: \ \theta < \theta_0.$$

4.21 Repita el Ejemplo 4.6.1 usando la prueba exacta. Compare el resultado obtenido con el del ejemplo.

4.22 Para los datos del Ejercicio 3.16, plantee un sistema de hipótesis para probar que desde el momento de contacto, la enfermedad demora menos de 20 horas en manifestar síntomas.

4.23 Se desea comparar dos marcas de bombillos, Philips y Filips, para eso se realiza un estudio donde se pone a funcionar bombillos de ambas marcas y se registra el tiempo que duran estos bombillos antes de fundirse, de lo cual se obtienen los siguientes datos. Para 100 bombillos de la marca Philips, el tiempo promedio de duración es de 11300 horas, mientras que para 100 bombillos de la marca Filips, el tiempo promedio de duración es de 8050 horas. ¿Se puede afirmar que las dos marcas tienen la misma calidad?

4.24 Demuestre la expresión (4.7.2), (4.7.3) y (4.7.5).

Parte II

Inferencia estadística multivariante

Capítulo 5

Distribuciones multivariantes

Hasta este punto, hemos estudiado situaciones donde se mide una variable aleatoria en diferentes individuos en una muestra. En este capítulo comenzamos a estudiar situaciones donde se mide más de una variable en diferentes individuos. En estos casos no solo estudiamos propiedades de cada variable sino también la posible relación que existe entre ellas. Comenzaremos repasando brevemente algunos conceptos relacionados. Para más detalles, el lector puede consultar Anderson (1984) y Johnson & Wichern (1998).

5.1 Vectores aleatorios

Uno de los conceptos más importantes de las distribuciones multivariantes es el concepto de vectores aleatorios que introducimos a continuación.

Definición 5.1.1. *Un vector $(X_1, \cdots, X_p)'$ cuyos componentes X_1, $\cdots$, X_p son variables aleatorias se llama un vector aleatorio. p es la dimensión del vector.*

Estudiar un vector aleatorio de dimensión p equivale a estudiar las p variables componentes X_1, $\cdots$, X_p, y la razón por la cual lo hacemos conjuntamente es porque posiblemente haya estructuras de dependencia entre estas variables de interés. Consideramos los siguientes ejemplos de vectores aleatorios.

Ejemplo 5.1.1. *El primer ejemplo se trata de los juegos de azar. Suponga que se lanza un dado 2 veces, y sea X_1 el número de veces que se obtiene el número 6, y X_2 el número de veces que se obtiene un número par. Entonces X_2 siempre toma valores mayores o iguales que la variable X_1. Entonces la distribución de X_1 y X_2 es como se ilustra en la Tabla 5.1.*

Utilizando las distribuciones de X_1 y X_2, tenemos que $E(X_1) = 7/36$ y $E(X_2) = 1$. Por otro lado, la variable X_1X_2 toma los valores 0, 1, 2 y 4 con las probabilidades $25/36$, $6/36$, $4/36$ y $1/36$, respectivamente, de donde $E(X_1X_2) = 1/2$. Y se tiene

que $Cov(X_1, X_2) = 29/36$, mostrando claramente que existe relación lineal positiva entre las variables X_1 y X_2.

X_1	0	1	2
$P(X_1 = x)$	30/36	5/36	1/36
$P(X_2 = x)$	9/36	18/36	9/36

Tabla 5.1: *Distribuciones de probabilidad de las variables X_1 y X_2 en el Ejemplo 5.1.1.*

Ejemplo 5.1.2. *Considere un centro de atención que atiende quejas y reclamos de los usuarios de alguna empresa. Y suponga que se interesa conocer los motivos del incremento del número de clientes que llegan a este centro con quejas y reclamos en los últimos meses. Y se sospecha que uno de los motivos puede ser que los empleados de este centro atienden a los clientes de forma ágil, además de ofrecer un servicio agradable de buena calidad, y por esta razón, los clientes prefieren este centro de atención a los demás. El supervisor del centro de atención instala un dispositivo que pide opinión del cliente acerca del servicio recibido (en una escala de 1 a 5), además de registrar el tiempo de duración de cada cliente atendido.*

Para una hora determinada del día, denotamos X_1 como el tiempo de duración promedio de cliente atendido durante esta hora por todos los trabajadores del centro, y X_2 como la calificación promedio obtenida durante el mismo periodo de tiempo, y X_3 como el número de clientes que llegan al centro dentro de esa hora. Entonces para conocer si X_3 es influenciado por X_1 y X_2 necesitamos saber si hay una estructura de relación lineal entre X_1, X_2 y X_3. Este análisis se puede llevar a cabo usando las herramientas que veremos a lo largo del presente capítulo.

En el caso de una variable aleatoria Y, para conocer acerca de su comportamiento, basta conocer su función de distribución $F_Y(y)$ o su función de densidad $f_Y(y)$. Cuando se dispone de un vector aleatorio $(X_1, \cdots, X_p)$, su función de distribución es simplemente la función de distribución conjunta de los componentes X_1, $\cdots$, X_p. Recordemos algunas definiciones y propiedades asociadas a las distribuciones conjuntas.

Definición 5.1.2. *Sea X_1, $\cdots$, X_p variables aleatorias, se define la función de distribución conjunta como*

$$F_{X_1,\cdots,X_p}(x_1, \cdots, x_p) = Pr(X_1 \leq x_1, \cdots, X_n \leq x_p). \tag{5.1.1}$$

Si las variables son discretas, la función de densidad conjunta se define como

$$f_{X_1,\cdots,X_p}(x_1, \cdots, x_p) = Pr(X_1 = x_1, \cdots, X_n = x_p). \tag{5.1.2}$$

Si las variables son continuas, la función de densidad conjunta se define como

$$f_{X_1,\cdots,X_p}(x_1, \cdots, x_p) = \frac{\partial^p F_{X_1,\cdots,X_p}(x_1, \cdots, x_p)}{\partial x_1 \cdots \partial x_p}. \tag{5.1.3}$$

Cuando los componentes del vector $\mathbf{X}$ son variables independientes, entonces podemos obtener la función de distribución conjunta y la función de densidad conjunta como

$$F_{X_1,\cdots,X_p}(x_1,\cdots,x_p)=\prod_{i=1}^{n}F_{X_i}(x_i).$$

y

$$f_{X_1,\cdots,X_p}(x_1,\cdots,x_p)=\prod_{i=1}^{n}f_{X_i}(x_i).$$

Dado un vector aleatorio o un conjunto de variables aleatorias, se puede encontrar la función de densidad marginal de una variable específica a partir de la función de densidad conjunta, como lo ilustra el siguiente resultado.

Resultado 5.1.1. *Dadas X_1, $\cdots$, X_p variables aleatorias, la función de densidad marginal de X_k con $k=1,\cdots,p$ está dada por:*

$$f_{X_k}(x)=\int_{\mathbb{R}}\cdots\int_{\mathbb{R}}f_{X_1,\cdots,X_p}(x_1,\cdots,x_p)dx_1\cdots dx_{k-1}dx_{k+1}dx_p, \tag{5.1.4}$$

si las variables son continuas. En el caso de que las variables sean discretas, la función de densidad marginal está dada por

$$f_{X_k}(x)=\sum_{x_1}\cdots\sum_{x_{k-1}}\sum_{x_{k+1}}\cdots\sum_{x_p}f_{X_1,\cdots,X_p}(x_1,\cdots,x_p). \tag{5.1.5}$$

Nota: dada la función de densidad de X_1, $\cdots$, X_p, también se puede obtener la función de densidad marginal de un subconjunto de variables integrando o sumando apropiadamente, similar a la definición anterior.

Al igual que las variables aleatorias, los vectores aleatorios tienen características que nos permiten conocer sus comportamientos.

Definición 5.1.3. *Dado un vector aleatorio $\mathbf{X}=(X_1,\cdots,X_p)'$, su esperanza se define como $\boldsymbol{\mu}=E(\mathbf{X})=(E(X_1),\cdots,E(X_p))$.*

Resultado 5.1.2. *Dado un vector aleatorio $\mathbf{X}$ de dimensión p, A una matriz de constantes de tamaño $r\times p$ y b un vector de constantes de tamaño $r\times 1$, se tiene que*

$$E(A\mathbf{X}+b)=AE(\mathbf{X})+b. \tag{5.1.6}$$

Demostración. Tenemos

$$\begin{aligned}
E(A\mathbf{X}+b) &= E\left(\begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{r1} & \cdots & a_{rp} \end{pmatrix}\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_r \end{pmatrix}\right) \\
&= E\begin{pmatrix} a_{11}X_1 + \cdots + a_{1p}X_p + b_1 \\ \vdots \\ a_{r1}X_1 + \cdots + a_{rp}X_p + b_r \end{pmatrix} \\
&= \begin{pmatrix} E(a_{11}X_1 + \cdots + a_{1p}X_p + b_1) \\ \vdots \\ E(a_{r1}X_1 + \cdots + a_{rp}X_p + b_r) \end{pmatrix} \\
&= \begin{pmatrix} a_{11}E(X_1) + \cdots + a_{1p}E(X_p) + b_1 \\ \vdots \\ a_{r1}E(X_1) + \cdots + a_{rp}E(X_p) + b_r \end{pmatrix}.
\end{aligned}$$

Y por otro lado

$$\begin{aligned}
AE(\mathbf{X})+b &= \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{r1} & \cdots & a_{rp} \end{pmatrix}\begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_r \end{pmatrix} \\
&= \begin{pmatrix} a_{11}E(X_1) + \cdots + a_{1p}E(X_p) + b_1 \\ \vdots \\ a_{r1}E(X_1) + \cdots + a_{rp}E(X_p) + b_r \end{pmatrix}.
\end{aligned}$$

De donde se concluye que $E(A\mathbf{X}+b) = AE(\mathbf{X})+b$. □

Resultado 5.1.3. *Dados dos vectores aleatorios* $\mathbf{X}$ *y* $\mathbf{Y}$ *de una misma dimensión* p*, se tiene que* $E(\mathbf{X}+\mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y})$.

Demostración. Tenemos:

$$\begin{aligned}
E(\mathbf{X}+\mathbf{Y}) &= E\begin{pmatrix} X_1+Y_1 \\ \vdots \\ X_p+Y_p \end{pmatrix} \\
&= \begin{pmatrix} E(X_1+Y_1) \\ \vdots \\ E(X_p+Y_p) \end{pmatrix} \\
&= \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix} + \begin{pmatrix} E(Y_1) \\ \vdots \\ E(Y_p) \end{pmatrix} \\
&= E(\mathbf{X}) + E(\mathbf{Y}).
\end{aligned}$$

□

Como consecuencia inmediata de los resultados anteriores, se tiene que

$$E(A\mathbf{X} + B\mathbf{Y} + b) = AE(\mathbf{X}) + BE(\mathbf{Y}) + b,$$

para matrices A, B y b de tamaños apropiados.

Al igual que con el concepto de vector aleatorio, también se puede definir una matriz aleatoria, la cual es una matriz cuyos elementos son variables aleatorias. También se puede definir la esperanza de una matriz aleatoria de manera análoga, esto es, la esperanza de una matriz aleatoria es una matriz de constantes donde cada elemento es la esperanza de la correspondiente variable aleatoria.

Ahora, se define un concepto análogo al concepto de la varianza para una variable aleatoria.

Definición 5.1.4. *Dado un vector aleatorio $\mathbf{X} = (X_1, \cdots, X_p)'$, la matriz de varianzas y covarianzas se define como*

$$\mathbf{\Sigma} = Var(\mathbf{X}) = E[(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})'] \tag{5.1.7}$$

En primer lugar, la anterior definición está bien dada. Nótese que la dimensión de $\mathbf{X} - E\mathbf{X}$ es $p \times 1$, y por consiguiente la dimensión de $(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})'$ es $p \times p$. Es decir, la matriz de varianzas y covarianzas de un vector aleatorio es una matriz cuadrada, cuyo número de filas (columnas) corresponde al número de variables de estudio.

Ahora, para entender la razón del nombre matriz de varianzas y covarianzas, se observa que:

$$\begin{aligned}
\mathbf{\Sigma} &= E[(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})'] \\
&= E\begin{pmatrix} X_1 - EX_1 \\ \vdots \\ X_p - EX_p \end{pmatrix} \begin{pmatrix} X_1 - EX_1, \cdots, X_p - EX_p \end{pmatrix} \\
&= E\begin{pmatrix} (X_1 - EX_1)(X_1 - EX_1) & \cdots & (X_1 - EX_1)(X_p - EX_p) \\ \vdots & \ddots & \vdots \\ (X_p - EX_p)(X_1 - EX_1) & \cdots & (X_p - EX_p)(X_p - EX_p) \end{pmatrix} \\
&= \begin{pmatrix} E(X_1 - EX_1)(X_1 - EX_1) & \cdots & E(X_1 - EX_1)(X_p - EX_p) \\ \vdots & \ddots & \vdots \\ E(X_p - EX_p)(X_1 - EX_1) & \cdots & E(X_p - EX_p)(X_p - EX_p) \end{pmatrix} \\
&= \begin{pmatrix} Var(X_1) & \cdots & Cov(X_1, X_p) \\ \vdots & \ddots & \vdots \\ Cov(X_p, X_1) & \cdots & Var(X_p) \end{pmatrix}.
\end{aligned}$$

En conclusión, el elemento ij, con $i, j = 1, \cdots, p$ corresponde a la covarianza entre las variables X_i y X_j. Recordando que $Cov(X, X) = Var(X)$, se tiene que los p elementos de la diagonal de $\mathbf{\Sigma}$ corresponden a las varianzas de las p variables de estudio, y los elementos fuera de la diagonal son las covarianzas de todas las combinaciones de distintas variables. Nótese que la matriz $\mathbf{\Sigma}$ es simétrica debido a la simetría del operador covarianza. Ahora denotando $Cov(X_i, X_j)$ con σ_{ij} para todo $i, j = 1, \cdots, p$, se tiene la siguiente notación que es común en la literatura estadística

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}.$$

Ahora, recordemos que la covarianza entre dos variables aleatorias mide la relación lineal que existe entre ellas, de tal forma que valores positivos de la covarianza indica que la relación lineal es proporcional, y valores negativos de la covarianza indican que la relación lineal es inversamente proporcional. Finalmente, si la covarianza es nula, entonces se concluye que no hay ninguna relación lineal entre las variables, lo cual no implica que las variables sean independientes, sino que son incorrelacionadas. De hecho, si dos variables son independientes, entonces son incorrelacionadas, pero el recíproco de esta afirmación no se tiene en general.

La matriz de varianzas y covarianzas tiene ciertas propiedades que son útiles en el desarrollo de la estadística multivariada. El siguiente resultado enuncia algunas.

Resultado 5.1.4. *Dado un vector aleatorio $\mathbf{X}$ de dimensión p con $\boldsymbol{\mu}$ la esperanza, y $\mathbf{\Sigma}$ la matriz de varianzas y covarianzas, entonces se tiene que*

(a) $\mathbf{\Sigma}$ *es simétrica,*

(b) $\mathbf{\Sigma}$ *es semidefinida positiva*

(c) $\mathbf{\Sigma} = E(\mathbf{X}\mathbf{X}') - \boldsymbol{\mu}\boldsymbol{\mu}'$

(d) dada A una matriz de constantes de tamaño $r \times p$ y b un vector de constantes de tamaño $r \times 1$, entonces $Var(A\mathbf{X} + b) = A\mathbf{\Sigma}A'$.

Demostración. (a) $\mathbf{\Sigma}$ es simétrica, puesto que su elemento ij-ésimo σ_{ij} corresponde a $Cov(X_i, X_j)$ que es igual a $Cov(X_j, X_i)$. Es decir, $\sigma_{ij} = \sigma_{ji}$ para todo $i, j = 1, \cdots, p$, de donde se completa la prueba.

(b) Para ver que $\mathbf{\Sigma}$ es semidefinida positiva, tomamos un vector u de dimensión $p \times 1$, y veamos que $u'\mathbf{\Sigma}u \geq 0$. Para eso, se define $Y = (\mathbf{X} - \boldsymbol{\mu})'u$ que es una variable aleatoria, y se tiene que $E(Y^2) \geq 0$. Pero

$$\begin{aligned} E(Y^2) &= E(Y'Y) \quad \text{pues } Y' = Y \text{ por } Y \text{ unidimensional} \\ &= E(u'(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'u) \\ &= u'E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'u \\ &= u'\mathbf{\Sigma}u. \end{aligned}$$

De donde se tiene que $u'\boldsymbol{\Sigma}u \geq 0$ para todo vector u, de donde se concluye que $\boldsymbol{\Sigma}$ es semidefinida positiva.

(c) Tenemos que:

$$\begin{aligned}\boldsymbol{\Sigma} &= E\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\right] \\ &= E(\mathbf{X}\mathbf{X}') - \boldsymbol{\mu}E(\mathbf{X}') - E(\mathbf{X})\boldsymbol{\mu}' - E(\boldsymbol{\mu}\boldsymbol{\mu}') \\ &= E(\mathbf{X}\mathbf{X}') - \boldsymbol{\mu}\boldsymbol{\mu}' - \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\mu}\boldsymbol{\mu}' \\ &= E(\mathbf{X}\mathbf{X}') - \boldsymbol{\mu}\boldsymbol{\mu}'\end{aligned}$$

(d) En primer lugar, se tiene que $E(A\mathbf{X}+b) = A\boldsymbol{\mu}+b$, ahora,

$$\begin{aligned}Var(A\mathbf{X}+b) &= E\left[(A\mathbf{X}+b)-(A\boldsymbol{\mu}+b)\right]\left[(A\mathbf{X}+b)-(A\boldsymbol{\mu}+b)\right]' \\ &= E\left[(A\mathbf{X}-A\boldsymbol{\mu})(A\mathbf{X}-A\boldsymbol{\mu})'\right] \\ &= E\left[A(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'A'\right] \\ &= AE\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\right]A' \\ &= A\boldsymbol{\Sigma}A'\end{aligned}$$

□

Los componentes de la matriz de varianzas y covarianzas permiten encontrar las posibles relaciones lineales entre las variables aleatorias. Entre más grande sea la covarianza en absoluto, más relación lineal existe, pero se sabe que la magnitud de una covarianza depende de la escala de medición de las variables, por eso es difícil determinar cuándo una covarianza es grande. Por esta razón, es frecuente el uso del coeficiente de correlación para examinar la relación lineal entre dos variables; en este sentido, se introduce el concepto de la matriz de correlaciones en la siguiente definición.

Definición 5.1.5. *Dado un vector aleatorio $\mathbf{X} = (X_1, \cdots, X_p)'$ con matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$, la matriz de correlación se define como*

$$\boldsymbol{\rho} = \mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2}, \tag{5.1.8}$$

donde $\mathbf{D}$ es la matriz diagonal que contiene las varianzas de las variables X_1, $\cdots$, X_n.

En la anterior expresión, el término $\mathbf{D}^{-1/2}$ se refiere a la inversa de la matriz raíz cuadrada de $\mathbf{D}$, esto es, $\mathbf{D}^{1/2}$. Por ser $\mathbf{D}$ diagonal, $\mathbf{D}^{1/2}$ es una matriz diagonal con elementos de la diagonal iguales a la raíz cuadrada de los elementos de la diagonal de $\mathbf{D}$. El cálculo de la raíz cuadrada de una matriz semidefinida positiva se presenta más adelante.

El elemento ij-ésimo de la matriz $\boldsymbol{\rho}$ corresponde a la correlación entre las variables X_i y X_j; de esta manera, los elementos de la diagonal son iguales a 1 puesto que para cualquier variable X se tiene que $Corr(X,X) = 1$. Para comprobar esto,

$$
\begin{aligned}
\boldsymbol{\rho} &= \mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2} \\
&= \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{pmatrix}^{-1/2} \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{pmatrix}^{-1/2} \\
&= \begin{pmatrix} \sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^{-1} \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix} \begin{pmatrix} \sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^{-1} \end{pmatrix} \\
&= \begin{pmatrix} \sigma_1 & \sigma_{12}\sigma_1^{-1} & \cdots & \sigma_{1p}\sigma_1^{-1} \\ \sigma_{21}\sigma_2^{-1} & \sigma_2 & \cdots & \sigma_{2p}\sigma_2^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1}\sigma_p^{-1} & \sigma_{p2}\sigma_p^{-1} & \cdots & \sigma_p \end{pmatrix} \begin{pmatrix} \sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^{-1} \end{pmatrix} \\
&= \begin{pmatrix} 1 & \sigma_{12}\sigma_1^{-1}\sigma_2^{-1} & \cdots & \sigma_{1p}\sigma_1^{-1}\sigma_p^{-1} \\ \sigma_{21}\sigma_2^{-1}\sigma_1^{-1} & 1 & \cdots & \sigma_{2p}\sigma_2^{-1}\sigma_p^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1}\sigma_p^{-1}\sigma_1^{-1} & \sigma_{p2}\sigma_p^{-1}\sigma_2^{-1} & \cdots & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}.
\end{aligned}
$$

Resultado 5.1.5. *Dado un vector aleatorio $\mathbf{X}$ de dimensión p y $\boldsymbol{\rho}$ la matriz de correlaciones, entonces se tiene que*

(a) $\boldsymbol{\rho}$ es simétrica,

(b) $\boldsymbol{\rho}$ es semidefinida positiva

Demostración. (a) $\boldsymbol{\rho}' = (\mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2})' = \mathbf{D}^{-1/2}\boldsymbol{\Sigma}'\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2} = \boldsymbol{\rho}$.

(b) Sea u un vector de dimensión $p \times 1$, se tiene que $u'\boldsymbol{\rho}u = u'\mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2}u = (\mathbf{D}^{-1/2}u)'\boldsymbol{\Sigma}(\mathbf{D}^{-1/2}u)$ la cual es no negativo, por ser $\boldsymbol{\Sigma}$ semidefinida positiva. Y el resultado queda demostrado.

□

Otro concepto importante con respecto a un vector aleatorio es la función generadora de momentos que, similar al caso univariado, caracteriza la distribución del vector aleatorio. Lo definimos a continuación.

Definición 5.1.6. *Dado un vector aleatorio* $\mathbf{X} = (X_1, \cdots, X_p)'$*, la función generadora de momentos de* $\mathbf{X}$ *es una función vectorial que se define como*

$$M_{\mathbf{X}}(\mathbf{t}) = E(e^{\mathbf{X}'\mathbf{t}}), \tag{5.1.9}$$

donde $\mathbf{t} \in \mathbb{R}^p$.

Análogo a la independencia entre variables aleatorias, también se puede definir la independencia entre dos vectores aleatorios. Diremos que dos vectores (no necesariamente del mismo tamaño) $\mathbf{X}_1$ y $\mathbf{X}_2$ son independientes si la función de densidad conjunta de $\mathbf{X}_1$ y $\mathbf{X}_2$ es igual al producto de las funciones de densidad marginales, esto es

$$f_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2) = f_{\mathbf{X}_1}(\mathbf{x}_1) f_{\mathbf{X}_2}(\mathbf{x}_2) \tag{5.1.10}$$

En términos de la función de densidad condicional, la anterior definición es equivalente a decir que $\mathbf{X}_1$ y $\mathbf{X}_2$ son independientes si

$$f_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1|\mathbf{x}_2) = f_{\mathbf{X}_1}(\mathbf{x}_1) \tag{5.1.11}$$

o

$$f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2|\mathbf{x}_1) = f_{\mathbf{X}_2}(\mathbf{x}_2). \tag{5.1.12}$$

Puesto que

$$\begin{aligned} f_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1|\mathbf{x}_2) &= \frac{f_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2)}{f_{\mathbf{X}_2}(\mathbf{x}_2)} \\ &= \frac{f_{\mathbf{X}_1}(\mathbf{x}_1) f_{\mathbf{X}_2}(\mathbf{x}_2)}{f_{\mathbf{X}_2}(\mathbf{x}_2)} \\ &= f_{\mathbf{X}_1}(\mathbf{x}_1) \end{aligned}$$

Y de manera completamente similar se puede ver (5.1.12). También la ecuación (5.1.10) se puede extender al caso de tener más de dos vectores aleatorios.

Vale la pena aclarar que la razón por la que se estudia conjuntamente p variables aleatorias es tener en cuenta la estructura de independencia que existen entre estas variables; por esta razón, en el caso de que existen dos vectores aleatorios independientes, o equivalentemente, dos conjuntos de variables aleatorias independientes, es más conveniente estudiarlos por separado.

Cuando un conjunto de vectores aleatorios son independientes, se puede intercambiar la esperanza y el producto tal como ilustra el siguiente resultado.

Resultado 5.1.6. *Sean* $\mathbf{X}_1, \cdots, \mathbf{X}_n$ *vectores aleatorios independientes, entonces*

$$E\left(\prod_{i=1}^{n} \mathbf{X}_i\right) = \prod_{i=1}^{n} E(\mathbf{X}_i)$$

Demostración.

$$\begin{aligned} E\left(\prod_{i=1}^{n}\mathbf{X}_i\right) &= \int\cdots\int \mathbf{x}_1\cdots\mathbf{x}_n f(\mathbf{x}_1\cdots\mathbf{x}_n)d\mathbf{x}_1\cdots d\mathbf{x}_n \\ &= \int\cdots\int \mathbf{x}_1\cdots\mathbf{x}_n f(\mathbf{x}_1)\cdots f(\mathbf{x}_n)d\mathbf{x}_1\cdots d\mathbf{x}_n \\ &= \int \mathbf{x}_1 f(\mathbf{x}_1)d\mathbf{x}_1\cdots\int \mathbf{x}_n f(\mathbf{x}_n)d\mathbf{x}_n \\ &= \prod_{i=1}^{n} E(\mathbf{X}_i) \end{aligned}$$

□

Utilizando el anterior resultado, podemos ver que la función generadora de momentos es muy útil para encontrar la distribución de una combinación lineal de un conjunto de vectores aleatorios independientes. Tenemos el siguiente resultado.

Resultado 5.1.7. *Sean $\mathbf{X}_1, \cdots, \mathbf{X}_n$ vectores aleatorios independientes, cada uno de dimensión p, y se denota la función generadora de momentos de $\mathbf{X}_i$ por $M_i(\mathbf{t})$ para todo $i = 1, \cdots, n$. Sea $\mathbf{Y} = \sum_{i=1}^{n} a_i\mathbf{X}_i$ donde $a_1, \cdots, a_n$ son constantes reales, entonces $\mathbf{Y}$ es un vector de dimensión p y su función generadora de momentos está dada por*

$$M_{\mathbf{Y}}(\mathbf{t}) = \prod_{i=1}^{n} M_i(a_i\mathbf{t}).$$

Demostración. Tenemos que

$$\begin{aligned} M_{\mathbf{Y}}(\mathbf{t}) = E(e^{\mathbf{Y}'\mathbf{t}}) &= E(e^{\sum_{i=1}^{n} a_i\mathbf{X}_i}) \\ &= E\left(\prod_{i=1}^{n} e^{a_i\mathbf{X}_i}\right) \\ &= \prod_{i=1}^{n} E(e^{a_i\mathbf{X}_i}) \\ &= \prod_{i=1}^{n} M_i(a_i\mathbf{t}) \end{aligned}$$

□

Ahora, cuando tenemos dos vectores aleatorios $\mathbf{X}_1$ y $\mathbf{X}_2$, y existe cierta relación lineal entre ellos, podemos pensar que los valores que toma $\mathbf{X}_1$ dependen de los valores de $\mathbf{X}_2$. De esta forma, podemos definir la esperanza de $\mathbf{X}_1$ condicionada a que $\mathbf{X}_2$ tome el valor $\mathbf{x}_2$. Esta esperanza se define como

$$E(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2) = \int \mathbf{x}_1 f_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1|\mathbf{x}_2)d\mathbf{x}_1$$

Y también podemos definir la varianza de $\mathbf{X}_1$ condicionada a que $\mathbf{X}_2$ tome el valor $\mathbf{x}_2$ como

$$Var(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2) = E\left[(\mathbf{X}_1 - E(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2))^2 |\mathbf{X}_2 = \mathbf{x}_2\right]$$

Resultado 5.1.8. *Dados dos vectores aleatorios $\mathbf{X}$ y $\mathbf{Y}$, de dimensión p y q, respectivamente. Y sean A y B matrices de constantes de dimensión $r \times p$ y $k \times q$, respectivamente. Entonces se tiene que*

$$Cov(A\mathbf{X}, B\mathbf{Y}) = ACov(\mathbf{X}, \mathbf{Y})B'. \tag{5.1.13}$$

Demostración. Se deja como ejercicio. □

Ejemplo 5.1.3. *Dado un vector aleatorio $\mathbf{X} = (X_1, X_2)'$ con $X_1 \sim N(1, 4)$ y X_2 con distribución exponencial de media 5, y $Cov(X_1, X_2) = 2$, encuentra*

(a) La esperanza, la matriz de varianzas y covarianzas y la matriz de correlaciones de $\mathbf{X}$.

(b) Encuentra la esperanza, la matriz de varianzas y covarianzas y la matriz de correlaciones del vector aleatorio conformado por las variables $X_1 - X2$, $(X_1 + X_2)/2$

5.2 Algunas distribuciones multivariantes

5.2.1 Distribución multinomial

La distribución multinomial es una generalización de la distribución binomial. En la distribución binomial, existe un número n de ensayos donde para cada ensayo existen dos posibles resultados que se etiquetan como éxito o fracaso. En la distribución multinomial, para cada uno de los n ensayos, existen r tipos de posibles resultados. De esta manera, las variables $X_1, \cdots, X_r$ que denotan el número de resultados tipo $1, \cdots, r$ en los n ensayos tiene distribución multinomial. La definición formal de esta distribución se da a continuación.

Definición 5.2.1. *Dado un vector aleatorio $\mathbf{X} = (X_1, \cdots, X_r)'$, se dice que el vector aleatorio $\mathbf{X}$ tiene distribución multinomial (o equivalentemente que las variables $X_1, \cdots, X_r$ tienen distribución multinomial) si su función de densidad de probabilidad está dada por*

$$f_{X_1,\cdots,X_r}(x_1, \cdots, x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r}, \tag{5.2.1}$$

para $x_i = 0, \cdots, n$, $\sum_{i=1}^r x_i = n$ y $\sum_{i=1}^r p_i = 1$. La notación que usaremos es $\mathbf{X} \sim Multi(n, p_1, \cdots, p_r)$.

La distribución multinomial tiene las siguientes propiedades

Resultado 5.2.1. *Si el vector aleatorio $\mathbf{X}$ tiene distribución multinomial con la función de densidad dada por (5.2.1), entonces*

1. $E(\mathbf{X}) = n\mathbf{p}$, *donde* $\mathbf{p} = (p_1, \cdots, p_r)'$
2. $Var(\mathbf{X}) = n[diag(\mathbf{p}) - \mathbf{pp}']$, *con* $diag(\mathbf{p})$ *denotando la matriz diagonal que contiene los elementos de* $\mathbf{p}$.

Gelman, Carlin, Stern & Rubin (2004) utilizan la distribución multinomial para investigar acerca de preferencias de los votantes en elecciones presidenciales.

Ejemplo 5.2.1. *Suponga que para las elecciones presidenciales hay en total tres candidatos, A, B y C. En una encuesta de opinión realizada por una compañía de investigación, se encontró que de las 2436 personas entrevistadas, 810 apoyan al candidato A, es decir, votarían por A, 654 al candidato B y los restantes 972 al candidato C. Si denotamos el número de personas que votarían por los tres candidatos A, B y C como* X_1, X_2 *y* X_3, *respectivamente, entonces el vector* (X_1, X_2, X_3) *se distribuye como* $Multi(n, p_1, p_2, p_3)$ *donde* n *denota el número de votantes, y* p_i *denota la probabilidad de que un votante vote por el candidato* i *con* $i = 1, 2, 3$, *o también se puede interpretar a* p_i *como el porcentaje de votos que obtiene el candidato* i *en las elecciones.*

Suponiendo que se conoce el número total de votantes n, *si antes de la realización de las elecciones, se puede obtener una estimación de las probabilidades* p_i, *se puede calcular la esperanza de* X_1, X_2 *y* X_3, *es decir, el número de votos que se espera obtener para cada uno de los tres candidatos. En el siguiente capítulo, se discutirá sobre cómo obtener estas estimaciones utilizando los datos de una muestra.*

5.2.2 Distribución normal multivariante

Una de las distribuciones multivariantes más importantes es la distribución normal multivariante, cuya definición se enuncia a continuación.

Definición 5.2.2. *Dado un vector aleatorio* $\mathbf{X} = (X_1, \cdots, X_p)'$, *se dice que* $\mathbf{X}$ *tiene distribución normal* p*-dimensional si su función de densidad está dada por*

$$f(\mathbf{x}) = |\mathbf{\Sigma}|^{-1/2}(2\pi)^{-p/2}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}, \tag{5.2.2}$$

donde $\boldsymbol{\mu}$ *es un vector de* $\mathbb{R}^p$ *y* $\mathbf{\Sigma}$ *es una matriz semidefinida positiva y* $|\mathbf{\Sigma}|$ *denota el determinante de* $\mathbf{\Sigma}$. *La notación para este caso es* $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$.

En los métodos estadísticos multivariados en muchos casos es necesario medir la distancia entre datos multivariados y su esperanza, pero teniendo en cuenta la estructura de varianza del vector aleatorio. Esta distancia se conoce como la *distancia de Mahalanobis* que se define como

$$D = \sqrt{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

De esta forma podemos observar que la función de densidad de una distribución normal multivariante se puede escribir como

$$f(\mathbf{x}) = |\mathbf{\Sigma}|^{-1/2}(2\pi)^{-p/2}\exp\{-\frac{1}{2}D^2\}$$

Luego, $f(\mathbf{x})$ es grande para valores de $\mathbf{x}$ cercanos a la media $\boldsymbol{\mu}$, esto se puede confirmar en la gráfica de la función de densidad de una distribución normal bivariada en la Figura 5.1. En el siguiente ejemplo consideramos dos casos particulares de la distancia de Mahalanobis.

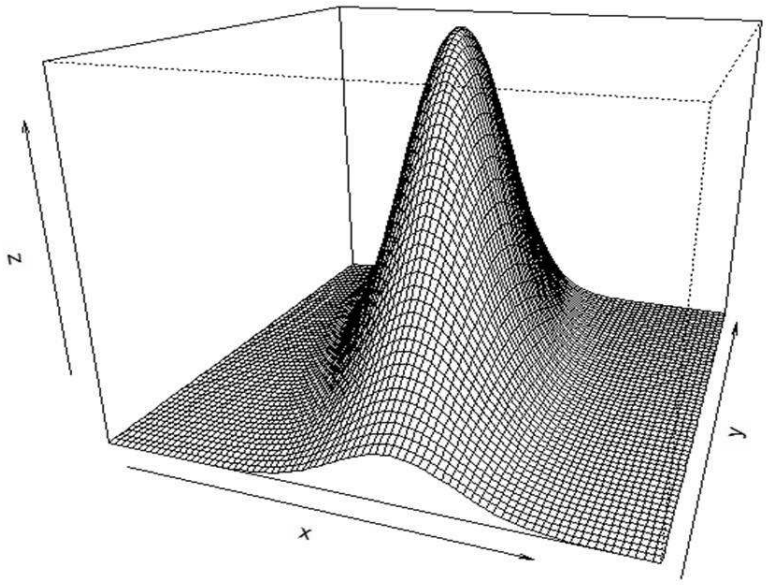

Figura 5.1: *Función de densidad de un vector aleatorio con distribución normal multivariante.*

Ejemplo 5.2.2. *Para un vector aleatorio* $\mathbf{X}$ *con distribución* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, tenemos que*

1. *Si* $\boldsymbol{\Sigma}$ *es de la forma* $\sigma^2\mathbf{I}$*, y* $\boldsymbol{\mu}$ *es de la forma* $\mathbf{1}_p\mu$*, esto es, si las variables son independientes e idénticamente distribuidas como* $N(\mu, \sigma^2)$*, entonces la distancia de Mahalanobis entre* $\mathbf{X}$ *y* $\boldsymbol{\mu}$ *toma la forma de una la distancia euclidiana, y está dada por*

$$D = \sqrt{(\mathbf{X}-\boldsymbol{\mu})'(\sigma^2\mathbf{I})^{-1}(\mathbf{X}-\boldsymbol{\mu})} = \sqrt{\sum_{i=1}^{p}\frac{(X_i-\mu)^2}{\sigma^2}}$$

2. *Si* $\boldsymbol{\Sigma}$ *es digonal con* $\sigma_1^2, \cdots, \sigma_p^2$ *los elementos de la diagonal, entonces la distancia de Mahalanobis entre* $\mathbf{X}$ *y* $\boldsymbol{\mu}$ *está dada por*

$$D = \sqrt{\sum_{i=1}^{p}\frac{(X_i-\mu_i)^2}{\sigma_i^2}}$$

es decir, se toma la distancia entre cada variable X_i *y su esperanza* μ_i*, pero ponderada por la respectiva varianza; de esta forma, variables con grandes varianzas aportan poco en el cómputo de la distancia de Mahalanobis.*

En la definición anterior, cuando $p = 2$, $\mathbf{X} = (X_1, X_2)'$, si denotamos $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ y $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$, tenemos que $|\Sigma| = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 = \sigma_1^2\sigma_2^2(1-\rho^2)$ donde $\rho = Cor(X_1, X_2)$, entonces $\boldsymbol{\Sigma}^{-1} = (\sigma_1^2\sigma_2^2(1-\rho^2))^{-1} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}$, y podemos escribir su función de densidad como:

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \\ &\quad \times \exp\left\{ -\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)} \left[\sigma_2^2(x_1-\mu_1)^2 - 2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2) + \sigma_1^2(x_2-\mu_2)^2 \right] \right\} \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ &\quad \times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1^2\sigma_2^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right] \right\}. \end{aligned}$$

En la Figura 5.1, se observa la función de densidad de una distribución normal bivariada que es obtenida mediante el siguiente código

```
> mu1<-1
> mu2<-2
> sig1<-1
> sig2<-4
> sig12<--1
> rho<-sig12/(sqrt(sig1*sig2))
>
> x<-seq(-2.5,5.5,0.1)
> y<-seq(-2,4,0.1)
>
> f<-function(x,y){
+ fun<-exp(-(((x-mu1)^2)/sig1+((y-mu2)^2)/
+ sig2-2*rho*(x-mu1)*(y-mu2)/sqrt(sig1*sig2))/(2*(1-rho^2)))
+ fun/(2*pi*sqrt(sig1*sig2)*sqrt(1-rho^2))
+ }
>
> z<- outer(x, y, f)
> op <- par(bg = "white")
> persp(x, y, z, theta = 20, phi = 20, expand = 0.8)
```

Para observar mejor la forma de la función de densidad en una distribución normal bivariada, en la Figura 5.2 presentamos diferentes gráficas de contorno de esta función de densidad con $\sigma_1 = 2$, $\sigma_2 = 5$ y diferentes valores de ρ. Podemos observar que el contorno de la función es elíptico cuando $\rho \neq 0$. Cuando ρ toma valores extremos hacia 1 ó -1, la elipse se hace cada vez más prominente. Y también podemos ver que cuando $\rho > 0$, valores grandes de X_1 se asocian con valores grandes de X_2, lo cual coincide con la interpretación del coeficiente de correlación ρ, y de manera análoga cuando $\rho < 0$, valores grandes de X_1 se asocian con valores pequeños de X_2.

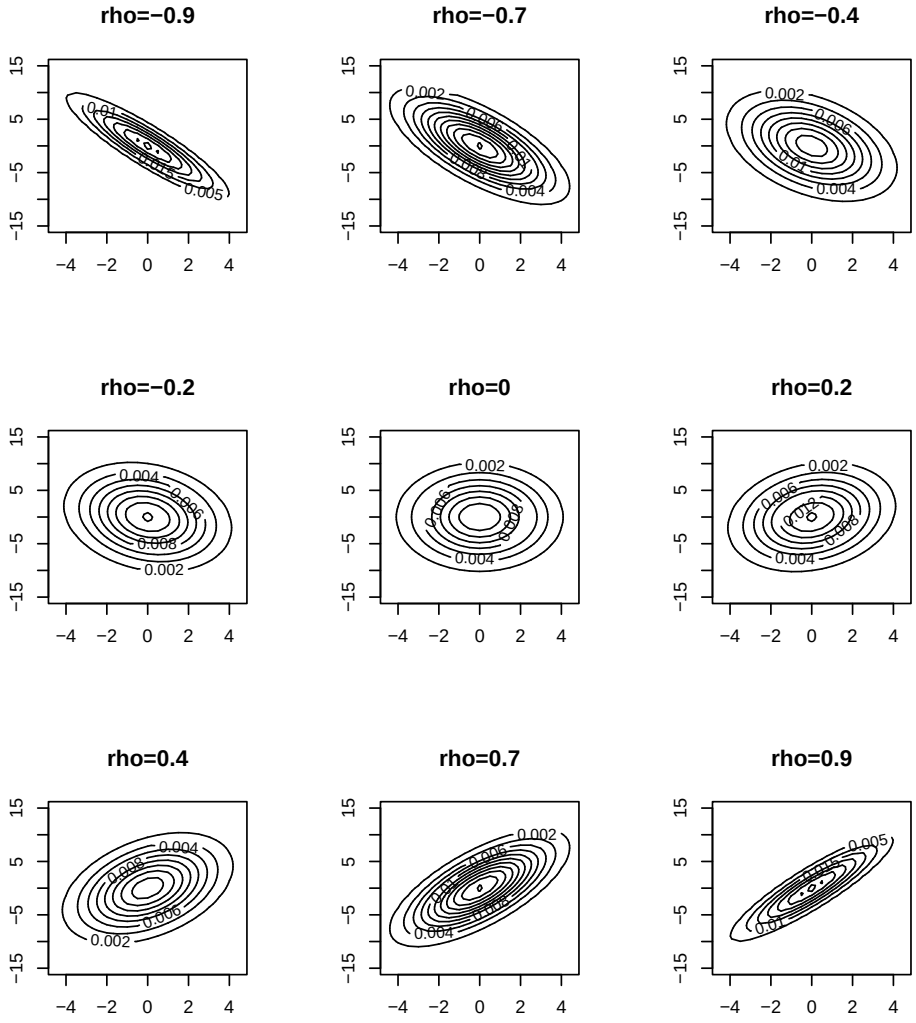

Figura 5.2: *Gráfica de contorno para la distribución normal bivariante con* $\sigma_1 = 2$, $\sigma_2 = 5$ *y diferentes valores de* ρ.

Un ejemplo de la distribución normal bivariada que es muy común en la vida real son las variables peso y estatura en individuos de una ciudad. En la Figura 5.3 se muestra el histograma y gráfica de dispersión de estas dos variables medidas en 200 habitantes de una ciudad. De los histogramas de las dos variables, podemos observar que cada variable puede ser descrita con la distribución normal; y de la gráfica de dispersión se observa una dependencia lineal entre estas dos variables dando indicio de

una estructura de correlación, y dado lo anterior, es necesario analizar las dos variables conjuntamente y la distribución apropiada será la distribución normal bivariada. También es necesario señalar que esta gráfica es sólo de carácter exploratorio, más adelante se introducirán pruebas estadísticas acerca de si un conjunto de datos multivariados proviene de una distribución normal multivariante.

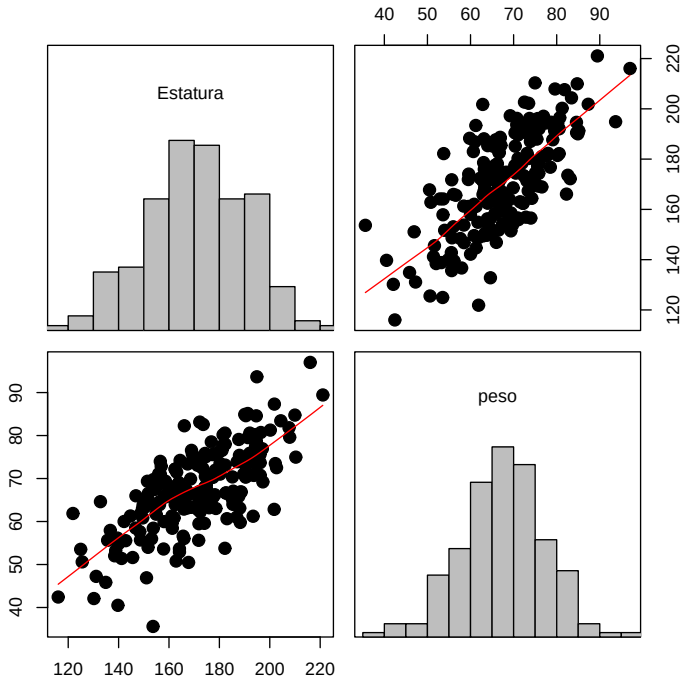

Figura 5.3: *Histograma y gráfica de dispersión de las variables estatura y peso de 200 habitantes de una ciudad.*

Ahora, consideramos los datos de Student (1908) donde se registran los incrementos de sueño (en horas) en 10 pacientes con dos tipos de sedantes de hidrobromuro. Estos datos se muestran en la Tabla 5.2, donde los datos con signo negativo indican que se obtuvo una disminución de sueño en vez del incremento. Dado que los datos solo corresponden a 10 pacientes que constituyen una muestra pequeña, podemos utilizar las gráficas QQ plot para ver si la distribución normal se ajusta bien a los dos grupos de datos. Esta gráfica se muestra en la Figura 5.4.

Nota: cuando el vector aleatorio $\mathbf{X}$ es unidimensional, se reduce a una variable aleatoria, y su esperanza y matriz de varianzas y covarianzas se reducen a constantes, y si las denotan por μ y σ^2, la función de densidad en (5.2.2) se reduce a la función de densidad de una distribución normal univariada.

Paciente	Sedante A	Sedante B
1	0.7	1.9
2	-1.6	0.8
3	-0.2	1.1
4	-1.2	0.1
5	-1.0	0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	1.4

Tabla 5.2: *Incrementos de sueño (en horas) en 10 pacientes con dos tipos de sedantes de hidrobromuro.*

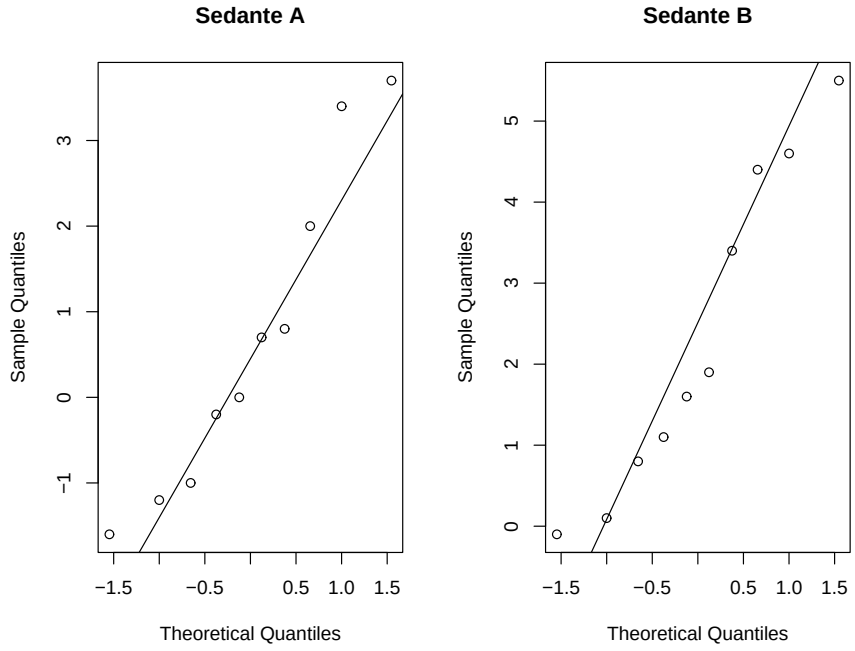

Figura 5.4: *Gráficas QQ plot para los datos de la Tabla 5.2.*

Dado un vector aleatorio $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, en el caso cuando $\boldsymbol{\mu} = \mathbf{0}$ y $\boldsymbol{\Sigma} = \mathbf{I}_p$, se dice que $\mathbf{X}$ tiene distribución normal estándar multivariante. En este los componentes de $\mathbf{X}$ son variables aleatorias independientes, cada una con distribución normal estándar.

Característica de la función de densidad de la distribución N_p

Dado que la distribución normal multivariante es una generalización de la distribución normal univariada, su función de densidad también tiene comportamientos similares a los del caso univariado, como lo presenta el siguiente resultado.

Resultado 5.2.2. *Dado un vector aleatorio $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, se tiene que*

1. *la función de densidad es simétrica con respecto a $\boldsymbol{\mu}$, esto es, $f(\boldsymbol{\mu}+\mathbf{a}) = f(\boldsymbol{\mu}-\mathbf{a})$ para todo vector $\mathbf{a} \in \mathbb{R}^p$.*
2. *la función de densidad tiene un máximo en $\boldsymbol{\mu}$.*

Demostración. 1. Tenemos que

$$\begin{aligned} f(\boldsymbol{\mu}+\mathbf{a}) &= |\boldsymbol{\Sigma}|^{-1/2}(2\pi)^{-p/2}\exp\{-\frac{1}{2}(\mathbf{a})'\boldsymbol{\Sigma}^{-1}(\mathbf{a})\} \\ &= |\boldsymbol{\Sigma}|^{-1/2}(2\pi)^{-p/2}\exp\{-\frac{1}{2}(-\mathbf{a})'\boldsymbol{\Sigma}^{-1}(-\mathbf{a})\} \\ &= f(\boldsymbol{\mu}-\mathbf{a}) \end{aligned}$$

2. La matriz $\boldsymbol{\Sigma}$ es semidefinida positiva, y por consiguiente sus valores propios son no negativos; más aun, son todos positivos, pues $\boldsymbol{\Sigma}$ es invertible. De esta manera, los valores propios de $\boldsymbol{\Sigma}^{-1}$ también son todos positivos, de donde se concluye que también $\boldsymbol{\Sigma}^{-1}$ es semidefinida positiva; por consiguiente, para cualquier vector $\mathbf{x}$, se tiene que $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \geq 0$, entonces la función de densidad de $\mathbf{X}$ tiene un máximo cuando $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = 0$, y esto se tiene cuando $\mathbf{x} = \boldsymbol{\mu}$, es decir, la función de densidad de $\mathbf{X}$ tiene un máximo en $\boldsymbol{\mu}$.

□

Momentos y función generadora de momentos de un vector aleatorio con distribución N_p

Resultado 5.2.3. *Si $\mathbf{Z} = (Z_1, \cdots, Z_p)' \sim N_p(\mathbf{0}, \mathbf{I}_p)$, entonces las variables $Z_1, \cdots, Z_p$ son independientes con $Z_i \sim N(0,1)$. Y $E(\mathbf{Z}) = \mathbf{0}$, $Var(\mathbf{Z}) = \mathbf{I}_p$ y $M_{\mathbf{Z}}(\mathbf{t}) = e^{\mathbf{t}'\mathbf{t}/2}$*

Demostración. Entonces por la definición de distribución multivariante, se tiene que la función de densidad $\mathbf{Z}$ está dada por

$$\begin{aligned} f(\mathbf{z}) = f(z_1, \cdots, z_p) &= (2\pi)^{-p/2}\exp\left\{-\frac{1}{2}\mathbf{z}'\mathbf{z}\right\} \\ &= (2\pi)^{-p/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{p} z_i^2\right\} \\ &= \prod_{i=1}^{P} r(2\pi)^{-1/2}\exp\left\{-\frac{1}{2}z_i^2\right\} \end{aligned}$$

donde cada término $(2\pi)^{-1/2}\exp\left\{-\frac{1}{2}z_i^2\right\}$ corresponde a la función de densidad de una distribución $N(0,1)$. Así, podemos ver que los componentes de $\mathbf{Z}$, $Z_1, \cdots, Z_p$ son variables aleatorias independientes e idénticamente distribuidas con distribución normal estándar. Y dadas las propiedades en términos de la esperanza y la varianza de esta distribución, es fácil ver que $E(\mathbf{Z}) = \mathbf{0}$ y $Var(\mathbf{Z}) = \mathbf{I}_p$. Adicionalmente, la función generadora de momentos de $\mathbf{Z}$ se puede calcular como

$$M_{\mathbf{Z}}(\mathbf{t}) = E(e^{\mathbf{Z}'\mathbf{t}}) = \prod_{i=1}^{p} E(e^{Z_i t_i}) = \prod_{i=1}^{p} M_{Z_i}(t_i) = e^{\mathbf{t}'\mathbf{t}/2}$$

□

Utilizando el anterior resultado con respecto a la distribución normal estándar multivariante, tenemos las siguientes propiedades para cualquier distribución normal multivariante.

Resultado 5.2.4. *Si $\mathbf{X}$ es un vector aleatorio con distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces tenemos que*

1. *Para cualquier matriz A de dimensión $r \times p$ y vector b de dimensión $r \times 1$, se tiene que $A\mathbf{X} + b \sim N_r(A\boldsymbol{\mu} + b, A\boldsymbol{\Sigma}A)'$. En particular, para cualesquiera constantes $c_1, \cdots, c_p$, se tiene que $\sum_{i=1}^{p} c_i X_i$ tiene distribución normal.*
2. *$E(\mathbf{X}) = \boldsymbol{\mu}$ y $Var(\mathbf{X}) = \boldsymbol{\Sigma}$.*
3. *La función generadora de momentos de $\mathbf{X}$ está dada por*

$$M_{\mathbf{X}}(t) = \exp\{\boldsymbol{\mu}'\mathbf{t} + \frac{\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}{2}\}. \tag{5.2.3}$$

   Además la función generadora de momentos caracteriza la distribución de $\mathbf{X}$.

4. *Por ser la matriz $\boldsymbol{\Sigma}$ semidefinida positiva, existe una matriz A simétrica cuadrada con $\boldsymbol{\Sigma} = AA$, y el vector aleatorio $A^{-1}(\mathbf{X} - \boldsymbol{\mu})$ tiene distribución $N_p(\mathbf{0}, \mathbf{I}_p)$.*
5. *Cualquier subvector de $\mathbf{X}$ también tiene distribución normal multivariante.*
6. *se tiene que la variable aleatoria $(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ es decir, la distancia de Mahalanobis entre $\mathbf{X}$ y $\boldsymbol{\mu}$, tiene distribución Ji-cuadrado con p grados de libertad.*

Demostración. La demostración de estas propiedades se basa principalmente en el hecho de que dado un vector $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_p)$, entonces mediante el teorema de transformación, se puede encontrar que la función de densidad del vector $\boldsymbol{\Sigma}^{1/2}\mathbf{Z} + \boldsymbol{\mu}$ está dada por

$$f(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2}(2\pi)^{-p/2}\exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}$$

de donde podemos ver que $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, y en conclusión $\mathbf{X}$ se puede escribir como $\mathbf{X} = \boldsymbol{\Sigma}^{1/2}\mathbf{Z} + \boldsymbol{\mu}$ donde $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_p)$. Usando este hecho podemos probar las propiedades enunciadas como sigue:

1. Para demostrar este resultado, se usa el hecho de que la función generadora de momentos caracteriza la distribución. Veamos que la función generadora de momentos de $A\mathbf{X}+b$ es de la forma de (5.2.3). Tenemos que

$$\begin{aligned} M_{A\mathbf{X}+b}(\mathbf{t}) &= E(e^{(A\mathbf{X}+b)'\mathbf{t}}) = E(e^{\mathbf{X}'A'\mathbf{t}})e^{b'\mathbf{t}} = M_{\mathbf{X}}(A't)e^{b'\mathbf{t}} \\ &= \exp\{\boldsymbol{\mu}'A'\mathbf{t} + \frac{(A'\mathbf{t})'\boldsymbol{\Sigma}A'\mathbf{t}}{2}\}\exp\{b'\mathbf{t}\} \\ &= \exp\{(\boldsymbol{\mu}'A' + b')t + \frac{\mathbf{t}'A\boldsymbol{\Sigma}A'\mathbf{t}}{2}\}. \end{aligned}$$

   Esta función es de la forma de (5.2.3), de donde se concluye que $A\mathbf{X}+b \sim N_r(A\boldsymbol{\mu}+b, A\boldsymbol{\Sigma}A)'$.

2. Se tiene trivialmente usando conjuntamente la parte 1, los Resultados 5.1.2, 5.2.3 y la parte (d) del Resultado 5.1.4.

3. Tenemos que

$$M_{\mathbf{X}}(\mathbf{t}) = E(e^{(\boldsymbol{\Sigma}^{1/2}\mathbf{Z}+\boldsymbol{\mu})'\mathbf{t}}) = M_{\mathbf{Z}}(\boldsymbol{\Sigma}^{1/2}\mathbf{t})e^{\boldsymbol{\mu}'\mathbf{t}} = e^{(\boldsymbol{\Sigma}^{1/2}\mathbf{t})'\boldsymbol{\Sigma}^{1/2}\mathbf{t}/2}e^{\boldsymbol{\mu}'\mathbf{t}} = e^{\boldsymbol{\mu}'\mathbf{t}+\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2}$$

4. Trivial usando el resultado anterior.

5. Ilustrando la demostración suponiendo que si $\mathbf{X}$ se puede dividir en dos subvectores, entonces cada subvector tiene distribución normal. Esto es, si

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

   entonces particionando adecuadamente $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}$ y $\mathbf{Z}$, $\mathbf{X} = \boldsymbol{\Sigma}^{1/2}\mathbf{Z} + \boldsymbol{\mu}$ se puede escribir como

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{1/2} & \boldsymbol{\Sigma}_{12}^{1/2} \\ \boldsymbol{\Sigma}_{21}^{1/2} & \boldsymbol{\Sigma}_{22}^{1/2} \end{pmatrix} \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

   Y por consiguiente $\mathbf{X}_1 = (\boldsymbol{\Sigma}_{11}^{1/2}, \boldsymbol{\Sigma}_{12}^{1/2})(\mathbf{Z}_1, \mathbf{Z}_2)' + \boldsymbol{\mu}_1$. La aplicación del numeral 3 lleva a la conclusión de que el subvector $\mathbf{X}_1$ tiene distribución normal. De manera análoga se encuentra que el subvector $\mathbf{X}_2$ también tiene distribución normal.

6. Utilizando, de nuevo, $\mathbf{X} = \boldsymbol{\Sigma}^{1/2}\mathbf{Z} + \boldsymbol{\mu}$, tenemos que

$$(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) = \mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{p} Z_i^2$$

   Utilizando la definición de la distribución χ^2 y el hecho de que las variables Z_1, $\cdots$, Z_p son independientes, cada uno con distribución normal estándar, tenemos que $(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim \chi_p^2$.

□

En la propiedad 4, se debe hallar la raíz cuadrada de la matriz $\mathbf{\Sigma}$. Este cálculo se puede llevar a cabo usando la descomposición $\mathbf{\Sigma} = QDQ'$, siendo Q la matriz que contiene los vectores propios ortonormales de $\mathbf{\Sigma}$, y $\mathbf{D}$ la matriz diagonal que contiene los valores propios de $\mathbf{\Sigma}$. Al definir $A = QD^{1/2}Q'$, se tiene que $\mathbf{\Sigma} = AA$. El siguiente código en R permite encontrar la matriz A

```
> p<-3
> gama<-matrix(c(6,1,-1,1,5,0,-1,0,4),p,p)
> Q<-eigen(gama)$vectors
> D<-diag(eigen(gama)$values)
> A<-Q%*%sqrt(D)%*%t(Q)
```

La propiedad 5 del resultado anterior, establece que si un vector $(X_1, X_2)'$ tiene distribución normal bivariada, entonces tanto X_1 como X_2 tienen distribución normal univariante. Sin embargo, el recíproco no es cierto, es decir, dadas dos variables X_1 y X_2, cada una con distribución normal, el vector $(X_1, X_2)'$ no necesariamente tiene distribución normal bivariada. Pero si X_1 y X_2 son variables independientes, entonces el vector $(X_1, X_2)'$ sí tiene distribución normal bivariada (Ejercicio 5.8).

Resultado 5.2.5. *Sean vectores aleatorios $\mathbf{X}_1, \cdots, \mathbf{X}_n$ independientes y distribuidos como $N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$, entonces, se tiene que para constantes $c_1, \cdots, c_n$, se tiene que el vector aleatorio $\mathbf{Y} = \sum_{i=1}^n c_i\mathbf{X_i}$ se distribuye como*
$N_p(\sum_{i=1}^n c_i\boldsymbol{\mu}_i, \sum_{i=1}^n c_i\mathbf{\Sigma}_i)$

Demostración. Se hace uso de la función generadora de momentos, tenemos que

$$m_{\mathbf{Y}}(\mathbf{t}) = m_{\sum_{i=1}^n c_i\mathbf{X_i}}(\mathbf{t}) \tag{5.2.4}$$

$$= \prod_{i=1}^n m_{\mathbf{X}_i}(c_i\mathbf{t}) \qquad \text{Resultado 5.1.7.} \tag{5.2.5}$$

$$= \prod_{i=1}^n \exp\{c_i\boldsymbol{\mu}_i'\mathbf{t} + \frac{c_i^2\mathbf{t}'\mathbf{\Sigma}_i\mathbf{t}}{2}\} \tag{5.2.6}$$

$$= \exp\{\sum_{i=1}^n \boldsymbol{\mu}_i'\mathbf{t} + \frac{\mathbf{t}'\sum_{i=1}^n c_i\mathbf{\Sigma}_i\mathbf{t}}{2}\}, \tag{5.2.7}$$

de donde se concluye que $\mathbf{Y} \sim N_p(\sum_{i=1}^n c_i\boldsymbol{\mu}_i, \sum_{i=1}^n c_i\mathbf{\Sigma}_i)$ □

Nótese que en el anterior resultado, cuando son los vectores aleatorios $\mathbf{X}_1, \cdots, \mathbf{X}_n$ son idénticamente distribuidos, con distribución $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, se tiene que el promedio definido como

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i \tag{5.2.8}$$

tiene distribución $N_p(\boldsymbol{\mu}, \frac{1}{n}\mathbf{\Sigma})$.

Finalmente, consideramos la distribución condicional dentro de la distribución normal multivariante. Suponga que $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, y dividimos el vector $\mathbf{X}$ como

$\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')'$ donde $\mathbf{X}_1$ y $\mathbf{X}_2$ son vectores de dimensión p_1 y p_2, respectivamente, con $p_1 + p_2 = p$. Esta división también induce a una división en el vector de medias $\boldsymbol{\mu}$ y la matriz de varianzas $\boldsymbol{\Sigma}$ como $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')'$ y

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

donde $\boldsymbol{\mu}_1$ y $\boldsymbol{\mu}_2$ son los vectores de medias de $\mathbf{X}_1$ y $\mathbf{X}_2$; $\boldsymbol{\Sigma}_{11}$ y $\boldsymbol{\Sigma}_{22}$ son las matrices de varianzas y covarianzas de $\mathbf{X}_1$ y $\mathbf{X}_2$; y $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}'$ es la matriz que contiene las covarianzas entre cada componente de $\mathbf{X}_1$ y cada componente de $\mathbf{X}_2$. En el siguiente resultado mostramos la distribución de $\mathbf{X}_1$ dado $\mathbf{X}_2$ cuando el vector completo $\mathbf{X}$ tiene distribución normal multivariante.

Resultado 5.2.6. *Dada la división enunciada anteriormente, si* $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, entonces la distribución de* $\mathbf{X}_1$ *condicionada a que* $\mathbf{X}_2$ *tome el valor* $\mathbf{x}_2$ *es normal con esperanza*

$$E(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

y matriz de varianzas y covarianzas

$$Var(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Demostración. La prueba del resultado consiste en encontrar la función de densidad de $\mathbf{X}_1$ dado $\mathbf{X}_2 = \mathbf{x}_2$. Tenemos que

$$\begin{aligned} f(\mathbf{x}_1|\mathbf{x}_2) &= \frac{f_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2)}{f_{\mathbf{X}_2}(\mathbf{x}_2)} \\ &= \frac{|\boldsymbol{\Sigma}|^{-1/2}(2\pi)^{-p/2}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}}{|\boldsymbol{\Sigma}_{22}|^{-1/2}(2\pi)^{-p_2/2}\exp\{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\}} \end{aligned}$$

Usando la propiedad de una matriz particionada con respecto al determinante, tenemos que $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{22}||\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}|$ y $p_1 + p_2 = p$, tenemos que

$$\begin{aligned} f(\mathbf{x}_1|\mathbf{x}_2) = {} & |\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}|^{-1/2}(2\pi)^{-p_1/2} \\ & \times \exp\left\{-\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - (\mathbf{x}_2-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right]\right\} \end{aligned}$$

Ahora consideramos el término dentro del exponente $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - (\mathbf{x}_2 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)$, usando la propiedad de una matriz particionada con respecto a la inversa, tenemos que

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \mathbf{B}^{-1} & -\mathbf{B}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{B}^{-1} & \boldsymbol{\Sigma}_{22}^{-1} + \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{B}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix}$$

con $\mathbf{B} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Desarrollando los productos matriciales, se puede ver que

$$
\begin{aligned}
&(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - (\mathbf{x}_2-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\\
&= (\mathbf{x}_1-\boldsymbol{\mu}_1)'\mathbf{B}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1) - (\mathbf{x}_1-\boldsymbol{\mu}_1)'\mathbf{B}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\\
&\qquad - (\mathbf{x}_2-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{B}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)\\
&\qquad\qquad + (\mathbf{x}_2-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{B}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\\
&= \left[(\mathbf{x}_1-\boldsymbol{\mu}_1)' - (\mathbf{x}_2-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right]\mathbf{B}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)\\
&\qquad - \left[(\mathbf{x}_1-\boldsymbol{\mu}_1)' - (\mathbf{x}_2-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right]\mathbf{B}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\\
&= \left[(\mathbf{x}_1-\boldsymbol{\mu}_1)' - (\mathbf{x}_2-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right]\mathbf{B}^{-1}\left[\mathbf{x}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right]\\
&= \left[\mathbf{x}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right]'\mathbf{B}^{-1}\left[\mathbf{x}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right].
\end{aligned}
$$

Utilizando lo anterior, tenemos que la función de densidad de $\mathbf{X}_1$ dado $\mathbf{X}_2=\mathbf{x}_2$ es

$$
\begin{aligned}
&f(\mathbf{x}_1|\mathbf{x}_2) = |\mathbf{B}|^{-1/2}(2\pi)^{-p_1/2}\\
&\exp\left\{-\frac{1}{2}\left(\mathbf{x}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right)'\mathbf{B}^{-1}\left(\mathbf{x}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right)\right\}
\end{aligned}
$$

con $\mathbf{B}=\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Esta función coincide con la función de densidad de una distribución normal con media y $\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)$ y matriz de varianzas $\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. □

Observamos que cuando no existe ninguna relación lineal entre $\mathbf{X}_1$ y $\mathbf{X}_2$, $\boldsymbol{\Sigma}_{12}=0$ y como consecuencia $E(\mathbf{X}_1|\mathbf{X}_2=\mathbf{x}_2)=E(\mathbf{X}_1)$ y $Var(\mathbf{X}_1|\mathbf{X}_2=\mathbf{x}_2)=Var(\mathbf{X}_1)$, esto es, el condicionamiento sobre $\mathbf{X}_2=\mathbf{x}_2$ no afecta la distribución del vector $\mathbf{X}_1$.

5.2.3 Distribución Wishart

Al igual que la definición de un vector aleatorio, se puede definir una matriz aleatoria como sigue

Definición 5.2.3. *Una matriz A de dimensión $p\times q$, cuyos elementos son variables aleatorias se llama una matriz aleatoria.*

La teoría asociada a las matrices aleatorias es bastante complicada, los lectores interesados pueden consultar Gupta & Nagar (2000) para una revisión detallada de este aspecto. En el presente libro sólo se introducirá la distribución Wishart con el fin de desarrollar la teoría de inferencia acerca de la matriz de varianzas y covarianzas de una distribución normal multivariante[1], mas no estamos considerando datos en forma matricial que puedan ser descritos con la distribución Wishart.

[1]Cuando se desea estimar la matriz de varianzas y covarianzas de una distribución normal multivariante, el estimador también debe ser de forma matricial. En el próximo capítulo se verá que la distribución de este estimador está asociada con la distribución Wishart, y evaluaremos la calidad del estimador usando las propiedades de esa distribución.

Definición 5.2.4. *Una matriz aleatoria A cuadrada de dimensión $p \times p$ tiene distribución Wishart con matriz de parámetros Σ y grados de libertad n, si la función de densidad de A está dada por*

$$f(A) = \frac{|A|^{(n-p-1)/2}}{2^{np/2}|\Sigma|^{n/2}\mathbf{\Sigma}_P r(\frac{n}{2})} \exp\left\{-\frac{1}{2}tr(\Sigma^{-1}A)\right\} \tag{5.2.9}$$

donde Σ es una matriz $p \times p$ definida positiva, n es un entero positivo con $n \geq p$, y $\mathbf{\Sigma}_P r(\cdot)$ es la función gamma multivariada definida como

$$\mathbf{\Sigma}_P r(k) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \mathbf{\Sigma}(\frac{2k+1-i}{2}).$$

La notación para esta distribución es $A \sim W(n, \Sigma)$.

La distribución Wishart es la versión multivariante de la distribución Ji-cuadrado, cuando $p = 1$, A se reduce a un vector aleatorio y cuando $\Sigma = 1$, la anterior expresión se reduce a la función de densidad de una distribución χ^2_n.

Existe otra definición equivalente para la distribución Wishart, que afirma:

Definición 5.2.5. *Sea $\mathbf{X}_1, \cdots, \mathbf{X}_n$ vectores aleatorios independientes e idénticamente distribuidos como $N_p(\mathbf{0}, \mathbf{\Sigma})$, entonces la matriz $\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'$ tiene distribución Wishart con matriz de parámetros Σ y grados de libertad n.*

En el siguiente resultado se enuncian algunas propiedades de la distribución Wishart.

Resultado 5.2.7. *Si $A \sim W(n, \Sigma)$, entonces*

1. $E(A) = n\Sigma$
2. *para cualquier matriz B de constantes de dimensión $r \times p$, se tiene que $BAB' \sim W(n, B\mathbf{\Sigma}B')$*
3. *la función característica está dada por $C_A(\Theta) = |I_p - 2i\Theta\Sigma|^{-n/2}$ para Θ matriz de dimensión $p \times p$*

Demostración. La demostración hará uso de la definición 5.3.3.

1. Tenemos que
$$E(A) = E(\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i') = \sum_{i=1}^n E(\mathbf{X}_i\mathbf{X}_i') = \sum_{i=1}^n Var(\mathbf{X}_i) = \sum_{i=1}^n \mathbf{\Sigma} = n\mathbf{\Sigma}$$
2. Tenemos que $BAB' = B\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'B' = \sum_{i=1}^n B\mathbf{X}_i\mathbf{X}_i'B' = \sum_{i=1}^n B\mathbf{X}_i(B\mathbf{X}_i)'$. Por el Resultado 5.2.4 parte 1, se tiene que $B\mathbf{X}_i \sim N_p(\mathbf{0}, B\Sigma B')$; además, $Cov(B\mathbf{X}_i, B\mathbf{X}_j) = BCov(\mathbf{X}_i, \mathbf{X}_j)B' = B\mathbf{0}B' = \mathbf{0}$, de donde se concluye que los vectores $B\mathbf{X}_1, \cdots, B\mathbf{X}_n$ son independientes. Entonces por la definición 5.3.3, se tiene que $\sum_{i=1}^n B\mathbf{X}_i(B\mathbf{X}_i)' \sim W(n, B\Sigma B')$, y el resultado queda demostrado.

□

Resultado 5.2.8. *Sean $A_1, \cdots, A_n$ matrices aleatorias independientes donde $A_i \sim W(n_i, \Sigma)$ para $i = 1, \cdots, n$, entonces $\sum_{i=1}^n A_i \sim W(\sum_{i=1}^n n_i, \Sigma)$.*

Demostración. La demostración hará uso de la función característica de la distribución Wishart. Luego, tenemos que

$$\begin{aligned} C_{\sum_{i=1}^n A_i}(\Theta) &= \prod_{i=1}^n C_{A_i}(\Theta) \qquad \text{(por la independencia)} \\ &= \prod_{i=1}^n |I_p - 2i\Theta\Sigma|^{-n_i/2} = |I_p - 2i\Theta\Sigma|^{-\sum_{i=1}^n n_i/2}, \end{aligned}$$

de donde se concluye que $\sum_{i=1}^n A_i \sim W(\sum_{i=1}^n n_i, \Sigma)$. □

5.2.4 Distribución T^2 de Hotelling

La distribución T^2 de Hotelling debe su nombre al estadístico americano Harold Hotelling y es muy útil en la teoría de la inferencia estadística multivariada. Aunque esta distribución se define para variables aleatorias y no para vectores aleatorios, la razón por la que presentamos esta distribución en este capítulo del libro es que la distribución T^2 de Hotelling está íntimamente ligada con la distribución normal multivariante, como se enuncia a continuación.

Figura 5.5: *Harold Hotelling (1895-1973).*

Definición 5.2.6. *Dado un vector aleatorio $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, y sea $\mathbf{W}$ una matriz aleatoria con distribución $W(\boldsymbol{\Sigma}, n-1)$ para algún entero positivo n, si $\mathbf{W}$ es independiente de $\mathbf{X}$, entonces*

$$T^2 = (\mathbf{X} - \boldsymbol{\mu})' \left(\frac{\mathbf{W}}{n-1} \right)^{-1} (\mathbf{X} - \boldsymbol{\mu})$$

tiene distribución T^2 de Hotelling de grados de libertad p y $n-1$.

5.3 Ejercicios

5.1 Verificar las probabilidades y las esperanzas del Ejemplo 5.1.1.

5.2 Demuestre el Resultado 5.1.8.

5.3 Diga dos ejemplos en la vida real de variables que deben ser analizadas conjuntamente debido a la presencia de estructuras de dependencia lineal.

5.4 Dado un vector aleatorio $\mathbf{X} = (X_1, X_2)'$ con $X_1 \sim N(1, 4)$ y X_2 con distribución exponencial de media 5, y $Cov(X_1, X_2) = 2$, encuentre

(a) La esperanza, la matriz de varianzas y covarianzas y la matriz de correlaciones de $\mathbf{X}$.

(b) La esperanza, la matriz de varianzas y covarianzas y la matriz de correlaciones del vector aleatorio conformado por las variables $X_1 - X2$, $(X_1 + X_2)/2$

5.5 En una urna con 15 bolas negras, 10 rojas y 20 verdes, se extraen aleatoriamente con reemplazo 18 bolas, y sea X_1, X_2 y X_3 que denotan el número de bolas negras, rojas y verdes extraídas, respectivamente.

(a) ¿Qué distribución tiene el vector $(X_1, X_2, X_3)'$? Escriba su función de densidad.

(b) ¿Cuál es la probabilidad de que de las 18 bolas extraídas, 4 sean negras, 8 sean rojas y 6 sean verdes?

(c) ¿Cuál es la probabilidad de que de las 18 bolas extraídas no haya ninguna roja?

(d) ¿Cuál es la probabilidad de que todas las 18 bolas extraídas sean verdes?

5.6 Dado un vector aleatorio $\mathbf{X} = (X_1, X_2, X_3)'$ con $N_3\left(\begin{pmatrix} 1 \\ 3 \\ -1 \end{pmatrix}, \begin{pmatrix} 6 & 0 & -1 \\ 0 & 5 & 2 \\ -1 & 2 & 4 \end{pmatrix}\right)$

(a) ¿Qué distribución tiene el subvector $(X_1, X_3)'$?

(b) ¿Cuáles variables de $\mathbf{X}$ son independientes?

(c) Con la ayuda de R, encuentra la matriz $\mathbf{\Sigma}^{-1/2}$ y luego estandariza $\mathbf{X}$.

(d) ¿Qué distribución tiene X_1 dado que $X_2 = x_2$?

(e) ¿Qué distribución tiene X_1 y X_2 dado que $X_3 = x_3$?

(f) Escriba la función generadora de momentos de $\mathbf{X}$.

(g) Escriba la función generadora de momentos de $(X_1, X_3)'$.

5.7 Sea $\mathbf{X}$ un vector aleatorio con distribución $N(\boldsymbol{\mu}, \mathbf{\Sigma})$ con $\boldsymbol{\mu} = (1, 3, -2)'$, y

$$\mathbf{\Sigma} = \begin{pmatrix} 7 & 3 & -3 \\ 3 & 6 & 0 \\ -3 & 0 & 5 \end{pmatrix}$$

(a) Encuentre la distribución de $\mathbf{Y} = (X_1, X_3)'$ y $Cov(\mathbf{Y}, X_2)$.

(b) Encuentre la función generadora de momentos y la matriz de correlaciones de $\mathbf{X}$ y $(X_2, X_3)'$.

(c) Se define $Y_1 = -X_1 - 2X_2 + 3$ y $Y_2 = X_1 + X_2 + 3X_3 - 1$. Encuentre la distribución de $(Y_1, Y_2)'$.

5.8 Sea $\mathbf{X}$ y $\mathbf{Y}$ vectores aleatorios independientes con distribución $N_2\left(\begin{pmatrix}1\\-1\end{pmatrix}, \begin{pmatrix}6 & 0\\0 & 5\end{pmatrix}\right)$ y $N_2\left(\begin{pmatrix}2\\-3\end{pmatrix}, \begin{pmatrix}3 & -2\\-2 & 4\end{pmatrix}\right)$.

(a) Escriba la función generadora de momentos de $(\mathbf{X} + \mathbf{Y})/2$ y $2\mathbf{X} + \mathbf{Y}$.

(b) Encuentre la distribución de $(\mathbf{X} + \mathbf{Y})/2$ y $2\mathbf{X} + \mathbf{Y}$.

(c) Escriba explícitamente la función de densidad de $(\mathbf{X} + \mathbf{Y})/2$.

5.9 Sean $X_i \sim N(\mu_1, \sigma_i^2)$ con $i = 1, 2$ variables independientes, encuentre la función de densidad conjunta del vector aleatorio $(X_1, X_2)'$ y vea que éste tiene distribución normal bivariante.

5.10 Demuestre la propiedades 2 y 4 del Resultado 5.2.4.

Capítulo 6

Inferencia multivariante

En el ámbito de la inferencia univariada, se definió una muestra aleatoria como un conjunto de variables aleatorias independientes e idénticamente distribuidas. En la inferencia multivariante, se define análogamente a una muestra aleatoria como un conjunto de vectores aleatorios independientes e idénticamente distribuidos.

En este parte del libro estudiamos tópicos de inferencia para los parámetros de una distribución multivariante basada en una muestra aleatoria $\mathbf{X}_1, \cdots, \mathbf{X}_n$. Consideraremos el método de máxima verosimilitud para encontrar los estimadores puntuales, y posteriormente el tema de pruebas de hipótesis. Es claro que dentro del ámbito de la inferencia multivariante, la distribución teórica también tiene más de un parámetro, y en este caso se hablará de regiones de confianza para el vector de parámetros. Sin embargo, en muchos casos, este tema no es de tanto interés como lo es en la inferencia univariada puesto que no es posible visualizar las regiones de confianza cuando éstas son subconjuntos de $\mathbb{R}^p$ con $p > 2$. Cuando se estudia la distribución normal multivariante éstas regiones de confianza son útiles para juzgar un sistema de hipótesis utilizando la dualidad que existe entre estos dos métodos.

Primero consideramos el método de máxima verosimilitud, para lo cual se define la función de verosimilitud.

Definición 6.0.1. *Dado un conjunto de vectores aleatorios $\mathbf{X}_1, \cdots, \mathbf{X}_n$ con función de densidad $f_{\mathbf{X}}(\mathbf{x}, \boldsymbol{\theta})$, donde $\boldsymbol{\theta}$ es el vector de los parámetros de la distribución, se define la función de verosimilitud como la función de densidad conjunta de los n vectores aleatorios, y se denota como $L(\mathbf{x}_1, \cdots, \mathbf{x}_n, \boldsymbol{\theta})$.*

Es claro que cuando los vectores aleatorios constituyen una muestra aleatoria, se tiene que

$$L(\mathbf{x}_1, \cdots, \mathbf{x}_n, \boldsymbol{\theta})) = \prod_{i=1}^{n} f_{\mathbf{X}_i}(\mathbf{x}_i, \boldsymbol{\theta})$$

El estimador de máxima verosimilitud de $\boldsymbol{\theta}$ se define como el valor de $\boldsymbol{\theta}$ que maximiza la función de verosimilitud, y se denota como $\hat{\boldsymbol{\theta}}_{MV}$.

6.1 Inferencia en la distribución multinomial

6.1.1 Una muestra

Estimador de máxima verosimilitud y sus propiedades

En esta parte retomamos el Ejemplo 5.2.1 del capítulo anterior, donde en un problema de investigación de intención de voto en una elección presidencial se observa la necesidad de estimar los parámetros $p_1, \cdots, p_r$ en una distribución multinomial $(n, p_1, \cdots, p_r)$ donde n se asume conocido. Tal como se describió en el Ejemplo 5.2.1 en una muestra aleatoria, se observa el valor que toma el vector $(X_1, \cdots, X_r)$, y estamos interesados en encontrar $\hat{p}_{iMV}$ para $i = 1, \cdots, r$. Para eso, primero encontramos la función de verosimilitud, que en este caso concuerda con la función de densidad conjunta del vector $(X_1, \cdots, X_r)$. Tenemos que

$$L(x_1, \cdots, x_r, p_1, \cdots, p_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r}.$$

Con el fin de encontrar los valores de p_i que maximiza L, calculamos $\ln L$ como es de costumbre, y tenemos que

$$\ln L(x_1, \cdots, x_r, p_1, \cdots, p_r) = \ln n! - \ln x_1! - \cdots - \ln x_r! + x_1 \ln p_1 + \cdots + x_r \ln p_r$$

Entonces nuestro objetivo es maximizar la función $\ln L$, pero teniendo en cuenta que las probabilidades p_i deben cumplir la restricción de que $p_1 + \cdots + p_r = 1$, es decir, tenemos un problema de maximización con restricciones, y recurrimos a la técnica del multiplicador de Lagrange. Tenemos que la función de Lagrange está dada por

$$\Lambda = \ln n! - \ln x_1! - \cdots - \ln x_r! + x_1 \ln p_1 + \cdots + x_r \ln p_r + \lambda(p_1 + \cdots + p_r - 1)$$

Y derivamos Λ con respecto a $p_1, \cdots, p_r$ y λ, tenemos que para cada $i = 1, \cdots, r$

$$\frac{\partial \Lambda}{\partial p_i} = \frac{x_i}{p_i} + \lambda$$

y

$$\frac{\partial \Lambda}{\partial \lambda} = p_1 + \cdots + p_r - 1$$

Igualando estas derivadas a cero, tenemos que $p_i = -x_i/\lambda$ para todo $i = 1, \cdots, r$ y $\sum_{i=1}^r p_i = 1$. De donde tenemos que $-\sum_{i=1}^r x_i/\lambda = 1$, esto es $\lambda = -\sum_{i=1}^r x_i = -n$. Y finalmente tenemos que los p_i que maximizan a $\ln L$ y que cumplen con la restricción de $p_1 + \cdots + p_r = 1$ están dados por

$$\hat{p}_{i,MV} = \frac{X_i}{n} = \bar{X}_i$$

para todo $i = 1, \cdots, r$. Esto es, los estimadores de máxima verosimilitud son simplemente las proporciones muestrales.

Aplicando el anterior resultado al problema de elecciones presidenciales del Ejemplo 5.2.1, podemos calcular las estimaciones de p_1, p_2 y p_3 como $810/2436 \approx 0.3325 = 33.25\,\%$, $654/2436 \approx 0.2684 = 26.84\,\%$ y $972/2436 \approx 0.399 = 39.9\,\%$, respectivamente. Y si se supone que 10 millones 500 mil personas se inscribieron para participar en la votación, podemos afirmar que se espera que el candidato A obtenga $33.25\,\%$ de las votaciones, es decir, $10500000 * 0.3325 = 3491250$ votos, casi 3 millones y medio de votos.

Propiedad del estimador de máxima verosimilitud

Se vio que el estimador de máxima verosimilitud del vector $(p_1, \cdots, p_r)$ es el vector de proporciones muestrales $(\bar{X}_1, \cdots, \bar{X}_r)$, ahora estudiamos este estimador en términos del sesgo y la varianza.

El concepto del sesgo para el caso multivariado es una extensión natural del sesgo del caso univariado, y en este caso, estamos interesados en ver si $E(\bar{X}_1, \cdots, \bar{X}_r)$ es igual al vector de parámetros $(p_1, \cdots, p_r)$. Es fácil verificar que eso es cierto, puesto que si $\mathbf{p} = (p_1, \cdots, p_r)$, tenemos que

$$E(\bar{X}_1, \cdots, \bar{X}_r) = \frac{1}{n} E(X_1, \cdots, X_r) = \frac{1}{n} n\mathbf{p}$$

usando el hecho de que el vector $(X_1, \cdots, X_r)$ es $Multi(n, p_1, \cdots, p_r)$ y el Resultado 5.2.1.

Ahora para calcular la matriz de varianzas y covarianzas del estimador $(\bar{X}_1, \cdots, \bar{X}_r)$ utilizamos de nuevo el Resultado 5.2.1, y tenemos que

$$Var(\bar{X}_1, \cdots, \bar{X}_r) = \frac{1}{n^2} Var(X_1, \cdots, X_r) = \frac{1}{n}[diag(\mathbf{p}) - \mathbf{p}\mathbf{p}']$$

y naturalmente al incrementar el tamaño la estimación de máxima verosimilitud se hace más precisa.

Prueba de hipótesis

En esta parte consideramos el problema de prueba de hipótesis sobre el vector de probabilidades $\mathbf{p} = (p_1, \cdots, p_r)$. Dado que se trata de un vector de parámetros, puede haber una variedad grande de sistemas; por ejemplo, puede haber un sistema con H_0 simple de la forma

$$H_0 : \ \mathbf{p} = (p_1^*, \cdots, p_r^*)' \quad \text{vs.} \quad H_1 : \ \mathbf{p} \neq (p_1^*, \cdots, p_r^*)'. \tag{6.1.1}$$

También puede haber hipótesis sobre sólo algunas de estas probabilidades; por ejemplo, sólo es de interés saber si se puede asumir $p_1 = p_2$, mas no son de interés las otras probabilidades p_3, $\cdots$, p_k. En este caso el sistema de hipótesis es

$$H_0 : \ p_1 = p_2 \quad \text{vs.} \quad H_1 : \ p_1 \neq p_2.$$

En cualquier caso utilizaremos la prueba de razón de verosimilitud. Primero consideramos el sistema de H_0 simple (6.1.1). Para este sistema, la estadística de razón generalizada de verosimilitud está dada por

$$\lambda = \frac{sup_{\mathbf{\Theta}_0 \bigcup \mathbf{\Theta}_1} L(p_1, \cdots, p_r)}{sup_{\mathbf{\Theta}_0} L(p_1, \cdots, p_r)}. \tag{6.1.2}$$

Ahora, en el sistema de hipótesis (6.1.1) $\mathbf{\Theta}_0 \bigcup \mathbf{\Theta}_1 = \mathbf{\Theta}$, el espacio paramétrico completo del vector $(p_1, \cdots, p_r)$, entonces el numerado de la estadística λ se convierte en la función de verosimilitud evaluada en el estimador de máxima verosimilitud. Por otro lado, como H_0 es una hipótesis simple, entonces el denominador de la estadística λ es la función de verosimilitud evaluada en los valores especificados por H_0, esto es, $(p_1^*, \cdots, p_r^*)$. De lo anterior, tenemos que

$$\lambda = \frac{\bar{x}_1^{x_1} \cdots \bar{x}_r^{x_r}}{(p_1^*)^{x_1} \cdots (p_r^*)^{x_r}}$$

donde $\bar{x}_i$ es la proporción observada de la categoría i, para $i = 1, \cdots, r$. Y se rechaza H_0 para valores grandes de la estadística λ. Para establecer exactamente cuánto un valor de λ se puede considerar como grande, es necesario conocer la distribución nula de λ, pero por la forma de esta estadística será muy difícil encontrar tal distribución. Por consiguiente, se recurre a la distribución asintótica de la estadística $2 \ln \lambda$ que en este caso está dada por

$$2 \ln \lambda = 2 \sum_{i=1}^{r} x_i \ln \left(\frac{\bar{x}_i}{p_i^*} \right)$$

cuya distribución asintótica es $\chi^2_{v_1 - v_0}$ donde v_1 y v_0 son el número de parámetros libres bajo H_1 y H_0, respectivamente. En el sistema (6.1.1), H_0 es una hipótesis simple para el vector $(p_1, \cdots, p_r)$, por consiguiente ninguna p_i se puede cambiar libremente, y por consiguiente $v_0 = 0$. Por otro lado, en el espacio paramétrico completo $\mathbf{\Theta}$, los parámetros deben cumplir la restricción de $p_1 + \cdots + p_r = 1$, por lo tanto, el valor p_r está determinado de manera única mediante $p_r = 1 - p_1 - \cdots - p_{r-1}$ y por consiguiente solo hay $r - 1$ parámetros libres en $\mathbf{\Theta}$ y por consiguiente $v_1 = k - 1$. Y finalmente que

$$2 \ln \lambda = 2 \sum_{i=1}^{r} x_i \ln \left(\frac{\bar{x}_i}{p_i^*} \right) \sim_{asym} \chi^2_{k-1}$$

bajo H_0. Y por consiguiente se rechaza H_0 si $-2 \ln \lambda > \chi^2_{r-1,1-\alpha}$.

También podemos calcular el p valor teniendo en cuenta la forma de región de rechazo. Éste se calcula como

$$p \text{ valor} = 1 - F_{\chi^2_{r-1}}(v)$$

donde v es el valor de la estadística $2 \ln \lambda$ y $F_{\chi^2_{r-1}}$ denota la función de distribución de la distribución χ^2_{r-1}. Aplicamos el anterior resultado en el siguiente ejemplo.

Ejemplo 6.1.1. *Retomando el Ejemplo 5.2.1 donde estudia el favoritismo en una elección presidencial con tres candidatos A, B y C. Si se cree que A puede obtener 43 % de*

los votos, mientras que B y C pueden obtener 25 % y 32 %, respectivamente. Si en una muestra de 2436, el número de personas que apoyan a los tres candidatos son 810, 654 y 972, respectivamente. Entonces para ver si esta hipótesis $H_0 : \mathbf{p} = (0.43, 0.25, 0.32)$ *es apoyada por los datos, calculamos la estadística de razón de verosimilitud como*

$$2\ln\lambda = 2\left(810 * \ln\left(\frac{810/2436}{0.43}\right) + 654 * \ln\left(\frac{654/2436}{0.25}\right) + 972 * \ln\left(\frac{972/2436}{0.32}\right)\right)$$
$$= 105.72$$

el cual comparando en el percentil $\chi^2_{2,0.95} = 5.99$ *conduce al rechazo de* H_0*. Es decir, los datos sugieren que el vector de parámetros* (p_1, p_2, p_3) *no toma el valor de* $(0.43, 0.25, 0.32)$*, pero eso no necesariamente implica que los tres valores supuestos son todos equivocados, sino que por lo menos uno de estos valores no es apropiado para el parámetro teórico.*

El anterior cálculo también se puede llevar a cabo usando el siguiente código que calcula conjuntamente las estimaciones muestrales, el valor de la estadística $2\ln\lambda$ *y el* p *valor.*

```
> multi<-function(x,p0){
+ if(length(x)!=length(p0))
+ stop("X y P0 deben tener el mismo tamaño")
+ r<-length(x)
+ est<-x/sum(x)
+ estad<-2*sum(x*log(est/p0))
+ p<-pchisq(estad,r-1,lower.tail = F)
+ list(estima=est,estadistica=estad,p.valor=p)
+ }
>
> x<-c(810,654,972)
> p_0<-c(0.43,0.25,0.32)
>
> multi(x,p_0)
$estima
[1] 0.3325123 0.2684729 0.3990148

$estadistica
[1] 105.7276

$p.valor
[1] 1.100361e-23
```

Nótese que obtenemos la misma decisión de rechazar H_0*.*

Como se comentó al principio del capítulo, también podemos utilizar la estadística de razón de verosimilitud para probar otros sistemas acerca de $\mathbf{p}$ donde H_0 no es necesariamente simple. Ilustramos con el siguiente ejemplo.

Ejemplo 6.1.2. *Continuando con el ejemplo anterior, suponga que se cree que el nivel de popularidad del candidato A y C son aproximadamente iguales, y por consiguiente se sospecha que obtenga el mismo porcentaje de votos. En este caso el sistema de hipótesis que se desea probar es*

$$H_0: \ p_1 = p_3 \quad \textit{vs.} \quad H_1: \ p_1 \neq p_3$$

Si hacemos uso de la estadística (6.1.2), el numerador de λ es de nuevo la función de verosimilitud evaluada en los estimadores de máxima verosimilitud de p_1, p_2 y p_3; mientras que para encontrar el denominador de λ, es necesario encontrar el estimador de máxima verosimilitud de la función de verosimilitud bajo H_0.

Bajo H_0 $p_1 = p_3$, la función de verosimilitud toma la forma de

$$L_0 = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_1^{x_3} = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1+x_3} p_2^{x_2}.$$

Podemos maximizar la función $\ln L$ sujeto a la restricción $p_1 + p_2 + p_3 = 1$ que ahora toma la forma de $2p_1 + p_2 = 1$ bajo H_0. De nuevo, utilizando el multiplicador de Lagrange tenemos que los estimadores de máxima verosimilitud de p_1 y p_2 bajo H_0 están dados por

$$\hat{p}_1 = \frac{x_1 + x_3}{2(x_1 + x_2 + x_3)} = \frac{1 - \bar{x}_2}{2}$$

y

$$\hat{p}_2 = \frac{x_2}{x_1 + x_2 + x_3} = \bar{x}_2.$$

Y la estadística de razón de verosimilitud λ está dada por

$$\begin{aligned} \lambda &= \frac{\bar{x}_1^{x_1} \bar{x}_2^{x_2} \bar{x}_3^{x_3}}{(\frac{1-\bar{x}_2}{2})^{x_1+x_3} \bar{x}_2^{x_2}} \\ &= \frac{\bar{x}_1^{x_1} \bar{x}_3^{x_3}}{(\frac{1-\bar{x}_2}{2})^{x_1+x_3}} \end{aligned}$$

Con el fin de utilizar la distribución nula asintótica de la estadística $2 \ln \lambda$ para encontrar una regla de decisión, tenemos que

$$2 \ln \lambda = 2 \left(x_1 \ln \left(\frac{2\bar{x}_1}{1 - \bar{x}_2} \right) + x_3 \ln \left(\frac{2\bar{x}_3}{1 - \bar{x}_2} \right) \right)$$

que se distribuye bajo H_0 como $\chi^2_{v_1 - v_0}$ asintóticamente. Bajo la hipótesis nula $p_1 = p_2$, entonces una vez se conoce el valor de p_2, necesariamente $p_1 = p_3 = (1 - p_2)/2$ de donde podemos ver que solo hay un parámetro libre bajo H_0, de donde $v_0 = 1$; por otro lado, en $\boldsymbol{\Theta}$ el número de parámetros libres es $k - 1$, en este caso $v_1 = 2$. Y tenemos que la distribución nula asintótica de $2 \ln \lambda$ es χ^2_1 y tenemos que la regla de decisión será rechazar H_0 si $2 \ln \lambda$ es mayor que el percentil $\chi^2_{1,1-\alpha}$.

Para los datos del ejemplo anterior, la estadística de prueba se calcula como

$$2\ln\lambda = 2\left(810\ln\left(\frac{2*810/2436}{1-654/2436}\right)+972\ln\left(\frac{2*972/2436}{1-654/2436}\right)\right)=14.75$$

el cual comparado con el percentil $\chi^2_{1,0.95}=3.84$ *conduce a la decisión de rechazar* H_0*, es decir, los datos muestran evidencia en contra de la hipótesis de que los candidatos A y C obtendrán el mismo porcentaje de votos.*

6.1.2 Dos muestras

Una aplicación muy importante de la distribución multinomial se encuentra en la rama de la investigación de mercados que entre todos sus tópicos estudia la percepción de los clientes con respecto a un producto. Una herramienta fundamental para conocer la opinión de los clientes en la investigación de mercados es por medio de las encuestas donde puede haber preguntas del tipo "¿Le gusta el empaque del producto?", donde las posibles respuestas pueden ser Mucho, Poco o Nada. De lo anterior, si en una muestra de n personas, definimos X_1, X_2 y X_3 como número de personas que respondieron Mucho, Poco y Nada sobre el empaque, entonces claramente (X_1, X_2, X_3) se distribuye como $Multi(n, p_1, p_2, p_3)$.

Ahora, cada producto según su presentación, empaque y precio puede cambiar su grado de aceptación en diferentes grupos de clientes, por ejemplo los diferentes estratos socioeconómicos, o hombres y mujeres, y dependiendo del perfil de estos grupos, las estrategias de promoción del producto van dirigiendo más a estos grupos. Por consiguiente, es importante para los directivos de la marca saber si efectivamente hay diferencia en los distintos grupos poblacionales con respecto a la percepción de la marca.

Por ejemplo, para saber si un nuevo sabor de café tiene mejor acogida entre los hombres o entre las mujeres, se realiza un estudio por encuesta donde se pregunta a los entrevistados cómo les ha parecido el producto. Si las posibles respuestas sobre el grado de gusto son bueno, regular y malo, entonces podemos en primer lugar calcular las estimaciones puntuales acerca de qué porcentajes de hombres les parecen bueno, regular y malo, y también los mismos porcentajes entre las mujeres, y posteriormente formular el siguiente sistema de hipótesis para ver si los hombres y las mujeres tienen diferencias significativas con respecto a la percepción que tienen acerca del nuevo producto.

$$H_0:\ (p_{11},p_{21},p_{31})=(p_{12},p_{22},p_{32}) \quad \text{vs.} \quad H_1:\ (p_{11},p_{21},p_{31})\neq(p_{12},p_{22},p_{32})$$

donde p_{11}, p_{21} y p_{31} corresponden a porcentajes de hombres a los que les ha parecido bueno, regular y malo el producto, y p_{12}, p_{22} y p_{32} los porcentajes correspondientes entre las mujeres.

Estimador de máxima verosimilitud y sus propiedades

Suponemos que tenemos dos vectores de la misma dimensión $(X_1,\cdots,X_r)$ y $(Y_1,\cdots,Y_r)$ que se distribuyen $Multi(n_X, p_{11},\cdots,p_{r1})$ y $Multi(n_Y, p_{12},\cdots,p_{r2})$,

respectivamente. Al suponer que estos dos vectores son independientes, tenemos que la función de verosimilitud conjunta de las dos muestras está dada por

$$L = \frac{n_X!}{x_1! \cdots x_r!} p_{11}^{x_1} \cdots p_{r1}^{x_r} \frac{n_Y!}{y_1! \cdots y_r!} p_{12}^{y_1} \cdots p_{r2}^{y_r}.$$

Para encontrar los estimadores de máxima verosimilitud, simplemente utilizamos el multiplicador de Lagrange para maximizar $\ln L$ dada por

$$\ln L = x_1 \ln p_{11} + \cdots + x_r \ln p_{r1} + y_1 \ln p_{12} + \cdots + y_r \ln p_{r2} + \ln \frac{n_X!}{x_1! \cdots x_r!} + \ln \frac{n_Y!}{y_1! \cdots y_r!} \quad (6.1.3)$$

sujeto a las dos restricciones $p_{11} + \cdots + p_{r1} = 1$ y $p_{12} + \cdots + p_{r2} = 1$, y es fácil ver que los estimadores de máxima verosimilitud están dados por (Ejercicio 6.3)

$$\hat{p}_{i1,MV} = \frac{X_i}{n_X} = \bar{X}_i \quad (6.1.4)$$

y

$$\hat{p}_{i2,MV} = \frac{Y_i}{n_Y} = \bar{Y}_i \quad (6.1.5)$$

para $i = 1, \cdots, r$. Además también es fácil ver que cada uno de estos estimadores es insesgado.

Prueba de hipótesis

Dada la motivación en el campo de investigación de mercados, estamos interesados en el siguiente sistema de hipótesis

$$H_0 : \ (p_{11}, \cdots, p_{r1}) = (p_{12}, \cdots, p_{r2}) \quad \text{vs.} \quad H_1 : \ (p_{11}, \cdots, p_{r1}) \neq (p_{12}, \cdots, p_{r2}). \quad (6.1.6)$$

Utilizaremos la prueba de razón generalizada de verosimilitud dada en (6.1.2) para este sistema. De nuevo $\boldsymbol{\Theta}_0 \bigcup \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}$ es el espacio paramétrico completo de los vectores $(p_{11}, \cdots, p_{r1})$ y $(p_{12}, \cdots, p_{r2})$, por consiguiente el numerador de λ será la función de verosimilitud L evaluada en los estimadores de máxima verosimilitud encontrados en (6.1.4) y (6.1.5), $L(\hat{p}_{11}, \cdots, \hat{p}_{r1}, \hat{p}_{12}, \cdots, \hat{p}_{r2})$.

Por otro lado, para encontrar el denominador de λ, es necesario maximizar L bajo H_0. Bajo H_0, tenemos que $p_{i1} = p_{i2} = p_i$ para todo $i = 1, \cdots, r$, y L se convierte en

$$\begin{aligned} L &= \frac{n_X!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r} \frac{n_Y!}{y_1! \cdots y_r!} p_1^{y_1} \cdots p_r^{y_r}. \\ &= p_1^{x_1+y_1} \cdots p_r^{x_r+y_r} \frac{n_X!}{x_1! \cdots x_r!} \frac{n_Y!}{y_1! \cdots y_r!}. \end{aligned}$$

Tomando logaritmo a L, tenemos que

$$\ln L = (x_1 + y_1)\ln p_1 + \cdots + (x_r + y_r)\ln p_r + \ln \frac{n_X!}{x_1! \cdots x_r!} + \ln \frac{n_Y!}{y_1! \cdots y_r!}.$$

Maximizando $\ln L$ sujeto a la restricción de que $p_1 + \cdots + p_r = 1$, tenemos que el estimador de máxima verosimilitud de p_i bajo H_0 es

$$\hat{p}_i = \frac{x_i + y_i}{n_X + n_Y}.$$

para todo $i = 1, \cdots, r$. Y por consiguiente el denominador de λ será $L(\hat{p}_1, \cdots, \hat{p}_r)$. Y

$$\lambda = \frac{L(\hat{p}_{11}, \cdots, \hat{p}_{r1}, \hat{p}_{12}, \cdots, \hat{p}_{r2})}{L(\hat{p}_1, \cdots, \hat{p}_r)}$$

de donde la estadística $2\ln\lambda$ está dada por

$$\begin{aligned} 2\ln\lambda &= 2\left(\ln L(\hat{p}_{11}, \cdots, \hat{p}_{r1}, \hat{p}_{12}, \cdots, \hat{p}_{r2}) + \ln L(\hat{p}_1, \cdots, \hat{p}_r)\right) \\ &= 2(x_1 \ln \bar{x}_1 + \cdots + x_r \ln \bar{x}_r + y_1 \ln \bar{y}_1 + \cdots + y_r \ln \bar{y}_r \\ &\qquad - (x_1 + y_1)\ln \frac{x_1 + y_1}{n_X + n_Y} - \cdots - (x_r + y_r)\ln \frac{x_r + y_r}{n_X + n_Y}). \end{aligned}$$

La distribución nula asintótica de esta estadística es $\chi^2_{v_1 - v_0}$ donde v_1 y v_0 son los grados de libertad de los parámetros bajo $\boldsymbol{\Theta}$ y $\boldsymbol{\Theta}_0$, respectivamente. Bajo $\boldsymbol{\Theta}$, la única restricción que deben cumplir los parámetros es que $p_{11} + \cdots + p_{r1} = 1$ y $p_{12} + \cdots + p_{r2} = 1$, por lo tanto $v_1 = 2(r-1)$; por otro lado, bajo $\boldsymbol{\Theta}_0$, los parámetros de las dos muestras son iguales, entonces $v_0 = r - 1$, de donde $v_1 - v_0 = r - 1$, y tenemos que

$$2\ln\lambda \sim_{asym} \chi^2_{r-1}$$

Y la regla de decisión para el sistema (6.1.6) es rechazar H_0 si $2\ln\lambda > \chi^2_{r-1,1-\alpha}$.

Teniendo en cuenta que H_0 se rechaza para valores grandes de $2\ln\lambda$ podemos calcular el p valor como

$$p \text{ valor} = 1 - F_{\chi^2_{r-1}}(v)$$

donde v es el valor de la estadística $2\ln\lambda$ y $F_{\chi^2_{r-1}}$ denota la función de distribución de la distribución χ^2_{r-1}. Aplicamos la anterior teoría en el siguiente ejemplo.

Ejemplo 6.1.3. *Suponga que se desea conocer la opinión que tienen los consumidores acerca del nuevo café con sabor de vainilla que una compañía lanzó hace dos meses al mercado. A cada persona entrevistada se le pregunta cómo le ha parecido el producto, y la respuesta puede ser Bueno, Regular o Malo. La entrevista se realizó con 586 personas, donde 349 son hombres y 237 son mujeres. Estamos interesados en saber si hay alguna diferencia significativa entre los dos géneros con respecto a la percepción que tienen acerca del producto. Si se detecta que el producto es débil entre las mujeres, se puede diseñar estrategias para traer más consumidores femeninos y así obtener mayor ganancia para la compañía.*

En la Tabla 6.1 se muestran los resultados de la encuesta donde cada entrada representa el número de personas del género correspondiente que respondieron Bueno, Regular y Malo. Y el sistema de que desea probar es

$$H_0: (p_{11}, p_{21}, p_{31}) = (p_{12}, p_{22}, p_{32}) \quad \textit{vs.} \quad H_1: (p_{11}, p_{21}, p_{31}) \neq (p_{12}, p_{22}, p_{32})$$

donde p_{11}, p_{21} y p_{31} corresponden a porcentajes de hombres a los que les han parecido bueno, regular y malo el producto, y p_{12}, p_{22} y p_{32} los porcentajes correspondientes entre las mujeres.

	Hombre	Mujer
Bueno	192	127
Regular	140	90
Malo	17	20
Total	349	237

Tabla 6.1: *Datos del ejemplo 6.1.3*

En este ejemplo, $k = 3$, $n_X = 349$ y $n_Y = 237$, y el siguiente código calcula las estimaciones puntuales, la estadística $2 \ln \lambda$ y el correspondiente p valor

```
> multi_2_muestra<-function(x,y){
+ if(length(x)!=length(y))
+ stop("X y Y deben tener el mismo tamaño")
+ r<-length(x)
+ est.X<-x/sum(x)
+ est.Y<-y/sum(y)
+ l1<-sum(x*log(x/sum(x)))+sum(y*log(y/sum(y)))
+ l2<-sum((x+y)*log((x+y)/sum(x+y)))
+ estad<-2*(l1-l2)
+ p<-pchisq(estad,r-1,lower.tail = F)
+ list(estima.X=est.X,estima.Y=est.Y,estadistica=estad,p.valor=p)
+ }
> hombre<-c(192,140,17)
> mujer<-c(127,90,20)
> multi_2_muestra(hombre,mujer)
$estima.X
[1] 0.5501433 0.4011461 0.0487106

$estima.Y
[1] 0.53586498 0.37974684 0.08438819

$estadistica
[1] 2.99966

$p.valor
[1] 0.2231681
```

Del p valor podemos ver que no hay una diferencia significativa entre los hombres y mujeres con respecto a la percepción del nuevo café con sabor a vainilla. De lo anterior, según la posición que tiene el producto frente a los competidores, las estrategias de mercadeo de este producto pueden seguir sin alterarse (si el producto ya tiene buena posición en el mercado) o en el otro caso, incorporar mejoras para atraer consumidores de ambos géneros sin enfocarse en un género en particular.

6.1.3 k muestras

La teoría expuesta anteriormente se puede extender para el caso de k muestras, que en el caso de investigación de mercados, puede ser utilizado para investigar si un producto tiene diferentes posiciones en k subgrupos poblacionales con $k > 2$. Por ejemplo, cómo es el hábito o frecuencia de consumo de espaguetis en estratos bajos, medios y altos[1]

En general, se dispone de una muestra aleatoria para cada subgrupo poblacional donde tenemos k vectores de parámetros $\mathbf{p}_1, \cdots, \mathbf{p}_k$ con $\mathbf{p}_j = (p_{1j}, \cdots, p_{rj})$ con p_{ij} denotando la proporción teórica de individuos de la población j que dieron la respuesta i, para $j = 1, \cdots, k$ y $i = 1, \cdots, r$. Y denotamos la muestra observada de tamaño n_j en la población j como $X_{1j}, \cdots, X_{rj}$, esto es, X_{ij} denota el número de individuos de la población j que dieron la respuesta i.

Es claro que teniendo el supuesto de que las k poblaciones son independientes, los estimadores de máxima verosimilitud de los porcentajes p_{ij} son simplemente las proporciones muestrales, es decir

$$\hat{p}_{ij} = \frac{X_{ij}}{n_i} = \bar{X}_{ij} \tag{6.1.7}$$

para $j = 1, \cdots, k$ y $i = 1, \cdots, r$. Y para ver si hay diferencia en las k poblaciones, planteamos el sistema

$$H_0 : \ \mathbf{p}_1 = \cdots = \mathbf{p}_k \quad \text{vs.} \quad H_1 : \ \text{Existen } \mathbf{p}_i \neq \mathbf{p}_j \text{ para algún } i, j \tag{6.1.8}$$

Y para calcular la estadística $2 \ln \lambda$, primero obtenemos el estimador de máxima verosimilitud de los parámetros bajo H_0 que establece que $p_{i1} = \cdots = p_{ij} = p_i$ para todo $i = 1, \cdots, r$ y $j = 1, \cdots, k$. Los estimadores de p_i están dados por

$$\hat{p}_i = \frac{x_{i1} + \cdots + x_{ik}}{n_1 + \cdots + n_k}. \tag{6.1.9}$$

Usando la anterior expresión y (6.1.7), tenemos que la estadística $2 \ln \lambda$ está dada por

$$2 \ln \lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{k} x_{ij} \ln \frac{\bar{x}_{ij}}{\hat{p}_i}.$$

[1] En el caso de Colombia, los estratos 1 y 2 se pueden clasificar como los bajos, 3 y 4 como medios, y 5 y 6 como altos.

Para encontrar el grado de libertad de la distribución nula asintótica de $2\ln\lambda$, observamos que en el espacio paramétrico completo, en cada una de las k poblaciones hay r parámetros que deben cumplir la condición de que la suma de las proporciones teóricas sea igual a 1; de esta forma, en cada población hay $r-1$ parámetros libres y en total $v_1 = k(r-1)$. Por otro lado, bajo H_0 las k poblaciones tienen el mismo vector de parámetros, es decir r parámetros, y al tener en cuenta la restricción, tenemos que $v_0 = r-1$, de esta forma tenemos que

$$2\ln\lambda \sim_{asym} \chi^2_{(r-1)*(k-1)}.$$

Rechazamos H_0 si $2\ln\lambda > \chi^2_{(r-1)*(k-1),1-\alpha}$ y el p valor correspondiente se puede calcular como

$$p \text{ valor} = 1 - F_{\chi^2_{(r-1)*(k-1)}}(v)$$

donde v es el valor de $2\ln\lambda$ en la muestra. El lector puede ver que las fórmulas desarrolladas para el caso de dos muestras es simplemente un caso particular de las fórmulas anteriores cuando $k = 2$.

Ejemplo 6.1.4. *Con el fin de conocer el perfil de los diferentes estratos con respecto a la frecuencia de consumo de espaguetis, se analizan los datos de la Tabla 6.2 obtenidos en una encuesta realizada a hogares de diferentes estratos con respecto al consumo de espaguetis.*

	Estrato bajo	Estrato medio	Estrato alto
Más de tres veces a la semana	95	53	33
Entre una y tres veces a la semana	120	97	79
De vez en cuando	35	45	19
Nunca consume espaguetis	6	13	3
Total entrevistados	256	208	134

Tabla 6.2: *Datos del Ejemplo 6.1.4*

La siguiente función nos permite calcular las estimaciones puntuales, el valor de la estadística de prueba y el correspondiente p valor.

```
> multi_k_muestra<-function(x){
+ # columna j de x debe corresponder a los conteos en la muestra j
+ r<-dim(x)[1]
+ k<-dim(x)[2]
+ est<-matrix(NA,r,k)
+ for(j in 1:k){
+ est[,j]<-x[,j]/colSums(x)[j]} ## estimación MV
+ p_0<-rowSums(x)/(sum(x)) ## estimación MV bajo H0
+ l1<-sum(x*log(est))
+ l2<-sum(rowSums(x)*log(p_0))
+ estad<-2*(l1-l2)
+ p<-pchisq(estad,(r-1)*(k-1),lower.tail = F)
```

```
+ list(est=est,estadistica=estad,p.valor=p)
+ }
>
> x<-matrix(c(95,120,35,6,53,97,45,13,33,79,19,3),4,3)
> multi_k_muestra(x)
$est
          [,1]      [,2]       [,3]
[1,] 0.3710938 0.2548077 0.24626866
[2,] 0.4687500 0.4663462 0.58955224
[3,] 0.1367188 0.2163462 0.14179104
[4,] 0.0234375 0.0625000 0.02238806

$estadistica
[1] 20.04436

$p.valor
[1] 0.002719489
```

Podemos observar que hay una diferencia significativa entre los estratos con respecto al consumo de espaguetis que indica que al momento de promocionar el producto, se debe tener en cuenta el grupo poblacional que más consume espaquetis y que, dadas las estimaciones puntuales, pueden ser los consumidores de estratos bajos.

6.2 Inferencia en la distribución normal multivariante

6.2.1 Estimador de máxima verosimilitud

En una muestra aleatoria proveniente de la distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, los parámetros son $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, la función de verosimilitud está dada por:

$$\begin{aligned} L(\mathbf{x}_1, \cdots, \mathbf{x}_n, \boldsymbol{\theta})) &= \prod_{i=1}^{n} f_{\mathbf{X}_i}(\mathbf{x}_i, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} |\boldsymbol{\Sigma}|^{-1/2} (2\pi)^{-p/2} \exp\left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right\} \\ &= |\boldsymbol{\Sigma}|^{-n/2} (2\pi)^{-pn/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right\} \end{aligned}$$

Recordando que maximizar una función f es equivalente a maximizar la función $\ln f$, tenemos que

$$\ln L = \frac{n}{2} \ln |\boldsymbol{\Sigma}^{-1}| - \frac{pn}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}). \tag{6.2.1}$$

La forma clásica para maximizar $\ln L$ consiste en sumar y restar el término $\bar{\mathbf{x}}$, el promedio de $\mathbf{x}_1, \cdots, \mathbf{x}_n$. Tenemos que:

$$\begin{aligned}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) &= \sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\boldsymbol{\mu}) \\ &= \underbrace{\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})}_{A} + n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}).\end{aligned}$$

En la anterior expresión A es un escalar, y por consiguiente, $A = tr(A)$, donde $tr(\cdot)$ denota el operador traza, esto es

$$\begin{aligned}A &= \sum_{i=1}^{n} tr((\mathbf{x}_i-\bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})) \\ &= \sum_{i=1}^{n} tr(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}}))') \\ &= tr(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}}))') \\ &= tr(\boldsymbol{\Sigma}^{-1}n\mathbf{S}_n) \\ &= ntr(\boldsymbol{\Sigma}^{-1}\mathbf{S}_n),\end{aligned}$$

con $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})$. Sustituyendo la anterior expresión en (6.2.1), se tiene que

$$\ln L = \frac{n}{2}\ln|\boldsymbol{\Sigma}^{-1}| - \frac{pn}{2}\ln 2\pi - \frac{n}{2}tr(\boldsymbol{\Sigma}^{-1}\mathbf{S}_n) - \frac{n}{2}(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}). \quad (6.2.2)$$

Para encontrar el valor de $\boldsymbol{\mu}$ que maximiza esta función, se observa que es el mismo valor de $\boldsymbol{\mu}$ que minimiza el término $(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})$, pero la matriz $\boldsymbol{\Sigma}^{-1}$ es semidefinida positiva, es decir, $(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \geq 0$, por lo tanto el valor más pequeño que puede tomar es el valor 0, y esto ocurre cuando $\boldsymbol{\mu} = \bar{\mathbf{x}}$. Por lo tanto, se concluye que $\hat{\boldsymbol{\mu}}_{MV} = \bar{\mathbf{X}}$.

Ahora, para encontrar el valor de $\boldsymbol{\Sigma}$ que maximiza a $\ln L$, en primer lugar se sustituye $\boldsymbol{\mu}$ por su estimador, lo cual conduce a maximizar la función $\boldsymbol{\mu}$, agregando el término $\frac{n}{2}\ln|\mathbf{S}_n|$ y eliminando el término $-\frac{pn}{2}\ln 2\pi$ que no dependen de $\boldsymbol{\Sigma}$, se tiene que la función que se debe maximizar es

$$L^* = -\frac{n}{2}\ln|\boldsymbol{\Sigma}| + \frac{n}{2}\ln|\mathbf{S}_n| - \frac{n}{2}tr(\boldsymbol{\Sigma}^{-1}\mathbf{S}_n). \quad (6.2.3)$$

Aplicando propiedades del determinante, se tiene que:

$$\begin{aligned} L^* &= \frac{n}{2}\ln\frac{1}{|\boldsymbol{\Sigma}|} + \frac{n}{2}\ln|\mathbf{S}_n| - \frac{n}{2}tr(\boldsymbol{\Sigma}^{-1}\mathbf{S}_n) \\ &= \frac{n}{2}\ln|\boldsymbol{\Sigma}^{-1}| + \frac{n}{2}\ln|\mathbf{S}_n| - \frac{n}{2}tr(\boldsymbol{\Sigma}^{-1}\mathbf{S}_n) \\ &= \frac{n}{2}\ln|\boldsymbol{\Sigma}^{-1}\mathbf{S}_n| - \frac{n}{2}tr(\boldsymbol{\Sigma}^{-1}\mathbf{S}_n). \end{aligned}$$

Ahora, para cualquier matriz cuadrada A, se tiene que $|A| = \prod\lambda_i$ y $tr(A) = \sum\lambda_i$ donde los valores λ_i denotan los valores propios de A. Aplicando el anterior resultado a la matriz $\boldsymbol{\Sigma}^{-1}\mathbf{S}_n$, se tiene que

$$\begin{aligned} L^* &= \frac{n}{2}\ln\prod_{i=1}^{p}\lambda_i - \frac{n}{2}\sum_{i=1}^{p}\lambda_i \\ &= \frac{n}{2}\sum_{i=1}^{p}\ln\lambda_i - \frac{n}{2}\sum_{i=1}^{p}\lambda_i \\ &= \frac{n}{2}\sum_{i=1}^{P} r(\ln\lambda_i - \lambda_i) \end{aligned}$$

donde $\lambda_1, \cdots, \lambda_p$ son los valores propios de $\boldsymbol{\Sigma}^{-1}\mathbf{S}_n$.

Ahora, considere la función $f(x) = \ln x - x$. Esta función tiene un máximo en $x = 1$, entonces la función L^* tiene un máximo cuando $\lambda_i = 1$ para todo $i = 1, \cdots, p$. Es decir, L^* tiene un máximo cuando los valores propios de $\boldsymbol{\Sigma}^{-1}\mathbf{S}_n$ son iguales a 1. Esto implica que $\boldsymbol{\Sigma}^{-1}\mathbf{S}_n$ es la matriz identidad, de donde se concluye que el valor de $\boldsymbol{\Sigma}$ que maximiza la función $\ln L$ es $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})$. En conclusión $\hat{\boldsymbol{\Sigma}}_{MV} = \mathbf{S}_n$.

Existe otra forma de encontrar los estimadores $\hat{\boldsymbol{\mu}}_{MV}$ y $\hat{\boldsymbol{\Sigma}}_{MV}$ utilizando derivadas, análogamente al caso de inferencia univariada bajo distribución normal, pero en este caso se deriva con respecto a vector y matriz. En el apéndice F, se hace un resumen de las derivadas matriciales que se utilizará a continuación.

Primero escribimos a la función $\ln L$ dada en (6.2.2) como

$$\ln L = \frac{n}{2}\ln|\boldsymbol{\Sigma}^{-1}| - \frac{pn}{2}\ln 2\pi - \frac{n}{2}tr(\boldsymbol{\Sigma}^{-1}\mathbf{S}_n) - \frac{n}{2}tr\left\{\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'\right\}.$$

Las derivadas de $\ln L$ con respecto a los parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ se calculan como

$$\begin{aligned} \frac{\partial \ln L}{\partial \boldsymbol{\mu}} &= -\frac{n}{2}\frac{\partial tr\left\{\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'\right\}}{\partial \boldsymbol{\mu}} \\ &= -\frac{n}{2}\frac{\partial tr\left\{\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'\right\}}{\partial(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'}\frac{(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'}{\partial \boldsymbol{\mu}} \\ &= -\frac{n}{2}\boldsymbol{\Sigma}^{-1}(-2)(\bar{\mathbf{x}} - \boldsymbol{\mu}), \end{aligned}$$

De donde se tiene que $\hat{\boldsymbol{\mu}}_{MV} = \bar{\mathbf{x}}$.

Ahora, para encontrar $\hat{\boldsymbol{\Sigma}}_{MV}$, no derivamos $\ln L$ con respecto a $\boldsymbol{\Sigma}$, sino con respecto a $\boldsymbol{\Sigma}^{-1}$, apoyándonos en el argumento de que el valor de $\boldsymbol{\Sigma}$ que satisface $\frac{\partial \ln L}{\partial \boldsymbol{\Sigma}^{-1}} = 0$ también satisface $\frac{\partial \ln L}{\partial \boldsymbol{\Sigma}} = 0$. Tenemos

$$\begin{aligned}
\frac{\partial \ln L}{\partial \boldsymbol{\Sigma}^{-1}} &= \frac{n}{2}\frac{\partial \ln |\boldsymbol{\Sigma}^{-1}|}{\partial \boldsymbol{\Sigma}^{-1}} - \frac{n}{2}\frac{\partial tr\left\{\boldsymbol{\Sigma}^{-1}(\mathbf{S}_n + (\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})')\right\}}{\partial \boldsymbol{\Sigma}^{-1}} \\
&= \frac{n}{2}\frac{\partial \ln |\boldsymbol{\Sigma}^{-1}|}{\partial |\boldsymbol{\Sigma}^{-1}|}\frac{\partial |\boldsymbol{\Sigma}^{-1}|}{\partial \boldsymbol{\Sigma}^{-1}} - \frac{n}{2}\frac{\partial tr\left\{\boldsymbol{\Sigma}^{-1}(\mathbf{S}_n + (\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})')\right\}}{\partial \boldsymbol{\Sigma}^{-1}} \\
&= \frac{n}{2}\left\{2\boldsymbol{\Sigma} - diag(\boldsymbol{\Sigma})\right\} \\
&- \frac{n}{2}\left\{2(\mathbf{S}_n + (\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})') - diag(\mathbf{S}_n + (\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})')\right\} \\
&= \frac{n}{2}\left\{2(\boldsymbol{\Sigma} - \mathbf{S}_n - (\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})') - diag(\boldsymbol{\Sigma} - \mathbf{S}_n - (\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})')\right\}
\end{aligned}$$

De esta forma, se debe cumplir que

$$2(\boldsymbol{\Sigma} - \mathbf{S}_n - (\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})') - diag(\boldsymbol{\Sigma} - \mathbf{S}_n - (\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})') = \mathbf{0}$$

Es fácil verificar que para una matriz $\mathbf{A}$, si $2\mathbf{A} - diag(\mathbf{A}) = \mathbf{0}$, entonces necesariamente $\mathbf{A} = \mathbf{0}$. Por lo tanto, tenemos que

$$\boldsymbol{\Sigma} = \mathbf{S}_n + (\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'$$

Reemplazando $\boldsymbol{\mu}$ por su estimador de máxima verosimilitud $\bar{\mathbf{X}}$, tenemos que

$$\hat{\boldsymbol{\Sigma}}_{MV} = \mathbf{S}_n$$

Ahora, usando $\mathbf{S}_n$ como estimador de $\boldsymbol{\Sigma}$, podemos obtener un estimador de la matriz de correlaciones $\boldsymbol{\rho}$ como

$$\hat{\rho}_{MV} = \hat{\mathbf{D}}^{-1/2}\mathbf{S}_n\hat{\mathbf{D}}^{-1/2}, \tag{6.2.4}$$

donde $\hat{\mathbf{D}} = diag(\mathbf{S}_n)$.

Ilustramos el cálculo de estas matrices con los datos de Student (1908) correspondientes a horas de incremento de sueño debido al uso de dos tipos de sedantes. Estos datos se encuentran en la Tabla 5.2. En R, las funciones que calculan la matriz de varianzas y covarianzas muestrales y la matriz de correlaciones son `var` y `cor`.

```
> a<-c(0.7,-1.6,-0.2,-1.2,-1,3.4,3.7,0.8,0,2)
> b<-c(1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,1.4)
> x<-matrix(c(a,b),10,2)
```

```
> var(x)*9/10
        [,1]   [,2]
[1,] 3.1064 2.7822
[2,] 2.7822 3.6081

> cor(x)
       [,1]      [,2]
[1,] 1.000000 0.831037
[2,] 0.831037 1.000000
```

De donde tenemos que la matriz de varianzas y covarianzas y la matriz de correlaciones muestrales están dadas por

$$\hat{\mathbf{S}}_n = \begin{pmatrix} 3.11 & 2.78 \\ 2.78 & 3.61s \end{pmatrix}$$

y

$$\hat{\rho} = \begin{pmatrix} 1 & 0.83 \\ 0.83 & 1 \end{pmatrix}$$

También podemos encontrar el estimador de mínimos cuadrados para el vector de medias teóricas $\boldsymbol{\mu}$. Siguiendo el mismo razonamiento en la estimación de mínimos cuadrados en el caso univariado, se espera que el el vector $\boldsymbol{\mu}$ esté cercano a las observaciones $\mathbf{X}_1, \cdots, \mathbf{X}_n$. En este caso, necesitamos medir la distancia entre $\boldsymbol{\mu}$ y cada $\mathbf{X}_i$ con $i = 1, \cdots, n$, una de las distancias más comunes entre vectores es la distancia euclidiana. Utilizando esta distancia, debemos encontrar el valor de $\boldsymbol{\mu}$ que minimiza la cantidad

$$Q = \sum_{i=1}^{n} (\mathbf{X}_i - \boldsymbol{\mu})'(\mathbf{X}_i - \boldsymbol{\mu}). \tag{6.2.5}$$

Y lo presentamos a continuación.

Resultado 6.2.1. *Sea $\mathbf{X}_1, \cdots, \mathbf{X}_n$ una muestra aleatoria proveniente de una distribución con media teórica $\boldsymbol{\mu}$, entonces el estimador de mínimos cuadrados de $\boldsymbol{\mu}$ es $\bar{\mathbf{X}}$.*

Demostración. Para encontrar el estimador de mínimos cuadrados de $\boldsymbol{\mu}$ derivamos la expresión (6.2.5) con respecto a $\boldsymbol{\mu}$, tenemos que

$$\frac{\partial Q}{\partial \boldsymbol{\mu}} = -2 \sum_{i=1}^{n} (\mathbf{X}_i - \boldsymbol{\mu}).$$

Igualando la anterior expresión a cero, tenemos que $\boldsymbol{\mu} = \sum_{i=1}^{n} \mathbf{X}_i / n = \bar{X}$, el cual coincide con el estimador de máxima verosimilitud encontrado anteriormente. □

Adicionalmente, para el desarrollo del anterior estimador, no se tuvo en cuenta la distribución teórica; de esta forma, en muestras provenientes de cualquier distribución

multivariante podemos estimar el vector de medias teóricas con la media muestral $\bar{X}$. Los lectores pueden ver que bajo la distribución multinomial el estimador de máxima verosimilitud de $\mathbf{p}$ también coincide con el estimador de mínimos cuadrados $\bar{X}$.

6.2.2 Propiedades de los estimadores de máxima verosimilitud

En el ámbito de la inferencia multivariante, la calidad de un estimador se mide a través de su esperanza y su matriz de varianzas y covarianzas.

Para el estimador de máxima verosimilitud de $\boldsymbol{\mu}$: $\bar{\mathbf{X}}$, se ha visto que la distribución del estimador es $N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$ (5.2.8), de donde se concluye que $E(\bar{\mathbf{X}}) = \boldsymbol{\mu}$, esto es, $\bar{\mathbf{X}}$ es un estimador insesgado de $\boldsymbol{\mu}$. Por otro lado, $Var(\bar{\mathbf{X}}) = \frac{1}{n}\boldsymbol{\Sigma}$, esto implica que al aumentar el tamaño de la muestra n, los componentes de la matriz de varianzas y covarianzas de $\bar{\mathbf{X}}$ disminuyen, y se puede estimar con más precisión al vector $\boldsymbol{\mu}$.

Ahora, con respecto al estimador de máxima verosimilitud de $\boldsymbol{\Sigma}$, notado como $\mathbf{S}_n$, primero se estudia su sesgo. Luego, se tiene que

$$\begin{aligned}
E(\mathbf{S}_n) &= \frac{1}{n}E(\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})) \\
&= \frac{1}{n}E(\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})') - E((\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})') \\
&= \frac{1}{n}\sum_{i=1}^{n}E((\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})') - E((\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})') \\
&= \boldsymbol{\Sigma} - Var(\bar{\mathbf{x}} - \boldsymbol{\mu}) \\
&= \boldsymbol{\Sigma} - \frac{1}{n}\boldsymbol{\Sigma} \\
&= \frac{n-1}{n}\boldsymbol{\Sigma},
\end{aligned}$$

de donde se concluye que el estimador de máxima verosimilitud $\mathbf{S}_n$ no es insesgado para $\boldsymbol{\Sigma}$ (la misma situación ocurrió en la inferencia univariada). Pero una simple modificación de $\mathbf{S}_n$ multiplicando por $\frac{n}{n-1}$ nos lleva a un estimador insesgado. En conclusión, un estimador insesgado para $\boldsymbol{\Sigma}$ es

$$\mathbf{S}_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'.$$

Resultado 6.2.2. *Se tiene que* $(n-1)\mathbf{S}_{n-1}$ *tiene distribución* $W(n-1, \boldsymbol{\Sigma})$.

Demostración. En primer lugar, tenemos que

$$\begin{aligned}
(n-1)\mathbf{S}_{n-1} &= \sum_{i=1}^{n}(\mathbf{X}_i-\bar{\mathbf{X}})(\mathbf{X}_i-\bar{\mathbf{X}})' \\
&= \sum_{i=1}^{n}(\mathbf{X}_i-\boldsymbol{\mu}+\boldsymbol{\mu}-\bar{\mathbf{X}})(\mathbf{X}_i-\boldsymbol{\mu}+\boldsymbol{\mu}-\bar{\mathbf{X}})' \\
&= \sum_{i=1}^{n}(\mathbf{X}_i-\boldsymbol{\mu})(\mathbf{X}_i-\boldsymbol{\mu})' + \sum_{i=1}^{n}(\mathbf{X}_i-\boldsymbol{\mu})(\boldsymbol{\mu}-\bar{\mathbf{X}})' \\
&+ (\boldsymbol{\mu}-\bar{\mathbf{X}})\sum_{i=1}^{n}(\mathbf{X}_i-\boldsymbol{\mu})' + n(\boldsymbol{\mu}-\bar{\mathbf{X}})(\boldsymbol{\mu}-\bar{\mathbf{X}})' \\
&= \sum_{i=1}^{n}(\mathbf{X}_i-\boldsymbol{\mu})(\mathbf{X}_i-\boldsymbol{\mu})' + n(\bar{\mathbf{X}}-\boldsymbol{\mu})(\boldsymbol{\mu}-\bar{\mathbf{X}})' \\
&+ n(\boldsymbol{\mu}-\bar{\mathbf{X}})(\bar{\mathbf{X}}-\boldsymbol{\mu})' + n(\boldsymbol{\mu}-\bar{\mathbf{X}})(\boldsymbol{\mu}-\bar{\mathbf{X}})' \\
&= \sum_{i=1}^{n}(\mathbf{X}_i-\boldsymbol{\mu})(\mathbf{X}_i-\boldsymbol{\mu})' - n(\bar{\mathbf{X}}-\boldsymbol{\mu})(\bar{\mathbf{X}}-\boldsymbol{\mu})' \\
&- n(\bar{\mathbf{X}}-\boldsymbol{\mu})(\bar{\mathbf{X}}-\boldsymbol{\mu})' + n(\bar{\mathbf{X}}-\boldsymbol{\mu})(\bar{\mathbf{X}}-\boldsymbol{\mu})' \\
&= \sum_{i=1}^{n}(\mathbf{X}_i-\boldsymbol{\mu})(\mathbf{X}_i-\boldsymbol{\mu})' - n(\bar{\mathbf{X}}-\boldsymbol{\mu})(\bar{\mathbf{X}}-\boldsymbol{\mu})'
\end{aligned}$$

En el primer término los vectores $\mathbf{X}_i-\boldsymbol{\mu}$ con $i=1,\cdots,n$ son independientes e idénticamente distribuidos como $N_p(\mathbf{0},\boldsymbol{\Sigma})$, entonces por la definición de la distribución Wishart, se tiene que el término $\sum_{i=1}^{n}(\mathbf{X}_i-\boldsymbol{\mu})(\mathbf{X}_i-\boldsymbol{\mu})'$ tiene distribución $W(n,\boldsymbol{\Sigma})$, y su función característica está dada por $|I_p-2i\Theta\Sigma|^{-n/2}$.

Por otro lado, $\bar{\mathbf{X}}\sim N_p(\boldsymbol{\mu},\frac{1}{n}\boldsymbol{\Sigma})$, de donde se concluye que $\bar{\mathbf{X}}-\boldsymbol{\mu}\sim N_p(\mathbf{0},\frac{1}{n}\boldsymbol{\Sigma})$, y por consiguiente $\sqrt{n}(\bar{\mathbf{X}}-\boldsymbol{\mu})\sim N_p(\mathbf{0},\boldsymbol{\Sigma})$, entonces se concluye que $n(\bar{\mathbf{X}}-\boldsymbol{\mu})(\bar{\mathbf{X}}-\boldsymbol{\mu})'\sim W(1,\boldsymbol{\Sigma})$, y su función característica está dada por $|I_p-2i\Theta\Sigma|^{-1/2}$.

Usando la igualdad $(n-1)\mathbf{S}_{n-1}=\sum_{i=1}^{n}(\mathbf{X}_i-\boldsymbol{\mu})(\mathbf{X}_i-\boldsymbol{\mu})'-n(\bar{\mathbf{X}}-\boldsymbol{\mu})(\bar{\mathbf{X}}-\boldsymbol{\mu})'$, se tiene que la función característica de $(n-1)\mathbf{S}_{n-1}$ está dada por $|I_p-2i\Theta\Sigma|^{-n/2}/|I_p-2i\Theta\Sigma|^{-1/2}=|I_p-2i\Theta\Sigma|^{-(n-1)/2}$, la cual corresponde a una distribución $W(n-1,\boldsymbol{\Sigma})$. □

Del anterior resultado, se tiene que $E((n-1)\mathbf{S}_{n-1})=(n-1)\boldsymbol{\Sigma}$, de donde $E(\mathbf{S}_{n-1})=\boldsymbol{\Sigma}$, indicando una vez más que $\mathbf{S}_{n-1}$ es un estimador insesgado para $\boldsymbol{\Sigma}$.

Para los datos de incremento de sueños utilizados anteriormente, podemos calcular la matriz $\mathbf{S}_{n-1}$ usando

```
> a<-c(0.7,-1.6,-0.2,-1.2,-1,3.4,3.7,0.8,0,2)
> b<-c(1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,1.4)
> x<-matrix(c(a,b),10,2)
```

```
> var(x)
         [,1]     [,2]
[1,] 3.451556 3.091333
[2,] 3.091333 4.009000
> cor(x)
       [,1]      [,2]
[1,] 1.000000 0.831037
[2,] 0.831037 1.000000
```

dando como resultado

$$\mathbf{S}_{n-1} = \begin{pmatrix} 3.45 & 3.09 \\ 3.09 & 4.01 \end{pmatrix}$$

Utilizando la anterior matriz $\mathbf{S}_{n-1}$ y el Resultado 5.2.6., podemos conocer acerca del incremento en horas de sueño debido al sedante A y la necesidad de aplicar este sedante a un paciente si tenemos el resultado en el paciente utilizando el sedante B.

Si denotamos X_A y X_B como los incrementos en sueño debido a los sedantes A y B, y μ_A y μ_B sus medias teóricas, respectivamente, entonces el Resultado 5.2.6. afirma que

$$E(X_A|X_B = b) = \mu_A + \sigma_{AB}(\sigma_B)^{-2}(b - \mu_B)$$

y

$$Var(X_A|X_B = b) = \sigma_A^2 - \sigma_{AB}^2(\sigma_B)^{-2}$$

donde σ_A^2, σ_B^2 y σ_{AB} denotan $Var(X_A)$, $Var(X_B)$ y $Cov(X_A, X_B)$, respectivamente. Es claro que $E(X_A|X_B = b)$ y $Var(X_A|X_B = b)$ dependen de los parámetros teóricos y por consiguiente son desconocidos. Sin embargo, al utilizar las estimaciones de estos parámetros teóricos, podemos obtener las estimaciones de esta esperanza condicional y esta varianza condicional. Al tener en cuenta que μ_A y μ_B se estiman con los promedios muestrales dados por 0.66 y 2.33, respectivamente, tenemos que

$$\hat{E}(X_A|X_B = b) = 0.66 + 3.09 * (b - 2.33)/4.01$$

y

$$\hat{Var}(X_A|X_B = b) = 3.45 - 3.09^2/4.01 = 1.069$$

De esta forma, si en un paciente particular, el sedante B produjo un incremento de 3 horas de sueño, tenemos que $0.66+3.09*(3-2.33)/4.01 = 1.176$ y podemos concluir que se espera que el sedante A produzca aproximadamente 1 hora y 10 minutos de sueño, con una desviación estándar de aproximadamente $\sqrt{1.069} = 1.03$ horas.

6.3 Región de confianza y pruebas de hipótesis para el vector de medias

Se necesita hallar un $\mathfrak{S}$ subconjunto de $\mathbb{R}^p$ de tal forma que $Pr(\boldsymbol{\mu} \in \mathfrak{S}) = 1 - \alpha$. Un acercamiento a este problema puede ser encontrar intervalos de confianza de $100 \times (1-$

$\alpha)\,\%$ para cada componente de $\boldsymbol{\mu}$, esto es: $Pr(\mu_i \in S_i) = 1-\alpha$ para todo $i = 1, \cdots, p$. Y proponer como candidato para $\mathfrak{S}$ el producto cartesiano $\mathfrak{S}^* = S_1 \times \cdots \times S_p$ que es un rectángulo en $\mathbb{R}^p$. Pero el nivel de confianza de $\mathfrak{S}^*$ no necesariamente es $100 \times (1-\alpha)\,\%$. Lo único que se puede afirmar es que

$$\begin{aligned} Pr(\boldsymbol{\mu} \in S_1 \times \cdots \times S_p) &= Pr(\mu_1 \in S_1 \times \mu_p \in S_p) \\ &= Pr(\bigcap_{i=1^p \mu_i \in S_i}) \\ &= Pr((\bigcup_{i=1}^{p} \mu_i \notin S_i)^c) \\ &= 1 - Pr(\bigcup_{i=1}^{p} \mu_i \notin S_i) \\ &\geq 1 - \sum_{i=1}^{p} Pr(\mu_i \notin S_i) \\ &= 1 - p\alpha. \end{aligned}$$

De esta forma, si $\boldsymbol{\mu} = (\mu_1, \mu_2)'$, entonces para $\alpha = 0.05$, se tiene que $Pr(\boldsymbol{\mu} \in S_1 \times S_2) \geq 0.9$, es decir, el nivel de confianza de $S_1 \times \cdots \times S_p$ puede ser inferior a los 95 %. Más aún, cuando el número de variables de estudio p aumenta, $1 - p\alpha$ se torna muy pequeño, y así la región de confianza $S_1 \times \cdots \times S_p$ no será muy útil en la práctica. La única alternativa para que $1 - p\alpha$ no disminuya demasiado es disminuir el valor de α, pero de esta manera, cada S_i tendrá una longitud muy grande (ver (3.2.6)), es decir, cada S_i será casi todo el eje real, y el producto cartesiano $S_1 \times \cdots \times S_p$ se convertirá en $\mathbb{R}^p$ y una región de confianza de esta magnitud es poco precisa y por consiguiente, no es muy útil en la práctica.

6.3.1 $\boldsymbol{\Sigma}$ conocida

En el caso de la inferencia univariada, cuando la varianza es conocida, se encuentra un intervalo de confianza para μ usando la variable pivote $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$. En la inferencia multivariada, se puede encontrar una variable pivote para $\boldsymbol{\mu}$ con distribución univariante. Aplicando la propiedad 6 del Resultado 5.2.4. al vector aleatorio $\bar{\mathbf{X}}$ cuya distribución es $N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$, se tiene que

$$n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim \chi^2_p.$$

De esta forma, $n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$ es una variable pivote para $\boldsymbol{\mu}$ y se puede construir un intervalo de confianza para esta variable pivote[2]. Usando el hecho de que

$$Pr(n(\bar{\mathbf{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) < \chi^2_{p,1-\alpha}) = 1 - \alpha. \tag{6.3.1}$$

[2]Nótese que esta variable pivote es simplemente la distancia de Mahalanobis entre el promedio muestral $\bar{\mathbf{X}}$ y el media teórica $\boldsymbol{\mu}$.

El método de la variable pivote expuesta en capítulos anteriores sugiere que el siguiente paso es despejar el parámetro de interés, en este caso el vector $\boldsymbol{\mu}$, pero claramente esto no se puede llevar a cabo por la complejidad de la variable pivote. Por lo tanto, sencillamente se afirma que una región de confianza para $\boldsymbol{\mu}$ es el conjunto

$$\mathfrak{S}_1(\boldsymbol{\mu}) = \{\mathbf{v} \in \mathbb{R}^p : n(\bar{\mathbf{X}} - \mathbf{v})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \mathbf{v}) < \chi^2_{p,1-\alpha}\}. \tag{6.3.2}$$

Para explorar acerca de la forma de la anterior región, tomamos $p = 2$, y denotando la matriz $\boldsymbol{\Sigma}^{-1}$ como A tenemos que

$$\begin{aligned}(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) &= (\boldsymbol{\mu} - \bar{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}}) \\ &= (\mu_1 - \bar{x}_1, \mu_2 - \bar{x}_2)\begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22}\end{pmatrix}\begin{pmatrix}\mu_1 - \bar{x}_1 \\ \mu_2 - \bar{x}_2\end{pmatrix} \\ &= a_{11}(\mu_1 - \bar{x}_1)^2 + 2a_{12}(\mu_1 - \bar{x}_1)(\mu_2 - \bar{x}_2) + a_{22}(\mu_2 - \bar{x}_2)^2.\end{aligned}$$

Vista como una función de μ_1 y μ_2, la anterior función describe un elipse con centro en las estimaciones $(\bar{x}_1, \bar{x}_2)$. En la Figura 6.1, se muestra algunas gráficas de esta función con $\bar{x}_1 = \bar{x}_2 = 0$, $n = 10$ y $\alpha = 0.05$, con diferentes valores de $\boldsymbol{\Sigma}$.

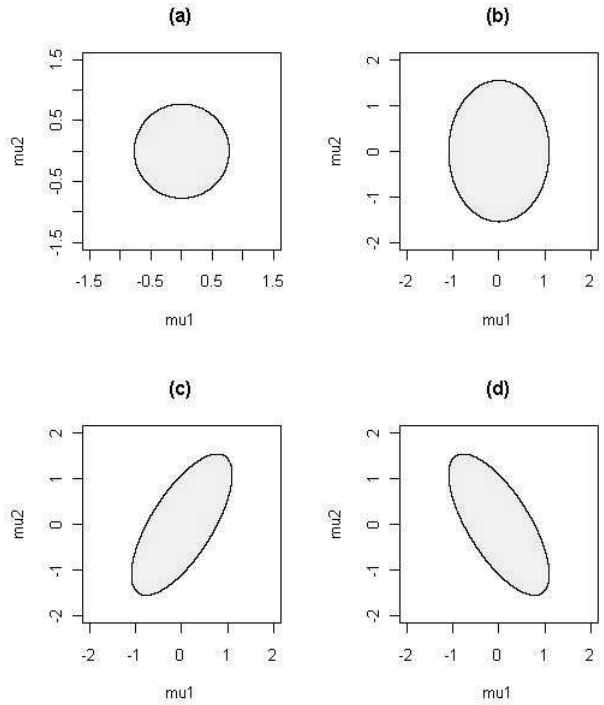

Figura 6.1: *Región de confianza* $\mathfrak{S}_1(\boldsymbol{\mu})$ *con* $p = 2$ *y diferentes valores de* $\boldsymbol{\Sigma}$.

Observamos que

1. La gráfica (a) corresponde al caso cuando $\boldsymbol{\Sigma} = \mathbf{I}_2$, observa que en este caso la región de confianza es un círculo.

2. La gráfica (b) corresponde al caso cuando $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$, en este caso la región de confianza es la parte que encierra una elipse. Además nótese que como la varianza de la segunda variable es más grande, entonces la proyección del elipse sobre el eje de μ_2 es también más amplio.

3. Ahora, la gráfica (c) corresponde al caso cuando $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix}$, donde la elipse está inclinada hacia la derecha asemejandose a una recta con tendencia positiva, puesto que una covarianza positiva entre las dos variables indica que cuando μ_1 toma valores grandes, también lo hace μ_2.

4. Finalmente, la gráfica (d) corresponde al caso cuando $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix}$. En este caso, la gráfica muestra una tendencia negativa entre las dos variables, consecuencia de que la covarianza es negativa.

En el anterior procedimiento para encontrar región de confianza para $\boldsymbol{\mu}$, se utilizó el intervalo de confianza para la variable pivote $n(\bar{\mathbf{X}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu})$ dado en (6.3.1), el cual es un intervalo unilateral superior. Se puede pensar en hallar un intervalo unilateral inferior o bilateral para esta variable pivote para encontrar otras regiones de confianza para $\boldsymbol{\mu}$.

Primero consideramos un intervalo unilateral inferior. Dada la distribución de $n(\bar{\mathbf{X}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu})$, se tiene que

$$Pr(\chi^2_{p,\alpha} < n(\bar{\mathbf{X}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu})) = 1-\alpha.$$

De esta manera, otra región de confianza para $\boldsymbol{\mu}$ estaría dada por

$$\mathfrak{S}_2(\boldsymbol{\mu}) = \{\mathbf{v} \in \mathbb{R}^p : \chi^2_{p,\alpha} < n(\bar{\mathbf{X}}-\mathbf{v})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}-\mathbf{v})\}. \tag{6.3.3}$$

Cuando $p = 2$, la función $(\bar{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})$ vista como función de μ_1 y μ_2 es una elipse como se vio anteriormente, entonces la región $\mathfrak{S}_2(\boldsymbol{\mu})$ describe la parte fuera de una elipse como lo ilustra la Figura 6.2 con las mismas especificaciones que la Figura 6.1. Aunque estas regiones tienen un nivel de confianza del 95 %, no son usadas en la práctica. Nótese que la estimación puntual $(\bar{x}_1, \bar{x}_2)$, se encuentra en el centro de las elipses, y por consiguiente, queda excluida de la región de confianza $\mathfrak{S}_2$, algo contradictorio, sin duda.

Por otro lado, un intervalo bilateral para la variable pivote $n(\bar{\mathbf{X}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu})$, nos conduce a la siguiente región de confianza para $\boldsymbol{\mu}$:

$$\mathfrak{S}_3(\boldsymbol{\mu}) = \{\mathbf{v} \in \mathbb{R}^p : \chi^2_{p,\alpha/2} < n(\bar{\mathbf{X}}-\mathbf{v})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}-\mathbf{v}) < \chi^2_{p,1-\alpha/2}\}. \tag{6.3.4}$$

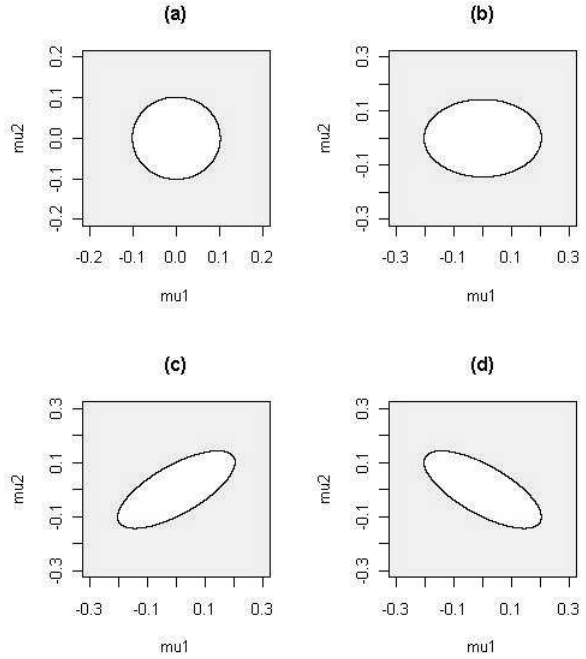

Figura 6.2: *Región de confianza* $\mathfrak{S}_2(\boldsymbol{\mu})$ *con* $p = 2$ *y diferentes valores de* $\boldsymbol{\Sigma}$.

La expresión (6.3.4) representa un disco como lo ilustra la Figura 6.3. En la práctica tampoco resulta útil $\mathfrak{S}_3$ puesto que la estimación puntual de $\boldsymbol{\mu}$ también se encuentra excluida. En conclusión, la región de confianza más apropiada para $\boldsymbol{\mu}$ es $\mathfrak{S}_1(\boldsymbol{\mu})$, y será usada de ahora en adelante.

Retomando la dualidad que existe entre la estimación por intervalo de confianza y las pruebas de hipótesis, podemos establecer que para el sistema:

$$H_0 : \ \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad vs. \quad H_1 : \ \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

H_0 será rechazada cuando $\boldsymbol{\mu}_0$ no se encuentra en la región de confianza $\mathfrak{S}_1(\boldsymbol{\mu})$, esto es, cuando $n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) > \chi^2_{p,1-\alpha}$.

6.3.2 $\boldsymbol{\Sigma}$ desconocida

Cuando la matriz de varianzas y covarianzas es desconocida, es natural usar su estimador insesgado $\mathbf{S}_{n-1}$, en este caso la estadística pivote de la sección anterior se convierte en

$$(\bar{\mathbf{x}} - \boldsymbol{\mu})'(\mathbf{S}_{n-1})^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}). \tag{6.3.5}$$

Para conocer si la anterior estadística sigue siendo una variable pivote para $\boldsymbol{\mu}$, es necesario encontrar su distribución.

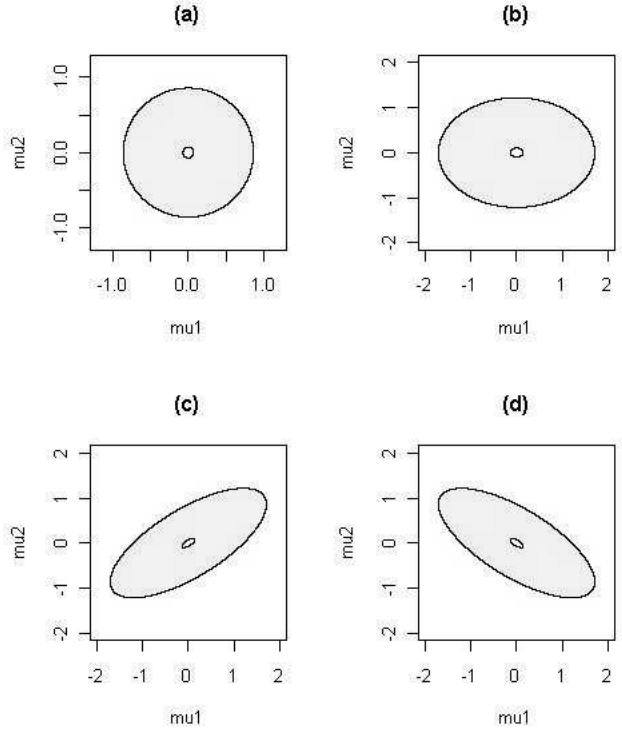

Figura 6.3: *Región de confianza* $\mathfrak{S}_3(\boldsymbol{\mu})$ *con* $p = 2$ *y diferentes valores de* $\boldsymbol{\Sigma}$.

Luego, observamos que

1. $\bar{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$ y
2. $(n-1)\mathbf{S}_{n-1} \sim W(n-1, \boldsymbol{\Sigma})$ de donde $\frac{(n-1)}{n}\mathbf{S}_{n-1} \sim W(n-1, \boldsymbol{\Sigma}/n)$

Usando las anteriores propiedades y la definición de la distribución T^2 de Hotelling, tenemos que

$$(\bar{\mathbf{x}} - \boldsymbol{\mu})' \left(\frac{\mathbf{S}_{n-1}}{n}\right)^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) = n(\bar{\mathbf{x}} - \boldsymbol{\mu})'(\mathbf{S}_{n-1})^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim T^2(p, n-1)$$

Nótese que cuando $p = 1$, la estadística T^2 se convierte en $\frac{n(\bar{X}-\mu)^2}{S_{n-1}}$ la cual es el cuadrado de la estadística $\frac{\sqrt{n}(\bar{X}-\mu)}{S_{n-1}}$ cuya distribución corresponde a t_{n-1}.

Para obtener calcular percentiles y/o probabilidades de una distribución T^2, Bowker (1960) encontró el siguiente resultado.

Resultado 6.3.1. *Dada una variable con distribución T^2 con grados de libertad p y $n-1$, se tiene que*

$$\frac{T^2}{n-1}\left(\frac{n-p}{p}\right) \sim F^p_{n-p}.$$

Con lo anterior, se concluye que $n(\bar{\mathbf{x}}-\boldsymbol{\mu})'(\mathbf{S}_{n-1})^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})$ es una variable pivote para $\boldsymbol{\mu}$, de esta forma, se tiene que una región de confianza para $\boldsymbol{\mu}$ es

$$\mathfrak{S}(\mathbf{v}) = \{\mathbf{v} \in \mathbb{R}^p : n(\bar{\mathbf{X}}-\mathbf{v})'(\mathbf{S}_{n-1})^{-1}(\bar{\mathbf{X}}-\mathbf{v}) < c\}, \tag{6.3.6}$$

donde c es el percentil $1-\alpha$ de la distribución T^2 con grados de libertad p y $n-1$. Aquí se toma el límite superior para la variable pivote puesto que se vio en el caso cuando $\boldsymbol{\Sigma}$ es conocida, que las regiones de confianza obtenidas usando el límite inferior o ambos superior e inferior no son adecuadas en la práctica.

Ahora, usando la relación que existe entre la distribución T^2 y la distribución F, se tiene que la región de confianza $\mathfrak{S}$ se puede escribir como

$$\mathfrak{S}(\mathbf{v}) = \left\{\mathbf{v} \in \mathbb{R}^p : \frac{n(n-p)}{p(n-1)}(\bar{\mathbf{X}}-\mathbf{v})'(\mathbf{S}_{n-1})^{-1}(\bar{\mathbf{X}}-\mathbf{v}) < f^p_{n-p,1-\alpha}\right\}. \tag{6.3.7}$$

Esta región de confianza, igual que el caso cuando $\boldsymbol{\Sigma}$ es conocida, consta del área interior de una elipse.

Dada la anterior región de confianza para $\boldsymbol{\mu}$, se puede obtener una regla de decisión para el sistema

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad vs. \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

dada por: Rechazar H_0, si $\dfrac{n(n-p)}{Pr(n-1)}(\bar{\mathbf{X}}-\mu_0)'(\mathbf{S}_{n-1})^{-1}(\bar{\mathbf{X}}-\mu_0) > f^p_{n-p,1-\alpha}$

Ahora, volvemos al principio del capítulo 6.2, donde se consideró la propuesta de usar $S_1 \times \cdots \times S_p$ como región de confianza para $\boldsymbol{\mu}$ donde S_i es un intervalo de confianza para μ_i con $i = 1, \cdots, p$.

6.4 Región de confianza y pruebas de hipótesis para una combinación lineal de medias

En algunas situaciones, el vector de parámetros de interés no es el vector de medias $\boldsymbol{\mu}$, sino alguna función de él, o a veces se desea confirmar o refutar alguna relación que puede existir entre los componentes de $\boldsymbol{\mu}$. Por ejemplo, un laboratorio médico fabrica un nuevo tipo de medicamentos para bajar el nivel de colesterol, y confía en que con dos meses de tratamiento, puede lograr una reducción de un 20 %.

Para confirmar o refutar esta afirmación, se ensaya el medicamento con n pacientes, se les toma el nivel de colesterol antes de iniciar el tratamiento, y denotamos estas observaciones como $x_{11}, \cdots, x_{1n}$; y después de dos meses del tratamiento con el nuevo medicamento, se vuelve a tomar el nivel del colesterol, denotando estas observaciones

como $x_{21}, \cdots, x_{2n}$. En este caso, se dispone de dos variables de estudio: el nivel de colesterol sin tratamiento y el nivel de colesterol después del tratamiento; si denotamos sus respectivas esperanzas como μ_1 y μ_2, lo que se quiere probar es que $0.8\mu_1 = \mu_2$.

Muchas relaciones entre los componentes de $\boldsymbol{\mu}$ como la de la anterior situación, pueden ser descritas como $C\boldsymbol{\mu} = \mathbf{v}$, donde C es de dimensión $k \times p$, y $\mathbf{v}$ de $k \times 1$. Por ejemplo, la relación $0.8\mu_1 = \mu_2$ es equivalente a $(0.8, -1)\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} = 0$, en donde $C = (0.8, 1)$, y $\boldsymbol{v} = 0$. Y al encontrar una regla de decisión para aceptar o rechazar $C\boldsymbol{\mu} = \mathbf{v}$, se habrá resuelto el problema.

En conclusión, el marco de trabajo es: dada una muestra aleatoria $\mathbf{X}_1, \cdots, \mathbf{X}_n$ con distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, se quiere probar la hipótesis H_0: $C\boldsymbol{\mu} = \mathbf{v}$. Utilizando la propiedad 1 del resultado 5.2.4, tenemos que cada uno de los vectores aleatorios $C\mathbf{X}_1, \cdots, C\mathbf{X}_n$ tiene distribución $N_p(C\boldsymbol{\mu}, C\boldsymbol{\Sigma}C')$, además usando la propiedad (5.1.13), se tiene que estos son independientes.

De esta forma, al definir nuevos vectores aleatorios $\mathbf{Y}_i = C\mathbf{X}_i$ con $i = 1, \cdots, n$, el problema se convierte en probar la hipótesis $H_0 : \boldsymbol{\mu}_Y = \mathbf{v}$ para una muestra aleatoria $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$, y este problema ya ha sido tratado en la sección anterior tanto para el caso cuando la matriz de varianzas y covarianzas es conocida como cuando no lo es. A continuación se adaptará este procedimiento para el caso cuando la matriz de varianzas y covarianzas es desconocida, el otro caso se deja como ejercicio.

De acuerdo a la teoría desarrollada en la sección anterior, se rechaza $H_0 : \boldsymbol{\mu}_Y = \mathbf{v}$ si

$$\frac{n(n-k)}{k(n-1)}(\bar{\mathbf{Y}} - \mathbf{v})'(\mathbf{S}_{n-1,Y})^{-1}(\bar{\mathbf{Y}} - \mathbf{v}) > f^k_{n-k,1-\alpha} \tag{6.4.1}$$

donde k es la dimensión de los vectores $\mathbf{Y}_i$, y $\mathbf{S}_{n-1,Y}$ es la matriz de varianzas y covarianzas muestrales de $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$.

En la práctica, se observan los valores que toman los vectores aleatorios originales: $\mathbf{x}_1, \cdots, \mathbf{x}_n$, y para verificar si se cumple o no la regla de decisión (6.4.1), se puede adoptar cualquiera de los dos siguientes procedimientos:

1. Calcular las respectivas observaciones $\mathbf{y}_1, \cdots, \mathbf{y}_n$, premultiplicando cada uno de $\mathbf{x}_1, \cdots, \mathbf{x}_n$ por la matriz C. Nótese que una vez especificamos la relación que se desea probar, la matriz C es conocida (más adelante, se consideran diferentes tipos de relaciones y las formas de la matriz C). Al calcular $\bar{\mathbf{y}}$ y $\mathbf{S}_{n-1,Y}$, se puede determinar fácilmente la aceptación o el rechazo de H_0. Sin embargo, este procedimiento representa bastantes cálculos, puesto que para obtener los $\mathbf{y}_1, \cdots, \mathbf{y}_n$, se realiza una multiplicación matricial n veces.

2. El segundo procedimiento consiste en escribir la regla de decisión (6.4.1) en términos de las observaciones $\mathbf{x}_1, \cdots, \mathbf{x}_n$. Para eso se observa que en primer lugar,

$$\bar{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i = \frac{1}{n}\sum_{i=1}^{n}C\mathbf{X}_i = C(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i) = C\bar{\mathbf{X}}.$$

Por otro lado, se tiene que

$$\begin{aligned}\mathbf{S}_{n-1,Y} &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{Y}_i-\bar{Y})(\mathbf{Y}_i-\bar{Y})' \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(C\mathbf{X}_i-C\bar{X})(C\mathbf{X}_i-\bar{X})' \\ &= \frac{1}{n-1}\sum_{i=1}^{n}C(\mathbf{X}_i-\bar{X})(\mathbf{X}_i-\bar{X})'C' \\ &= C(\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i-\bar{X})(\mathbf{X}_i-\bar{X})')C' = C\mathbf{S}_{n-1,X}C'.\end{aligned}$$

Usando estas propiedades, (6.4.1) se convierte en: Se rechaza $H_0 : \boldsymbol{\mu}_Y = \mathbf{v}$ si

$$\frac{n(n-k)}{k(n-1)}(C\bar{\mathbf{X}}-\mathbf{v})'(C\mathbf{S}_{n-1,X}C')^{-1}(C\bar{\mathbf{X}}-\mathbf{v}) > f^{k}_{n-k,1-\alpha} \qquad (6.4.2)$$

Ahora, retomamos el ejemplo del medicamento para bajar el nivel de colesterol. Suponga que el tratamiento fue aplicado a 20 pacientes, y los resultados se muestran en la Tabla 6.3. Los siguientes códigos nos permiten determinar si se cumple o no la regla de decisión.

```
> ante<-c(230,245,220,250, 260,250,220,300,310,290,260,240,210,
+         220,250,245,274,230,285,275)
> desp<-c(210,230,215,220,240,220,210,260,280,270,230,235,200,
+         200,210,230,250,210,260,230)
> X<-data.frame(cbind(ante,desp))
> bar<-mean(X)
> bar
ante  desp
253.2 230.5
> S2<-var(X)
> n<-length(ante)
> C<-matrix(c(0.8,-1),1,2)
> k<-nrow(C)
> f<-(C%*%mean(X))^2*(C%*%S2%*%t(C))^{-1}*n*(n-k)/(k*(n-1))
> f
         [,1]
[1,] 194.4231
> alpha<-0.05
> qf(1-alpha,k,n-k)
[1] 4.38075
```

Se observa que el valor de la estadística F es mucho mayor comparando con el percentil de la distribución, de donde se concluye que los datos muestran una fuerte

evidencia en contra de la hipótesis nula, y el medicamento no está disminuyendo el nivel de colesterol en un 20 %. Nótese que en la muestra observada, el nivel promedio de colesterol antes y después del tratamiento es 253.2 y 230.5, respectivamente, esto indica que el medicamento disminuyó el nivel de colesterol en un 9 %, muy lejano al 20 % que esperaba el laboratorio.

Paciente	Antes	Después
1	230	210
2	245	230
3	220	215
4	250	220
5	260	240
6	250	220
7	220	210
8	300	260
9	310	280
10	290	270
11	260	230
12	240	235
13	210	200
14	220	200
15	250	210
16	245	230
17	274	250
18	230	210
19	285	260
20	275	230

Tabla 6.3: *El nivel de colesterol de 20 pacientes antes y después del tratamiento medido en mg/dL.*

Ahora la pregunta es, ¿este 9 % de disminución que se observó en la muestra es realmente significativo, o se puede considerar despreciable? En este caso, la hipótesis planteada se cambia a $H_0 : \mu_1 = \mu_2$, la cual puede escribir como $H_0 : (1, -1)\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix} = 0$. Al utilizar el anterior código de R modificando la matriz C, se tiene que el valor de la estadística es 80.9, de tal manera que comparado con el percentil 4.38, se concluye que el nivel promedio de colesterol con el tratamiento sí presenta una disminución significativa.

6.5 Prueba de hipótesis para la matriz de varianzas y covarianzas

Cuando el parámetro de interés es la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$, comúnmente se trabaja sólo con la prueba de hipótesis cerca de $\boldsymbol{\Sigma}$, mas no con la región de confianza,

puesto que ésta pertenece a un espacio de dimensión muy grande. Aun cuando el número de variables de estudio es dos, la dimensión de $\boldsymbol{\Sigma}$ es de 2×2, y su respectiva región de confianza sería un subconjunto de $\mathbb{R}^2 \times \mathbb{R}^2$, lo cual dificulta la visualización y la correspondiente interpretación. Por lo tanto, sólo estamos enfocados en el problema de prueba de hipótesis para $\boldsymbol{\Sigma}$.

Suponga habitualmente que se dispone de una muestra aleatoria $\mathbf{X}_1, \cdots, \mathbf{X}_n$ proveniente de la distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, y se desea probar el sistema de hipótesis

$$H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \quad vs. \quad H_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0. \tag{6.5.1}$$

En el caso univariado, cuando se trató el tema de pruebas de hipótesis acerca de la varianza teórica, se consideró el caso cuando la media teórica es conocida y posteriormente cuando ésta no es conocida. Sin embargo, en la práctica, en muchas ocasiones, se carece del conocimiento acerca de la media teórica. En el caso de pruebas de hipótesis acerca de la matriz de varianzas y covarianzas, es muy difícil tener conocimiento del vector de medias teóricas $\boldsymbol{\mu}$, puesto que se deben conocer las p medias teóricas de las p variables de estudio. Por esta razón, no asumiremos ningún conocimiento previo de $\boldsymbol{\mu}$, sino que lo estimaremos usando $\bar{\mathbf{X}}$ en caso de ser requerido.

Usaremos el método de la prueba de de razón generalizada de verosimilitud presentada anteriormente para el sistema de hipótesis (6.5.1). Recordemos que

$$\lambda = \frac{\sup_{\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1} L(\theta, x_1, \cdots, x_n)}{\sup_{\boldsymbol{\Theta}_0} L(\theta, x_1, \cdots, x_n)}, \tag{6.5.2}$$

y rechaza H_0 para valores grandes de λ. Ahora, usando como la función logarítmica es monótona, se tiene que se rechaza H_0 para valores grandes de $\ln \lambda$, la cual es igual a

$$\ln \lambda = \sup_{\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1} \ln L(\theta, x_1, \cdots, x_n) - \sup_{\boldsymbol{\Theta}_0} \ln L(\theta, x_1, \cdots, x_n),$$

recurriendo nuevamente a que la función logarítmica es monótona.

Ahora, retomando la ecuación (6.2.2)

$$\ln L = -\frac{n}{2} \ln |\boldsymbol{\Sigma}| - \frac{pn}{2} \ln 2\pi - \frac{n}{2} tr(\boldsymbol{\Sigma}^{-1} \mathbf{S}_n) - \frac{n}{2} (\bar{\mathbf{x}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}),$$

y teniendo en cuenta que $\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}$, y $\hat{\boldsymbol{\mu}}_{MV} = \bar{\mathbf{X}}$, tenemos que:

$$\begin{aligned} \sup_{\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1} \ln L(\theta, x_1, \cdots, x_n) &= -\frac{n}{2} \ln |\mathbf{S}_n| - \frac{pn}{2} \ln 2\pi - \frac{n}{2} tr((\mathbf{S}_n)^{-1} \mathbf{S}_n) \\ &= -\frac{n}{2} \ln |\mathbf{S}_n| - \frac{pn}{2} \ln 2\pi - \frac{n}{2} tr(I_p) \\ &= -\frac{n}{2} \ln |\mathbf{S}_n| - \frac{pn}{2} \ln 2\pi - \frac{np}{2}. \end{aligned} \tag{6.5.3}$$

Y por otro lado,

$$\begin{aligned} \sup_{\boldsymbol{\Theta}_0} \ln L(\theta, x_1, \cdots, x_n) &= \ln L(\boldsymbol{\Sigma}_0, x_1, \cdots, x_n) \\ &= -\frac{n}{2} \ln |\boldsymbol{\Sigma}_0| - \frac{pn}{2} \ln 2\pi - \frac{n}{2} tr(\boldsymbol{\Sigma}_0^{-1} \mathbf{S}_n). \end{aligned}$$

De esta forma, se tiene que

$$\begin{aligned}\ln\lambda &= \frac{n}{2}\ln|\boldsymbol{\Sigma}_0| + \frac{n}{2}tr(\boldsymbol{\Sigma}_0^{-1}\mathbf{S}_n) - \frac{n}{2}\ln|\mathbf{S}_n| - \frac{np}{2} \\ &= \frac{1}{2}\left\{n(\ln|\boldsymbol{\Sigma}_0| - \ln|\mathbf{S}_n|) + ntr(\boldsymbol{\Sigma}_0^{-1}\mathbf{S}_n) - np\right\}. \end{aligned} \tag{6.5.4}$$

Se tiene que la distribución de $2\ln\lambda$ es χ^2_v, donde $v = v_1 - v_0$ con v_1 denotando el número de parámetros bajo la hipótesis alterna y v_0 el número de parámetros bajo la hipótesis nula, en el sistema de hipótesis (6.5.1), $v_0 = 0$, puesto que bajo H_0, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$, y no hay ningún parámetro desconocido; por otro lado, bajo H_1, $\boldsymbol{\Sigma}$ puede tomar cualquier forma, y no se conoce ningún valor específico que toma, por lo tanto, v_1 es el número de componentes distintos de $\boldsymbol{\Sigma}$, esto es, $v_1 = p(p+1)/2$. Por lo tanto, $2\ln\lambda \sim \chi^2_{p(p+1)/2}$, y se rechaza H_0 cuando $2\ln\lambda > \chi^2_{Pr(p+1)/2,1-\alpha}$.

Otra forma de expresar la estadística de prueba es usando los valores propios de la matriz $\boldsymbol{\Sigma}_0^{-1}\mathbf{S}_n$. Supongamos que estos se denotan por $\lambda_1, \cdots, \lambda_p$, tenemos que:

$$\begin{aligned} 2\ln\lambda &= n(\ln|\boldsymbol{\Sigma}_0| - \ln|\mathbf{S}_n|) + ntr(\boldsymbol{\Sigma}_0^{-1}\mathbf{S}_n) - np \\ &= -n\ln|\boldsymbol{\Sigma}_0^{-1}\mathbf{S}_n| + ntr(\boldsymbol{\Sigma}_0^{-1}\mathbf{S}_n) - np \\ &= -n\ln\prod_{i=1}^{p}\lambda_i + n\sum_{i=1}^{p}\lambda_i - np \\ &= n(\sum_{i=1}^{P} r(\lambda_i - \ln\lambda_i) - p) \end{aligned}$$

El sistema de hipótesis (6.5.1), incluye un gran serie de estructuras dependiente de la forma que toma la matriz $\boldsymbol{\Sigma}_0$. A continuación, se presentan algunos casos particulares que son de utilidad en la práctica.

Prueba de independencia de un conjunto de variables aleatorias

Un problema común en un estudio estadístico es determinar, en un conjunto de variables de estudio, cuáles son independientes. Bajo el supuesto de que la muestra aleatoria proviene de una distribución normal, una covarianza nula implica la independencia. Por lo tanto, para probar que un conjunto de variables son independientes, basta probar que la matriz de varianzas y covarianzas es una matriz diagonal. De esta manera, se plantea el siguiente sistema de hipótesis.

$$H_0\text{: } \boldsymbol{\Sigma} \text{ es diagonal} \quad \text{vs.} \quad H_1\text{: } \boldsymbol{\Sigma} \text{ no es diagonal.}$$

Se usa la prueba de razón generalizada de verosimilitud para encontrar la estadística de prueba para este sistema. Para calcular la estadística de prueba $2\ln\lambda$, recordamos la ecuación (6.5.3),

$$\sup_{\boldsymbol{\Theta}_0\cup\boldsymbol{\Theta}_1} \ln L(\theta, x_1, \cdots, x_n) = -\frac{n}{2}\ln|\mathbf{S}_n| - \frac{pn}{2}\ln 2\pi - \frac{np}{2}.$$

Por otro lado, el estimador MV de $\mathbf{\Sigma}$ bajo H_0 es la matriz $\hat{\mathbf{D}} = diag(\mathbf{S}_n)$, puesto que las covarianzas son restringidas a tomar el valor 0, entonces se tiene que

$$\sup_{\mathbf{\Theta}_0} \ln L(\theta, x_1, \cdots, x_n) = -\frac{n}{2} \ln |\hat{\mathbf{D}}| - \frac{pn}{2} \ln 2\pi - \frac{n}{2} tr(\hat{\mathbf{D}}^{-1} \mathbf{S}_n).$$

Por lo tanto, se tiene que

$$2 \ln \lambda = n(\ln |\hat{\mathbf{D}}| - \ln |\mathbf{S}_n|) + ntr(\hat{\mathbf{D}}^{-1} \mathbf{S}_n) - np.$$

Aunque dado un conjunto de valores observados de los vectores aleatorias de la muestra aleatoria, se puede calcular el valor de la anterior estadística, existe una expresión equivalente, pero más sencillo. Tenemos que

$$\begin{aligned} \ln |\mathbf{S}_n| - \ln |\hat{\mathbf{D}}| &= \ln \frac{|\mathbf{S}_n|}{|\hat{\mathbf{D}}|} \\ &= \ln \frac{|\mathbf{S}_n|}{|\hat{\mathbf{D}}^{1/2}||\hat{\mathbf{D}}^{1/2}|} \\ &= \ln |\hat{\mathbf{D}}^{-1/2}||\mathbf{S}_n||\hat{\mathbf{D}}^{-1/2}| \\ &= \ln |\hat{\boldsymbol{\rho}}|. \end{aligned}$$

Por otro lado,

$$\begin{aligned} tr(\hat{\mathbf{D}}^{-1} \mathbf{S}_n) &= tr(\hat{\mathbf{D}}^{-1/2} \hat{\mathbf{D}}^{-1/2} \mathbf{S}_n) \\ &= tr(\hat{\mathbf{D}}^{-1/2} \mathbf{S}_n \hat{\mathbf{D}}^{-1/2}) \\ &= tr(\hat{\boldsymbol{\rho}}) \\ &= p \end{aligned}$$

pues $\hat{\boldsymbol{\rho}}$ es una matriz de dimensión $p \times p$ donde los elementos de la diagonal son 1. Usando las dos anteriores expresiones, se tiene que

$$2 \ln \lambda = -n \ln |\boldsymbol{\rho}|.$$

Con lo desarrollado anteriormente, se puede calcular el valor de la estadística $2 \ln \lambda$ en una muestra observada, pero para decidir sobre la aceptación o el rechazo de la hipótesis nula, se necesita saber la distribución de la estadística, la cual es χ^2_v, con $v = v_1 - v_0$. v_1 es el número de parámetros bajo la hipótesis alterna, la cual especifica que $\mathbf{\Sigma}$ no es una matriz diagonal, así que su estimación se lleva a cabo de la forma habitual, mediante la matriz $\mathbf{S}_n$, tal que el número de parámetros que se estiman es $p(p+1)/2$, el número de elementos diferentes en $\mathbf{\Sigma}$. Por otro lado, bajo la hipótesis nula, $\mathbf{\Sigma}$ es una matriz diagonal, así que el número de parámetros que se estiman es el número de elementos en la diagonal de $\mathbf{\Sigma}$, esto es $v_0 = p$. En conclusión, $2 \ln \lambda \sim \chi^2_{p(p+1)/2-p}$ y en una muestra observada, se rechaza H_0 cuando el valor de la estadística es mayor a $\chi^2_{p(p+1)/2-p,1-\alpha}$. El p-valor se puede calcular fácilmente como $1 - F_{\chi^2_{p(p+1)/2-p}}(v)$ donde v denota el valor observado de la estadística $2 \ln \lambda$.

Ejemplo 6.5.1. *Consideramos los datos de la Tabla 5.2 acerca del incremento de sueño producidos por dos tipos de sedantes, al final de la sección 6.2.2 se calculó el tiempo esperado de incremento con el sedante A usando lo observado con respecto al sedante B. La correlación muestral de estos dos tiempos es de 0.83, lo cual sugiere que las variables son dependientes, y tiene sentido usar una variable para pronosticar la otra. Aquí usaremos la prueba desarrollada anteriormente para corroborar que las dos variables son efectivamente dependientes.*

```
> a<-c(0.7,-1.6,-0.2,-1.2,-1,3.4,3.7,0.8,0,2)
> b<-c(1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,3.4)
> x<-matrix(c(a,b),10,2)
> alpha<-0.5
> p<-2
> n<-10
> esta<--n*log(det(cor(x)))
> perce<-qchisq(1-alpha,p*(p+1)/2-p)
> p.val<-pchisq(esta,p*(p+1)/2-p,lower.tail=F)
> esta
[1] 11.73193
> perce
[1] 0.4549364
> p.val
[1] 0.0006143678
```

De donde podemos ver que el valor de la estadística $2\ln\lambda$ *es muy grande comparado con el percentil* $\chi^2_{p(p+1)/2-p,1-\alpha}$*, produciendo un* p*-valor muy pequeño indicando que las variables son dependientes.*

Prueba de independencia entre conjuntos de variables aleatorias

En algunas situaciones, se puede clasificar las variables de estudio en varios grupos, y podemos estar interesados en saber si estos grupos de variables son independientes. Por ejemplo, con respecto al estudio sobre regiones en términos de calidad de vida, podemos preguntarnos si las variables asociadas con la tasa de natalidad y mortalidad son independientes de las variables de esperanza de vida tanto de hombres como de mujeres.

Supongamos que se dispone una muestra aleatoria $\mathbf{X}_1, \cdots, \mathbf{X}_n$ proveniente de una distribución normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, y que las p variables se clasifican en k grupos, con p_1, $\cdots$, p_k variables, respectivamente, donde $p_1+\cdots+p_k = p$. Sin pérdida de generalidad, suponga que para cada $\mathbf{X}_i$, las primeras p_1 variables corresponden a las del primer grupo, las siguientes p_2 corresponden a las del segundo grupo, y así sucesivamente; en caso contrario, se reordenan los componentes para que el anterior supuesto se cumpla.

Con el anterior supuesto, verificar que los k grupos de variables sean independientes es equivalente al hecho de que la matriz de varianzas y covarianzas sea diagonal por bloques, esto es:

$$H_0\colon \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_k \end{pmatrix} \quad \text{vs.} \quad H_1\colon \boldsymbol{\Sigma} \text{ es de otra forma.}$$

donde $\boldsymbol{\Sigma}_i$ es de dimensión $p_i \times p_i$ y representa la matriz de varianzas y covarianzas de las variables del grupo i, con $i = 1, \cdots, k$. Para juzgar el anterior sistema de hipótesis, volvemos a utilizar la prueba de razón generalizada de verosimilitudes. De nuevo,

$$\sup_{\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1} \ln L(\theta, x_1, \cdots, x_n) = -\frac{n}{2} \ln |\mathbf{S}_n| - \frac{pn}{2} \ln 2\pi - \frac{np}{2}.$$

Por otro lado, el estimador de $\boldsymbol{\Sigma}$ bajo la hipótesis H_0 es la matriz diagonal por bloques:

$$\mathbf{S}_D = \begin{pmatrix} \mathbf{S}_{1,n} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{2,n} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{S}_{k,n} \end{pmatrix}, \tag{6.5.5}$$

donde $\mathbf{S}_{i,n}$ es la matriz de varianzas y covarianzas muestrales de las variables aleatorias del grupo i, con $i = 1, \cdots, k$. Entonces se tiene que

$$\sup_{\boldsymbol{\Theta}_0} \ln L(\theta, x_1, \cdots, x_n) = -\frac{n}{2} \ln |\mathbf{S}_D| - \frac{pn}{2} \ln 2\pi - \frac{n}{2} tr((\mathbf{S}_D)^{-1} \mathbf{S}_n).$$

De esta forma, la estadística de prueba está dada por:

$$2 \ln \lambda = n \ln |\mathbf{S}_D| + ntr((\mathbf{S}_D)^{-1} \mathbf{S}_n) - n \ln |\mathbf{S}_n| - np.$$

Para simplificar la anterior expresión, nótese que

$$\begin{aligned} (\mathbf{S}_D)^{-1} \mathbf{S}_n &= \begin{pmatrix} \mathbf{S}_{1,n} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{2,n} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{S}_{k,n} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{S}_{1,n} & \mathbf{S}_{12,n} & \cdots & \mathbf{S}_{1k,n} \\ \mathbf{S}_{21,n} & \mathbf{S}_{2,n} & \cdots & \mathbf{S}_{2k,n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{k1,n} & \mathbf{S}_{k2,n} & \cdots & \mathbf{S}_{k,n} \end{pmatrix} \\ &= \begin{pmatrix} (\mathbf{S})_{1,n}^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & (\mathbf{S})_{2,n}^{-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & (\mathbf{S})_{k,n}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{S}_{1,n} & \mathbf{S}_{12,n} & \cdots & \mathbf{S}_{1k,n} \\ \mathbf{S}_{21,n} & \mathbf{S}_{2,n} & \cdots & \mathbf{S}_{2k,n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{k1,n} & \mathbf{S}_{k2,n} & \cdots & \mathbf{S}_{k,n} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I}_{p_1} & (\mathbf{S})_{1,n}^{-1} \mathbf{S}_{12,n} & \cdots & (\mathbf{S})_{1,n}^{-1} \mathbf{S}_{1k,n} \\ (\mathbf{S})_{2,n}^{-1} \mathbf{S}_{21,n} & \mathbf{I}_{p_2} & \cdots & (\mathbf{S})_{2,n}^{-1} \mathbf{S}_{2k,n} \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{S})_{k,n}^{-1} \mathbf{S}_{k1,n} & (\mathbf{S})_{k,n}^{-1} \mathbf{S}_{k2,n} & \cdots & \mathbf{I}_{p_k} \end{pmatrix}. \end{aligned}$$

De donde se concluye que $tr((\mathbf{S}_D)^{-1} \mathbf{S}_n) = tr(\mathbf{I}_{p_1}) + tr(\mathbf{I}_{p_2}) + \cdots + \mathbf{I}_{p_k} = p$. De esta forma, se tiene que:

$$2 \ln \lambda = n \ln |\mathbf{S}_D| - n \ln |\mathbf{S}_n|.$$

Ahora, la distribución de $2\ln\lambda$ es $\chi^2_{v_1-v_0}$, donde $v_1 = p(p+1)/2$ como se ha visto anteriormente; mientras que v_0 corresponde al número de parámetros que se debe estimar bajo H_0, entonces v_0 corresponde a la suma del número de parámetros diferentes que contienen las matrices $\mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_k$, esto es, $v_0 = \frac{p_1(p_1+1)}{2} + \cdots + \frac{p_k(p_k+1)}{2}$. En conclusión, se rechaza la independencia entre los k grupos de variables si el valor de la estadística de prueba $2\ln\lambda$ es mayor al percentil $\chi^2_{v,1-\alpha}$.

Ejemplo 6.5.2. *Tomamos los datos MUNDODES de Peña (2002), y usaremos las variables* **Tasa Nat***: Razón de natalidad por 1000 habitantes,* **Tasa Mort***: Razón de mortalidad por 1000 habitantes,* **Mort. Inf***: mortalidad infantil (por debajo de un año),* **Esp. Hom***: Esperanza de vida en hombres y* **Esp. Muj***: Esperanza de vida en mujeres, y nos centramos en los datos correspondientes a los países suramericanos que se muestran en la Tabla 6.4.*

	Tasa Nat	Tasa Mort	Mort. Inf	Esp. Hom	Esp. Muj
1	20.7	8.4	25.7	65.5	72.7
2	46.6	18.0	51.0	51.0	55.4
3	28.6	7.9	63.0	62.3	67.6
4	23.4	5.8	17.1	68.1	75.1
5	27.4	6.1	40.0	63.4	69.2
6	32.9	7.4	63.0	63.4	67.6
7	28.3	7.3	56.0	60.4	66.1
8	34.8	6.6	42.0	64.4	68.5
9	32.9	8.3	59.9	56.8	66.5
10	18.0	9.6	31.0	68.4	74.9
12	23.2	7.9	43.2	63.7	71.8
13	17.5	6.4	35.3	54.2	72.1
14	20.1	5.4	36.3	60.9	62.3
15	19.3	10.4	63.8	56.7	52.9

Tabla 6.4: *Indicadores de desarrollo de quince países.*

Supongamos que se desea probar que la variable esperanza de vida es independiente de los indicadores de natalidad y mortalidad; para eso, el sistema de hipótesis apropiado es

$$H_0\text{: } \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \end{pmatrix} \quad \textit{vs.} \quad H_1\text{: } \mathbf{\Sigma} \textit{ es de otra forma.}$$

Es decir, las cinco variables de estudio se dividen en 2 grupos, con $p_1 = 3$ y $p_2 = 2$. Los códigos de R para efectuar el procedimiento de esta prueba son los siguientes:

```
> Tasa.Nat<-c(20.7,46.6,28.6,23.4,27.4,32.9,28.3,34.8,32.9,18,27.5)
> Tasa.Mort<-c(8.4,18,7.9,5.8,6.1,7.4,7.3,6.6,8.3,9.6,4.4)
> Mort.Inf<-c(25.7,111,63,17.1,40,63,56,42,109.9,21.9,23.3)
> Esp.H<-c(65.5,51,62.3,68.1,63.4,63.4,60.4,64.4,56.8,68.4,66.7)
> Esp.M<-c(72.7,55.4,67.6,75.1,69.2,67.6,66.1,68.5,66.5,74.9,72.8)
```

```
> X<-data.frame(cbind(Tasa.Nat,Tasa.Mort,Mort.Inf,Esp.H,Esp.M))
> n<-dim(X)[1]
> p1<-3
> p2<-2
> p<-p1+p2

> S2_n<-var(X)*(n-1)/n
> S2_d<-S2_n
> S2_d[-(1:p1),1:p1]<-0
> S2_d[1:p1,-(1:p1)]<-0
> estadistica<-n*log(det(S2_d))-n*log(det(S2_n))
> estadistica
[1] 41.01620
> v1<-p*(p+1)/2
> v0<-p1*(p1+1)/2+p2*(p2+1)/2
> qchisq(0.95,v1-v0)
[1] 12.59159
```

Como el valor de la estadística es mayor al percentil de la distribución, se concluye que los dos grupos de variables son dependientes.

6.6 Ejercicios

6.1 Para los datos del Ejemplo 6.1.1

(a) Plantee un sistema de hipótesis para probar que el candidato B obtendrá la mitad de votos del candidato A.

(b) Desarrolle la prueba de razón generalizada de verosimilitudes para este sistema. Debe encontrar la estadística de prueba, su distribución nula, y la fórmula para calcular el p valor.

(c) Aplique la prueba encontrada a los datos del Ejemplo 6.1.1.

6.2 Para conocer lo que harán los egresados del bachillerato en Bogotá justo después de su grado, se entrevistó a 1200 graduandos tanto en colegios públicos como en colegios privados, donde cada graduando entrevistado escogió una de las siguientes opciones

	Colegios privados	Colegios públicos	
Estudiar en una universidad	210	306	
Estudiar una carrera técnica	150	246	
Trabajar	90	198	
Total	450	750	1200

(a) ¿Se puede afirmar que la proyección hacia el futuro de los graduandos es la mismo en los colegios privados que públicos?

(b) Si se cree que en los colegios públicos la mitad de los graduandos buscan trabajar, un tercio estudian una carrera técnica y el resto estudian una carrera profesional ¿los datos apoyan esta afirmación?

(c) ¿Se puede afirmar que en los colegios privados, el porcentaje de estudiantes que estudian una carrera profesional y el porcentaje de estudiantes que estudian una carrera pública son iguales?

6.3 Verifique las expresiones (6.1.4) y (6.1.5) de los estimadores de máxima verosimilitud bajo distribuciones multinomiales en un problema de dos muestras.

6.4 Verifique la expresión (6.1.9).

6.5 Para el conjunto de datos de la Tabla 6.4,

(a) Encuentre una estimación para el vector de medias, la matriz de varianzas y covarianzas y la matriz de correlaciones de las variables Esperanza de vida de hombre y Esperanza de vida de mujeres

(b) Plantee un sistema para probar que las dos esperanzas de vida promedio son iguales, y decida si esta afirmación es verdadera.

(c) Plantee un sistema para probar que las mujeres en promedio viven 5 años más que los hombres, y decida si esta afirmación es verdadera.

(d) Plantee un sistema para probar que la esperanza de vida de los hombres es independiente de la esperanza de vida de las mujeres, y decida si esta afirmación es verdadera.

6.6 Dada una muestra aleatoria $\mathbf{X}_1, \cdots, \mathbf{X}_n$ con distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $\boldsymbol{\Sigma}$ conocida, desarrolle una regla de decisión para la hipótesis $C\boldsymbol{\mu} = \mathbf{v}$.

6.7 Dada una muestra aleatoria $\mathbf{X}_1, \cdots, \mathbf{X}_n$ con distribución $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, escriba las siguientes hipótesis en forma de $C\boldsymbol{\mu} = \mathbf{v}$,

(a) H_0: $\mu_1 + \mu_3 = \mu_2 + \mu_4$,

(b) H_0: $\mu_1 = \mu_2$ y $\mu_3 = \mu_4$

(c) H_0: $\mu_1 - \mu_2 = \mu_3 - \mu_4 + 5$

(d) H_0: $\mu_1 = \mu_2 + \mu_3 - 1$ y $\mu_4 = \mu_1 + \mu_2$

6.8 Dada una muestra aleatoria $\mathbf{X}_1, \cdots, \mathbf{X}_n$ con distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

(a) Plantee un sistema de hipótesis para probar que las p variables de estudio son independientes y tienen la misma varianza.

(b) Escriba la forma de la estadística de prueba de la razón generalizada de verosimilitudes $2 \ln \lambda$.

(c) Encuentre la distribución de la estadística $2 \ln \lambda$.

Apéndice A

Breve historia del desarrollo estadístico

1650-1700

En este periodo se encuentran los orígenes de la probabilidad y de la estadística mediante el tratamiento matemático del juego y el estudio sistemático de las cifras de mortalidad. Esta época es conocida como la era de la revolución científica en donde grandes nombres como Galileo y Newton dieron algunas ideas de la probabilidad sin influenciar su desarrollo teórico.

Antes de este periodo, hubo algunas contribuciones a la probabilidad, en tanto que Cardano (1501-76) dio algunas probabilidades asociadas al lanzamiento de los dados. Sin embargo, una masa crítica de investigadores y resultados fue alcanzada solamente después de las discusiones entre Pascal y Fermat.

Las estadísticas poblacionales surgen mediante el trabajo de Graunt. William Petty (amigo de Graunt) creó el término Política Aritmética refiriéndose al estudio cuantitativo de la demografía y de la economía. Gregory King fue una importante figura de la siguiente generación. Sin embargo, la línea econométrica no surgió de la manera adecuada. De hecho, el economista más influyente del siglo XVIII, Adam Smith, escribió, ≪Yo no tengo ninguna esperanza en la política aritmética≫.

Una nueva forma de matemáticas de seguros de vida es creada a partir del trabajo de Graunt por los matemáticos Halley, Hudde y de Witt. Mucho después, algunos probabilistas escribirían acerca de temas actuariales, entre ellos de Moivre, Simpson, Price, De Morgan, Gram, Thiele, Cantelli, Cramér y de Finneti. En el siglo XX algunos temas actuariales mas la motivación de G. J. Lidstonen, estimularon a E. T. Whittaker y A. C. Aitken en la contribución del desarrollo estadístico y el análisis numérico.

En nuevas instituciones, además de las universidades tradicionales, se apuntalan estos desarrollos. En Paris y Londres se crean grupos privados de discusión, entre ellos

el de Mersenne, desde donde se crean la Academia de Ciencias y la Sociedad Real de Londres (archivos). En Philosophical Transactions se publican muchas contribuciones a la probabilidad y a la estadística, incluyendo artículos escritos por Halley, de Moivre, Bayes, Pearson, Fisher, Jeffreys y Neyman. Las academias de Berlin y St. Petersburg se formaron poco después.

Siglo XVIII

Hald (1990) llamó a la primera parte de esta época el gran salto (1708-1718), pues hubo contribuciones muy importantes en distintos temas de la probabilidad. Aunque las raíces de la probabilidad y de la estadística son muy distintas, en los comienzos del siglo XVIII se entendía que los dos temas estaban cercanamente relacionados.

Jakob Bernoulli (Ars Conjectandi) y Arnauld (Logique) sugieren una concepción de la probabilidad un poco más amplia que la asociada a los juegos, chances y oportunidades. La ley de los grandes números de Bernoulli establece una teoría que vincula la probabilidad con los datos.

Montmort (*Essay dánalyse sur les jeux de hazard* (1708)) y deMoivre (*Doctrine of Chances* (1718)) son autores que producen nuevos resultados de la teoría de los juegos extendiendo el trabajo de Pascal y de Huygens.

El artículo de Arbuthnot en 1710 (*An Argument for Divine Providence, taken from the Constant Regularity Observed in the Births of Both Sexes*) usa una prueba de significación (la prueba del signo) para establecer que la probabilidad de nacimiento de un varón no es de un medio. Estos cálculos fueron refinados por Gravesande y por Nicolás Bernoulli. Aparte de haber sido una de las primeras aplicaciones de la probabilidad a las estadísticas sociales, el artículo de Arbuthnot ilustra una conexión cercana entre la teología y la probabilidad en la literatura de la época. El trabajo de John Craig establece otro ejemplo de esta situación.

La consideración de la evaluación de riesgos, dramatizada por la Paradoja de San Petersburgo (formulada por Nicolás Bernoulli en 1713 y discutida por Gabriel Cramer) guió la teoría de la esperanza moral (o utilidad esperada) formulada por Daniel Bernoulli (1737).

La probabilidad se establece en la ciencia de la Física, mientras que en la astronomía muestra una influencia. La aplicación más duradera en la astronomía trata acerca de la combinación de observaciones. La teoría resultante de los errores es el ancestro más importante de la inferencia estadística moderna, particularmente en el campo de la teoría de estimación.

Los más importantes astrónomos y matemáticos, incluidos Daniel Bernoulli, Boscovich, Euler, Lambert, Mayer y Lagrange, trataron el problema de la combinación de observaciones astronómicas, ≪para minimizar los errores surgidos de las imperfecciones de los instrumentos y de los órganos de los sentidos≫, en palabras de Thomas Simpson. Simpson introdujo la idea de postular una distribución para los errores. Se desarrollaron algunas pruebas de significación, la mayoría de ellas aplicadas en astronomía. Daniel

Bernoulli, John Michell (1767) y Crossley calcularon las chances (odds) de que el sistema de Pléyades (siete cabrillas) fuera un sistema de estrellas y no un conglomerado aleatorio.

Se realizan afirmaciones en forma de intervalo para el parámetro de la distribución Binomial (ancestros de los intervalos de confianza modernos). Estos fueron propuestos por Lagrange y por Laplace en la década de 1780.

En 1770 Condorcet empieza una publicación acerca de matemáticas sociales, para la aplicación de la teoría de probabilidad en las decisiones de jurados y otras asambleas. Su trabajo tuvo una fuerte influencia en Laplace y Poisson. Otros autores franceses de este periodo son D´Alembert y Buffon; el primero es recordado por sus comentarios críticos en la teoría de probabilidad y el último por el experimento de la aguja.

Un desarrollo importante en la teoría de la probabilidad fue el trabajo de probabilidad condicional con aplicaciones a la probabilidad inversa o Inferencia Bayesiana propuesto por Bayes y Laplace.

Siglo XIX

La característica fundamental son sus fuertes cambios anunciados y gestados en el pasado pero que se efectuarían, de hecho, en este siglo. Cambios en todos los ámbitos de la vida y el conocimiento. Revoluciones de todas las índoles tendrían su lugar. La ciencia y la economía se retroalimentarían, el término ≪científico≫, acuñado en 1833 por William Whewell, sería parte fundamental del lenguaje de la época

1800-1830

Este periodo se encuentra dominado por las figuras de Laplace y Gauss. Laplace estudió en su totalidad la probabilidad y la estadística; Gauss se enfocó solamente en la teoría de los errores.

El trabajo en la teoría de los errores alcanzó un clímax con la introducción del método de los mínimos cuadrados que fue publicado por Legendre en 1805. Durante veinte años hubo tres razonamientos basados en la teoría de la probabilidad: el argumento bayesiano de Gauss (con una distribución a priori uniforme), el argumento de Laplace basado en el teorema central del límite y el argumento de Gauss que se basó en el teorema de Gauss-Markov. El trabajo de investigación continuó a través del siglo XIX con la ayuda y contribución de numerosos astrónomos y matemáticos; entre ellos Cauchy, Poisson, Fourier, Bessel, Encke, Chauvenet y Newcomb (aparece la distribución de Cauchy como un caso poco elegante de la teoría de los errores). Pearson, Fisher[1] y Jeffreys aprenden la teoría de los errores desde la perspectiva de los

[1]Salsburg (2002) afirma que algunos de los primeros artículos de R. A. Fisher son altamente matemáticos. El artículo del coeficiente de correlación, que K. Pearson publicó en *Biometrika*, es denso con respecto a la notación matemática. Una página típica de esta teoría está llena de fórmulas, al menos en un 50 %. Sin embargo, hubo artículos en los que ninguna fórmula matemática aparecía

astrónomos.

Gauss encontró una segunda aplicación de los mínimos cuadrados en la geodesia. Los geodesistas hicieron importantes contribuciones a los mínimos cuadrados, particularmente desde la perspectiva computacional. Los epónimos, Gauss-Jordan y Cholesky, son puestos en honor a posteriores geodesistas. Helmert (la transformada de Helmert) fue un geodesista que contribuyó a la teoría de los errores. Nótese que el topógrafo Frank Yates contribuyó enormemente a la estadística siendo colega y sucesor de Fisher en Rothamsted.

En Gran Bretaña se llevó a cabo el primer censo poblacional en 1801. Éste terminó la controversia acerca del tamaño de la población que empezó con Price, amigo de Bayes, quien argumentaba que la población había decrecido en el siglo XVIII. Numerosos escritores lanzaron estimaciones, incluyendo a Eden.

William Playfair encontró nuevas formas de representación gráfica de los datos. Sin embargo, nadie le prestó atención. La teoría estadística que ganó terreno en los siguientes 150 años no tuvo en cuenta la idea de la graficación de los datos. Esta idea es reciente y se asocia con Tukey.

Concluye la era de las academias y los mayores avances se dan en las universidades. El sistema de educación francesa fue transformado gracias a la revolución y el siglo XIX vio el surgimiento de la universidad alemana.

1830-1860

Este periodo vio el surgimiento de de la sociedad estadística, la cual ha estado activa en la escena científica desde entonces, aunque el significado de la palabra ≪Estadística≫ ha cambiado desde el principio de la literatura filosófica de la probabilidad. En este periodo, también se dio la más glamorosa rama del análisis empírico de las series temporales, el llamado ≪ciclo de las manchas solares≫.

Desde 1830 han existido varias sociedades estadísticas, incluyendo la London (Royal) Statistical Society y la American Statistical Association (ahora la más grande del mundo). El International Statistical Institute fue fundado en 1885 aunque ha organizado congresos internacionales desde 1853. Las estadísticas estuvieron basadas en las poblaciones humanas y en Francia André-Michel Guerry mapeó una clase de estadísticas morales. Quetelet fue un catalizador en la formación de la London Society.

Desde 1840, existe la literatura filosófica de probabilidad. La literatura inglesa

entre líneas. Por ejemplo, en uno de ellos, se discuten las distintas formas en las que la teoría de Darwin, de adaptación aleatoria, se adecuaba a las estructuras anatómicas más adecuadas. En otro artículo, se especula sobre la evolución de la preferencia sexual. Fisher se unió al movimiento de la Eugenesia y en 1917 una editorial en *Eugenics Review*, en donde hacía un llamado para la creación de una política nacional para incrementar la tasa de natalidad de las clases profesionales y entre los artistas más hábiles y desalentar los nacimientos entre las clases bajas.

Su argumento era que las políticas gubernamentales que ayudaban a las personas pobres ayudaban a que estas clases procrearan y pasaran sus genes a la siguiente generación, mientras que las preocupaciones de la clase media, en términos de seguridad económica, hacían que los matrimonios se postergaran y las familias no fueran grandes en número.

empieza con la discusión de probabilidad de John Stuart Mill (1843). Este fue seguido por John Venn, W. Stanley Jevons y Karl Pearson. Hubo un traslape en la literatura de lógica y de probabilidad. De Morgan y Boole también aportaron exhaustivas y largas discusiones acerca de la probabilidad.

En 1843, Schwabe observa que la actividad de las manchas solares (sunspot) era periódica, después de décadas de investigación, no sólo en la física solar sino en el magnetismo terrestre, meteorología e incluso economía, donde se examinaban las series para ver si su periodicidad coincidía con la de las manchas solares. Incluso antes de la manía o moda de las manchas solares hubo un interés intenso en la periodicidad en la meteorología, en el estudio de las mareas y otras ramas de la física observacional.

Juntos, Laplace y Quetelet, habían analizado datos meteorológicos y Herschel había escrito un libro al respecto. Las técnicas en uso variaban desde las más simples, como la tabla de Buys Ballot, a formas más sofisticadas como el análisis armónico. Al final del siglo, el físico Arthur Schuster introdujo el periodograma. Sin embargo, por ese entonces, una forma rival del análisis de series temporales, basada en la correlación y promovida por Pearson, Yule, Hooker y otros, fue tomando forma.

1860-1880

Dos importantes campos de aplicación se abrieron en este periodo. La probabilidad encontró una aplicación más profunda en la física, particularmente en la teoría de gases, naciendo así la mecánica estadística. Los problemas de la mecánica estadística estaban detrás del alcance de los avances de la probabilidad a comienzos del siglo XX. El estudio estadístico de la herencia, desarrollado dentro de la biometría, tuvo lugar. Al mismo tiempo, el mundo sufrió importantes cambios geográficos. Un trabajo importante en la teoría de la probabilidad venía desarrollándose en Rusia, mientras que el trabajo estadístico venía de Inglaterra.

En 1860, James Clerk Maxwell usó la curva del error (distribución normal) en la teoría de los gases; parece que él estaba influenciado por Quetelet. Boltzmann y Gibbs desarrollaron la teoría de gases dentro de la mecánica estadística.

Galton inaugura el estudio estadístico de la herencia, trabajo continuado en el siglo XX por Pearson y Fisher. La correlación fue una de las más distintivas contribuciones de la escuela inglesa.

En contraste, la llamada ≪dirección continental≫ investigaba qué tan apropiado era el uso de los modelos para el tratamiento de las tasas de nacimientos y defunciones por considerar la estabilidad de las series sobre el tiempo.

Hubo una mayor penetración de la estadística en la psicología y en la economía. Se tuvo en cuenta el trabajo de los políticos aritméticos de 1650. El trabajo de Jevons sobre números índice fue inspirado por la teoría de los errores. La investigación en series temporales económicas fue inspirada por el trabajo de lo meteorólogos acerca de la variación estacional de los ciclos solares y sus correlaciones en la tierra.

1880-1900

En este periodo la escuela inglesa estadística tomó forma. Pearson fue el personaje dominante hasta que Fisher lo desplazó en la década de 1920.

Galton introdujo la correlación y una teoría basada en el anterior concepto fue rápidamente desarrollada por Pearson, Edgeworth, Sheppard y Yule. La correlación fue la mayor salida desde el trabajo estadístico de Laplace y Gauss. Empezó a ser ampliamente aplicada en biología, psicología y ciencias sociales.

En economía, Edgeworth siguió algunas ideas de Jevons sobre números índice. Sin embargo, en Inglaterra la economía estadística era más cercana al trabajo en estadísticas oficiales o periodismo financiero. En Italia, Vilfredo Pareto descubrió una regularidad estadística en la distribución del ingreso (distribución de Pareto).

Siglo XX

El siglo XX se ha caracterizado por los avances de la tecnología, medicina y ciencia en general, fin de la esclavitud (al menos nominalmente), liberación de la mujer en la mayor parte de los países, como también por crisis y despotismos humanos, que causaron efectos tales como las guerras mundiales, el genocidio y el etnocidio, las políticas de exclusión social y la generalización del desempleo y de la pobreza. Como consecuencia, se profundizaron las inequidades en cuanto al desarrollo social, económico y tecnológico y en cuanto a la distribución de la riqueza y la calidad de vida entre los países y los habitantes de las distintas regiones del mundo. En los últimos años del siglo, especialmente a partir de 1989-1991 con el derrumbe de los regímenes colectivistas de Europa, comenzó el fenómeno llamado globalización o mundialización.

1900-1920

En los años de la gran guerra (Primera Guerra Mundial, entre 1914 y 1918) la probabilidad y la estadística se esparcieron por todos lados. Durante la guerra, la investigación en probabilidad casi se detiene por causa de que la gente se enlistaba en los servicios armados. Pearson, Lévy y Wiener trabajaron en balística, Jeffreys en meteorología y Yule en administración.

En 1900, David Hilbert propuso un conjunto de problemas para el siglo XX, de los cuales el sexto problema fue ≪tratar... por medio de axiomas, aquellas ciencias físicas en las cuales las matemáticas juegan un papel importante; en primer lugar están la teoría de la probabilidad y la mecánica≫. La teoría de la medida, que jugaría un papel muy importante en la axiomatización de la probabilidad, fue creada por Borel y Lebesgue, entre otros.

Desde diferentes campos surgieron contribuciones que eventualmente encontraron lugar en la teoría de los procesos estocásticos. En física, Einstein y Smoluchovski trabajaron en el movimiento browniano. Bachelier desarrolló un modelo similar aplicado

a la especulación financiera; alternamente, Lundberg desarrolló una teoría de riesgo colectivo. La enfermedad de la malaria y la migración de los mosquitos fueron el foco principal de la investigación de Pearson que originó el problema de la caminata aleatoria. Ronald Ross y A. G.McKendrick, sin la referencia del anterior trabajo de Daniel Bernoulli, crearon modelos matemáticos de epidemias.

Aunque Mendel no usó la probabilidad en su trabajo sobre genética (publicado en 1866), sus ideas fueron probabilizadas cuando Pearson, Yule y Fisher investigaron si los principios de la genética podrían racionalizar los hallazgos de los biometristas.

Charles Spearman (1863-1945) impulsó la correlación y esta empezó a ser parte importante de la sicología. Entre las contribuciones a la estadística estuvieron la correlación de rangos y el análisis factorial. Godfrey Thomson fue un crítico severo del análisis factorial de la inteligencia basado en el trabajo de Spearman. En la década de 1930, Louis L. Thurstone desarrolló el análisis factorial múltiple.

En economía, especialmente en los Estados Unidos, los métodos cuantitativos empezaron a ser más prominentes. Las figuras más importantes fueron Warren Persons, Irving Fisher, Wesley Mitchell y H. L. Moore. La mayoría de su trabajo se clasificaría en el análisis de series de tiempo.

Las aplicaciones industriales en probabilidad empezaron con el trabajo de Erlang sobre congestión de sistemas telefónicos, el ancestro de la teoría de colas.

Los desarrollos institucionales incluyen, en 1911, la creación del Departamento de Estadísticas Aplicadas en UCL encabezado por Pearson. Yule podría ser llamado ≪el primer estadístico moderno≫.

1920-1930

La mayoría de las personas que dominaron la probabilidad y la estadística tuvieron un impacto temprano. De ellos, el individuo que tuvo un mayor impacto fue Fisher en estadística. El alemán era el idioma tradicional en la literatura científica de la época. Sin embargo, Fisher escribía en inglés, pues creía que la época de escritura en alemán había terminado con Gauss. Los avances en probabilidad incluyeron refinamientos del teorema central del límite (Lindeberg hizo una muy importante contribución) y de la ley fuerte de los grandes números junto con nuevos resultados que incluían la ley del algoritmo dominado. Hubo contribuciones de la mayoría de los países del continente europeo; por ejemplo, Mazurkiewicz de Polonia y en 1935, Turing quien repitió el trabajo de Lindeberg sin saber de su publicación.

Los fundamentos de la probabilidad recibieron mucha atención y ciertas posiciones encontraron expresiones clásicas: la interpretación lógica de la probabilidad (grado de creencia razonable) fue propuesta por los filósofos de Cambridge, W. E. Johnson, J. M. Keynes y C. D. Broad, y presentada a una audiencia científica por Jeffreys; el punto de vista frecuentista fue desarrollado por von Mises.

En estadística, R. A. Fisher generó nuevas ideas sobre estimación y prueba de hipótesis y su trabajo de diseño experimental movió este tópico desde los linderos

hasta el centro de la estadística. Sus métodos estadísticos para investigadores (1925) constituyeron el libro más influyente del siglo XX. W. A. Shewhart (ASQ) fue el pionero del control de calidad, que se convirtió en una aplicación muy importante de la estadística en la industria.

1930-1940

En contra de una economía en recession y de una política desastrosa, hubo importantes desarrollos en probabilidad, teoría estadística y sus aplicaciones. En la Unión Soviética, a los matemáticos les iba mejor que a los economistas o a los genetistas y pudieron salir de su país y publicar en revistas internacionales. De esta manera, Kolmogorov y Khinchin publicaron en Alemania, donde precisamente los judíos fueron expulsados de la academia desde 1934.

En probabilidad, los principales desarrollos fueron la axiomatización de la probabilidad por Kolmogorov y el desarrollo de la teoría de los procesos estocásticos por él y por Khinchin. Su trabajo es usualmente visto como el comienzo de la probabilidad moderna.

En los fundamentos de la probabilidad, Bruno de Finetti y Frank Ramsey (1903-1930) (St. Andrews, N.-E. Sahlin) trabajaron en la probabilidad subjetiva. Ramsey empezó con el criticismo de la escuela de lógica de Cambridge, en particular Keynes. Una superestructura estadística no se dio sino años después, Jeffreys dio un tratamiento completo a la estadística fundamentado en su noción lógica de la probabilidad, aunque la forma prevaleciente era la clásica. *Biometrika* detuvo la publicación de la investigación biológica y se enfocó en la estadística teórica. El Instituto de Estadística Matemática fue fundado en 1930 y su revista, *The Annals of Mathematical Statistics*, apareció en 1933. El primer laboratorio de estadística en los Estados Unidos fue creado en Iowa por Snedecor en 1933. Snedecor fue fuertemente influenciado por Fisher.

En el campo de la inferencia estadística, el mayor desarrollo fue la teoría del prueba de hipótesis de Neyman-Pearson. El análisis multivariado se convirtió en una material identificable, formado por contribuciones como la distribución Wishart (1928), los componentes principales de Harold Hotelling (1933) y la correlación canónica (1936) y el análisis discriminante de Fisher (1936).

Las aplicaciones de las matemáticas y estadísticas a la economía se juntaron en el movimiento econométrico. Entre los líderes de la década de 1930 estuvieron Jan Tinbergen y Ragnar Frisch. Los econometristas que ganaron el premio Nobel en economía son Engle, Granger, Haavelmo, Heckman, Klein y McFadden.

1940-1950

Entre los millones de muertos de la Segunda Guerra Mundial se contaron algunos matemáticos y estadísticos. Doeblin es el más famoso de los finados, y uno de los libros de Neyman está dedicado a la memoria de diez colegas y amigos. Esta guerra

incentivó el estudio de la probabilidad y la estadística. Al final de la guerra, muchas personas se encontraron trabajando como estadísticos, hubo nuevas aplicaciones y la importancia de esta material fue más ampliamente reconocida.

Las persecuciones nazis y la Segunda Guerra Mundial empujaron la migración de muchos estadísticos a los Estados Unidos. Algunas de las más importantes figuras de la probabilidad en la postguerra en Estados Unidos son: Feller, M. Kac (MGP), Wald, G. E. P. Box (MGP), W. G. Cochran (ASA) (MGP), W. Hoeffding (MGP), H. O. Hartley (MGP), F. J.Anscombe (Obit. p. 17) (MGP), Z. W. Birnbaum (MGP) y O. Kempthorne (MGP).

Los métodos no-paramétricos empezaron a ser sistemáticamente estudiados, usando técnicas de la teoría de la inferencia estadística; E. J. G. Pitman fue un pionero importante. Las pruebas estadísticas para la prueba de hipótesis vinieron de no-estadísticos como Spearman (rangos) o Wilcoxon (prueba de Wilcoxon). El repertorio conocido de las pruebas del signo, pruebas de permutación y la prueba de Kolmogorov-Smirnov se expandió rápidamente en el medio.

El análisis moderno de series de tiempos vino de la unión de la teoría de los procesos estocásticos, la teoría de la predicción y la teoría de la inferencia estadística. Uno de los principales pioneros de esta década fue M. S. Bartlett. En la década de 1950 Tukey fue una figura importante. En la década de 1960, Kalman (filtro de Kalman) y los sistemas de ingeniería hicieron importantes contribuciones y en la década de 1970, los métodos de G. E. P. Box y G. M. Jenkins (Box-Jenkins) fueron adoptados en la economía y los negocios.

1950-1980

Este es un periodo de expansión hacia más países, más gente, más departamentos, más libros, más revistas. Los computadores empiezan a tener un gran impacto.

Los departamentos existentes de estadística se expanden. Nuevas instituciones son creadas, entre ellas el Laboratorio Estadístico en Cambridge en 1947 y el Departamento de Estadística en Harvard en 1958.

El alcance de la teoría de la probabilidad se incrementa con el nacimiento de nuevos subcampos como la teoría de colas y la teoría de la renovación. El libro de Feller, *Introduction to Probability Theory* hizo un impacto muy grande en el mundo de habla inglesa, pues promovió el estudio de tópicos más avanzados como las cadenas de Markov.

En materia estadística hubo un renacimiento Bayesiano. En Estados Unidos, la teoría de decisión Bayesiana reflejó la influencia de la teoría de la decisión de Wald. W. Edwards Deming continúo el trabajo de Shewhart en control de calidad y fue muy efectivo a la hora de adoptar estos métodos en la industria. Laplace y Quetelet vieron el trabajo de los censos como posibles aplicaciones de la probabilidad pero el uso de la teoría estadística para recopilación de información oficial llegó sólo después de las actividades de Morris Hansen en la oficina de censos de Estados Unidos.

1980-Presente (los efectos del computador)

Este periodo describe el efecto impactante de los ordenadores en el desarrollo de métodos estadísticos desde su advenimiento, en la década de 1950 y el dramático cambio en la historia de la probabilidad y la estadística en las recientes décadas. Al final del siglo XIX, las máquinas mecánicas calculadoras proveyeron el material para la investigación de Pearson y Fisher y la construcción de sus tablas estadísticas. Con la disponibilidad de los computadores, las viejas actividades tomaron menos tiempo y nuevas actividades fueron posibles.

Las tablas estadísticas de números aleatorios fueron mucho más fáciles de producir y luego desaparecieron, pues su función fue sometida a los paquetes estadísticos.

Una gran masa de datos, más grande que en épocas pasadas, puede ser analizada. El Data mining exhaustivo es posible.

Modelos y métodos más complejos pueden ser usados. Los nuevos métodos se han diseñado con la idea de la implementación computacional. Por ejemplo, la familia de los modelos lineales generalizados vinculada al programa GLIM.

En el siglo XX, cuando Student (1908) escribió sobre la media normal y Yule (1926) escribió sobre las correlaciones sin sentido, se usaron experimentos basados en muestras y en la década de 1920 se comenzó producir tablas de números aleatorios. Esto cambió con la introducción de los métodos asistidos por el computador para la generación de números pseudo-aleatorios; más aún, los métodos de Monte-Carlo (introducidos por von Neumann y Ulam) fueron posibles.

Desde 1980 los métodos de Monte Carlo han sido estudiados y usados directamente en el análisis de datos. En la inferencia clásica, el bootstrap ha sido prominente.

Apéndice B

Herramientas de bondad de ajuste

En este apéndice discutiremos herramientas estadísticas que nos pueden dar una idea acerca de qué distribución probabilística puede ser apropiada para un conjunto de datos observados, es decir, qué tan bien se ajusta una distribución teórica a los datos, de allí el nombre de ≪bondad de ajuste≫. Entre estas herramientas se encuentran herramientas gráficas como el QQ plot y pruebas estadísticas de uso común.

Gráficas QQ plot

La gráfica QQ plot es una gráfica muy sencilla y útil que compara la similitud entre los percentiles de una distribución probabilística y los percentiles muestrales de un conjunto de datos[1]. La idea de esta gráfica consiste en que si la distribución teórica ajusta bien a los datos, entonces los percentiles teóricos deben ser similares a los percentiles muestrales.

Para ilustrar el método, consideramos un conjunto de datos $x_1, \cdots, x_n$. Aunque no sabemos qué distribución es la que mejor se ajusta a estos datos, podemos calcular la función de distribución empírica $\hat{F}_n(x)$ dada por

$$\hat{F}_n(x) = \frac{\text{Número de datos que es menor ó igual a } x}{n} \tag{B.0.1}$$

De esta forma, la función de distribución empírica de un conjunto de datos será una función escalonada tal como en la Figura B.1, donde los valores $x_{(1)}, \cdots, x_{(n)}$ son los datos ordenados. En caso de que los datos muestrales provienen de una variable continua, podemos volver continua a la función $\hat{F}_n(x)$ uniendo los puntos medios de cada segmento, en la Figura B.1, ésta es la función con línea discontinua. Y al recordar

[1] El nombre de QQ plot viene del inglés *Quantile-Quantile plot*.

que la definición de un percentil en una distribución continua, tenemos que el dato $x_{(j)}$ es el percentil muestral de probabilidad aproximada igual $\left(\frac{j-1}{n}+\frac{j}{n}\right)/2=\frac{j-0.5}{n}$.

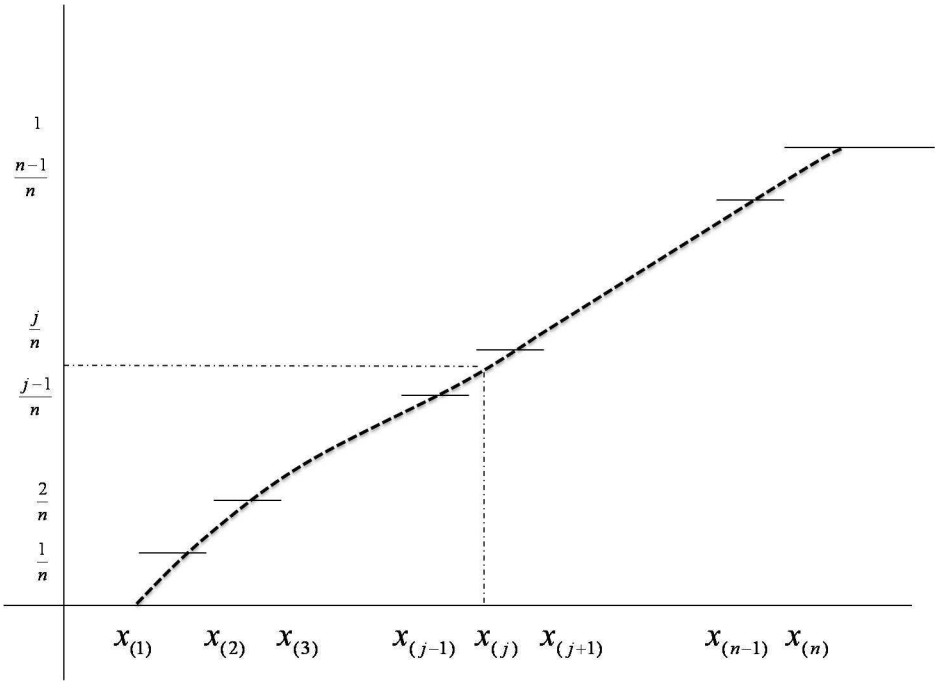

Figura B.1: *Función de distribución empírica y percentil muestral.*

Dado lo anterior, si una distribución teórica es adecuada para los datos, entonces $x_{(j)}$ debe ser similar al percentil teórico $\frac{j-0.5}{n}$, esto es, $x_{(j)}\approx F^{-1}\left(\frac{j-0.5}{n}\right)$ donde F denota la función de distribución de la distribución teórica. De esta forma, la gráfica QQ plot consiste en graficar los puntos $\left(F^{-1}\left(\frac{j-0.5}{n}\right),x_{(j)}\right)$ para todo $j=1,\cdots,n$.

Ahora, en la mayoría de las distribuciones, la función F y F^{-1} depende del parámetro de la distribución, y este parámetro no es conocido, y por consiguiente no hay forma de calcular exactamente $F^{-1}\left(\frac{j-0.5}{n}\right)$. Una forma de solucionar esto es tratar de encontrar una relación entre los datos ordenados y la parte de $F^{-1}\left(\frac{j-0.5}{n}\right)$ que no involucre a los parámetros desconocidos, y mirar si en la gráfica aparece tal relación. A continuación, ilustramos el cómputo y aplicación de esta gráfica para diferentes distribuciones.

QQ plot para una distribución exponencial

Suponga que tenemos un conjunto de datos positivos que son realizaciones de una variable continua, y nos interesa saber si la distribución exponencial puede ser apropiada para estos datos utilizando la gráfica QQ plot. Para eso recordamos que la función de distribución de una distribución exponencial está dada por $F_X(x)=1-e^{-x/\theta}$ para

$x > 0$, de donde tenemos que la inversa de F está dada por $F^{-1}(x) = -\theta \ln(1-x)$, y de esta forma, $x_{(j)}$ debe ser similar a $-\theta \ln\left(1 - \frac{j-0.5}{n}\right)$, lo cual es equivalente a que debe haber una relación lineal entre $x_{(j)}$ y $-\ln\left(1 - \frac{j-0.5}{n}\right)$. Por lo tanto, en la práctica, se grafican los puntos $\left(-\ln\left(1 - \frac{j-0.5}{n}\right), x_{(j)}\right)$ para $j = 1, \cdots, n$ y se observa si hay una tendencia lineal. En caso afirmativo, se puede concluir que la distribución exponencial es apropiada para los datos. Además, podemos estimar el parámetro θ como la pendiente de la recta (sin intercepto) que mejor se ajusta a los puntos $\left(-\ln\left(1 - \frac{j-0.5}{n}\right), x_{(j)}\right)$ para $j = 1, \cdots, n$.

Para mostrar la efetividad de esta gráfica para identificar una distribución exponencial, simulamos dos conjuntos de datos de la distribución $Exp(0.2)$ de tamaño 10 y 50. En cada conjunto de datos se grafica el QQ plot junto con la recta sin intercepto que mejor se ajusta a estos puntos. Este procedimiento se lleva a cabo con los siguiente códigos, donde adicionalmente se calcula la estimación de θ según el método descrito anteriormente. Y los resultados se muestran en la Figura B.2.

```
> qq.exp<-function(y){
+ y<-sort(y)
+ n<-length(y)
+ j<-c(1:n)
+ percen<--log(1-(j-0.5)/n)
+ plot(percen,y,xlab="",ylab="",main="QQ plot exponencial")
+ }
>
> qq.exp.line<-function(y){
+ y<-sort(y)
+ n<-length(y)
+ j<-c(1:n)
+ percen<--log(1-(j-0.5)/n)
+ abline(0,1/lm(percen ~ y-1)$coef)
+ return(1/lm(percen ~ y-1)$coef)
+ }
>
> set.seed(1234)
> x1<-rexp(10,5)
> x2<-rexp(50,5)
> par(mfrow=c(1,2))
> qq.exp(x1)
> qq.exp.line(x1)
        y
0.1663394
> qq.exp(x2)
> qq.exp.line(x2)
        y
0.2251790
```

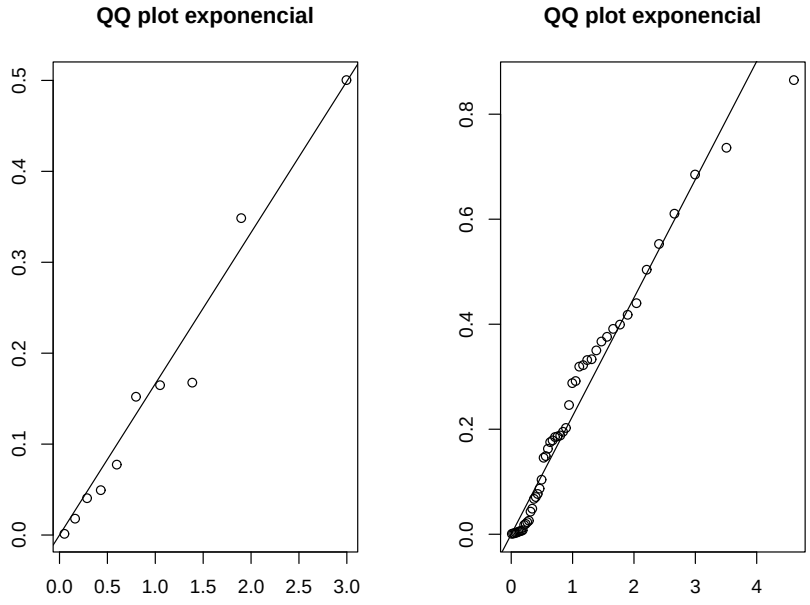

Figura B.2: *Gráfica QQ para verificar la distribución exponencial para dos conjuntos de datos generados de* $Exp(0.2)$ *con* $n = 10$ *y* $n = 50$.

En ambos conjuntos de datos, la gráfica muestra la tendencia de una línea recta, indicando que los percentiles muestrales son parecidos a los percentiles teóricos, y por consiguiente, la distribución exponencial puede ser apropiada para los datos. Además podemos ver que la estimación de θ cuando $n = 50$ es de 0.225, que es muy similar a la estimación de máxima verosimilitud $\bar{x} = 0.224$.

Ahora, también podemos simular datos de otras distribuciones, y ver qué tan buena es la gráfica QQ para detectar distribuciones que es no exponencial. Simulamos 40 datos de las distribuciones $Unif(3, 5)$, $Gamma(3, 0.2)$ y $N(10, 4)$, y en cada uno de estos conjuntos vemos el ajuste de una distribucion exponencial usando QQ plot. La gráfica resultante se muestra en la Figura B.3, donde podemos ver que se pueden detectar fácilmente las distribuciones que no corresponden a la distribución exponencial.

QQ plot para una distribución normal

Ahora, discutimos sobre el QQ plot para ver si una distribución normal puede ser apropiada para los datos. La idea sigue consistiendo en comparar a $x_{(j)}$ que el percentil muestral $\frac{j-0.5}{n}$ con el percentil teórico $\frac{j-0.5}{n}$ de una distribución normal. Ahora, la distribución normal también tiene dos parámetros μ y σ excepto en la distribución normal estándar donde $\mu = 0$ y $\sigma = 1$. Entonces la idea es estandarizar las

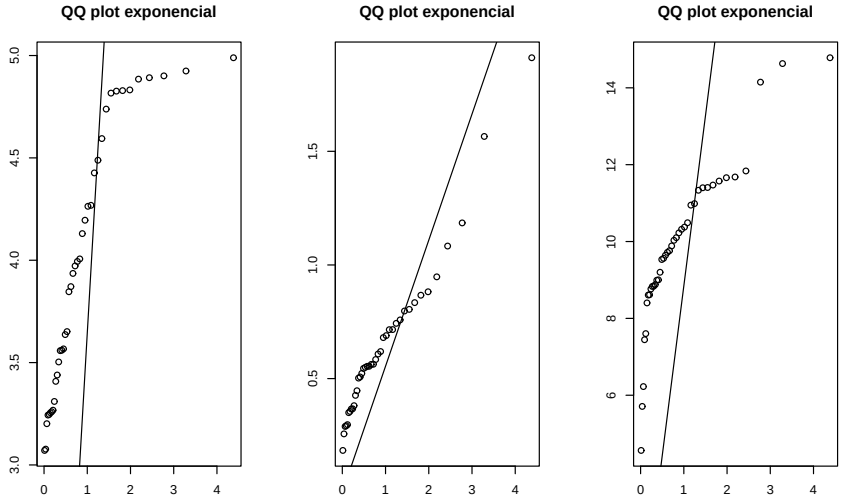

Figura B.3: *Gráfica QQ para verificar la distribución exponencial para tres conjuntos de datos generados de* $Unif(3,5)$, $Gamma(3,0.2)$ *y* $N(10,4)$ *con* $n = 40$.

observaciones $x_1, \cdots, x_n$ y comparar los datos estandarizados con los percentil de la distribución normal estándar. Si la distribución normal es apropiada para estos datos, entonces los datos estandarizados deben ser aproximadamente iguales a los percentiles teóricos. Por tanto, entonces en la práctica graficamos una diagrama de dispersión de los puntos $(z_{\frac{j-0.5}{n}}, z_{(j)})$ para $j = 1, \cdots, n$, donde $z_{\frac{j-0.5}{n}}$ es el percentil $\frac{j-0.5}{n}$ de la distribución normal estándar, y $z_{(j)}$ es el dato $x_{(n)}$ estandarizado, y la nube de puntos debe estar alrededor de una línea recta de 45° de inclinación.

Para ver la efectividad de este método, simulamos dos conjuntos de datos de la distribución $N(5,4)$ con $n = 10$ y $n = 50$, y en cada muestra se grafica el anterior QQ plot. Utilizamos los siguientes códigos, y la gráfica resultante se muestra en la Figura B.4, donde podemos ver que los percentiles muestrales son muy parecidos a los percentiles teóricos y por consiguiente, llegamos a la conclusión correcta de que la distribución normal es apropiada para los datos.

```
> set.seed(123)
> n1<-10
> n2<-50
> x1<-rnorm(n1,5,2)
> z1<-(x1-mean(x1))/sd(x1)
>
> j1<-c(1:n1)
> percen1<-(j1-0.5)/n1
>
```

```
> x2<-rnorm(n2,5,2)
> z2<-(x2-mean(x2))/sd(x2)
>
> j2<-c(1:n2)
> percen2<-(j2-0.5)/n2
>
> par(mfrow=c(1,2))
> plot(qnorm(percen1),sort(z1))
> abline(0,1)
> plot(qnorm(percen2),sort(z2))
> abline(0,1)
```

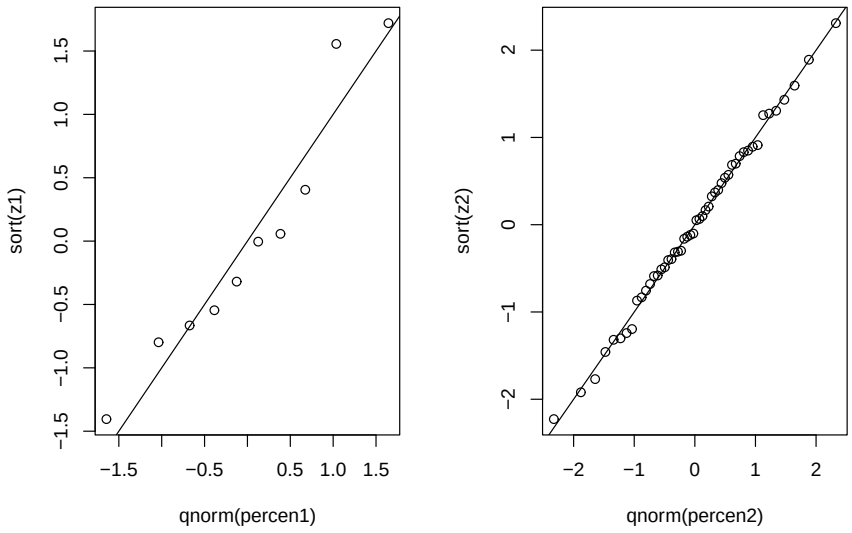

Figura B.4: *Gráfica QQ para verificar la distribución normal para dos conjuntos de datos generados de* $N(5,4)$ *con* $n = 10$ *y* $n = 50$.

En R, la gráfica QQ plot se realiza directamente sobre los datos $x_1, \cdots, x_n$ sin necesidad de estandarizar. La gráfica se produce con la funciones `qqnorm` y `qqline`. El lector puede ejecutar el comando para los objetos `x1` y `x2` creados anteriormente, y ver que se producen gráficas muy similares a las de la Figura B.4.

```
> par(mfrow=c(1,2))
> qqnorm(x1)
> qqline(x1)
> qqnorm(x2)
> qqline(x2)
```

QQ plot para una distribución Weibull

La distribución Weibull es común en la práctica para modelar el tiempo transcurrido hasta el suceso de algún evento, y en estas prácticas, podemos utilizar la gráfica QQ plot para ver la validez de suponer esta distribución para un conjunto de datos. Para desarrollar esta gráfica es necesario conocer la función de distribución de la distribucion Weibull, la cual podemos obtener fácilmente de la función de densidad dada en (1.1.11). Ésta está dada por

$$F_X(x) = 1 - \exp\left\{-\frac{x^k}{\theta^k}\right\}$$

para $x > 0$, y podemos obtener la inversa de F dada por

$$F^{-1}(x) = \theta\left[-\ln(1-x)\right]^{1/k}$$

De esta forma, los datos ordenados $x_{(j)}$ deben ser parecidos al percentil teórico $F^{-1}(\frac{j-0.5}{n}) = \theta\left[-\ln\left(1-\frac{j-0.5}{n}\right)\right]^{1/k}$. Y por consiguiente,

$$\ln x_{(j)} \approx \ln\left\{\theta\left[-\ln\left(1-\frac{j-0.5}{n}\right)\right]^{1/k}\right\} = \ln\theta + \frac{1}{k}\ln\left[-\ln\left(1-\frac{j-0.5}{n}\right)\right]$$

Es decir, si graficamos los puntos $\left(\ln\left[-\ln\left(1-\frac{j-0.5}{n}\right)\right], \ln x_{(j)}\right)$, la nube de puntos debe ser de forma lineal donde el intercepto es cercano a $\ln\theta$ y la pendiente $1/k$. Entonces en la práctica se grafica el diagrama de dispersión de estos puntos, y el intercepto y la pendiente de la línea que mejor se ajusta a estos puntos provee estimaciones de los parámetros k y θ.

Para verificar la efectividad de esta gráfica, simulamos dos conjuntos de datos de la distribución $Weibull(2, 10)$ con $n = 15$ y $n = 50$, y para cada conjunto de datos, graficamos el QQ plot utilizando la función `qq.wei` descrita a continuacion.

```
> qq.wei<-function(y){
+ y<-sort(y)
+ n<-length(y)
+ j<-c(1:n)
+ percen<-log(-log(1-(j-0.5)/n))
+
+ inte<-lm(log(y)~percen)$coef[1]
+ pend<-lm(log(y)~percen)$coef[2]
+ plot(percen,log(y),xlab="",ylab="",main="QQ plot Weibull")
+ abline(inte,pend)
+ thet<-exp(inte)
+ k<-1/pend
+ list("theta"=thet, "k"=k)
+ }
```

```
> set.seed(1234)

> x1<-rweibull(15,shape=2,scale=10)
> x2<-rweibull(50,shape=2,scale=10)

> par(mfrow=c(1,2))
> qq.wei(x1)
$theta
(Intercept)
   9.831436
$k
  percen
2.538407

> qq.wei(x2)
$theta
(Intercept)
   10.97860
$k
  percen
2.293684
```

La gráfica resultante se encuentra en la Figura B.5, donde podemos ver que los puntos están alrededor de la recta, aunque en el conjunto de 15 datos, el ajuste no parece del todo apropiado. Por otro lado, también observamos que las estimaciones de k y θ obtenidas con la función `qq.wei` son cercanas a los valores verdaderos $k = 2$ y $\theta = 10$.

Ahora miramos qué tan buena es esta gráfica para identificar datos que provienen de una distribución diferente de Weibull. Simulamos conjuntos de datos de las distribuciones $Unif(3, 5)$, $Gamma(3, 0.2)$ y $N(10, 4)$ con $n = 10$, y gráficas del QQ plot para cada uno de estos datos. Las gráficas resultantes se encuentran en la Figura B.6 donde vemos que no se muestra ningún mal ajuste de estas distribuciones. Aumentamos el tamaño muestral a $n = 50$ (ver Figura B.7); sin embargo, en el caso de la distribución uniforme y normal el QQ plot tampoco fue capaz de descubrir que los datos no provienen de una distribución Weibull.

Para los datos simulados de una distribución normal, podemos ver que los puntos tienen una forma curva, que no se puede aproximar por una recta, y de donde podemos identificar que la distribución Weibull no es apropiada para los datos.

El hecho de que la gráfica QQ no sea capaz de identificar algunas distribuciones diferentes de la Weibull, es sin duda una falla de la gráfica QQ plot para el caso de la distribución Weibull; sin embargo, las Figuras C.6 y C.7 son obtenidas a partir de algunos datos simulados, y no podemos garantizar que esta QQ plot funcione mal en todos los casos.

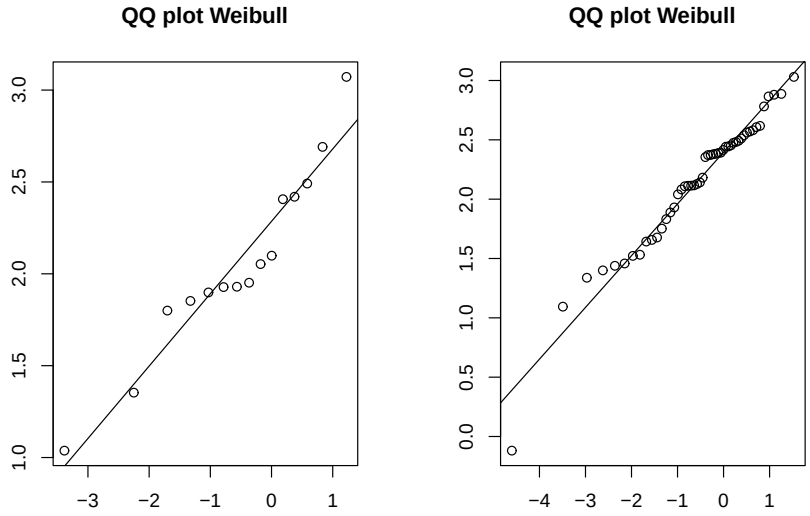

Figura B.5: *Gráfica QQ para verificar la distribución Weibull para dos conjuntos de datos generados de* $Weibull(2, 10)$ *con* $n = 15$ *y* $n = 50$.

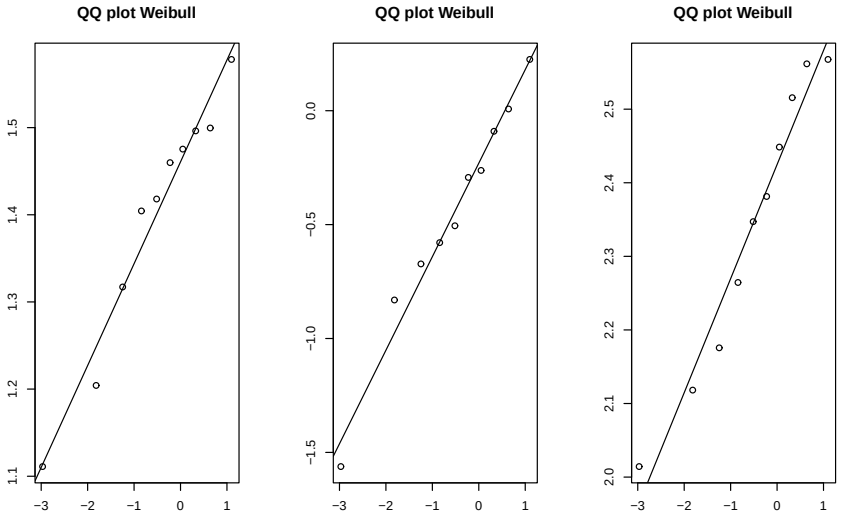

Figura B.6: *Gráfica QQ para verificar la distribución Weibull para tres conjuntos de datos generados de* $Unif(3, 5)$, $Gamma(3, 0.2)$ *y* $N(10, 4)$ *con* $n = 10$.

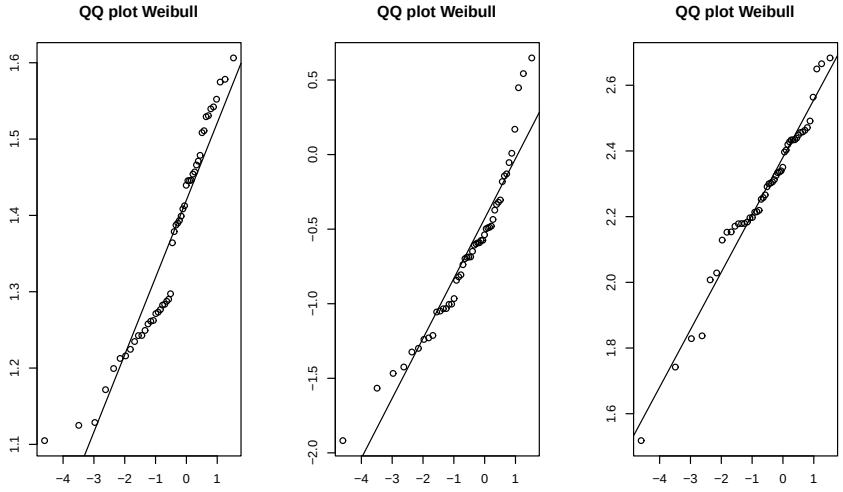

Figura B.7: *Gráfica QQ para verificar la distribución Weibull para tres conjuntos de datos generados de* $Unif(3,5)$*,* $Gamma(3,0.2)$ *y* $N(10,4)$ *con* $n = 50$.

QQ plot para una distribución Gamma pos-estimación

Cuando tenemos un conjunto de datos $x_1, \cdots, x_n$ que son positivos, altamente no simétricos y que parecen ser realizaciones de una variable continua, y pensamos ajustarles una distribución Gamma, lo ideal es poder hacer una gráfica QQ plot para ver si esta distribución puede ser apropiada para los datos, y en caso afirmativo, llevar a cabo procedimientos de inferencia acerca de los parámetros y/o otras cantidades de interés. Sin embargo, es muy difícil hallar la inversa de la función de distribución de la distribución Gamma, pues ésta es de una forma bastante complicada.

Sin embargo, podemos, mediante otras herramientas como el histograma, suponer la distribución Gamma para los datos, y al suponer la distribución Gamma, podemos utilizar los datos para estimar el parámetro de forma k y el parámetro de escala θ. Y podemos ver qué tan aproximados son los datos $x_{(j)}$ con el percentil teórico $\hat{F}^{-1}(\frac{j-0.5}{n})$ donde $\hat{F}$ denota la función de distribución $Gamma(\hat{k}, \hat{\theta})$. Es decir, se mira el ajuste de la distribución Gamma con los parámetros estimados (de allí, el nombre de pos-estimación) a los datos. Si la distribución Gamma es apropiada para los datos, y además la estimación de k y θ era buena, entonces se espera que $x_{(j)} \approx \hat{F}^{-1}(\frac{j-0.5}{n})$, y la nube de puntos se debe aproximar a una recta sin intercepto de 45°.

Ilustramos el anterior procedimiento con los datos en una encuesta a hogares bogotanos del estrato 3 que corresponden al gasto semanal en alimentación. La muestra está conformada por 389 hogares, y el histograma de los datos está dado en la Figura B.8 donde tenemos una no simetría muy marcada.

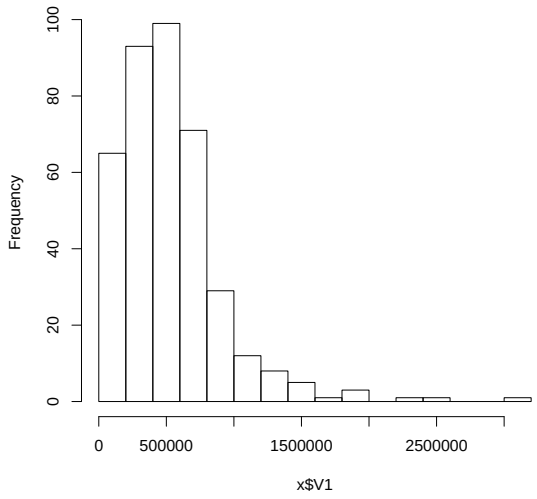

Figura B.8: *Histograma de los datos de gasto mensual en alimentación de 389 hogares bogotanos del estrato 3.*

Si consideramos ajustar una distribución Gamma a los datos, primero estimamos los parámetros k y θ vía el método de los momentos. Estas estimaciones son $\hat{k} = 1.89$, $\hat{\theta} = 278111.2$, y podemos utilizar la gráfica QQ para ver si la distribución $Gamma(1.89, 278111.2)$ se ajusta bien a los datos. Podemos utilizar el siguiente código en R para obtener la gráfica QQ.

```
> ## en y contiene los 389 datos
> n<-length(y)
> plot(qgamma((c(1:n)-0.5)/n,shape=1.89,scale=278111.2),sort(y),
+ xlab="Percentiles muestrales", ylab="percentiles teóricos")
> abline(0,1)
```

La gráfica resultante se encuentra en la Figura B.9, donde observamos que la mayoría de los percentiles muestrales son cercanos a los percentiles teóricos, indicando que la distribución Gamma es apropiada para estos datos. En el Ejemplo 2.3.15, también se discutió otra forma para ver el ajuste de la distribución de los datos que consiste en trazar la función de densidad sobre el histograma de los datos. Aquí también podemos optar por este enfoque, la gráfica se encuentra en la Figura B.10, donde vemos el buen ajuste de la distribución $Gamma(1.89, 278111.2)$ a los datos.

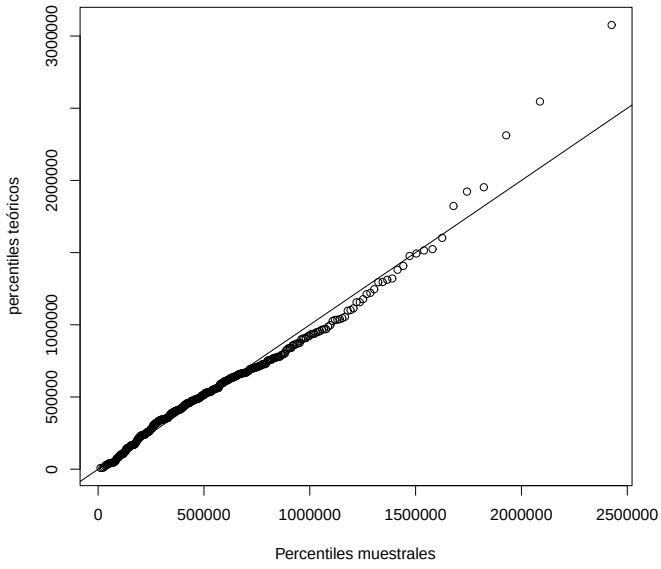

Figura B.9: *Gráfica QQ plot para ver el ajuste de la distribución* $Gamma(1.89, 278111.2)$ *a los datos de gasto mensual en alimentación de 389 hogares bogotanos del estrato 3.*

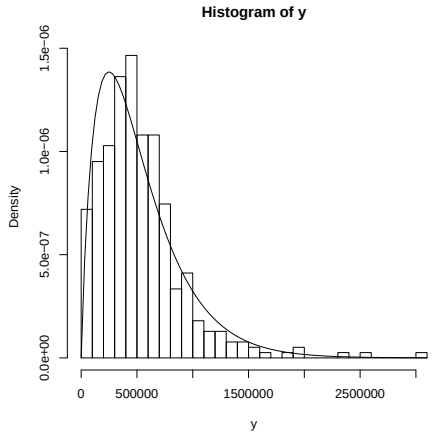

Figura B.10: *Histograma y función de densidad* $Gamma(1.89, 278111.2)$ *de los datos de gasto mensual en alimentación de 389 hogares bogotanos del estrato 3.*

QQ plot para una distribución Beta pos-estimación

La gráfica QQ para verificar si una distribución Beta puede ser apropiada para un conjunto de valores entre 0 y 1 se puede obtener de manera análoga al caso de una distribución Gamma. Es decir, primero estimamos los parámetros a y b de la distribución Beta usando los estimadores expuestos en el Ejemplo 2.3.16, y posteriormente comparamos los percentiles muestrales $x_{(j)}$ con los percentiles teóricos de la distribución $Beta(\hat{a}, \hat{b})$. Y si la distribución Beta es apropiada para los datos, y las estimaciones de a y b fueron buenas, entonces los percentiles muestrales deben ser aproximadamente iguales a los percentiles teóricos.

Ilustramos el anterior procedimiento para los datos del Ejemplo 2.3.16. El siguiente código calcula $\hat{a}$ y $\hat{b}$, además de graficar el QQ plot. Las estimaciones fueron $\hat{a} = 0.5447$ y $\hat{b} = 9.8$. Y el ajuste de la distribución $Beta(0.5447, 9.8)$ se puede observar en la Figura B.11, donde se puede observar un mejor ajuste que con la distribución exponencial (ver Figura 2.10 del Ejemplo 2.3.16).

```
> x<-c(0.7, 1.4, 19.7, 0.1, 12.4, 1.1, 0.5, 18.9, 5.0, 0.3,
+ 0.6, 5.4, 6.7, 0.9)/100
> n<-length(x)
> va<-var(x)*(n-1)/n
> bar<-mean(x)
> a<-bar^2*(1-bar)/va-bar
> b<-(1-bar)*(bar*(1-bar)/va-1)
> plot(qbeta((c(1:n)-0.5)/n,a,b),sort(x),xlab="Percentiles
+ muestrales", ylab="percentiles teóricos")
> abline(0,1)
```

Ahora, aunque la distribución Beta se caracteriza en que las realizaciones oscilan entre 0 y 1, esto no implica que cada vez que se trate de datos entre 0 y 1, necesariamente se haga referencia a la distribución Beta, puesto que una distribución normal con media 0.5 y una varianza pequeña también puede producir valores entre 0 y 1. A continuación simulamos 30 datos de la distribución $N(0.5, 0.3^2)$ y graficamos el QQ plot para la distribución Beta y otro para verificar la distribución normal. El código utilizado se encuentra a continuación, y podemos ver la gráfica resultante en la Figura B.12, donde claramente con el QQ plot para la distribución Beta no se logra un buen ajuste comparado con el QQ plot para la distribución normal.

```
> set.seed(12345678)
> x<-rnorm(30,0.5,0.3)
> n<-length(x)
> va<-var(x)*(n-1)/n
>  bar<-mean(x)
>  a<-bar^2*(1-bar)/va-bar
>  b<-(1-bar)*(bar*(1-bar)/va-1)
>  par(mfrow=c(1,2))
```

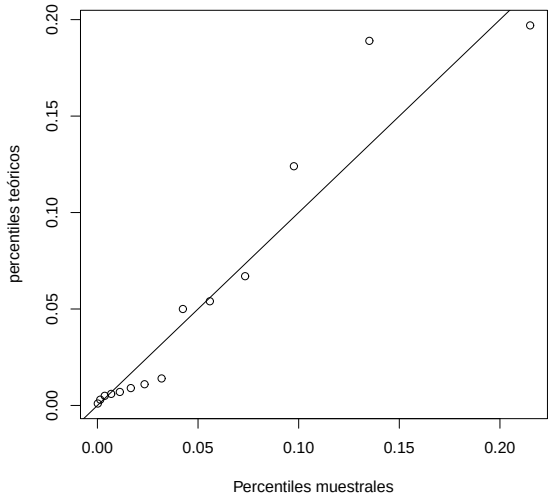

Figura B.11: *Gráfica QQ plot para verificar el ajuste de la distribución Beta a los datos del Ejemplo 2.3.16.*

```
> plot(qbeta((c(1:n)-0.5)/n,a,b),sort(x),xlab="Percentiles
+ muestrales", ylab="percentiles teóricos",main="QQ plot Beta")
> abline(0,1)
> qqnorm(x)
> qqline(x)
```

QQ plot para Normal multivariante

Ahora supongamos que tenemos observaciones de p variables continuas sobre n individuos $\mathbf{x}_1, \cdots, \mathbf{x}_n$, y estamos interesados en saber si es apropiado ajustarles una distribución normal p variante. Para ello, podemos razonar de la siguiente forma, si los vectores observados son realizaciones de una muestra de vectores aleatorios $\mathbf{X}_1, \cdots, \mathbf{X}_n$ con distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces por la propiedad 6 del Resultado 5.2.4, la distancia de Mahalanobis entre $\mathbf{X}_j$ y $\boldsymbol{\mu}$ dada por $(\mathbf{X}_j - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_j - \boldsymbol{\mu})$ tiene distribución χ^2_p para todo $j = 1, \cdots, n$. Al reemplazar $\boldsymbol{\mu}$ por su estimador $\bar{\mathbf{X}}$, y $\boldsymbol{\Sigma}$ por $\mathbf{S}_{n-1}$, podemos suponer que la distribución de $(\mathbf{X}_j - \bar{\mathbf{X}})'\mathbf{S}^{-1}_{n-1}(\mathbf{X}_j - \bar{\mathbf{X}})$ todavía no se aleja mucho de χ^2_p para todo $j = 1, \cdots, n$. Entonces en la práctica con los datos observados podemos pensar que las distancias

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}_{n-1}(\mathbf{x}_j - \bar{\mathbf{x}})$$

son realizaciones de una variable χ^2_p. Por lo tanto podemos graficar el QQ plot para el conjunto de distancias $d_1^2, \cdots, d_n^2$, es decir, creamos las distancias ordenadas

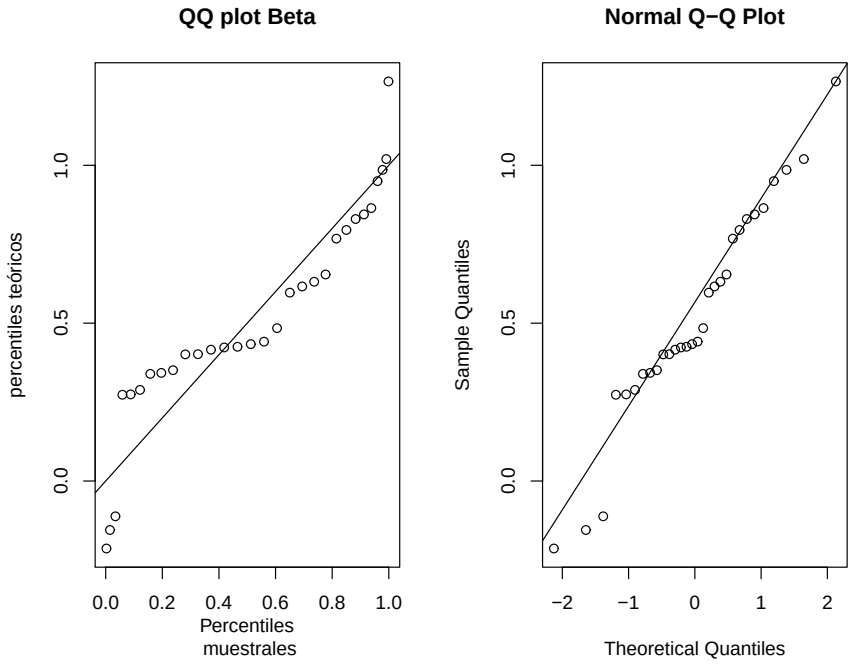

Figura B.12: *Gráfica QQ plot para verificar la distribución Beta y la distribución normal a 30 datos simulados de la distribución* $N(0.5, 0.3^2)$.

$d^2_{(1)}, \cdots, d^2_{(n)}$, y las comparamos con los percentiles $\chi^2_{p,\frac{j-0.5}{n}}$ con $j = 1, \cdots, n$. Si la distribución normal multivariante es apropiada para los datos, entonces $d^2_{(j)} \approx \chi^2_{p,\frac{j-0.5}{n}}$, y la nube de puntos debe estar cercana a una recta de 45° sin intercepto. Esta grafica se conoce también como el plot Ji-cuadrado.

Aplicamos esta herramienta gráfica a los datos de la Tabla 6.3 que corresponden a nivel de colesterol de 20 pacientes antes y después de un tratamiento. Para verificar que los datos pueden ser modelados por una distribución normal bivariante, utilizamos la función `chi.plot` de la siguiente forma

```
> chi.plot<-function(x){
+ x<-data.frame(x)
+ n<-dim(x)[1]
+ p<-dim(x)[2]
+ d2<-rep(NA,n)
+ bar<-mean(x)
+ S<-var(x)
+ for(i in 1:n){
+ d2[i]<-mahalanobis(x[i,],bar,S)
```

```
+ }
+ j<-c(1:n)
+ percen<-qchisq((j-0.5)/n,p)
+ plot(percen,sort(d2),main="plot Ji-cuadrado")
+ abline(0,1)
+ }
>
> ante<-c(230,245,220,250, 260,250,220,300,310,290,260,240,210,220,
+ 250,245,274,230,285,275)
> desp<-c(210,230,215,220,240,220,210,260,280,270,230,235,200,200,
+ 210,230,250,210,260,230)
> X<-data.frame(cbind(ante,desp))
> chi.plot(X)
```

La gráfica resultante se encuentra en la Figura B.13, donde podemos ver que excepto por los dos últimos puntos, los datos parece que pueden ser bien descritos por la distribución normal bivariante.

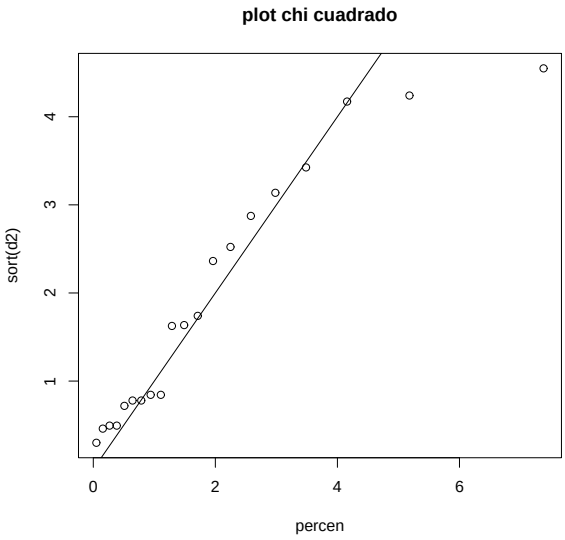

Figura B.13: *Gráfica QQ plot para verificar la distribución normal bivariante para los datos de la Tabla 6.3.*

Pruebas de bondad de ajuste

En este apartado consideramos varias pruebas estadísticas de uso común para verificar si cierta distribución puede ser apropiada para describir un conjunto de datos.

Prueba de normalidad de Shapiro-Wilk

Shapiro & Wilk (1965) propusieron una prueba estadística para verificar que la distribución normal describe bien a un conjunto de datos $x_1, \cdots, x_n$. El sistema de hipótesis de interés es

H_0: La distribución normal es adecuada para los datos

vs.

H_1: La distribución normal no es adecuada para los datos

Para este sistema, la estadística de prueba se define como

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

donde la constante a se define como

$$(a_1, \cdots, a_n) = \frac{\mathbf{m}'\mathbf{V}^{-1}}{\mathbf{m}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m}}^{1/2}$$

donde los componentes de $\mathbf{m}$: $m_1, \cdots, m_n$ son las esperanzas de las estadísticas de una muestra aleatoria con distribución normal estándar[2] y $\mathbf{V}$ es la matriz de varianzas y covarianzas de las n estadísticas de orden.

Para completar la regla de decisión, se necesita la distribución nula de W, que en este caso es muy complicado hallarla, y por consiguiente se halla la distribución nula asintótica de W. En Shapiro & Wilk (1965) se encuentran los percentiles de la distribución nula de W con diferentes valores de n que son obtenidos simulando repetidamente muestras de la distribución normal. Y en la práctica se rechaza H_0 para valores pequeños[3] de W.

En R, la prueba de Shapiro y Wilk se lleva a cabo usando la función `shapiro.test`. Ilustramos el uso de ésta con los datos de grosor de láminas de vidrios del Ejemplo 2.3.6. El código es como sigue

```
> vidrio<-c(3.56, 3.36, 2.99, 2.71, 3.31,3.68, 2.78, 2.95, 2.82,
+ 3.45, 3.42 ,3.15)
> shapiro.test(vidrio)

        Shapiro-Wilk normality test

data:  vidrio
W = 0.942, p-value = 0.5249
```

[2]Para calcular estas esperanzas, se debe calcular la función de densidad de las estadísticas de orden. El lector puede consultar Mayorga (2004, p. 27).

[3]Shapiro & Wilk (1965) mostraron que el valor máximo de W es 1, entonces en la práctica un valor cercano de W a 1 indica que la distribución normal es adecuada para los datos.

Tenemos que para este conjunto de datos, la estadística W tomó el valor de 0.942, lo cual es muy cercano al valor máximo 1, indicando que la distribución normal puede ser apropiada para los datos. Observando el p-valor, éste conduce a la misma decisión.

En la literatura estadística, existen numerosas pruebas estadísticas para verificar la normalidad de los datos, entre las más conocidas, la prueba de Anderson Darling, que es una prueba no paramétrica, y la prueba de Jarque Bera, que utiliza las propiedades acerca del tercer y cuarto momentos de una distribución normal.

Prueba de Kolmogorov Smirnov

La prueba de Kolmogorov Smirnov es una prueba que sirve para evaluar el ajuste de cualquier distribución a un conjunto de datos, y consiste en comparar la función de distribución de la distribución en referencia con la función de distribución empírica de los datos dada en (B.0.1). El sistema de hipótesis está dado por

$$H_0\text{: La distribución } F \text{ es adecuada para los datos}$$

vs.

$$H_1\text{: La distribución } F \text{ no es adecuada para los datos}$$

Y la estadística de prueba mide la diferencia máxima entre $F(x)$ y la distribución empírica $\hat{F}_n(x)$ y se define como

$$D = \sup_x |F(x) - \hat{F}_n(x)|$$

Dada la definición de D, se rechaza H_0 para valores grandes de D. Sin embargo, el desarrollo de la distribución nula de esta estadística no es fácil, y aquí no haremos los desarrollos teóricos. En la práctica calculamos el valor de la estadística D y lo comparamos con los percentiles provistos por Miller (1956) y podemos tomar una decisión sobre el sistema de hipótesis de interés.

En R esta prueba de hipótesis se lleva a cabo usando la función `ks.test`, donde debemos especificar la distribución de referencia F con opciones como `pgamma`, `pnorm`, pero teniendo en cuenta que se debe especificar los valores de los posibles parámetros de la distribución de referencia. En la mayoría de los casos, en la práctica, desconocemos los valores de los parámetros, y por consiguiente primero estimamos los parámetros usando los datos, y luego miramos si la distribución objetiva con estos parámetros estimados se ajusta bien a los datos o no.

Consideramos la aplicación de esta prueba a varios ejemplos considerados en el libro. Primero tomamos los datos del Ejemplo 2.3.6. Como la distribución de referencia es la distribución normal, entonces primero estimamos los parámetros μ y σ, y luego miramos si la distribución normal con estos parámetros es apropiada para los datos. El código utilizado es como sigue:

```
> vidrio<-c(3.56, 3.36, 2.99, 2.71, 3.31,3.68, 2.78, 2.95, 2.82,
+ 3.45, 3.42 ,3.15)
> mu=mean(vidrio)
> sd=sd(vidrio)
> mu
[1] 3.181667
> sd
[1] 0.3267702
> ks.test(vidrio,"pnorm",mu,sd)

        One-sample Kolmogorov-Smirnov test

data:  vidrio
D = 0.1527, p-value = 0.9029
alternative hypothesis: two-sided
```

Podemos ver que el valor de la estadística D es pequeño, y por consiguiente la distribución $N(3.18, 0.33^2)$ es adecuada para describir los datos.

Ahora consideramos los datos del Ejemplo 2.3.16 que corresponden a porcentajes, y en este ejemplo se creyó que la distribución Beta es apropiada para los datos. Utilizando el siguiente código, podemos ver que esta afirmación parece ser correcta.

```
> x<-c(0.7, 1.4, 19.7, 0.1, 12.4, 1.1, 0.5, 18.9, 5.0, 0.3,
+ 0.6, 5.4, 6.7, 0.9)/100
> n<-length(x)
> va<-var(x)*(n-1)/n
> bar<-mean(x)
> a<-bar^2*(1-bar)/va-bar
> a
[1] 0.5447021
> b<-(1-bar)*(bar*(1-bar)/va-1)
> b
[1] 9.80242
>  ks.test(x,"pbeta",a,b)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.2105, p-value = 0.4988
alternative hypothesis: two-sided
```

También podemos utilizar `ks.test` para ver el ajuste de una distribución discreta como la distribución Poisson. Utilizamos los datos del Ejemplo 2.3.2 para ilustrar este caso. Tenemos el siguiente código.

```
> x<-c(1, 1, 5, 5, 2, 3, 3, 6, 4, 3, 2, 3, 2, 3 ,4)
```

```
> mean(x)
[1] 3.133333
> ks.test(x,"ppois",mean(x))

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.2841, p-value = 0.1776
alternative hypothesis: two-sided
```

Con respecto a la anterior salida de R, podemos ver en primer lugar que la distribución $Pois(3.13)$ parece ser apropiada para los datos. Sin embargo, nos dice que debido a la presencia de datos empatados, no fue posible calcular el p-valor exacto. En este caso el valor de p-valor 0.1776 fue calculado usando la distribución asintótica de D, y por consiguiente puede no ser tan confiable en muestras pequeñas.

Prueba de Mardia

En una distribución normal, el coeficiente de simetría debe ser 0 y el curtosis debe ser 3, así que comparar estas cantidades con las estimaciones muestrales del coeficiente de simetría y curtosis también puede ser un criterio para ver si la distribución normal es apropiada para los datos. La prueba de normalidad multivariante de Mardia sigue este mismo razonamiento. Mardia (1970) define el coeficiente de simetría y curtosis multivariado de un vector aleatorio $\mathbf{X}$ de dimensión p con media $\boldsymbol{\mu}$ y la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ como

$$\boldsymbol{\beta}_{1p} = E\left\{\left[(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right]^3\right\}$$

y

$$\boldsymbol{\beta}_{2p} = E\left\{\left[(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right]^2\right\}.$$

Cuando la distribución de $\mathbf{X}$ es normal, debe suceder que $\boldsymbol{\beta}_{1p} = 0$ y $\boldsymbol{\beta}_{2p} = p(p+2)$.

En una muestra observada de vectores aleatorios de tamaño n, se puede estimar $\boldsymbol{\beta}_{1p}$ y $\boldsymbol{\beta}_{2p}$ con b_{1p} y b_{2p} definidos como

$$b_{1p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(\mathbf{x}_i-\bar{\mathbf{x}})'\mathbf{S}_n^{-1}(\mathbf{x}_j-\bar{\mathbf{x}})\right]^3$$

y

$$b_{2p} = \frac{1}{n}\sum_{i=1}^{n}\left[(\mathbf{x}_i-\bar{\mathbf{x}})'\mathbf{S}_n^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})\right]^2$$

Definidas b_{1p} y b_{2p} de esta forma, podemos ver que valores grandes de b_{1p} indican que el coeficiente de simetría teórico podría ser diferente de 0 y valores grandes de b_{2p} indican que el curtosis teórico puede ser diferente de $p(p+1)$.

El sistema de hipótesis de interés es

H_0: La distribución normal multivariante es adecuada para los datos

vs.

H_1: La distribución normal multivariante no es adecuada para los datos

Mardia (1970) demostró que bajo H_0, $B_1 = \frac{nb_{1p}}{6} \sim \chi^2_v$ con $v = p(p+1)(p+2)/6$ y $B_2 = \frac{b_{2p}-p(p+2)}{\sqrt{8p(p+2)/n}} \sim N(0,1)$. De esta forma, podemos afirmar que el coeficiente de simetría teórico es significativamente diferente de 0 si $B_1 > \chi^2_{v,1-\alpha}$ y que el curtosis es significativamente diferente de $p(p+1)$ si $|B_2| > z_{1-\alpha/2}$. Y los p-valores se pueden calcular como $1 - F_{\chi^2_v}(B_1)$ y $2(1-\Phi(|B_2|))$, respectivamente.

Aplicamos el anterior procedimiento a los datos de la Tabla 6.3 mediante la función `mardia` descrita a continuación.

```
> mardia<-function(x){
+ x<-data.frame(x)
+ n<-dim(x)[1]
+ p<-dim(x)[2]
+ bar<-mean(x)
+ S<-matrix(var(x)*(n-1)/n,2,2)
+ b1.ma<-matrix(NA,n,n)
+ b2.ma<-rep(NA,n)
+ for(i in 1:n){
+ for(j in 1:n){
+ b1.ma[i,j]<-(sum((x[i,]-bar)*(solve(S)%*%t(x[j,]-bar))))^3
+ }
+ b2.ma[i]<-(sum((x[i,]-bar)*(solve(S)%*%t(x[i,]-bar))))^2
+ }
+ b1<-sum(b1.ma)/(n^2)
+ b2<-sum(b2.ma)/n
+ B1<-n*b1/6
+ B2<-(b2-p*(p+2))/(sqrt(8*p*(p+2)/n))
+ v<-p*(p+1)*(p+2)/6
+ p.val1<-pchisq(B1,v,lower.tail=F)
+ p.val2<-2*pnorm(abs(B2),lower.tail=F)
+ list("B1"=B1,"pval1"=p.val1,"B2"=B2,"pval2"=p.val2)
+
+ }
>
> ante<-c(230,245,220,250, 260,250,220,300,310,290,260,240,210,220,
+ 250,245,274,230,285,275)
> desp<-c(210,230,215,220,240,220,210,260,280,270,230,235,200,200,
+ 210,230,250,210,260,230)
```

```
> x<-data.frame(cbind(ante,desp))
>
> mardia(x)
$B1
[1] 2.260101

$pval1
[1] 0.6880424

$B2
[1] -1.060516

$pval2
[1] 0.2889099
```

Podemos ver que ambos p-valores son grandes comparados con el nivel de significación, de donde podemos concluir que se puede asumir que el coeficiente de simetría es 0 y la curtosis es $p(p+1)$, y por consiguiente se acepta la distribución normal bivariante como una distribución apropiada para los datos.

Cabe resaltar que con que uno de los dos p-valores sea muy pequeño, ya se tiene un indicio de que la distribución normal no es adecuada para los datos, sea que los datos son no simétricos o tienen una curtosis muy grande o muy pequeña.

Apéndice C

Transformación de Box-Cox

En muchas técnicas de análisis estadístico de datos, se asume que un conjunto de datos proviene de una distribución normal con esperanza y varianza constante. Sin embargo, cuando enfrentamos datos que son altamente no simétricos[1] o con varianzas no constantes, este supuesto ya no es válido y los resultados de análisis pueden dejar de ser válidos.

Definición de la transformación Box-Cox

Box & Cox (1964) introdujeron una familia de transformaciones que puede inducir la distribución normal a un conjunto de datos. Suponga que el conjunto de datos se denota por $x_1, \cdots, x_n$, entonces podemos crear un nuevo conjunto de datos $x^{(\lambda)}$ definidos por

$$x^{(\lambda)} = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln x & \text{si } \lambda = 0 \end{cases} \tag{C.0.1}$$

La anterior transformación es válida solo para los datos positivos, pues la función logaritmo solo se define para estos valores. Sin embargo, si se presenta alguno o algunos datos negativos, podemos primero sumar una constante a todos los datos para garantizar que sean todos positivos. Por esta razón, asumimos sin pérdida de generalidad que todos los datos son positivos y aplicamos la transformación (C.0.1).

Casos particulares de la transformación Box-Cox

La transformación (C.0.1) es toda una familia de transformaciones según cambia el valor de λ, y por consiguiente, muchas transformaciones comunes en la práctica son

[1] Como las variables ingreso, gasto de personas o familias.

simplemente casos particulares de la transformación Box Cox. A continuación presentamos algunas de ellas.

Transformación logarítmica

La transformación logarítmica es una de las transformaciones más comunes en el área de la estadística llamada series de tiempo, que estudia la evolución de datos medidos a través del tiempo. La teoría básica de series de tiempo requiere que la variación de los datos no cambie a través del tiempo; sin embargo, algunos de estos datos presentan una varianza no constante. Un ejemplo típico son los datos de pasajeros que se pueden consultar con el comando `AirPassengers` en R. Estos datos registran un número total de pasajeros en aerolíneas internacionales, medidos mensualmente entre 1949 y 1960. Podemos visualizar estos datos en la parte (a) de la Figura C. 1, donde se observa que la variación de los datos cambia a través del tiempo de forma regular y creciente. En la parte (b) de la misma gráfica, observamos los datos transformados mediante la función logarítmica, podemos ver claramente que la varianza se logra estabilizar con esta transformación.

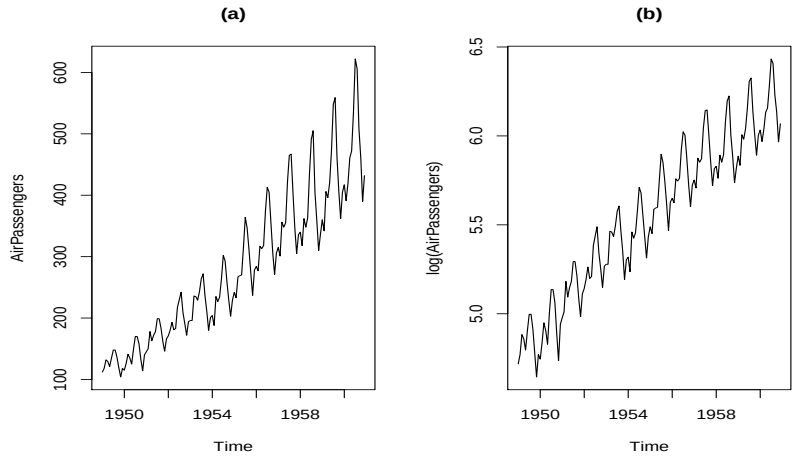

Figura C.1: *Datos del número total de pasajeros antes y después de aplicar la transformación logarítmica.*

Transformación raíz cuadrada

En (C.0.1), cuando $\lambda = 0.5$ tenemos el caso equivalente a una transformación raíz cuadrada. Esta transformación es común para datos que son enteros y que corresponden a conteo, pues en primer lugar, los datos de conteo son vistos como realización de una variable discreta y por consiguiente no es apropiado considerar una distribución

normal, mientras que al tomar la raíz cuadrada, los datos transformados ya no son enteros, y por ende pueden verse como una realización de una variable continua.

En la página de la Universidad de Delaware de los Estados Unidos[2], se muestra un conjunto de datos que describen el número de un tipo de peces en los ríos de Maryland. Estos datos son 38, 1, 13, 2, 13, 20, 50, 9, 28, 6, 4 y 43. Es claro que no es apropiado ver a estos datos como realizaciones de una distribución normal. Al tomar la raíz cuadrada de estos datos, se convierten en 6.16, 1.00, 3.61, 1.41, 3.61, 4.47, 7.07, 3.00, 5.29, 2.45, 2.00 y 6.56. Podemos aplicar la prueba de Shapiro Wilk y la prueba de Kolmogorov Smirnov para ver el efecto de la transformación para lograr la normalidad. Los resultados se muestran en la Tabla C.1., donde podemos observar con la prueba de Shapiro Wilk, que la estadística W es mayor para los datos transformados, indicando que hay mayor evidencia de los datos acerca de la distribución normal. Por otro lado, la estadística D de Kolmogorov Smirnov es menor para los datos transformados, indicando que en este caso, la función de distribución empírica es más cercana a la función de distribución normal.

Prueba	Datos	Estadística	p-valor
Shapiro Wilk	Datos originales	$W = 0.8874$	0.1091
	Datos transformados	$W = 0.9507$	0.6479
Kolmogorov Smirnov	Datos originales	$D = 0.2198$	0.6077
	Datos transformados	$D = 0.138$	0.9763

Tabla C.1: *Prueba de Shapiro Wilk y Kolmogorov Smirnov aplicadas a un conjunto de datos de conteo y datos transformados mediante la transformación raíz cuadrática.*

Estimación de máxima verosimilitud de λ

Es claro que la transformación de Box Cox depende del valor de λ, entonces una pregunta natural es ¿teniendo un conjunto de datos, cuál es el valor de λ que se debe utilizar? Una forma de responder esta pregunta es utilizando la estimación de máxima verosimilitud, es decir, al la función de verosimilitud de los datos originales x_1, $\cdots$, x_n como función de λ y se busca el valor de λ que maximiza esta función. Utilizando el teorema de transformación, tenemos que

$$\begin{aligned} L(x_1, \cdots, x_n, \lambda) = \prod_{i=1}^{n} f_{X_i}(x_i, \lambda) &= \prod_{i=1}^{n} f_{X_i^{(\lambda)}}(x_i^{(\lambda)}) \left| \frac{\partial x_i^{(\lambda)}}{\partial x_i} \right| \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x_i^{(\lambda)} - \mu)^2 \right\} x_i^{\lambda-1} \\ &= (2\pi\sigma^2)^{-n/2} \prod_{i=1}^{n} x_i^{\lambda-1} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i^{(\lambda)} - \mu)^2 \right\} \end{aligned}$$

[2]http://udel.edu/ mcdonald/stattransform.html

Tomando la función logaritmo, tenemos que

$$\begin{aligned}\ln L &= -\frac{n}{2}\ln(2\pi\sigma^2) + (\lambda-1)\sum_{i=1}^{n}\ln x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^{(\lambda)}-\mu)^2 \\ &= -\frac{n}{2}\ln(2\pi\sigma^2) + (\lambda-1)\sum_{i=1}^{n}\ln x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{x_i^{\lambda}-1}{\lambda}-\mu\right)^2\end{aligned}$$

Se debe tener en cuenta que esta función depende de μ, σ^2 y λ. Entonces primero reemplazamos μ y σ^2 por su estimación de máxima verosimilitud dadas por

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(\lambda)} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i^{\lambda}-1}{\lambda}$$

y

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^{(\lambda)}-\hat{\mu}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i^{\lambda}-1}{\lambda} - \frac{1}{n}\sum_{i=1}^{n}\frac{x_i^{\lambda}-1}{\lambda}\right)^2.$$

Y tenemos que

$$\ln L(\hat{\mu},\hat{\sigma}^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + (\lambda-1)\sum_{i=1}^{n}\ln x_i - \frac{n}{2}$$

De esta forma, la estimación de máxima verosimilitud de λ consiste en aquel valor de λ que maximiza la expresión

$$\begin{aligned}&-\frac{n}{2}\ln\hat{\sigma}^2 + (\lambda-1)\sum_{i=1}^{n}\ln x_i \\ &= -\frac{n}{2}\ln\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i^{\lambda}-1}{\lambda} - \frac{1}{n}\sum_{i=1}^{n}\frac{x_i^{\lambda}-1}{\lambda}\right)^2\right) + (\lambda-1)\sum_{i=1}^{n}\ln x_i\end{aligned}$$

Dado un conjunto de datos, la siguiente función `box.cox` calcula el valor de λ entre 0 y 100 que maximiza la anterior expresión, y una vez encontrada $\hat{\lambda}_{MV}$, transforma los datos de acuerdo a (C.0.1).

```
> box.cox<-function(x){
+ if(prod(x<=0)!=0){
+ y<-x-min(x)+1e-10
+ }
+ if(prod(x<=0)==0){y<-x}
+ n<-length(y)
+
+       L<-function(y,lambda){
```

```
+       if(lambda==0){
+       y.lam<-log(y) }
+       if(lambda!=0){
+       y.lam<-(y^lambda-1)/lambda  }
+
+       mu<-mean(y.lam)
+       sigma2<-var(y.lam)*(n-1)/n
+       L.lam<-(lambda-1)*sum(log(y))-log(sigma2)*n/2
+
+       list("L.lam"=L.lam,"xtrans"=y.lam)
+       }
+
+ lam<-seq(0,100,0.01)
+ L.val<-rep(NA,length(lam))
+ for(i in 1:length(lam)){
+ L.val[i]<-L(y,lam[i])$L.lam
+ }
+ lam.MV=lam[which(L.val==max(L.val))]
+ datos.tranf<-L(y,lam.MV)$xtrans
+ list("lambda"=lam.MV,"xtrans"=datos.tranf)
+ }
```

Para ilustrar el uso de la anterior función, simulamos un conjunto de datos de la distribución $Gamma(2, 0.2)$ que presenta un alto grado de asimetría. El código utilizado es

```
> set.seed(12345)
> a<-rgamma(50,2,5)
> transf<-box.cox(a)
> transf$lambda
[1] 0.32
> datos.trans<-transf$xtrans
> shapiro.test(a)
Shapiro-Wilk normality test

data:  a
W = 0.9184, p-value = 0.002064

> shapiro.test(datos.trans)

        Shapiro-Wilk normality test

data:  datos.trans
W = 0.9828, p-value = 0.6729

> ks.test(a,"pnorm",mean(a),sd(a))
```

```
        One-sample Kolmogorov-Smirnov test

data:  a
D = 0.1592, p-value = 0.1422
alternative hypothesis: two-sided


> ks.test(datos.trans,"pnorm",mean(datos.trans),sd(datos.trans))

        One-sample Kolmogorov-Smirnov test

data:  datos.trans
D = 0.0808, p-value = 0.8738
alternative hypothesis: two-sided
```

Podemos ver que la estimación de máxima verosimilitud de λ está dada por 0.32, y también podemos ver que con la transformación (C.0.1) con $\lambda = 0.32$, los datos transformados adquieren características de una distribución normal.

Apéndice D

Repaso matricial

En este parte, repasamos algunos de los conceptos acerca de matrices y vectores que se han utilizado en este libro.

Matriz y vector

Una matriz es simplemente un arreglo bidimensional de números que son las entradas de la matriz, estas dos dimensiones se conocen como filas y columnas. Por ejemplo, la matriz

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & -1 & -5 \end{bmatrix}$$

Esta matriz tiene 6 entradas, que se organizan en dos filas (cada una de 3 elementos) y 3 columnas (cada una de 2 elementos), y decimos que la dimensión de la matriz $\mathbf{A}$ es de 2×3. El número de filas y de columnas no son necesariamente iguales; en caso de que sean, se dice que la matriz es cuadrada. Por ejemplo

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & -1 & -5 \\ 0 & 7 & 3 \end{bmatrix}$$

es una matriz cuadrada de dimensión 3×3. Una matriz cuadrada muy particular es la matriz identidad $\mathbf{I}$ cuyos elementos en la diagonal son 1 y los elementos fuera de la diagonal son 0.

Un caso particular de las matrices son los vectores. Cuando el número de filas de una matriz es 1, la matriz es un vector fila, y su dimensión será de $1 \times q$, donde q denota el número de columnas. Cuando el número de columnas de una matriz es 1, la matriz es un vector columna, y su dimensión será de $p \times 1$ donde p denota el número

de filas. Así, un vector fila es de la forma $\mathbf{x} = [3, 4, -6]$ y un vector columna

$$\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

Con respecto a las matrices y los vectores, hay muchas propiedades y para su estudio detallado se necesita bastante tiempo, aquí solo hacemos repaso de las más básicas, además de las mencionadas a lo largo del texto.

Suma y producto entre matrices

La suma entre dos matrices $\mathbf{A}$ y $\mathbf{B}$ sólo es posible si las dos matrices tienen la misma dimensión, en este caso, la matriz $\mathbf{A} + \mathbf{B}$ tiene la misma dimensión y cada elemento de $\mathbf{A} + \mathbf{B}$ es la suma de los correspondientes elementos de $\mathbf{A}$ y $\mathbf{B}$, respectivamente. De esta forma

$$\begin{bmatrix} 1 & 2 & 0 \\ 2 & -1 & -5 \end{bmatrix} + \begin{bmatrix} 2 & 0 & -1 \\ -2 & -4 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 2 & -1 \\ 0 & -5 & -3 \end{bmatrix}$$

Como los vectores son caso particular de las matrices, entonces también la suma de vectores se rige por la anterior regla y solo podemos sumar vectores filas de la misma dimensión o vectores columnas de la misma dimensión. Y las reglas comutativas y asociativas son válidas para suma de matrices.

Dados dos vectores $\mathbf{x}$ y $\mathbf{y}$ con el mismo número de entradas, se define el producto punto como la suma de los productos de sus componentes. De esta forma

$$\begin{bmatrix} 1 & 2 & 0 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ -1 \end{bmatrix} = 1 * 3 + 2 * 1 + 0 * (-1) = 5$$

El producto entre dos matrices $\mathbf{A}$ y $\mathbf{B}$ dado por $\mathbf{AB}$ sólo es posible cuando el número de columnas de la matriz $\mathbf{A}$ es igual que el número de filas de $\mathbf{B}$. En este caso, el elemento en la fila i y columna j de $\mathbf{AB}$ será el producto punto entre la fila i de la matriz $\mathbf{A}$ y la columna j de la matriz $\mathbf{B}$. Por ejemplo

$$\begin{bmatrix} 1 & 2 & 0 \\ 2 & -1 & -5 \end{bmatrix} \begin{bmatrix} 4 & 2 & 0 & 2 \\ 2 & -1 & -5 & 0 \\ -2 & 1 & 3 & -2 \end{bmatrix} = \begin{bmatrix} 8 & 0 & -10 & 2 \\ 16 & 0 & -10 & 14 \end{bmatrix}$$

Las reglas asociativas y distributivas son válidas para producto entre dos matrices, pero la ley comutativa en general no es válida, es decir, $\mathbf{AB} \neq \mathbf{BA}$.

Transpuesta de una matriz

El operador transpuesto transforma una matriz de dimensión $p \times q$ a una matriz $q \times p$ invirtiendo las filas y las columnas, es decir, la fila i de la matriz original será la columna

i de la matriz transformada. Se acostumbra denotar la transpuesta de una matriz $\mathbf{A}$ con $\mathbf{A}^t$ o $\mathbf{A}'$, por ejemplo, si

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & -1 & -5 \end{bmatrix}$$

entonces

$$\mathbf{A}' = \begin{bmatrix} 1 & 2 \\ 2 & -1 \\ 0 & -5 \end{bmatrix}$$

Dada la definición de vector, podemos ver que la transpuesta de un vector fila es un vector columna, y viceversa. Con respecto a la operadora transpuesta, tenemos las siguientes propiedades

- $(\mathbf{A}')' = \mathbf{A}$
- $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$
- $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$

Hay una clase muy importante de matrices que se define en términos de la operadora transpuesta, decimos que una matriz $\mathbf{A}$ es simétrica si $\mathbf{A} = \mathbf{A}'$.

Determinante

El determinante es una propiedad asociada únicamente a las matrices diagonales, y se puede definir de muchas formas. Una forma es definirlo en términos de los *menores* de la matriz, y de forma recursiva, es decir, se define el determinante de una matriz de dimensión 1×1, y luego se define el determinante de una matriz de $n \times n$ en términos del determinante de matrices de $(n-1) \times (n-1)$. De esta forma, la definición del determinante se da en dos partes

- El determinante de una matriz de dimensión 1×1, $\mathbf{A} = [a]$ es simplemente la entrada a.
- Para una matriz cuadrada $\mathbf{A}$ de dimensión $n \times n$ se define el menor del elemento a_{ij} (el elemento en la fila i y columna j de $\mathbf{A}$) m_{ij} como el determinante de la matriz obtenida después de eliminar la fila i y la columna j. Y el determinante de $\mathbf{A}$ denotada como $|\mathbf{A}|$ se calcula como

  $$|\mathbf{A}| = \sum_{j=1}^{n} a_{ij}(-1)^{i+j} m_{ij}$$

  para cualquier fila i.

Y tenemos las siguientes propiedades básicas acerca del determinante

- $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$
- $|\mathbf{A}'| = |\mathbf{A}|$
- $|c\mathbf{A}| = c^n|\mathbf{A}|$ para toda constante c

Inversa

Una matriz cuadrada $\mathbf{A}$ tiene matriz inversa o equivalente $\mathbf{A}$ es invertible si existe una matriz cuadrada de la misma dimensión denotada por $\mathbf{A}^{-1}$ que satisface $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, donde $\mathbf{I}$ es la matriz identidad. Si una matriz $\mathbf{A}$ cuya transpuesta es igual a su inversa, entonces decimos que $\mathbf{A}$ es una matriz ortogonal, es decir, $\mathbf{A}\mathbf{A}' = \mathbf{I}$.

Tenemos las siguientes propiedades básicas acerca de una matriz inversa

- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ cuando las inversas existen
- $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$
- $\mathbf{A}$ es invertible si y solo si su determinante es diferente de 0
- $|\mathbf{A}|^{-1} = |\mathbf{A}^{-1}|$

Traza

La traza también se define para las matrices cuadradas, y se calcula simplemente como la suma de los elementos en la diagonal de la matriz cuadrada, y denotamos la traza de la matriz $\mathbf{A}$ como $tr(\mathbf{A})$.

Algunas propiedades con respecto a la traza son

- $tr(\mathbf{A} + \mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B})$
- $tr(c\mathbf{A}) = ctr(\mathbf{A})$ para toda constante c
- $tr(\mathbf{A}) = tr(\mathbf{A}')$
- $tr(\mathbf{AB}) = tr(\mathbf{BA})$

Valores y vectores propios

Dada una matriz cuadrada $\mathbf{A}$, si $\mathbf{v}$ es un vector diferente del vector $\mathbf{0}$ que satisface $\mathbf{Av} = \lambda\mathbf{v}$ para alguna constante λ, entonces decimos que $\mathbf{v}$ es un vector propio de $\mathbf{A}$ y λ es el valor propio asociado a $\mathbf{v}$. Los valores y vectores propios juegan un papel muy importante en la teoría con respecto a las matrices, ya que muchas de las anteriores propiedades se relacionan con estos conceptos. Tenemos las siguientes propiedades

- El determinante de una matriz es el producto de todos sus valores propios, de allí, vemos que una matriz es invertible si y solo si 0 no es un valor propio.
- La traza de una matriz es la suma de todos sus valores propios.
- Si λ es un valor propio de una matriz invertible $\mathbf{A}$, entonces $1/\lambda$ es un valor propio de $\mathbf{A}^{-1}$.

Formas cuadráticas y matices semidefinidas

En términos no muy complicados, una forma cuadrática se refiere al producto matricial de la forma $\mathbf{x}'\mathbf{Ax}$ donde $\mathbf{x}$ es un vector columna de dimensión n y $\mathbf{A}$ es una matriz cuadrada de dimensión $n \times n$. Dadas las dimensiones de $\mathbf{x}$ y $\mathbf{A}$, podemos ver que una forma cuadrática da como resultado un escalar.

Usando la definición de forma cuadrática, se definen matrices semidefinidas positivas y semidefinidas negativas. Decimos que $\mathbf{A}$ es semidefinida positiva si para todo vector $\mathbf{x}$ se tiene que $\mathbf{x}'\mathbf{Ax} \geq 0$, y es semidefinida negativa si para todo vector $\mathbf{x}$ se tiene que $\mathbf{x}'\mathbf{Ax} \leq 0$.

En la práctica es difícil verificar que una matriz es semidefinida positiva o semidefinida negativa usando directamente la definición. Como una alternativa, podemos utilizar los valores propios. Si los valores propios de una matriz son todos no negativos, entonces la matriz es semidefinida positiva; mientras que si los valores propios de una matriz son todos no positivos, entonces la matriz es semidefinida negativa.

Descomposición espectral y raíz cuadrada de una matriz

Dada una matriz cuadrada simétrica semidefinida positiva $\mathbf{A}$, se define la raíz cuadrada de $\mathbf{A}$ como $\mathbf{A}^{1/2}$ que satisface $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$. Cuando la matriz $\mathbf{A}$ es una matriz diagonal, entonces la raíz cuadrada de $\mathbf{A}$ se obtiene simplemente calculando la raíz de cada elemento en la diagonal. Por ejemplo

$$\begin{bmatrix} 4 & 0 \\ 0 & 9 \end{bmatrix}^{1/2} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

Sin embargo, en general los elementos de $\mathbf{A}^{1/2}$ no corresponden a la raíz de los elementos de $\mathbf{A}$. Su cálculo requiere de un resultado de descomposición llamada la descomposición espectral. La descomposición espectral afirma que para una matriz simétrica $\mathbf{A}$ se puede descomponer de la forma $\mathbf{A} = \mathbf{QDQ}'$ donde $\mathbf{D}$ es una matriz diagonal que contiene los valores propios de $\mathbf{A}$ y $\mathbf{Q}$ es una matriz ortogonal que contiene los vectores propios de $\mathbf{A}$. De esta forma, tenemos que

$$\mathbf{A} = \mathbf{QDQ}' = \mathbf{A} = \mathbf{QD}^{1/2}\mathbf{D}^{1/2}\mathbf{Q}' = \underbrace{\left(\mathbf{QD}^{1/2}\mathbf{Q}'\right)}_{\mathbf{A}^{1/2}}\underbrace{\left(\mathbf{QD}^{1/2}\mathbf{Q}'\right)}_{\mathbf{A}^{1/2}}$$

De donde calculamos la raíz cuadrada de $\mathbf{A}$ como $\mathbf{A}^{1/2} = \mathbf{QD}^{1/2}\mathbf{Q}'$.

Matriz particionada

Cuando tenemos una matriz $\mathbf{A}$, podemos agrupar u organizar los elementos de $\mathbf{A}$ de tal forma que $\mathbf{A}$ puede ser visto como una matriz cuyos elementos son a la vez matrices. Si $\mathbf{A}$ es de dimensión $p \times q$, entonces para $p_1 < p$ y $q_1 < q$, podemos tener la siguiente partición para $\mathbf{A}$,

$$\mathbf{A} = \left[\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1,q_1} & a_{1,q_1+1} & \cdots & a_{1q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{p_1,1} & \cdots & a_{p_1,q_1} & a_{p_1,q_1+1} & \cdots & a_{p_1,q} \\ \hline a_{p_1+1,1} & \cdots & a_{p_1+1,q_1} & a_{p_1+1,q_1+1} & \cdots & a_{p_1+1,q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{p,q_1} & a_{p,q1+1} & \cdots & a_{pq} \end{array}\right] = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

donde $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$ y $\mathbf{A}_{22}$ son matrices de dimensión $p_1 \times q_1$, $p_1 \times (q - q_1)$, $(p - p_1) \times q_1$ y $(p - p_1) \times (q - q_1)$, respectivamente.

Los resultados concernientes a las matrices particionadas que fueron utilizados en este libro hacen referencia a la partición de una matriz cuadrada e invertible. Suponga que $\mathbf{A}$ es una matriz $p \times p$ invertible y simétrica, y se tiene la siguiente partición para $\mathbf{A}$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

donde $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$ y $\mathbf{A}_{22}$ son matrices de dimensión $k \times k$, $k \times (p-k)$, $(p-k) \times k$ y $(p-k) \times (p-k)$, respectivamente, para algún $k < p$, además $\mathbf{A}_{11}$ y $\mathbf{A}_{22}$ son matrices invertibles. Entonces tenemos que

$$|\mathbf{A}| = |\mathbf{A}_{11}||\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{1}2| = |\mathbf{A}_{22}||\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}|$$

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{B}^{-1} & -\mathbf{B}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{pmatrix} \text{ donde } \mathbf{B} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}.$$

Derivadas matriciales

Cuando queremos derivar una cantidad escalar con respecto a un vector, esto puede verse simplemente como el gradiente de una función vectorial, por ejemplo, si deseamos derivar un producto punto entre dos vectores $\mathbf{a}$ y $\mathbf{x}$ con respecto a $\mathbf{x}$, entonces tenemos que

$$\frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \sum_{i=1}^{n} a_i x_i}{\partial (x_1, \cdots, x_n)} = (a_1, \cdots, a_n) = \mathbf{a}$$

De manera análoga, podemos ver que si $\mathbf{A}$ es una matriz cuadrada y $\mathbf{x}$ un vector columna, entonces

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}'$$

y

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

Por otro lado, también podemos derivar una cantidad escalar con respecto a una matriz $\mathbf{A}$ de tamaño $p \times q$. En este caso, la derivada resulta ser una matriz de $q \times p$ donde los elementos corresponden a la derivada de la escalar con respecto a los elementos de $\mathbf{A}'$. Por ejemplo

$$\begin{aligned}
\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{y}}{\partial \mathbf{A}} &= \frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j a_{ij}}{\partial \mathbf{A}} \\
&= \begin{bmatrix} \frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j a_{ij}}{\partial a_{11}} & \cdots & \frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j a_{ij}}{\partial a_{n1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j a_{ij}}{\partial a_{1n}} & \cdots & \frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j a_{ij}}{\partial a_{nn}} \end{bmatrix} \\
&= \begin{bmatrix} x_1 y_1 & \cdots & x_n y_1 \\ \vdots & \ddots & \vdots \\ x_1 y_n & \cdots & x_n y_n \end{bmatrix} \\
&= \mathbf{y}\mathbf{x}'
\end{aligned}$$

De forma análoga, también se puede ver

$$\frac{\partial \mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{y}}{\partial \mathbf{A}} = \left(\mathbf{x}\mathbf{y}' + \mathbf{y}\mathbf{x}'\right)\mathbf{A}'$$

Y tenemos las dos siguientes identidades que fueron utilizadas en la sección 6.2.1. Si $\mathbf{A}$ es simétrica, entonces

$$\frac{\partial |\mathbf{A}|}{\partial \mathbf{A}} = |\mathbf{A}| \left(2\mathbf{A}^{-1} - diag(\mathbf{A}^{-1})\right)$$

$$\frac{\partial tr(\mathbf{A}\mathbf{B}')}{\partial \mathbf{A}} = \mathbf{B} + \mathbf{B}' - diag(\mathbf{B})$$

Ahora suponga que queremos derivar un vector $\mathbf{y}$ de dimensión n con respecto a otro vector $\mathbf{x}$ de la misma dimensión, esta derivada es una matriz de dimensión $n \times n$, y está dada por

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix}$$

De donde podemos ver que $\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}'$.

Apéndice E

Inferencia en tablas de contingencia

La tabla de contingencia se refiere a una presentación de datos correspondientes a dos variables cualitativas, cada una de 2 categorías. Por ejemplo, los datos considerados en la Tabla 4.7, donde se recolecta la información de 124 solicitantes de empleo en una fábrica de placas metálicas acerca de la raza y el resultado de la solicitud de empleo.

Para estudiar si existe una asociación significativa entre el resultado de la solicitud y la raza del solicitante, suponga que los datos en general se pueden organizar como

Raza	Admitido	Rechazado	Total
Blanca	41	39	80
Negra	14	30	44
Total	55	69	124

Tabla E.1: *Los datos de la discriminación racial del Ejemplo 4.5.5.*

En términos genéricos , se tiene lo siguiente:

	Variable 2		
Variable1	Categoría A	Categoría B	Total
Categoría 1	a	b	$a+b$
Categoría 2	c	d	$c+d$
Total	$a+c$	$b+d$	n

Tabla E.2: *Tabla de contingencia* 2×2

Fisher desarrolló una prueba estadística para ver si las dos variables son independientes bajo el supuesto de que los totales por filas $a+b$, $c+d$ y los totales por columnas $a+c$, $b+d$ son fijados de antemano. Y en este caso, la probabilidad de observar los

datos de la Tabla E.2 se puede calcular mediante la distribución hipergeométrica como

$$p = \binom{a+b}{a}\binom{c+d}{c} \Big/ \binom{n}{a+c}$$

Si la hipótesis nula de interés es que las dos variables son independientes, entonces se puede calcular el p-valor como la suma de probabilidades p de todas las tablas iguales o más extremas de lo observado. En R, este p -valor se calcula como la suma de probabilidades de tablas con probabilidad menor o igual a la de la tabla observada, y podemos concluir que las dos variables son dependientes si el p-valor es pequeño comparado con el nivel de significación. El comando en R para este procedimiento es `fisher.test`, cuyo uso se ilustró en el Ejemplo 4.5.5.

Otra herramienta estadística para estudiar la independencia entre dos variables cualitativas es la prueba de Ji-cuadrado de Pearson. A diferencia de la prueba exacta de Fisher que solo se aplica para variables cualitativas con dos categorías, la prueba de Ji-cuadrado de Pearson puede ser utilizada cuando las variables tienen más de dos categorías. Si las dos variables tienen r y s categorías, entonces la estadística de prueba está definida por

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde O_{ij} denota el número de observaciones dentro de la fila i y la columna j; mientras que E_{ij} denota el número esperado de observaciones en la misma celda suponiendo que las variables cualitativas son independientes. Cuando la hipótesis nula no se tiene, se espera que las frecuencias O_{ij} sean muy distintas a las de E_{ij} induciendo un valor grande en la estadística X^2. Esta estadística tiene una distribución nula asintótica de $\chi^2_{(i-1)(j-1)}$, por consiguiente se rechaza la hipótesis de independencia entre las variables si $X^2 > \chi^2_{1-\alpha}$. El comando en R para este procedimiento es `chisq.test`, cuyo uso también se ilustró en el Ejemplo 4.5.5.

Dada que la distribución de la estadística X^2 es asintótica, se espera que para muestras pequeñas, el desempeño de esta prueba sea inferior al de la prueba exacta de Fisher. Simulamos tablas de contingencia donde las filas y las columnas son dependientes, es decir, las probabilidades marginales p_1 y p_2 son diferentes, y en cada tabla aplicamos las dos pruebas para mirar si se rechaza o no la hipótesis nula de independencia entre filas y columnas. Los resultados se muestran en la Tabla E.3, donde n denota los totales marginales por filas. Podemos ver que excepto el caso donde las probabilidades marginales son muy similares, es decir, cuando $p_1 - p_2 = 0.2$, la potencia de la prueba exacta de Fisher es siempre mayor que la prueba Ji-cuadrado, para $n = 5$ ó $n = 10$; mientras que para muestras más grandes, la potencia de estas dos pruebas es muy similar.

En el Ejemplo 6.1.4, se analizaron los datos acerca de la costumbre del consumo de espaguetis en estratos altos, medios y bajos, utilizando el siguiente comando en R,

```
> x<-matrix(c(95,120,35,6,53,97,45,13,33,79,19,3),4,3)
> chisq.test(x)
```

```
        Pearson's Chi-squared test

data:  x
X-squared = 20.7815, df = 6, p-value = 0.002008
```

Podemos ver que al igual que la prueba de razón de verosimilitud, se llega a la conclusión de dependencia entre las filas y las columnas, es decir, la costumbre de consumo de espaguetis no es lo mismo en los diferentes estratos socioeconómicos.

	$n=5$	$n=15$	$n=30$	$n=50$	$n=100$
	Prueba exacta de Fisher				
$p_1-p_2=0.2$	0.0177	0.1570	0.3668	0.6402	0.9372
$p_1-p_2=0.4$	0.1215	0.5360	0.9075	0.9934	1.0000
$p_1-p_2=0.6$	0.3743	0.8961	0.9988	1.0000	1.0000
$p_1-p_2=0.8$	0.7345	0.9988	1.0000	1.0000	1.0000
	Prueba Ji-cuadrado				
$p_1-p_2=0.2$	0.0924	0.1122	0.3668	0.6402	0.935
$p_1-p_2=0.4$	0.0367	0.5153	0.9075	0.9934	1.000
$p_1-p_2=0.6$	0.1044	0.8952	0.9988	1.0000	1.000
$p_1-p_2=0.8$	0.3470	0.9988	1.0000	1.0000	1.000

Tabla E.3: *Comparación de potencia para la prueba exacta de Fisher y la prueba de* χ^2 *en una tabla de contingencia* 2×2

Apéndice F

Tablas de percentiles de distribuciones

p	z_p	p	z_p
0.005	-2.58	0.995	2.58
0.01	-2.32	0.99	2.32
0.025	-1.96	0.975	1.96
0.05	-1.64	0.95	1.64
0.1	-1.28	0.9	1.28

Tabla F.1: *Algunos percentiles de la distribución normal estándar.*

					p					
gl	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	-63.66	-31.82	-12.71	-6.31	-3.08	3.08	6.31	12.71	31.82	63.66
2	-9.92	-6.96	-4.30	-2.92	-1.88	1.88	2.92	4.30	6.96	9.92
3	-5.84	-4.54	-3.18	-2.35	-1.64	1.64	2.35	3.18	4.54	5.84
4	-4.60	-3.75	-2.78	-2.13	-1.53	1.53	2.13	2.78	3.75	4.60
5	-4.03	-3.36	-2.57	-2.01	-1.47	1.47	2.01	2.57	3.36	4.03
6	-3.71	-3.14	-2.45	-1.94	-1.44	1.44	1.94	2.45	3.14	3.71
7	-3.50	-3.00	-2.36	-1.89	-1.41	1.41	1.89	2.36	3.00	3.50
8	-3.36	-2.90	-2.31	-1.86	-1.40	1.40	1.86	2.31	2.90	3.36
9	-3.25	-2.82	-2.26	-1.83	-1.38	1.38	1.83	2.26	2.82	3.25
10	-3.17	-2.76	-2.23	-1.81	-1.37	1.37	1.81	2.23	2.76	3.17
11	-3.11	-2.72	-2.20	-1.80	-1.36	1.36	1.80	2.20	2.72	3.11
12	-3.05	-2.68	-2.18	-1.78	-1.36	1.36	1.78	2.18	2.68	3.05
13	-3.01	-2.65	-2.16	-1.77	-1.35	1.35	1.77	2.16	2.65	3.01
14	-2.98	-2.62	-2.14	-1.76	-1.35	1.35	1.76	2.14	2.62	2.98
15	-2.95	-2.60	-2.13	-1.75	-1.34	1.34	1.75	2.13	2.60	2.95
16	-2.92	-2.58	-2.12	-1.75	-1.34	1.34	1.75	2.12	2.58	2.92
17	-2.90	-2.57	-2.11	-1.74	-1.33	1.33	1.74	2.11	2.57	2.90
18	-2.88	-2.55	-2.10	-1.73	-1.33	1.33	1.73	2.10	2.55	2.88
19	-2.86	-2.54	-2.09	-1.73	-1.33	1.33	1.73	2.09	2.54	2.86
20	-2.84	-2.53	-2.09	-1.72	-1.33	1.33	1.72	2.09	2.53	2.84
21	-2.83	-2.52	-2.08	-1.72	-1.32	1.32	1.72	2.08	2.52	2.83
22	-2.82	-2.51	-2.07	-1.72	-1.32	1.32	1.72	2.07	2.51	2.82
23	-2.81	-2.50	-2.07	-1.71	-1.32	1.32	1.71	2.07	2.50	2.81
24	-2.80	-2.49	-2.06	-1.71	-1.32	1.32	1.71	2.06	2.49	2.80
25	-2.79	-2.49	-2.06	-1.71	-1.32	1.32	1.71	2.06	2.49	2.79
26	-2.78	-2.48	-2.06	-1.71	-1.31	1.31	1.71	2.06	2.48	2.78
27	-2.77	-2.47	-2.05	-1.70	-1.31	1.31	1.70	2.05	2.47	2.77
28	-2.76	-2.47	-2.05	-1.70	-1.31	1.31	1.70	2.05	2.47	2.76
29	-2.76	-2.46	-2.05	-1.70	-1.31	1.31	1.70	2.05	2.46	2.76
30	-2.75	-2.46	-2.04	-1.70	-1.31	1.31	1.70	2.04	2.46	2.75
40	-2.70	-2.42	-2.02	-1.68	-1.30	1.30	1.68	2.02	2.42	2.70
50	-2.68	-2.40	-2.01	-1.68	-1.30	1.30	1.68	2.01	2.40	2.68
60	-2.66	-2.39	-2.00	-1.67	-1.30	1.30	1.67	2.00	2.39	2.66
70	-2.65	-2.38	-1.99	-1.67	-1.29	1.29	1.67	1.99	2.38	2.65
80	-2.64	-2.37	-1.99	-1.66	-1.29	1.29	1.66	1.99	2.37	2.64
90	-2.63	-2.37	-1.99	-1.66	-1.29	1.29	1.66	1.99	2.37	2.63

Tabla F.2: *Algunos percentiles de la distribución t student.*

	p									
gl	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43	104.21
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12	128.30

Tabla F.3: *Algunos percentiles de la distribución* χ^2.

	m									
n	1	2	3	4	5	6	7	8	9	10
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
70	2.78	2.38	2.16	2.03	1.93	1.86	1.80	1.76	1.72	1.69
80	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71	1.68
90	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67

Tabla F.4: *Algunos percentiles 0.9 de la distribución* F_n^m.

	m									
n	12	14	16	18	20	24	30	40	60	100
1	60.71	61.07	61.35	61.57	61.74	62.00	62.26	62.53	62.79	63.01
2	9.41	9.42	9.43	9.44	9.44	9.45	9.46	9.47	9.47	9.48
3	5.22	5.20	5.20	5.19	5.18	5.18	5.17	5.16	5.15	5.14
4	3.90	3.88	3.86	3.85	3.84	3.83	3.82	3.80	3.79	3.78
5	3.27	3.25	3.23	3.22	3.21	3.19	3.17	3.16	3.14	3.13
6	2.90	2.88	2.86	2.85	2.84	2.82	2.80	2.78	2.76	2.75
7	2.67	2.64	2.62	2.61	2.59	2.58	2.56	2.54	2.51	2.50
8	2.50	2.48	2.45	2.44	2.42	2.40	2.38	2.36	2.34	2.32
9	2.38	2.35	2.33	2.31	2.30	2.28	2.25	2.23	2.21	2.19
10	2.28	2.26	2.23	2.22	2.20	2.18	2.16	2.13	2.11	2.09
11	2.21	2.18	2.16	2.14	2.12	2.10	2.08	2.05	2.03	2.01
12	2.15	2.12	2.09	2.08	2.06	2.04	2.01	1.99	1.96	1.94
13	2.10	2.07	2.04	2.02	2.01	1.98	1.96	1.93	1.90	1.88
14	2.05	2.02	2.00	1.98	1.96	1.94	1.91	1.89	1.86	1.83
15	2.02	1.99	1.96	1.94	1.92	1.90	1.87	1.85	1.82	1.79
16	1.99	1.95	1.93	1.91	1.89	1.87	1.84	1.81	1.78	1.76
17	1.96	1.93	1.90	1.88	1.86	1.84	1.81	1.78	1.75	1.73
18	1.93	1.90	1.87	1.85	1.84	1.81	1.78	1.75	1.72	1.70
19	1.91	1.88	1.85	1.83	1.81	1.79	1.76	1.73	1.70	1.67
20	1.89	1.86	1.83	1.81	1.79	1.77	1.74	1.71	1.68	1.65
21	1.87	1.84	1.81	1.79	1.78	1.75	1.72	1.69	1.66	1.63
22	1.86	1.83	1.80	1.78	1.76	1.73	1.70	1.67	1.64	1.61
23	1.84	1.81	1.78	1.76	1.74	1.72	1.69	1.66	1.62	1.59
24	1.83	1.80	1.77	1.75	1.73	1.70	1.67	1.64	1.61	1.58
25	1.82	1.79	1.76	1.74	1.72	1.69	1.66	1.63	1.59	1.56
26	1.81	1.77	1.75	1.72	1.71	1.68	1.65	1.61	1.58	1.55
27	1.80	1.76	1.74	1.71	1.70	1.67	1.64	1.60	1.57	1.54
28	1.79	1.75	1.73	1.70	1.69	1.66	1.63	1.59	1.56	1.53
29	1.78	1.75	1.72	1.69	1.68	1.65	1.62	1.58	1.55	1.52
30	1.77	1.74	1.71	1.69	1.67	1.64	1.61	1.57	1.54	1.51
40	1.71	1.68	1.65	1.62	1.61	1.57	1.54	1.51	1.47	1.43
50	1.68	1.64	1.61	1.59	1.57	1.54	1.50	1.46	1.42	1.39
60	1.66	1.62	1.59	1.56	1.54	1.51	1.48	1.44	1.40	1.36
70	1.64	1.60	1.57	1.55	1.53	1.49	1.46	1.42	1.37	1.34
80	1.63	1.59	1.56	1.53	1.51	1.48	1.44	1.40	1.36	1.32
90	1.62	1.58	1.55	1.52	1.50	1.47	1.43	1.39	1.35	1.30

Tabla F.5: *Algunos percentiles 0.9 de la distribución* F_n^m.

	m									
n	1	2	3	4	5	6	7	8	9	10
1	161.5	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94

Tabla F.6: *Algunos percentiles 0.95 de la distribución* F_n^m.

	m									
n	12	14	16	18	20	24	30	40	60	100
1	243.9	245.4	246.5	247.3	248.0	249.1	250.1	251.1	252.2	253.0
2	19.41	19.42	19.43	19.44	19.45	19.45	19.46	19.47	19.48	19.49
3	8.74	8.71	8.69	8.67	8.66	8.64	8.62	8.59	8.57	8.55
4	5.91	5.87	5.84	5.82	5.80	5.77	5.75	5.72	5.69	5.66
5	4.68	4.64	4.60	4.58	4.56	4.53	4.50	4.46	4.43	4.41
6	4.00	3.96	3.92	3.90	3.87	3.84	3.81	3.77	3.74	3.71
7	3.57	3.53	3.49	3.47	3.44	3.41	3.38	3.34	3.30	3.27
8	3.28	3.24	3.20	3.17	3.15	3.12	3.08	3.04	3.01	2.97
9	3.07	3.03	2.99	2.96	2.94	2.90	2.86	2.83	2.79	2.76
10	2.91	2.86	2.83	2.80	2.77	2.74	2.70	2.66	2.62	2.59
11	2.79	2.74	2.70	2.67	2.65	2.61	2.57	2.53	2.49	2.46
12	2.69	2.64	2.60	2.57	2.54	2.51	2.47	2.43	2.38	2.35
13	2.60	2.55	2.51	2.48	2.46	2.42	2.38	2.34	2.30	2.26
14	2.53	2.48	2.44	2.41	2.39	2.35	2.31	2.27	2.22	2.19
15	2.48	2.42	2.38	2.35	2.33	2.29	2.25	2.20	2.16	2.12
16	2.42	2.37	2.33	2.30	2.28	2.24	2.19	2.15	2.11	2.07
17	2.38	2.33	2.29	2.26	2.23	2.19	2.15	2.10	2.06	2.02
18	2.34	2.29	2.25	2.22	2.19	2.15	2.11	2.06	2.02	1.98
19	2.31	2.26	2.21	2.18	2.16	2.11	2.07	2.03	1.98	1.94
20	2.28	2.22	2.18	2.15	2.12	2.08	2.04	1.99	1.95	1.91
21	2.25	2.20	2.16	2.12	2.10	2.05	2.01	1.96	1.92	1.88
22	2.23	2.17	2.13	2.10	2.07	2.03	1.98	1.94	1.89	1.85
23	2.20	2.15	2.11	2.08	2.05	2.01	1.96	1.91	1.86	1.82
24	2.18	2.13	2.09	2.05	2.03	1.98	1.94	1.89	1.84	1.80
25	2.16	2.11	2.07	2.04	2.01	1.96	1.92	1.87	1.82	1.78
26	2.15	2.09	2.05	2.02	1.99	1.95	1.90	1.85	1.80	1.76
27	2.13	2.08	2.04	2.00	1.97	1.93	1.88	1.84	1.79	1.74
28	2.12	2.06	2.02	1.99	1.96	1.91	1.87	1.82	1.77	1.73
29	2.10	2.05	2.01	1.97	1.94	1.90	1.85	1.81	1.75	1.71
30	2.09	2.04	1.99	1.96	1.93	1.89	1.84	1.79	1.74	1.70
40	2.00	1.95	1.90	1.87	1.84	1.79	1.74	1.69	1.64	1.59
50	1.95	1.89	1.85	1.81	1.78	1.74	1.69	1.63	1.58	1.52
60	1.92	1.86	1.82	1.78	1.75	1.70	1.65	1.59	1.53	1.48
70	1.89	1.84	1.79	1.75	1.72	1.67	1.62	1.57	1.50	1.45
80	1.88	1.82	1.77	1.73	1.70	1.65	1.60	1.54	1.48	1.43
90	1.86	1.80	1.76	1.72	1.69	1.64	1.59	1.53	1.46	1.41

Tabla F.7: *Algunos percentiles 0.95 de la distribución* F_n^m.

	m									
n	1	2	3	4	5	6	7	8	9	10
1	647.8	799.5	864.2	899.6	921.9	937.1	948.2	956.7	963.3	968.6
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27
70	5.25	3.89	3.31	2.97	2.75	2.59	2.47	2.38	2.30	2.24
80	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.28	2.21
90	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19

Tabla F.8: *Algunos percentiles 0.975 de la distribución* F_n^m.

	m									
n	12	14	16	18	20	24	30	40	60	100
1	976.7	982.5	986.9	990.4	993.1	997.3	1001	1006	1010	1013
2	39.41	39.43	39.44	39.44	39.45	39.46	39.46	39.47	39.48	39.49
3	14.34	14.28	14.23	14.20	14.17	14.12	14.08	14.04	13.99	13.96
4	8.75	8.68	8.63	8.59	8.56	8.51	8.46	8.41	8.36	8.32
5	6.52	6.46	6.40	6.36	6.33	6.28	6.23	6.18	6.12	6.08
6	5.37	5.30	5.24	5.20	5.17	5.12	5.07	5.01	4.96	4.92
7	4.67	4.60	4.54	4.50	4.47	4.41	4.36	4.31	4.25	4.21
8	4.20	4.13	4.08	4.03	4.00	3.95	3.89	3.84	3.78	3.74
9	3.87	3.80	3.74	3.70	3.67	3.61	3.56	3.51	3.45	3.40
10	3.62	3.55	3.50	3.45	3.42	3.37	3.31	3.26	3.20	3.15
11	3.43	3.36	3.30	3.26	3.23	3.17	3.12	3.06	3.00	2.96
12	3.28	3.21	3.15	3.11	3.07	3.02	2.96	2.91	2.85	2.80
13	3.15	3.08	3.03	2.98	2.95	2.89	2.84	2.78	2.72	2.67
14	3.05	2.98	2.92	2.88	2.84	2.79	2.73	2.67	2.61	2.56
15	2.96	2.89	2.84	2.79	2.76	2.70	2.64	2.59	2.52	2.47
16	2.89	2.82	2.76	2.72	2.68	2.63	2.57	2.51	2.45	2.40
17	2.82	2.75	2.70	2.65	2.62	2.56	2.50	2.44	2.38	2.33
18	2.77	2.70	2.64	2.60	2.56	2.50	2.44	2.38	2.32	2.27
19	2.72	2.65	2.59	2.55	2.51	2.45	2.39	2.33	2.27	2.22
20	2.68	2.60	2.55	2.50	2.46	2.41	2.35	2.29	2.22	2.17
21	2.64	2.56	2.51	2.46	2.42	2.37	2.31	2.25	2.18	2.13
22	2.60	2.53	2.47	2.43	2.39	2.33	2.27	2.21	2.14	2.09
23	2.57	2.50	2.44	2.39	2.36	2.30	2.24	2.18	2.11	2.06
24	2.54	2.47	2.41	2.36	2.33	2.27	2.21	2.15	2.08	2.02
25	2.51	2.44	2.38	2.34	2.30	2.24	2.18	2.12	2.05	2.00
26	2.49	2.42	2.36	2.31	2.28	2.22	2.16	2.09	2.03	1.97
27	2.47	2.39	2.34	2.29	2.25	2.19	2.13	2.07	2.00	1.94
28	2.45	2.37	2.32	2.27	2.23	2.17	2.11	2.05	1.98	1.92
29	2.43	2.36	2.30	2.25	2.21	2.15	2.09	2.03	1.96	1.90
30	2.41	2.34	2.28	2.23	2.20	2.14	2.07	2.01	1.94	1.88
40	2.29	2.21	2.15	2.11	2.07	2.01	1.94	1.88	1.80	1.74
50	2.22	2.14	2.08	2.03	1.99	1.93	1.87	1.80	1.72	1.66
60	2.17	2.09	2.03	1.98	1.94	1.88	1.82	1.74	1.67	1.60
70	2.14	2.06	2.00	1.95	1.91	1.85	1.78	1.71	1.63	1.56
80	2.11	2.03	1.97	1.92	1.88	1.82	1.75	1.68	1.60	1.53
90	2.09	2.02	1.95	1.91	1.86	1.80	1.73	1.66	1.58	1.50

Tabla F.9: *Algunos percentiles 0.975 de la distribución* F_n^m.

	m									
n	1	2	3	4	5	6	7	8	9	10
1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52

Tabla F.10: *Algunos percentiles 0.99 de la distribución* F_n^m.

	m									
n	12	14	16	18	20	24	30	40	60	100
1	6106	6143	6170	6192	6209	6235	6261	6287	6313	6334
2	99.42	99.43	99.44	99.44	99.45	99.46	99.47	99.47	99.48	99.49
3	27.05	26.92	26.83	26.75	26.69	26.60	26.50	26.41	26.32	26.24
4	14.37	14.25	14.15	14.08	14.02	13.93	13.84	13.75	13.65	13.58
5	9.89	9.77	9.68	9.61	9.55	9.47	9.38	9.29	9.20	9.13
6	7.72	7.60	7.52	7.45	7.40	7.31	7.23	7.14	7.06	6.99
7	6.47	6.36	6.28	6.21	6.16	6.07	5.99	5.91	5.82	5.75
8	5.67	5.56	5.48	5.41	5.36	5.28	5.20	5.12	5.03	4.96
9	5.11	5.01	4.92	4.86	4.81	4.73	4.65	4.57	4.48	4.41
10	4.71	4.60	4.52	4.46	4.41	4.33	4.25	4.17	4.08	4.01
11	4.40	4.29	4.21	4.15	4.10	4.02	3.94	3.86	3.78	3.71
12	4.16	4.05	3.97	3.91	3.86	3.78	3.70	3.62	3.54	3.47
13	3.96	3.86	3.78	3.72	3.66	3.59	3.51	3.43	3.34	3.27
14	3.80	3.70	3.62	3.56	3.51	3.43	3.35	3.27	3.18	3.11
15	3.67	3.56	3.49	3.42	3.37	3.29	3.21	3.13	3.05	2.98
16	3.55	3.45	3.37	3.31	3.26	3.18	3.10	3.02	2.93	2.86
17	3.46	3.35	3.27	3.21	3.16	3.08	3.00	2.92	2.83	2.76
18	3.37	3.27	3.19	3.13	3.08	3.00	2.92	2.84	2.75	2.68
19	3.30	3.19	3.12	3.05	3.00	2.92	2.84	2.76	2.67	2.60
20	3.23	3.13	3.05	2.99	2.94	2.86	2.78	2.69	2.61	2.54
21	3.17	3.07	2.99	2.93	2.88	2.80	2.72	2.64	2.55	2.48
22	3.12	3.02	2.94	2.88	2.83	2.75	2.67	2.58	2.50	2.42
23	3.07	2.97	2.89	2.83	2.78	2.70	2.62	2.54	2.45	2.37
24	3.03	2.93	2.85	2.79	2.74	2.66	2.58	2.49	2.40	2.33
25	2.99	2.89	2.81	2.75	2.70	2.62	2.54	2.45	2.36	2.29
26	2.96	2.86	2.78	2.72	2.66	2.58	2.50	2.42	2.33	2.25
27	2.93	2.82	2.75	2.68	2.63	2.55	2.47	2.38	2.29	2.22
28	2.90	2.79	2.72	2.65	2.60	2.52	2.44	2.35	2.26	2.19
29	2.87	2.77	2.69	2.63	2.57	2.49	2.41	2.33	2.23	2.16
30	2.84	2.74	2.66	2.60	2.55	2.47	2.39	2.30	2.21	2.13
40	2.66	2.56	2.48	2.42	2.37	2.29	2.20	2.11	2.02	1.94
50	2.56	2.46	2.38	2.32	2.27	2.18	2.10	2.01	1.91	1.82
60	2.50	2.39	2.31	2.25	2.20	2.12	2.03	1.94	1.84	1.75
70	2.45	2.35	2.27	2.20	2.15	2.07	1.98	1.89	1.78	1.70
80	2.42	2.31	2.23	2.17	2.12	2.03	1.94	1.85	1.75	1.65
90	2.39	2.29	2.21	2.14	2.09	2.00	1.92	1.82	1.72	1.62

Tabla F.11: *Algunos percentiles 0.99 de la distribución* F_n^m.

Bibliografía

Agresti, A. & Caffo, B. (2000), 'Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures.', *The American Statistician* **54**, 280–288.

Agresti, A. & Coull, B. (1998), 'Approximate is better than exact for interval estimation of binomial proportions', *The American Statistician* **52**, 119–126.

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*, john willey & sons edn, McGraw-Hill.

Ardilly, P. & Tillé, Y. (2006), *Sampling Methods: Exercises and Solutions.*, Springer.

Bartlett, M. S. (1937), 'Properties of sufficiency and statistical tests', *Proceedings of the Royal Society of London* **A**(160), 268–282.

Bickel, P. J. & Doksum, K. A. (2001), *Mathematical Statistics. Basic Ideas and Selected Topics*, Vol. 1, second edn, Prentice-Hall.

Blanco, L. (2004), *Probabilidad*, Universidad Nacional de Colombia, Unibiblos.

Box, G. E. P. & Cox, D. R. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society. Series B (Methodological)* **26**(2), 211–252.

Canavos, G. C. (1988), *Probabilidad y Estadística. Aplicaciones y Métodos.*, McGraw-Hill.

Carlin, B. P. & Louis, T. A. (1996), *Bayes and Empirical Bayes for Data Analysis*, 1 edn, Chapman and Hall/CRC.

Casella, G. & Berger, R. L. (2002), *Statitical Inference*, second edn, Duxbury.

Cepeda, E., Aguilar, W., Cervantes, V., Corrales, M., Díaz, I. & Rodríguez, D. (2008), 'Intervalos de confianza e intervalos de credibilidad para una proporción', *Revista Colombiana de Estadística* **31**(2), 211–228.

Conover, W. J. (1998), *Practical Nonparametric Statistics*, 3 edn, Wiley.

Garwood, F. (1936), 'Fiducial limits for the poisson distribution', *Biometrika* **28**, 437–442.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004), *Bayesian Data Analysis.*, Chapman & Hall/CRC.

Gupta, A. K. & Nagar, D. K. (2000), *Matrix variate distribucutions.*, Chapman & Hall/CRC.

Gutiérrez, H. A. (2009), *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*, Universidad Santo Tomas.

Gutiérrez, H. A. & Zhang, H. (2009), 'Análisis bayesiano para la diferencia de dos proporciones usando r', *Revista de métodos cuantitativos para la economía y la empresa* **9**, 50–70.

Johnson, R. A. & Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis*, Prentice Hall.

Jortiz, J. & Zhang, H. (2010), 'Inclusión de la igualdad en la hipótesis nula', *Comunicaciones en Estadística* **3**(2).

Leemis, L. M. & McQueston, J. T. (2008), 'Univariate distribution relationships', *The American Statistician* **62**(1), 45–53.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses*, second edn, Springer.

Lehmann, E. L. & Romano, J. P. (2005), *Testing Statistical Hypotheses*, third edn, Springer.

Mardia, K. V. (1970), 'Measures of multivariate skewness and kurtosis with applications', *Biometrika* **57**(3), 519–530.

Mayorga, J. H. (2004), *Inferencia estadística*, Universidad Nacional de Colombia, Unibiblos.

McCullagh, P. (1994), 'Does the moment-generating function characterize a distribution?', *The American Statistician* **48**(3), 208.

Miller, L. H. (1956), 'Table of percentage points of kolmogorov statistics', *Journal of the American Statistical Association* **51**, 111–121.

Mood, A., Graybill, F. A. & Boes, D. (1974), *Introduction to the Theory of Statistics*, third edn, McGraw-Hill.

Newcombe, R. (1998), 'Interval estimation for the difference between independent proportions: Comparison of eleven methods', *Statistics in Medicine* **17**, 873–890.

Peña, D. (2002), *Análisis de Datos Multivariantes*, McGraw-Hill.

Resnick, S. I. (1999), *A Probability Path*, Birhäuser.

Robert, C. & Casella, G. (1999), *Monte Carlo Statistical Methods*, Springer.

Salsburg, D. (2002), *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, Holt Paperbacks.

Schervish, M. J. (1996), '*P* values: What they are and what they are not', *The American Statistician* **50**(3), 203–206.

Shapiro, S. S. & Wilk, M. B. (1965), 'An analysis of variance test for normality (complete samples)', *Biometrika* **52**(3/4).

Stahl, S. (2008), 'La evolución de la distribución normal', *Comunicaciones en Estadística* **1**(1), 13–32.

Student (1908), 'The probable error of a mean', *Biometrika* **6**(1).

Welch, B. (1949), 'Further notes om mrs. aspin's tables', *Biometrika* **36**, 243–246.

Wichura, M. J. (1988), 'Algorithm as 241: The percentage points of the normal distribution', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **37**(3), 477–484.

Zhang, H. (2009), 'A note about maximum likelihood estimator in hypergeometric distribution', *Comunicaciones en Estadística.* **2**(2), 1.

Zhang, H., Gutiérrez, A. & Cepeda, E. (2009), Frequentist performance of some credibility and confidence intervals for the difference of two proportions. XIX simposio de estadística, Medellín, Colombia.

Índice de figuras

Índice de tablas

Índice alfabético